GRAPHTTS: GRAPH-TO-SEQUENCE MODELLING IN NEURAL TEXT-TO-SPEECH

이윤형

2020/06/19

Overview

- 기존 Tacotron2 모델에 Graph Auxiliary Encoder (GAE)를 추가하여 prosody 정보를 추출, 더욱 자연스러운 음성을 합성해낸다.

- GAE는 입력 text를 graph로 변환한 이후 이를 입력으로 받는다.

- 다양한 Graph Neural Network (GNN) 모듈 (GCN, GGNN-GRU, GGNN-LSTM)을 사용해 GAE를 구성하고 이들의 성능을 비교하였다.

Text-to-Graph Conversion

Hyper-parameter	Tensor dimension
Node embedding	512-D
Edge embedding	(E, 2, 3)
GraphTTS-encoder	512-D
GAE	128-D

Table 1. Model configuration

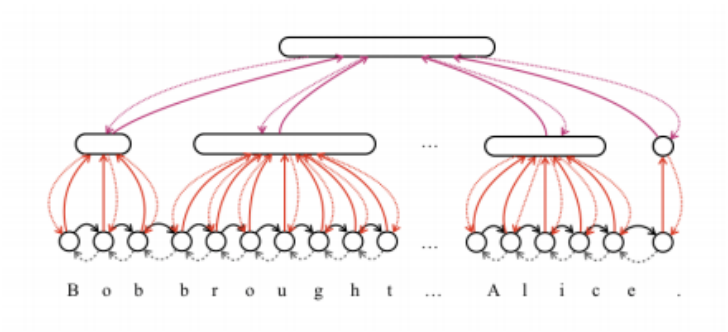
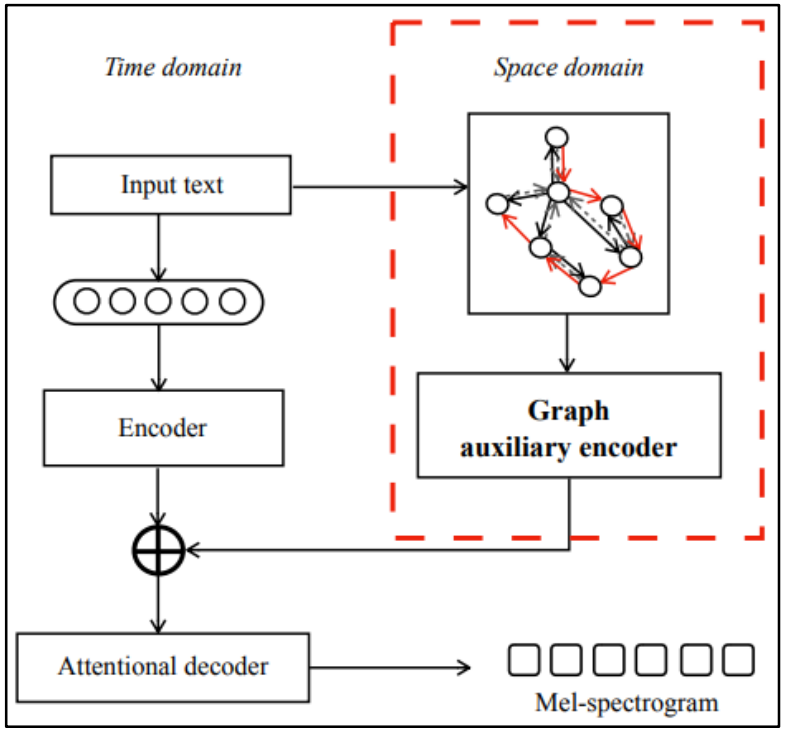


Fig. 3. Character-level text-to-graph module

- Characters of input texts are represented by nodes
- The adjacency connections among characters are modelled by edges.
  (solid: directed edges / dashed: reverse edges)
- The connection between characters in a word is identified with the strong connection
  (represented by word-level red lines)
- The connection between characters in different words is identified with the weak connection
  (represented by sentence-level purple lines)

Graph auxiliary encoder (GAE)



- The main information flow follows the original structure of Tacotron2
- It guarantees the monotonic alignments of text characters and speeches.
- GAE makes the prosody modelling of the speech generation as an automated procedure.

