Simple and Effective Unsupervised Speech Synthesis

Alexander H. Liu[1]* Cheng-I Jeff Lai[1]* Wei-Ning Hsu[2] Michael Auli[2] Alexei Baevski[2] James Glass[1]

[1] MIT CSAIL [2] Meta AI (* equal contribution)

Introduction

Background

- Supervised methods have achieved great success in text-to-speech (TTS) synthesis [1].
- However, the success relied heavily on the amount and the quality of labeled training data [2].

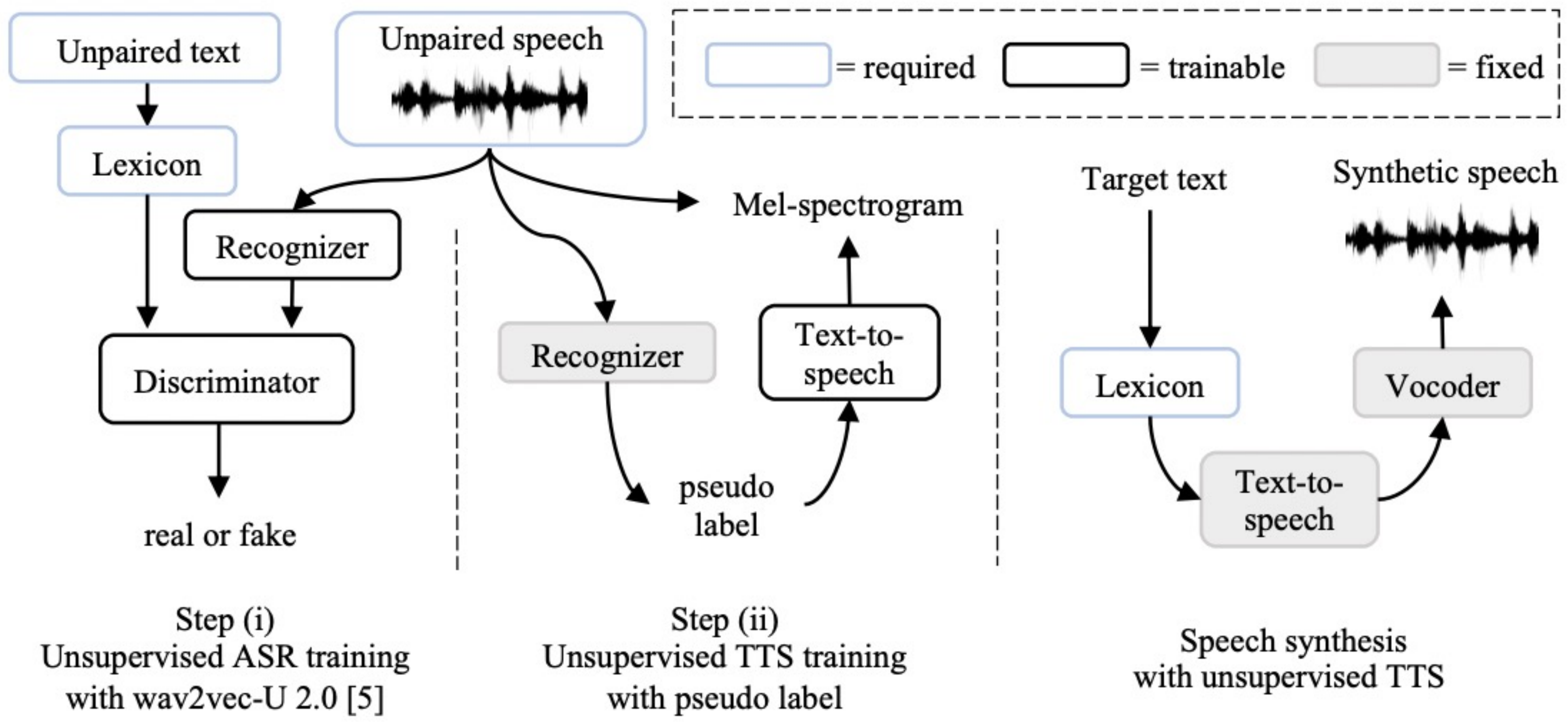
Our goal

Speech synthesis without labeled training data.

Key idea

Leveraging recent advance in unsupervised ASR to create pseudo-labeled training data.

Method



Step 1. Pre-train an unsupervised ASR with lexicon & unpaired speech and text [3].

Step 2. Train TTS model with ASR-labeled data that contains recognition error.

Inference: synthesize speech with phone sequence from target text.

Details

- ASR model: 2-layer CNN with input wav2vec2.0 feature.
- TTS model: 12-layer sequence-to-sequence Transformer.
- Training objective: L2 reconstruction error on Mel-spectrogram + guided attention loss [4].
- Lexicons defined by linguistics are widely used in existing TTS systems.
- Vocoder (spectrogram-to-waveform module) is trained with speech only, also widely used in TTS systems.

Results

Setup Experiment conducted on LJSpeech[5], a benchmark dataset with about 24 hours of read English speech from single female speaker. Text transcription is not used for unsupervised model. (Multi-speaker result on LibriTTS available in paper.)

Highlights

- The quality of synthetic speech from unsupervised model matches supervised method in human evaluation.
- Unsupervised model have slightly worse intelligibility when measured by machines.
- The TTS performs better with raw text input despite learning from imperfect pseudo-labeled data.

Preference Test (human evaluation)

	Preference over Supervised	
	Naturalness	Intelligibility
Unsupervised	50.2%	54.0%

Demo page (samples & code)

Mean Opinion Score (human evaluation)

Method	Input phone error rate suffered during training	MOS
Natural	-	4.05 ± 0.07
Supervised	0%	3.94 ± 0.08
Unsupervised	6.97%	3.91 ± 0.08

Intelligibility Test (commercial ASR evaluation)

Method	Source of input phone sequence for synthesize	Word error rate (%)
Natural	-	18.0
Supervised	text†	19.2
Unsupervised	text†	21.7
	ASR transcription‡	22.0

Conclusion

Key contribution

First unsupervised TTS: with simple and effective method, we show that training TTS without human-labeled data is feasible.

Future directions

- End-to-end training
- Generalize to low-resource languages where unsupervised methods are preferred.

Reference

[1] Natural TTS synthesis by conditioning Wavenet on Mel Spectrogram predictions, *Shen et al., 2021*
[2] Semi-supervised training for improving data efficiency in end-to-end speech synthesis, *Chung et al., 2017*
[3] Towards end-to-end unsupervised speech recognition, *Liu et al., 2022*
[4] Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention, *Tachibana et al., 2018*
[5] The LJ speech dataset, *Ito et al., 2017*

Acknowledgement We thank Tomoki Hayashi and Erica Cooper for their advice on TTS training and evaluation. This research was supported in part by the MIT-IBM Watson AI Lab.