Unsupervised Tex-to-Speech Synthesis by Unsupervised Automatic Speech Recognition

Junrui Ni, Liming Wang, Heting Gao

May 6, 2022

Outline

Motivation

- Text-to-speech (TTS) synthesis is an essential component of a spoken dialogue system
- Existing state-of-the-art TTS systems such as Tacotron 1&2[1], FastSpeech[2] and Transformer TTS[3] are trained with **paired** speech and text;
- Training a supervised text-to-speech (TTS) system requires **dozens of hours** of **single-speaker** high-quality recordings, which can be quite time-consuming and expensive to collect

[1] Yuxuan Wang et al. "Tacotron: Towards end-to-end speech synthesis". In: *arXiv*. 2017. URL: preprint%20arXiv:1703.10135, Jonathan Shen et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". In: *ICASSP*. 2018

[2] Yi Ren et al. "FastSpeech: Fast, Robust and Controllable Text to Speech". In: *Advances in Neural Information Processing Systems*. 2019

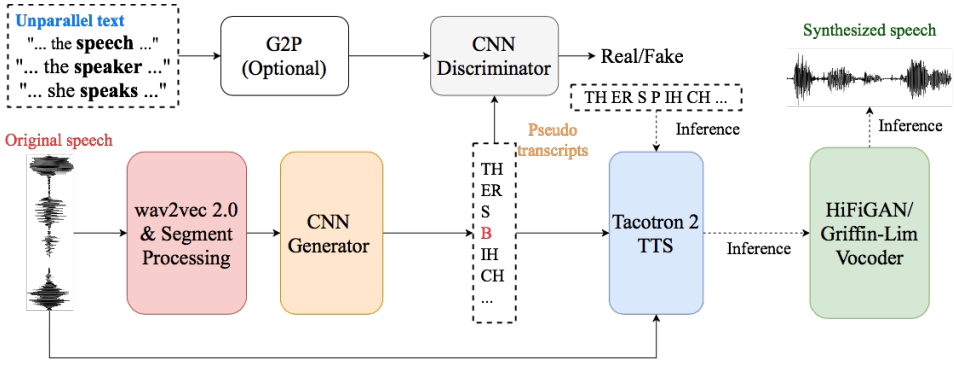
[3] N.Li et al. "Neural speech synthesis with transformer network". In: *AAAI*. vol. 33. 2019, pp. 6706–6713

Outline

Unsupervised TTS: problem formulation

- **Text input**: phoneme or character sequence $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_m]$;
- **Speech input**: **Unpaired** speech $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n], m \neq n$;
- **Output**: A **generator** function $G(\cdot)$ to map text into its corresponding speech waveform

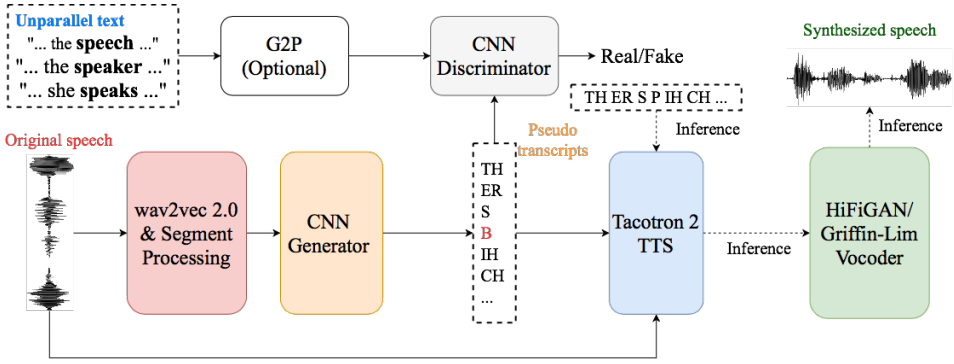
Proposed model for unsupervised TTS



- **Step 1**: Learn an unsupervised ASR (wav2vec-U[4]) to generate **pseudo-transcripts** for x_i's as $Y = [\tilde{y}_1, \cdots, \tilde{y}_n]$

