15th Dec. 2019

Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems

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

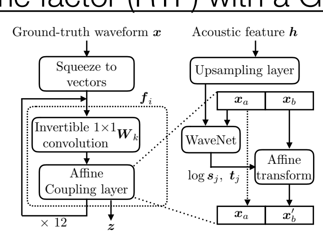
[1]National Institute of Information and Communications Technology, Japan, [2]Nagoya University, Japan

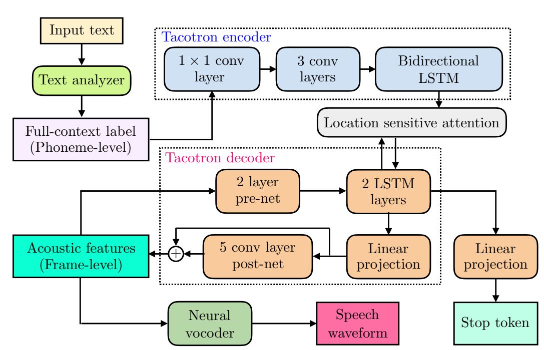
1. Introduction

- **Conventional text-to-speech (TTS) systems**
  - Duration and acoustic pipeline models with source-filter vocoders
    - Widely used in practical systems but not high quality synthesis
- **End-to-end neural TTS systems**
  - Sequence-to-sequence (seq2seq) model with neural vocoders
    - Jointly optimizing duration and acoustic models and directly converting character or phoneme sequences to acoustic features (mel-spectrogram)
  - State-of-the art end-to-end TTS models
    - Tacotron 2 with autoregressive WaveNet vocoder: Human quality synthesis
    - ClariNet (Deep voice 3 + parallel WaveNet): Entire end-to-end real-time neural TTS
    - Transformer-based TTS: Faster training than Tacotron 2
- **Problem of seq2seq models due to attention prediction error**
  - Speech samples sometimes cannot be successfully synthesized
    - Crucial problem for practical TTS systems
- **Real-time, high-fidelity, and stable neutral TTS systems with Tacotron structure**
  - Introducing conventional duration models to sophisticated seq2seq acoustic models
    - HMM-based forced alignment can be relatively easily obtained
    - Conventional duration model can estimate almost accurately predict phoneme durations

2. Seq2seq acoustic model with full-context label input

- **Tacotron 2 with full-context label input for pitch accent languages**
  - Input: Full-context label (130 dims)
  - Output: Mel-spectrogram (80 dims)
- **Real-time neural TTS with WaveGlow vocoder**
  - Real time factor (RTF) with a GPU: 0.16

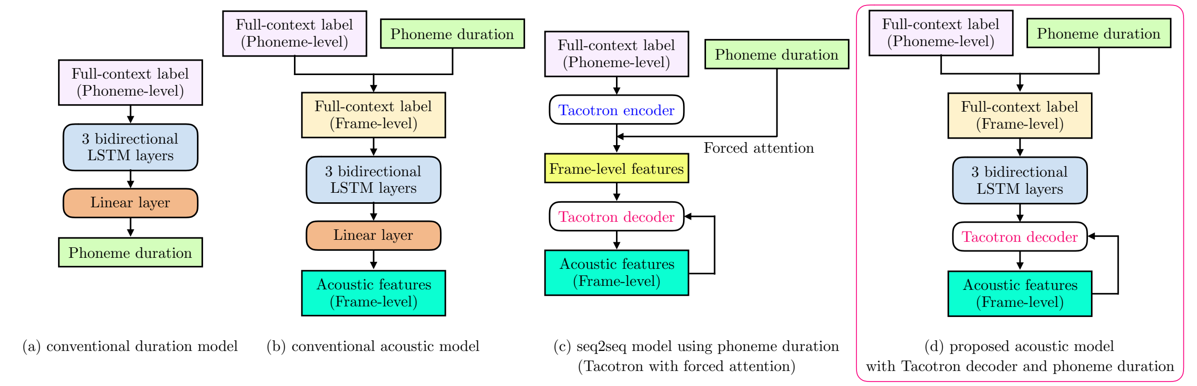




T. Okamoto *et al.*, Interspeech 2019

This model is also unstable due to attention-based seq2seq structure

3.Proposed method

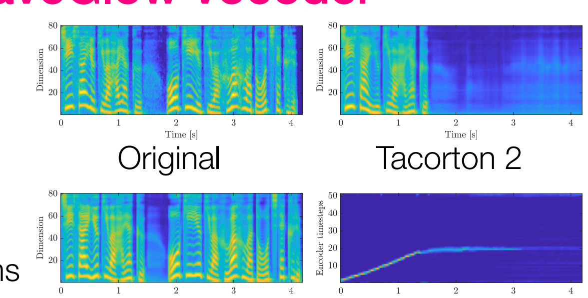


(a) conventional duration model (b) conventional acoustic model CAM (c) seq2seq model using phoneme duration (Tacotron with forced attention) FAT (d) proposed acoustic model with Tacotron decoder and phoneme duration PAM

- **Tacotron with forced attention (FAT)**
  - Encoded features are duplicated and redundant for decoder
    - FAT cannot outperform Tacotron (Y. Yasuda *et al.*, ICASSP 2019)
- **Proposed acoustic model with Tacotron decoder and phoneme duration (PAM)**
  - HMM-based forced alignment and bidirectional LSTM-based duration model
  - Acoustic model with bidirectional LSTM and decoder of Tacotron 2
    - Redundancy in FAT can be reduced

4. Experiments with WaveGlow vocoder

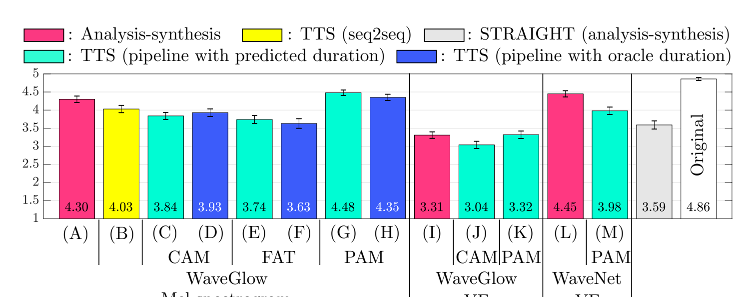
- **Simulation condition**
  - Japanese Female corpus: 18 h
  - Acoustic features
    - Mel-spectrogram: 80 dims, 12.5 ms
    - Vocoder features (fo, vuv, mel-cepstrum): 1 + 1 + 35 = 37 dims, 5 ms



Original Tacorton 2

PAM Predicted attention

- **Results**



MOS results with 15 listening subjects

RTF with an NVIDIA Tesla V100

Method	AM RTF	Total RTF
(A):WG-MELSPC-AS	-	0.066
(B):WG-MELSPC-TTS-seq2seq	0.063	0.13
(C):WG-MELSPC-TTS-CAM	0.015	0.08
(D):WG-MELSPC-TTS-CAM (OD)	0.015	0.08
(E):WG-MELSPC-TTS-FAT	0.049	0.12
(F):WG-MELSPC-TTS-FAT (OD)	0.049	0.12
(G):WG-MELSPC-TTS-PAM	0.061	0.13
(H):WG-MELSPC-TTS-PAM (OD)	0.061	0.13
(I):WG-VF-AS	-	0.06
(J):WG-VF-TTS-CAM	0.045	0.10
(K):WG-VF-TTS-PAM	0.138	0.20
(L):WN-VF-AS	-	200
(M):WN-VF-TTS-PAM	0.06	200

Real-time, high-fidelity, and stable neural TTS can be realized by PAM