Prosody-aware SpeechT5 For Expressive Neural TTS

- ICASSP 2023
- Microsoft, China / Microsoft Research Asia
- Paper, Demo, Github

최예린
서강대학교 인공지능학과
Email: lakahaga@u.sogang.ac.kr
2023.5.17

Overview

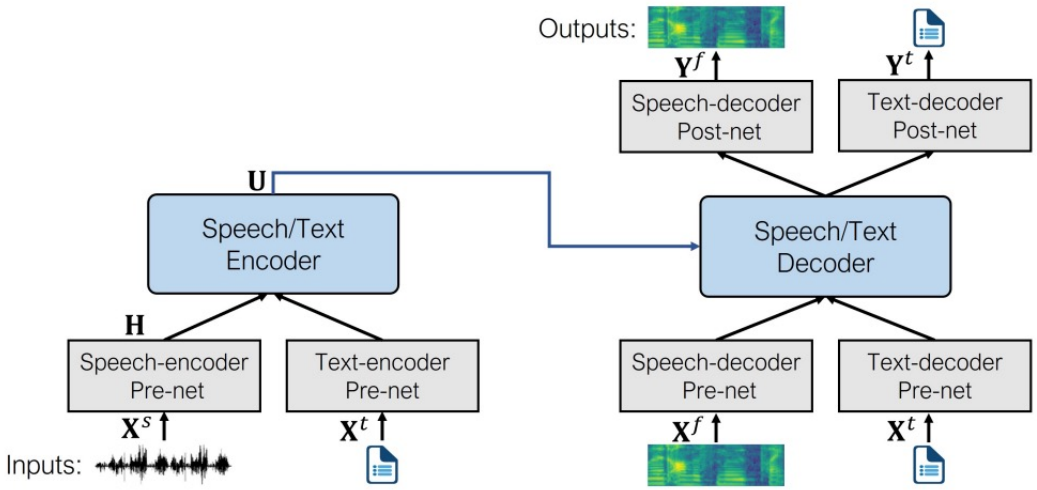
- 기존 SpeechT5 에 prosody modeling 구조를 추가
  - prosody 정보를 explicit하게 고려할 수 있도록 함
- 기존 prosody modeling 관련 pre-training
  - decoder만 혹은 prosody encoder만 사전 학습하는 방식으로 이루어짐

- SpeechT5
  - text와 speech 에 대해 unified modal representation을 학습하는 모델
  - 여기에 prosody 를 Explicit하게 고려할 수 있는 모듈을 추가하여
  - Prosody에 대해서 encoder, encoder, prosody encoder를 Jointly pretraining 하는 방법 제안