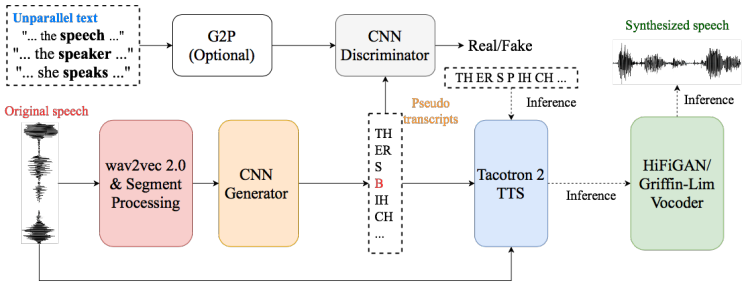
[4] Alexei Baevski et al. "Unsupervised Speech Recognition". In: *Neural Information Processing Systems*. 2021

Proposed model for unsupervised TTS



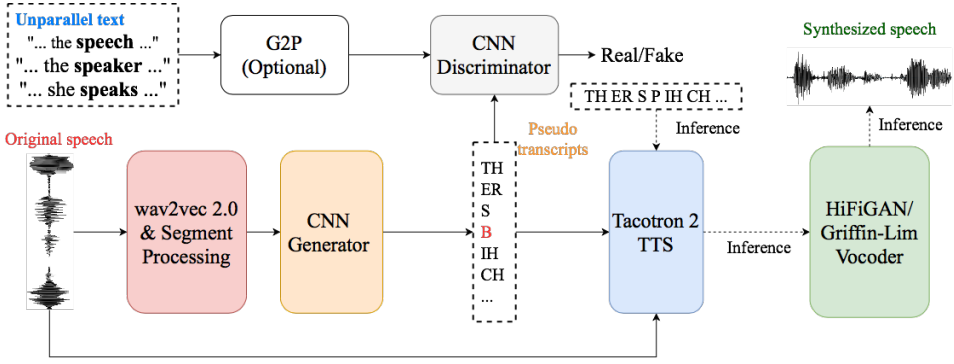
Text preprocessing: optionally apply **grapheme-to-phoneme (G2P)** converter on the character sequence

Proposed model for unsupervised TTS



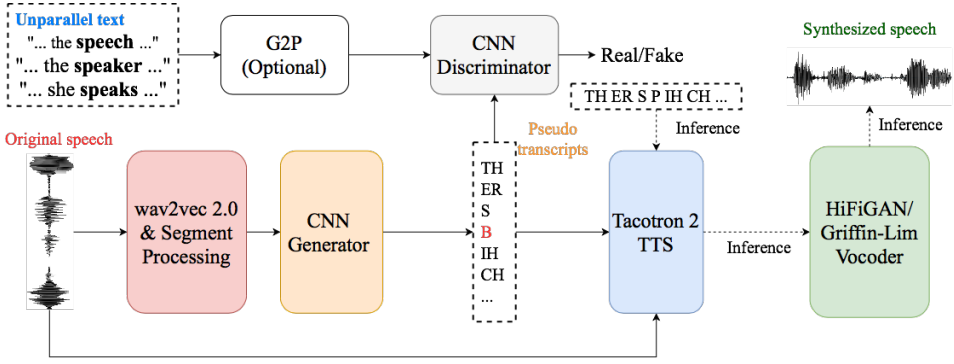
- **wav2vec 2.0 and segment processing:** wav2vec 2.0 trained with LibriLight + PCA + **average** over consecutive segments assigned to the same **K-means clusters**

Proposed model for unsupervised TTS



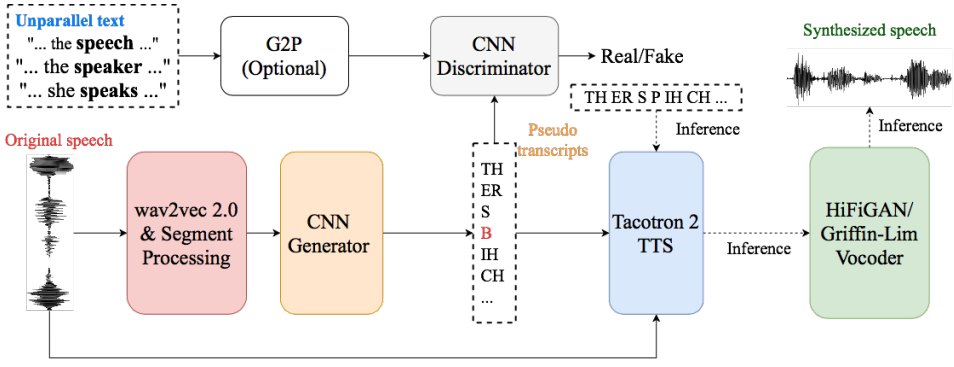
- **CNN generator:** 1-layer CNN that outputs a sequence of distributions over text units where consecutive segments with the same arg max value are **collapsed**

Proposed model for unsupervised TTS



- **CNN discriminator:** 4-layer CNN that tries to tell which source (real or generated) the input sequence is from against the generator

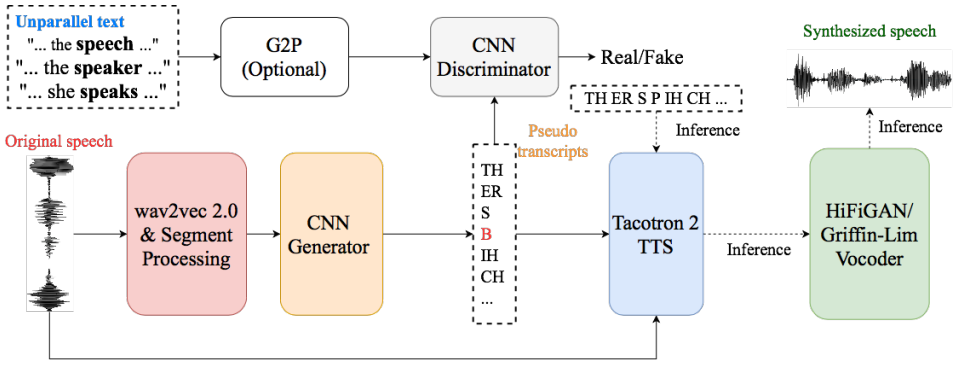
Proposed model for unsupervised TTS



- **Step 2**: Learn a **supervised** TTS (Tacotron 2[5]) to generate speech from **pseudo-transcripts**, $\tilde{x}_i = G(\tilde{y}_i), i = 1, \cdots, n$

[5] Jonathan Shen et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". In: *ICASSP*. 2018

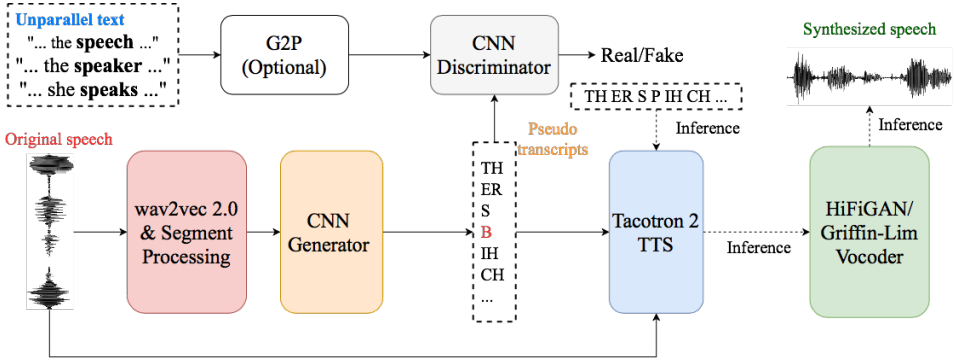
Proposed model for unsupervised TTS



- **Tacotron 2 TTS**: outputs **mel spectrograms**; follows the original Tacotron 2 model with additional guided attention loss [6] to ensure that the attention matrix close to diagonal

