Towards Unsupervised Speech Synthesis

Alexander H. Liu*, **Cheng-I Jeff Lai***, **James Glass***
MIT CSAIL
{alexhliu,clai24,glass}@mit.edu

Abstract

In this paper, we introduce the first unsupervised speech synthesis system that can be built with a simple recipe. The framework is based on a recently developed unsupervised speech recognition system and an existing neural-based speech synthesis paradigm. With unpaired audio, unpaired text, and lexicon, our method enables speech synthesis without the need for human-labeled corpus. Our preliminary result shows the unsupervised model achieved similar performance to its supervised counterpart in human opinion score.

1 Introduction

With the recent advance of deep learning, neural-based text-to-speech (TTS) systems have closed the gap between real and synthetic speech in terms of both intelligibility and naturalness (Shen et al., 2018). However, a sizable dataset composed by speech-text pairs is necessary in order to synthesize high-quality speech (Chung et al., 2019). Consequentially, commercialized speech synthesis systems are not available for the vast majority of languages (Tan et al., 2021).

In this preliminary study, we make the first attempt (to the best of our knowledge) to achieve unsupervised speech synthesis with the goal of addressing the aforementioned limitation of speech synthesis. We simulate an extreme situation where human annotated speech is unavailable by considering only unpaired audio, unpaired text, and a grapheme-to-phoneme lexicon. We proposed a simple two-step recipe to build TTS system under such condition: first, utilizing automatic speech recognition (ASR) models to provide pseudo label for untranscribed speech; second, training TTS models with machine-annotated speech only.

For the first step, we take advantage of the recent developments in unsupervised speech recognition (Baevski et al., 2021). We train wav2vec-U 2.0 (Anonymous, 2022), an ASR model that does not require speech data to be paired with text, to obtain pseudo speech-to-text annotation. To build speech synthesis systems as the second step, we follow the same learning paradigm of existing supervised TTS models (Wang et al., 2017) but use machine-annotated speech instead of human-annotated data.

In our preliminary experiment, we demonstrate the effectiveness of the simple method on clean, high-quality, single speaker dataset. We show that synthesizing natural speech is possible without the need of human-labeled data by performing subjective test. In-progress future works includes more comprehensive evaluation and more difficult setup on multi-speaker speech synthesis.

2 Background

2.1 Supervised Speech Synthesis

While there are different methods for speech synthesis, we focused on neural network-based TTS systems where text-to-speech transformation is modeled by deep neural networks. For example, sequence-to-sequence encoder-decoder architectures that are commonly used to for TTS can be based on recurrent neural networks (Wang et al., 2017), convolution neural networks (Tachibana et al., 2018), or transformers (Li et al., 2019). Under the sequence-to-sequence encoder-decoder paradigm, input text is first converted into phone sequence with the aid of lexicon then encoded into latent features with the encoder. The decoder then synthesizes mel spectrogram based on the encoder feature. The whole model can be trained by minimizing the reconstruction error between the output and target mel spectrogram.

Besides text-to-mel-spectrogram models described above, neural vocoders also played an important role in neural-based TTS systems (Ya-