Experiments

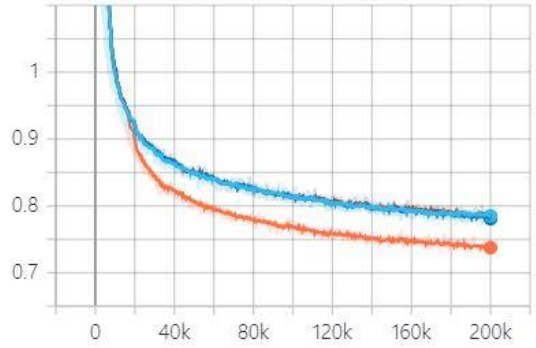
- Baseline: Transformer-TTS

- Experimental group: graph-tts, graph-tts-iter5, gae

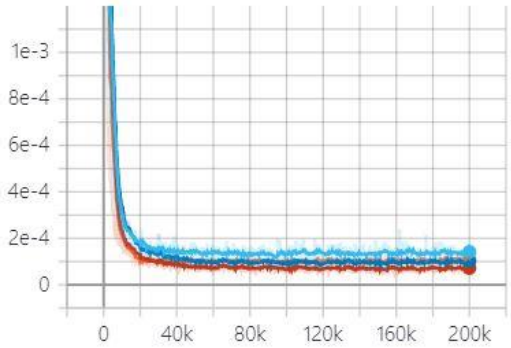
Experiments

- Loss curves
  (Orange: transformer-tts / Navy: graph-tts / Red: grap-tts-iter5 / Blue: gae)

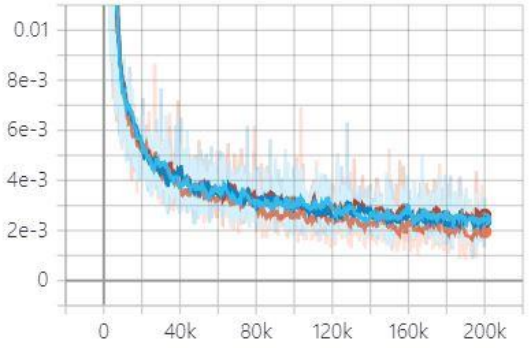
Train_mel_loss



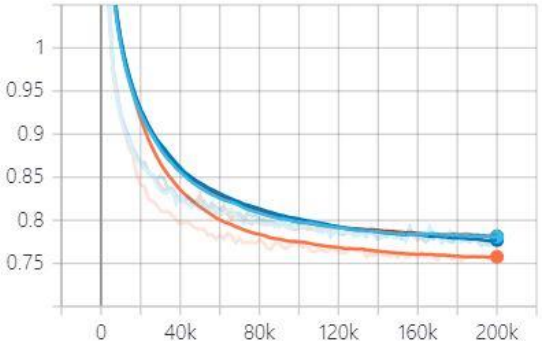
Train_guide_loss



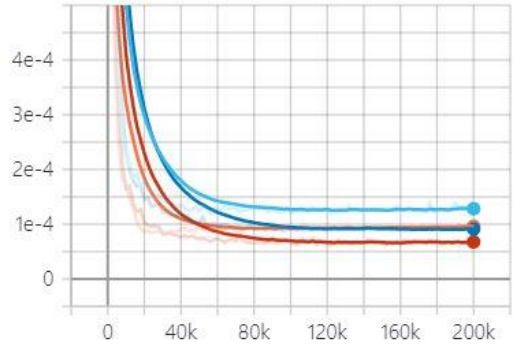
Train_bce_loss



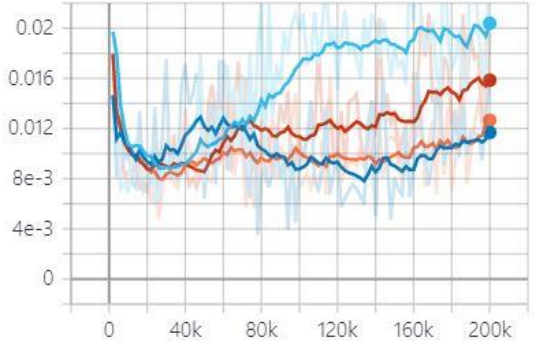
Validation_mel_loss



Validation_guide_loss



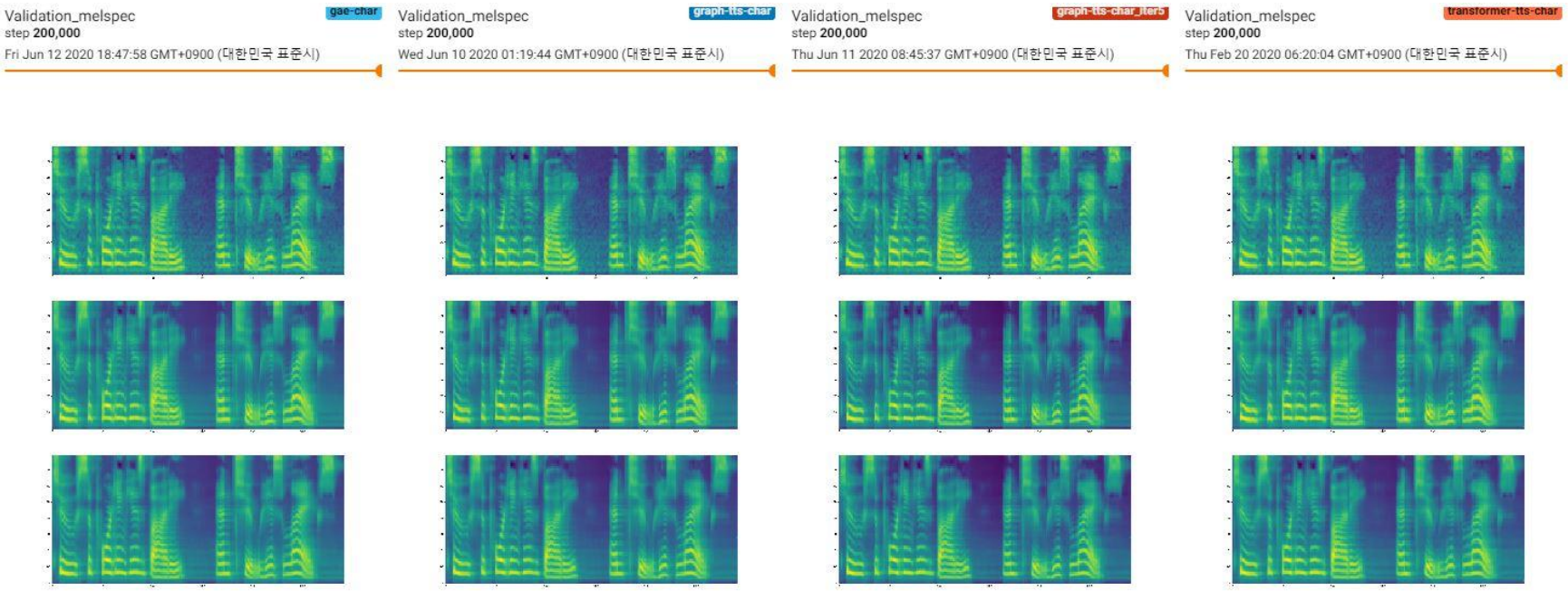
Validation_bce_loss



Experiments

- Outputs

1. Melspectrogram



Experiments

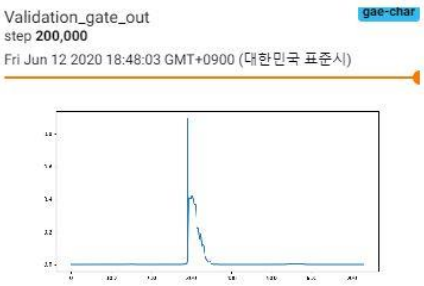
- Outputs

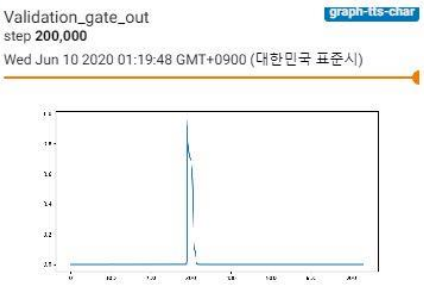
2. Enc-Dec alignments (6-layer, 4-head each)

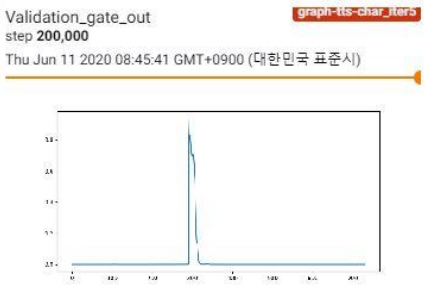
Experiments

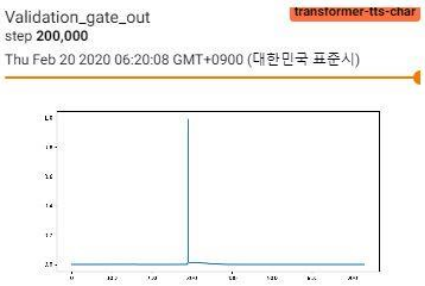
- Outputs

3. Stop prediction









Experiments

- Outputs

4. Audio Samples

GraphTTS / Transformer-TTS

Discussion

- Graph-TTS를 구현해보았으나 논문에서 주장하는 것과 달리 베이스라인 모델 대비 성능이 떨어지는 것을 확인하였다.
- 구현 상의 문제인지 제대로 된 소리를 들을 수 없었다.
- GNN의 message propagation 횟수를 늘릴 수록 성능 하락은 더욱 심해졌다.
- Train 시에는 제대로 학습이 이루어지고 음성을 생성해내는 것으로 보아 Graph Encoder가 실제로 align을 학습하는데 도움을 주는지 의심이 간다. (train-test gap)
- Baseline model을 Tacotron2가 아닌 TransformerTTS를 사용하였는데 여기서 오는 차이일 수도 있다고 생각이 든다.