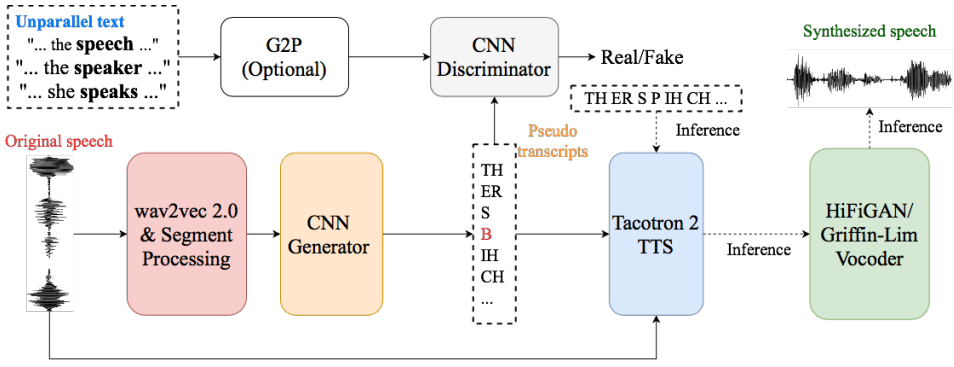
[6] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention". In: *ICASSP*. 2018, pp. 4784–4788

Proposed model for unsupervised TTS



- **Inference**: use **ground truth transcripts** as inputs

Proposed model for unsupervised TTS



- **Vocoders**: convert mel spectrogram into speech waveform; **HiFiGAN**[7] or **Griffin-Lim** vocoders both implemented in ESPnet

[7] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis". In: *Neural Information Processing Systems*. 2020

Outline

Experiment 1: Unsupervised TTS on English

- **Speech dataset**: 24-hour single-speaker LJSpeech corpus
- **Text dataset**: transcripts from the LibriSpeech corpus (unpaired with the speech, distribution mismatch)
- **Train-test split**: 300 utterances for validation and 500 utterances for testing; convert to phonemes using G2P [8]
- **wav2vec-U training**:
  1. **Grid search** the best weights for the auxiliary losses of the wav2vec-U system, i.e., code penalty, gradient penalty, and smoothness weight; 150k steps with a batch size of 160;
  2. **Self-training (ST) with a triphone HMM** using **Framewise** wav2vec 2.0+PCA features as input and pseudo phone sequences transcribed by the wav2vec-U generator as targets
  3. **Further ST with wav2vec 2.0 model** using the pseudo character targets obtained from the above step, and the Connectionist Temporal Classification (CTC) loss
- **TTS training**: trained for 80 epochs; HiFiGAN vocoder
- **Evaluation**: **character error rate (CER, lower is better)** and **word error rate (WER, lower is better)** by feeding the synthesized speech to a supervised ASR

[8] Kyubyong Park and Jongseok Kim. *g2pE*. https://github.com/Kyubyong/g2p, 2019

Results

Table: Unsupervised TTS results on the LJSpeech dataset

Language	Unsup ASR (PER)		Unsup TTS		Supervised TTS	
	No ST	ST	CER	WER	CER	WER
English	12.37	3.59	4.56	11.95	3.93	10.76

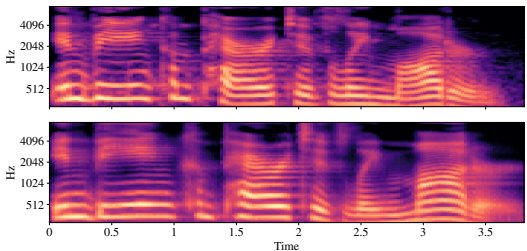


Figure: Mel-spectrograms for ground truth (upper) and synthetic speech by the unsupervised TTS model (lower) for the English sentence "in being comparatively modern."