* Equal contribution

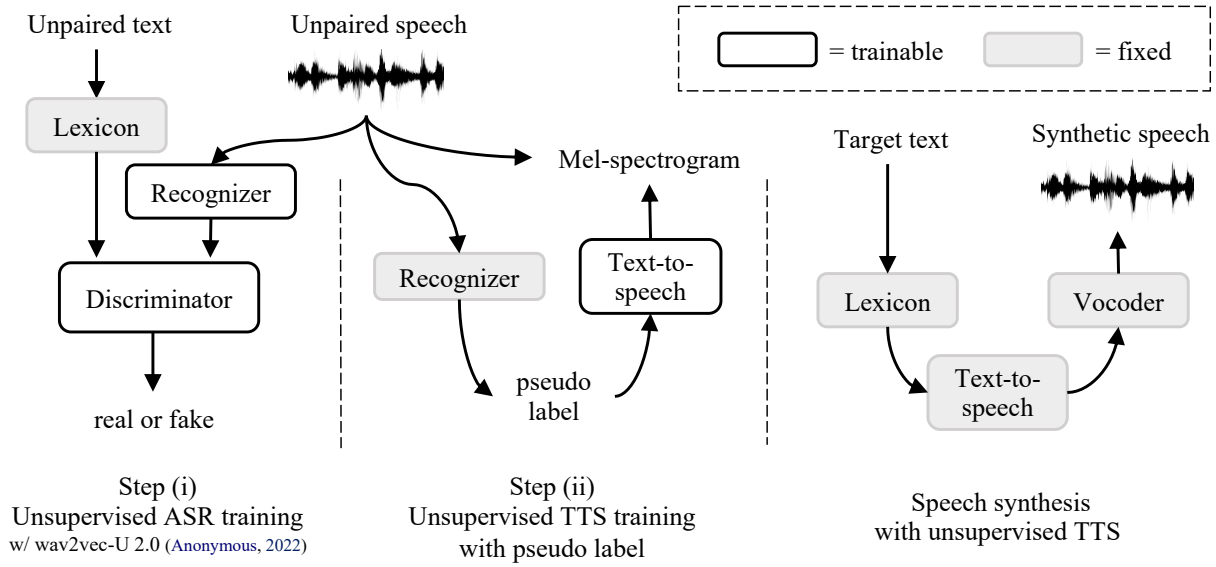

Figure 1: An overview of the proposed framework for unsupervised speech synthesis. Only unpaired text, unpaired audio, and lexicon are required. Step (i): build unsupervised ASR with existing method wav2vec-U 2.0 (Anonymous, 2022). Step (ii): train unsupervised TTS with unpaired speech and pseudo label provided by the ASR from previous step. Step (iii): Synthesize speech from text with lexicon, TTS model, and vocoder (trained with unpaired speech).

mamoto et al., 2020). Vocoders aim to generate waveform from synthetic mel spectorgram by learning from <waveform, mel spectrogram> pairs collected from audio data. Combining text-to-mel-spectorgram TTS models and neural vocoders, high-quality speech that is indistinguishable from real data can be synthesized (Van Den Oord et al., 2016; Shen et al., 2018). Nevertheless, considerable amount of human annotated speech is required behind the success of neural-based TTS systems.

2.2 Semi-supervised Speech Synthesis

To improve the data efficiency for neural-based TTS systems, methods utilizing unpaired training data have thrived. Unlike paired <audio, text> data, collecting unpaired text or unpaired audio is relatively easy and cheap. Prior works have found that semi-supervised learning using unpaired data jointly with paired data can benefit TTS systems and reduce the need of paired data in different ways. For example, pre-training encoder/decoder (Chung et al., 2019), improving input text representation (Wang et al., 2015; Jia et al., 2021), and data augmentation using the opposite natural of ASR and TTS (Tjandra et al., 2017; Ren et al., 2019; Liu et al., 2020).

However, existing semi-supervised methods are still bounded by the amount of paired training data (Chung et al., 2019; Ren et al., 2019; Liu et al., 2020). In this work, we seek to push the limit of neural-based TTS system to the extreme where *no* paired training data can be used.

2.3 Unsupervised Speech Recognition

Since the key to achieving unsupervised speech synthesis with the proposed method is to generate transcription of speech without human supervision, we first review prior works on unsupervised speech recognition. Wav2vec-U (Baevski et al., 2021) is the first existing method to achieve unsupervised speech recognition based on self-supervised speech representation of audio (Baevski et al., 2020). From a high-level prospective, the method can be viewed as a two-stage process: 1) a pre-processing stage performing series of feature engineering over the input audio representation to derive latent feature that is closely related to the underlying phone sequence; 2) the adversarial training stage to obtain speech recognizer which maps the pre-processed feature to phone units.

Following the prior work, wav2vec-U 2.0 (Anonymous, 2022) showed that the recognizer can be trained directly on audio representation without the pre-processing stage. Similar to the adversarial training in the prior work, the goal of the recognizer is to transcribe speech representation into phone sequence that is indistinguishable from the real phone sequence as illustrated in Fig. 1(i).While both methods showed that unsupervised ASR systems can perform

on par with supervised systems on benchmark recognition dataset[1], we select wav2vec-U 2.0 as the unsupervised ASR module in this work for simplicity.

3 Unsupervised TTS

3.1 Problem formulation

As the first study toward unsupervised TTS, our goal is to synthesize speech with the following resources:

- A clean read-speech audio corpus containing speech without paired text.
- A text corpus containing sentences where no exact match existed in the spoken corpus. Furthermore, there is no domain mismatch between the text and audio corpora.
- A lexicon providing the mapping between each word in the text corpora and its pronunciation representation, or the phone sequence. This is also known as the Grapheme-to-Phoneme (G2P) conversion in TTS.