- OOD 데이터에 대해서 좋은 성능을 보임

Remind) SpeechT5

- SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing
  - github

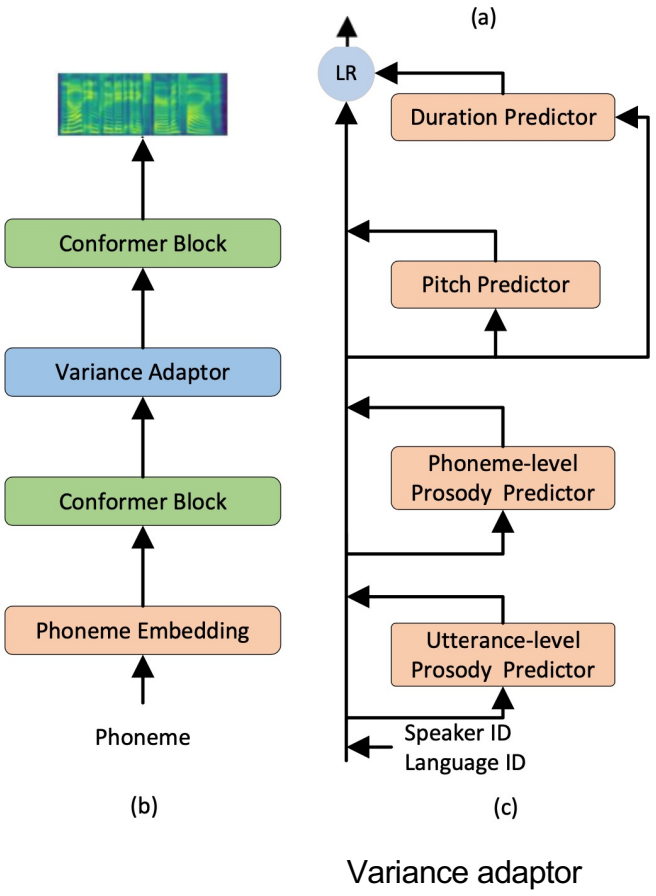


Encoder-decoder module and six modal-specific pre/post nets

- **Pre-net: raw input을 Shared encoder/ decoder에 알맞은 feature 로 바꿈**
  - **Text-encoder pre-net: / Speech-encoder pre-net: Convolution feature extractor**
  - **Speech-decoder pre-net: Tacotron2 의 decoder pre-net**
    - **stack of linear layer-relu-dropouts**
- **Post-net: decoder에서 생성된 것을 각 modality 에 맞게 변환**
  - **Speech-decoder post-net: linear projection**
  - **text-decoder post-net: linear projection to vocab size**

Backbone) DelightfulTTS

- DelightfulTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2021
  - Demo, Github(coqui, implementation in progress)



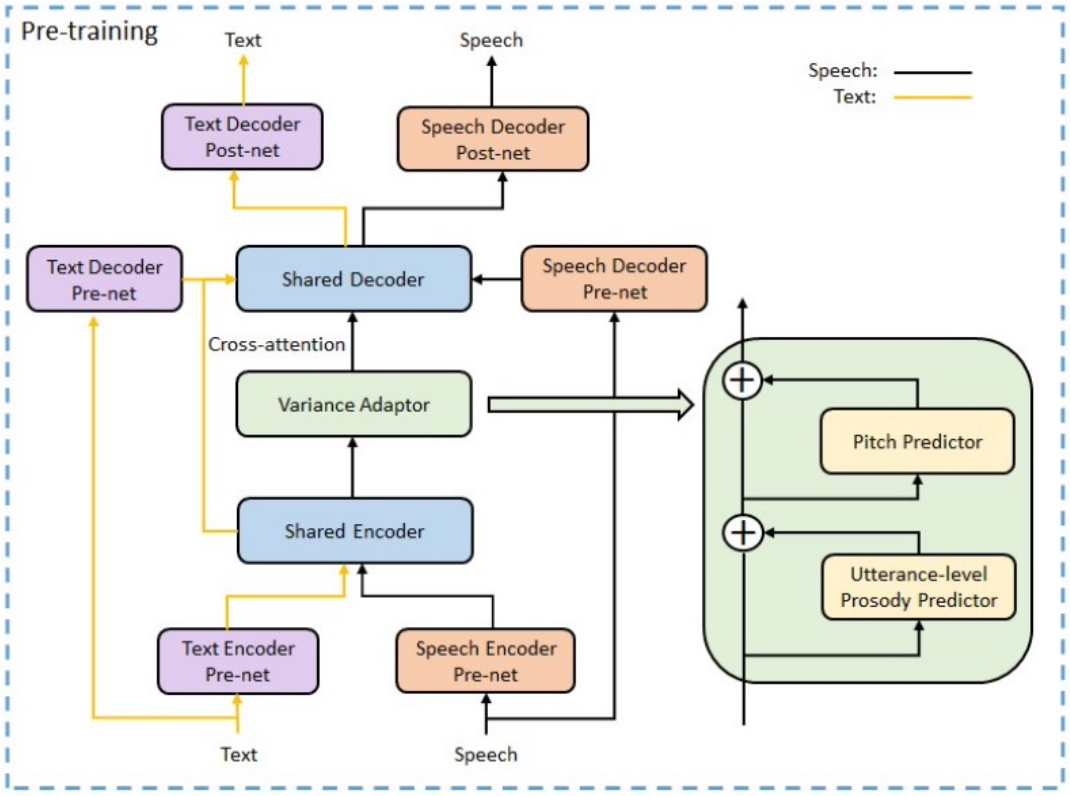
Variance adaptor

- 최대한 다양한 정보를 학습하는 variance adaptor 구성
  - Prosody predictor: GST의 reference encoder
  - Utterance-level prosody predictor
    - 기존 GST와 동일
    - CNN-GRU-Style token layer(Attention)
  - Phoneme-level prosody predictor
    - Input: utt-level prosody predictor의 output + phoneme encoder의 output
    - Attention
      - Q: phoneme encoder의 output
      - K,V: utt-level prosody predictor의 output

- Blizzard Challenge 2021 1등

Prosody-aware Speech T5

- 기존 SpeechT5의 encoder와 decoder 사이에 variance adaptor를 추가



(a) Prosody-aware SpeechT5

- Variance adaptor
  - Shared encoder의 last hidden state로부터 Prosody를 예측할 수 있도록 학습

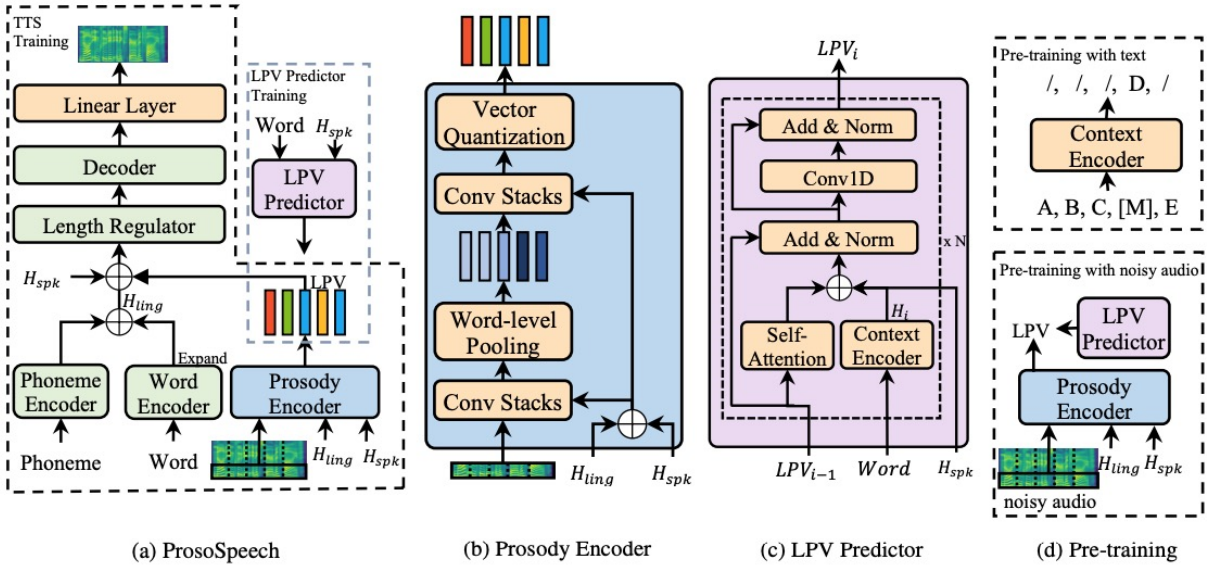
- pre-training에서는 phoneme label이 없기 때문에 duration predictor를 없애고,
- Use cross-attention to align mapping of encoder output to target frame-level features
  - Q: shared encoder output
  - K,V: variance adaptor output

- Text data는 prosodic information이 없기 때문에 variance adaptor를 지나지 않음

- Main block을 transformer에서 conformer로 변경
  - Shared encoder: 12 conformer blocks
  - Shared decoder: 6 conformer blocks

Prosody Encoder 구조

- [ProsoSpeech: Enhancing Prosody with Quantized Vector Pre-training In TTS]
  - ICASSP 2022, Zhejiang University, Speech Lab Alibaba



(a) ProsoSpeech (b) Prosody Encoder (c) LPV Predictor (d) Pre-training

- Noisy audio를 사용하여 prosody encoder를 pretrain
- Word-level prosody vector를 encoding

- ProsodyAwareSpeechT5에서 차용한 구조는 LPV Predictor
  - Encoder output으로부터 prosody vector를 예측하도록 훈련

- LPV Predictor
  - LPV : Latent Prosody Vector
  - Text input으로부터 word-level LPV sequence를 예측
  - 이때 정답값으로 prosody encoder에서 encoding한 LPV를 사용
  - Self-attention based autoregressive architecture

Prosody-aware Speech T5

- **Pretraining Loss**

$$\mathcal{L} = \mathcal{L}_{speecht5} + \mathcal{L}^{s}_{pitch} + \mathcal{L}^{s}_{utt} \quad (1)$$

$$\mathcal{L}_{speecht5} = \mathcal{L}^{s}_{mlm} + \mathcal{L}^{s}_{1} + \mathcal{L}^{s}_{bce} + \mathcal{L}^{t}_{mle} + \mathcal{L}^{joint}_{d} \quad (2)$$

- **Speech T5 Loss**
  - **Speech representation learning loss**
    - **Cross-entropy loss over masked time steps in acoustic unit sequence**
  - **Speech reconstruction loss**
    - **L1 loss between generated features and the GT features of input speech**
    - **Binary cross entropy loss for the stop token**
  - **Text reconstruction loss**
    - **Maximum likelihood estimation of reconstructing original text from corrupted text input**
  - **Diversity loss**
    - **Entropy loss of averaged softmax distribution of discrete representation in shared code-book**

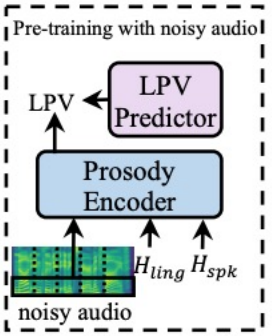
Prosody-aware Speech T5

- **Pretraining Loss**

$$\mathcal{L} = \mathcal{L}_{speecht5} + \mathcal{L}^{s}_{pitch} + \mathcal{L}^{s}_{utt} \quad (1)$$

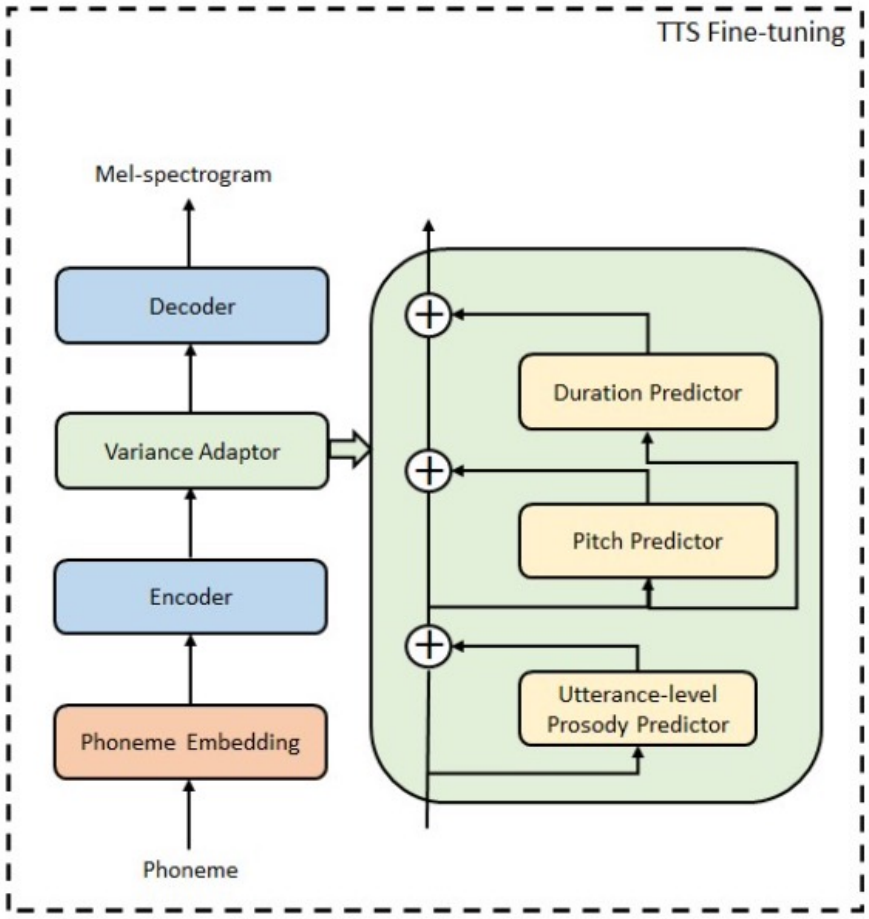
$$\mathcal{L}_{speecht5} = \mathcal{L}^{s}_{mlm} + \mathcal{L}^{s}_{1} + \mathcal{L}^{s}_{bce} + \mathcal{L}^{t}_{mle} + \mathcal{L}^{joint}_{d} \quad (2)$$

- **Pitch loss**
  - **L1 loss between predicted pitch and GT pitch**
- **Utterance-level prosody prediction loss**
  - **L1 loss between predicted utt-level prosody vector and the vector extracted from utt-level reference encoder**



(d) Pre-training

TTS Finetuning



(b) Acoustic Model in Neural TTS

- DelightfulTTS에서 phoneme-level prosody predictor를 제외한 구조
- Pretrain과 Finetune의 inconsistency
  - Sub-word vs Phoneme
  - Frame-level vs Phoneme-level
  - AR vs Parallel
- 해결 방법
  - Encoder의 첫 layer는 새로 initialize
  - Pretrained embedding을 모두 update
- 먼저 multi-speaker에 대해 tts-finetuning
  - 300 hours of English corpus (내부 코퍼스)
    - 29 speakers
- 이후, 특정 화자에 대해 adapt하는 과정 진행
  - 10 hours of single female speakers

Experiment

- 훈련 데이터
  - **Pretrain:**
    - **Speech : LibriSpeech 960 hours + LibriLigth audios**
    - **Text: LibriSpeeech의 transcript (40M sentences)**
  - **TTS finetune**
    - **300 hours of English corpus (내부 코퍼스) : 29 speakers**
    - 이후 10 시간 분량의 여성 단화자 코퍼스에 대해 학습
- 평가 데이터
  - **In-domain data**
    - 훈련에서 제외된 테스트셋
  - **OOD data**
    - **News, Audiobook에서 각각 30문장**
  - **Robustness test data**
    - 다른 도메인에서 랜덤하게 선정된 100문장 (1795 단어를 포함)

Result

Table 1. Subjective and objective results of different models. Issue rate is computed at word-level on the robustness test set. Prosody issues include word-level incorrect stress, intonation, tone and pause. Prosody statistics are calculated on the two OOD test sets (News and Audiobook). Positive CMOS scores indicate that the prosody-aware SpeechT5 is better than baseline.

Model	CMOS			Issue Rate	Prosody Statistics	
	In-domain	News (OOD)	Audiobook (OOD)	Prosody	Pitch Std	Pitch Range
Avg. Word Number per Sentence	12	23	31	-	-	-
Baseline	0.00	0.00	0.00	1.63%	33.687	166.488
+ prosody-aware SpeechT5	-0.016	**0.154**	**0.114**	**1.32%**	**37.378**	**177.503**
Relative Issue Rate Reduction	-	-	-	19.02%	-	-

- **Baseline: SpeechT5**
- **Subjective evaluation**
  - **CMOS : 긴 문장들에 대해서도 좋은 성능 (in domain은 baseline이 더 좋음**
  - **Prosody Issue Rate: word-level에서 Incorrect stress, intonation, tone, pause의 비율을 측정한 것**
    - **Robustness test data에 대해서 측정**
  - **Relative Issue Rate Reduction: baseline에 비해 감소한 비율**
- **Objective evaluation**
  - **Prosody statistics**
    - **Baseline에 비해 다양한 pitch 를 합성**

Ablation Study

- **Pretrain 데이터 및 structure에 대한 비교**

Table 2. Performance of training prosody-aware SpeechT5 with different strategies. Conformer is used in baseline neural TTS model.

Pre-training Method		CMOS	
Pre-training Data	Encoder-Decoder	News	Audiobook
LibriSpeech (960h)	Conformer	-0.090	-0.062
LibriLight (60kh)	Transformer	0.00	-0.006
LibriLight (60kh)	Conformer	0.154	0.114

- 동일 데이터셋에 대해서는 **Transformer** 보다는 **Conformer** 가 좋고
- 동일한 **Conformer**에 대해서는 대용량 데이터인 **LibriLight**를 사용하여 **Pre-train** 했을 때 더 성능이 좋음

Ablation Study

- **Pretraining module에 대한 ablation study**

Table 3. Comparison of pre-training different modules and their combinations. Encoder + Variance Adaptor + Decoder are all used in prosody-aware SpeechT5.

Pre-training Modules	CMOS	
	News	Audiobook
Prosody-aware SpeechT5	0.00	0.00
w/o Decoder and Variance Adaptor	0.027	-0.077
w/o Encoder and Variance Adaptor	0.033	-0.076

- **w/o Variance Adaptor = SpeechT5**

- **Decoder와 Variance adaptor는 pretrain하지 않고, TTS 학습했을 때**
  - **즉, encoder만 pretrain 했을 때**
- **Encoder와 Variance adaptor는 pretrain하지 않고, TTS 학습했을 때**
  - **즉, decoder만 pretrain 했을 때**

- 두가지 경우 모두 **News**에 대해서는 오히려 성능이 좋아지지만 **Audiobook**에 대해서는 성능이 떨어진다.
  - **Prosody**가 더 필요한 음성에서 **Variance adaptor**가 없을 때 성능이 떨어짐

Conclusion

- 기존 SpeechT5에 variance adptor를 추가하여 Prosody 정보를 포함한 pretrain 방법을 제시하여
- OOD data에 대해서 TTS의 성능을 높임

- Direct Speech-to-Speech modeling 방법에 적용
  - Audio 모델의 장점 -> prosody 등 text에 없는 정보를 활용 가능
  - Prosody 를 pretrain 과정에서 포함해야 할 것
  - 본 논문에서 제시한 pretrain 구조를 참고할 수 있을 것

끝

감사합니다.