- wav2vec-U sensitive to hyperparameters
- ST reduces the phone error rate on the test set by 70% relative
- UnsupTTS performed comparably with the supervised TTS

Outline

Experiment 2: Unsupervised TTS on six languages from CSS10

- **Speech dataset**: Japanese, Hungarian, Spanish, Finnish, German and Dutch from the CSS10 dataset, each with 15hr speech
- **Text dataset**: from the same CSS10 dataset with their paired relationship broken up
- **Data split**: 99 to 1, which gave about 50 to 100 utterances validation
- **wav2vec-U training**: the same English wav2vec 2.0 Large model to extract speech representations and the same training pipeline but with only **one ST stage**; experiment with **both** characters and phonemes
- **TTS training**: 80 epochs similar to the English system; results obtained using Griffin-Lim vocoder by default

Overall Results

Table: Unsupervised TTS results on the CSS10 dataset using English wav2vec 2.0 pretrained features

Language	Unsup TTS		Supervised TTS	
	CER	WER	CER	WER
Japanese	17.98	47.81	17.87	36.23
Hungarian	27.78	76.82	18.05	63.14
Spanish	23.03	55.52	18.19	36.74
Finnish	36.05	84.46	22.84	58.67
German	17.25	56.78	11.28	40.94
Dutch	53.01	89.41	34.53	76.71

- Self-training step still greatly reduces the error rates by 25% to 40% relative to all the languages
- The CERs of UnsupTTS **within 5% absolute** to the supervised TTS in all languages; Much larger gap for WER
- In the case of German, the TTS trained with pseudo transcripts achieves a **lower** CER compared to the unsupervised ASR system, suggesting the existence of internal mechanism by TTS to correct the noise in the pseudo-transcripts

Characters vs phonemes

Table: Effect of different text units on unsupervised TTS using Griffin-Lim vocoder

Language	Phoneme		Grapheme	
	CER	WER	CER	WER
Hungarian	**22.73**	**68.80**	27.78	76.82
Finnish	**27.58**	**67.87**	36.05	84.46
Dutch	**22.04**	**56.85**	53.01	89.41

- Use LanguageNet G2P*
- Both phoneme and character-based wav2vec-U can be unstable to train
- Phoneme-based system, when converged, achieves **lower** CER and WER than character-based system

*Mark Hasegawa-Johnson et al. "Grapheme-to-Phoneme Transduction for Cross-Language ASR". In: *SLSP*. 2020, pp. 3–19

Griffin-Lim vs HiFiGAN vocoders

Table: The effect of different pretrained vocoders (Griffin-Lim, HiFiGAN) on unsupervised TTS results for LJSpeech and various languages from CSS10

Language	Griffin-Lim		HiFiGAN	
	CER	WER	CER	WER
English	5.02	12.83	**4.56**	**11.95**
Japanese	**17.98**	**47.81**	20.58	54.09
Hungarian	27.78	76.82	**26.92**	**76.60**
Spanish	**23.03**	**55.52**	29.41	68.82
Finnish	**36.05**	**84.46**	37.66	87.48
German	**17.25**	**56.78**	18.45	59.90

- Griffin-Lim vocoder yields lower error rates than HiFiGAN on **4 out of 5** CSS10 languages
- HiFiGAN yields lower error rate in English and produces **more natural speech**, but tend to **skip phonemes** on unseen languages

Demo

TTS demo

Conclusion

- Propose UnsupTTS, an effective approach for unsupervised TTS
- Future direction: make wav2vec-U more stable; consider unmatched setting for multilingual data