We present a recipe for building unsupervised speech synthesis under these constraints. The training procedure is broken into two steps, described in the following subsections and illustrated in Fig. 1.

3.2 Pseudo labeling speech via unsupervised ASR

As illustrated in Fig. 1(i), the first step of the proposed method is to generate pseudo label for each utterance from the spoken corpus. To this end, we first train wav2vec-U 2.0 (Anonymous, 2022), an existing unsupervised speech recognition method described in Section 2.3 and detailed in Section 4.2, on with the unpaired audio and text corpus. After training, we decode the audio corpus with the resulting recognizer to obtain pseudo-labeled phone sequence for each utterance.

3.3 Unsupervised text-to-speech with pseudo label

For the second step, we train a sequence-to-sequence TTS with the pseudo-labeled audio corpus as shown in Fig. 1(ii) and detailed in Section 4.3. The goal of TTS module is to learn to recover mel-spectrogram that contains spoken content specified by the input (imperfect) phone annotation. During testing, mel-spectrogram can be synthesized by feeding the phone sequence representation of the desired sentence. To generate audible speech, a separate vocoder is used to convert mel-spectrograms into waveforms[2].

4 Experimental Setup

4.1 Dataset

Audio LJSpeech (Ito, 2017) is used as the unpaired audio source. It contains 13,100 utterances (≈24 hours) of read speech from a single female speaker. Following prior works on weakly-supervised TTS (Ren et al., 2019; Liu et al., 2020), 300 utterances are randomly selected for both validation and test set, leaving 12500 utterances for training. For the unsupervised ASR stage, the audio is downsampled from 22khz to 16khz to extract speech representation from wav2vec 2.0 (Baevski et al., 2020) to serve as the input. For the TTS training stage, target mel-spectrogram is extracted from the silence-removed audio with 80 mel filter banks.

Text The official text corpus provided by LibriSpeech (Panayotov et al., 2015) is used as the unpaired text source. Transcriptions of utterance from LJspeech is excluded[3] to ensure there is no exact match between text and audio data.

Lexicon Word-to-phones mapping is obtained through off-the-shelf phonemeizer (Park and Kim, 2019). To phonemize the text source, all punctuation marks are discarded and the difference between variants of the same phone is ignored. This results in a phoneme inventory with a size of 39 used for both ASR and TTS module.

4.2 Unsupervised ASR model

Training Wav2vec-U 2.0 (Anonymous, 2022) is used for pseudo-labeling unpaired speech. The recognizer is a 3-layered network taking 1024 dimensional speech representation as input. The input feature is first normalized and rescaled with batch normalization, followed by a linear projection, and a convolution neural network taking 9 frames at a time with a stride of 3 to predict the underlying

[1] Achieved with additional self-training techniques. In this work, the recognizer obtained with adversarial training is used for simplicity.

[2] Vocoder training is inherently unsupervised since no paired text is needed.

[3] https://github.com/flashlight/wav2letter/blob/main/recipes/sota/2019/raw_lm_corpus/README.md

phone. To output the sequence of phone prediction, consecutive predictions sharing the same most likely phone will be merged into a single prediction by random selection. The recognizer is trained with adversarial learning against a discriminator composed by a 2 layered convolution neural network with a receptive field of size 9. The discriminator is trained to distinguish between the recognizer output sequence predictions from the one-hot vector sequences representing phone sequences from the text corpora. Conversely, the recognizer is trained to mislead the discriminator into classifying its output as real phone sequence. Fairseq (Ott et al., 2019) implementation is used for training with default hyper-parameter[4].

Decoding To transcribe the unpaired audio after wav2vec-U 2.0 training, each utterance is decoded by the recognizer together with a phoneme-to-phoneme WFST (Mohri et al., 2002) (including a phone-based 6-gram language model trained from the lexicon-phomenized text corpora). Beam search with beam size of 15 is used for decoding with hyper-parameters selected with unsupervised metric (Baevski et al., 2021). As a reference, decoding the test set with this unsupervised ASR results in 6.97% phone error rate.

4.3 TTS model

Text-to-Speech Transformer-TTS (Li et al., 2019), a sequence-to-sequence encoder-decoder model, is selected as the phone-to-mel-spectrogram model used in our framework. Similar to Tacotron2, Transformer-TTS consists of an encoder, an autoregressive decoder, a pre-net, and a post-net. The encoder and decoder have 6 layers of transformer blocks. The training objective is to minimize L2 reconstruction error between the model output and the targeted mel-spectrogram corresponded to the input text. ESPNet (Hayashi et al., 2020) is used for training the model based on the default configuration[5]. We found that it is necessary to enforce guided attention loss in all the decoder attention heads (48 of them), while only increasing the guided attention weights does not help. In addition, we tripled the number of training epochs from 200 to 600. We hypothesized these modifications are due to the use of a reduced phone sets, which makes the text to mel-spectrogram learning more difficult.

Vocoder To decode the synthetic mel-spectrogram into waveform, we used the publicly available[6] Parallel WaveGAN (Yamamoto et al., 2020) that is trained on the same unpaired audio source for 3M iterations.

5 Preliminary result

To evaluate the proposed unsupervised TTS method, we compare the speech quality of the followings:

- **Unsupervised**: synthetic speech from the unsupervised TTS model.
- **Supervised** : synthetic speech from supervised topline – the same TTS model trained with ground truth transcription instead of machine annotation.
- **Natural**: original speech utterances from the dataset.

Preliminary result is collected through Mean Opinion Score (MOS), a subjective measuring quantifies naturalness, where workers are asked to rate each utterances from a 5-point (with 1-point increment) scale. We ran 150 HITs (crowdsourced tasks) to compare the synthesis quality between `Unsup`, `Sup` and `Nat`. 100 unique utterances from the LJspeech test set are selected randomly and designated as the MOS test set. In each HIT, 10 utterances from the MOS test set are chosen for `Nat` and synthesized for `Unsup` and `Sup`. The order presented to the workers are randomized.

Table 1: Speech naturalness measured by Mean Opinion Score (MOS) with 5-point scale on LJSpeech test split.

Method	MOS
Natural	4.05 ± 0.07
Supervised	3.94 ± 0.08
Unsupervised	3.91 ± 0.08

Results of evaluating the naturalness are listed in Table 1. With upper bound being the 4.05 scored by real data from LJSpeech, we can see the supervised model performed remarkably well by scoring 3.94 with a slightly higher variance. Comparing to the supervised topline, the proposed unsupervised method scored 3.91 with similar variance, degraded

[4] https://github.com/pytorch/fairseq/tree/main/examples/wav2vec/unsupervised

[5] https://github.com/espnet/espnet/tree/master/egs2/

[6] https://github.com/kan-bayashi/ParallelWaveGAN

by merely 0.03. Considering the 6.97% phone error rate the unsupervised model suffered during training, the small degradation suggested that the proposed pipeline method does not suffer much from error propagation. Even comparing against real speech, the MOS score of unsupervised TTS is only lower by 0.86, recovering over 96% of the score. This demonstrated the unexpected robustness of the proposed two-step pipeline method.

6 In-progress future work

In this work, we described a framework that first achieves unsupervised speech synthesis. The framework relied on the recent works in unsupervised speech recognition and the matured neural-based speech synthesis paradigm. As a preliminary study, we showed the proposed TTS system can match the performance of the supervised system in terms of MOS on English dataset without using human annotation. We are currently working on running more evaluations to show the (e.g., measuring intelligibility with ASR, preference test between supervised and unsupervised method). In addition, we would also like to conduct experiments in multi-speaker setup to validate the generalizability of our method.

References

Anonymous. 2022.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. In *Advances in Neural Information Processing Systems*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. of NeurIPS*.

Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan. 2019. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6940–6944. IEEE.

Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, and Xu Tan. 2020. Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7654–7658. IEEE.

Keith Ito. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. 2021. Png bert: augmented bert on phonemes and graphemes for neural tts. *arXiv preprint arXiv:2103.15060*.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713.

Alexander H Liu, Tao Tu, Hung-yi Lee, and Lin-shan Lee. 2020. Towards unsupervised speech recognition and synthesis with quantized speech representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7259–7263. IEEE.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proc. of ICASSP*, pages 5206–5210. IEEE.

Kyubyong Park and Jongseok Kim. 2019. g2pe. https://github.com/Kyubyong/g2p.

Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Almost unsupervised text to speech and automatic speech recognition. In *International Conference on Machine Learning*, pages 5410–5419. PMLR.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788. IEEE.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2017. Listening while speaking: Speech chain by deep learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 301–308. IEEE.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *SSW*, 125:2.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. Word embedding for recurrent neural network based tts synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE.