June 2023

Improving Neural Vocoder Stability Using Artificial Training Data

Lev Finkelstein

Vincent Wan

Rob Clark

Heiga Zen

Improving Neural Vocoder Stability Using Artificial Training Data

ABSTRACT

A text-to-speech (TTS) converter typically comprises a prosodic model that generates acoustic parameters from linguistic features paired with a neural vocoder. With such a configuration, some feature values can be difficult for the neural vocoder to process, resulting in audio artifacts. This disclosure describes techniques to improve neural vocoder performance, e.g., reduce audio artifacts, make the vocoder more robust to unusual acoustic feature variations, generally be more forgiving of errors made by the feature generator, etc. The techniques entail the use of an auxiliary training path that is driven by synthetic training examples generated by CHiVE inference with some random sampling far enough from the mean (zero).

KEYWORDS

- Clockwork hierarchical variational autoencoder (CHiVE)
- Text-to-speech (TTS)
- Neural vocoder
- Vocoder stability
- Training data
- Synthetic example
- Synthetic data
- Prosody
- Linguistic feature
- Audio artifacts

BACKGROUND

A text-to-speech (TTS) converter often comprises a prosodic model that generates acoustic parameters including prosodic parameters from linguistic features (e.g., CHiVE [4]) paired with a neural vocoder. The neural vocoder transfers acoustic features to audio after processing as necessary, e.g., by an autoencoder. Acoustic features can include fundamental frequency (F0), the zeroth cepstrum coefficient (C0), segment durations, energy, etc.

Such a TTS configuration has proved to be stable enough for moderate variations of acoustic features; however, some feature values continue to be difficult for a neural vocoder to process, resulting in audio artifacts. A reason for the appearance of artifacts is that the standard training dataset is based on the recorded audio samples, which generally lack unexpected acoustic feature fluctuations.

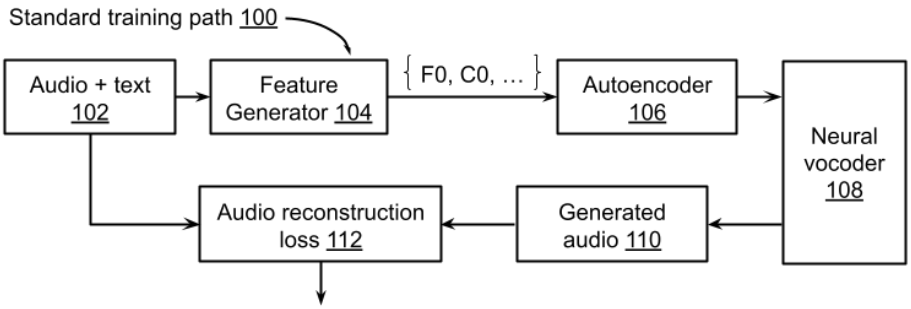

Fig. 1: Typical training path for a neural vocoder

Fig. 1 illustrates a standard training path (100) for a neural vocoder. Standard training examples are generated by passing audio and transcribed text (102) through a feature generator (104) that produces acoustic features such as{F0, C0, … }. After preprocessing (e.g., with an autoencoder 106) these examples are used for training the neural vocoder (108). An end-to-end speech synthesis technique (e.g., Tacotron [6]) or another technique (e.g., PnG NAT [7]) can be used as an alternative to autoencoder (106). Training is driven by the audio reconstruction loss

(112), which is a function of the difference between the audio signal generated by the neural vocoder (110) and the original audio signal.

Data augmentation [5] is a related technique that augments natural training data with synthetic training data.

DESCRIPTION

This disclosure describes techniques to improve neural vocoder performance, e.g., reduce audio artifacts, make the vocoder more robust to unusual acoustic feature variations, generally be more forgiving of errors made by the feature generator, etc. The techniques entail the augmentation of the standard training path with an auxiliary training path that is driven by synthetic training examples (e.g., generated by CHiVE inference or another system such as Tacotron) with some random sampling far enough from the mean (zero).

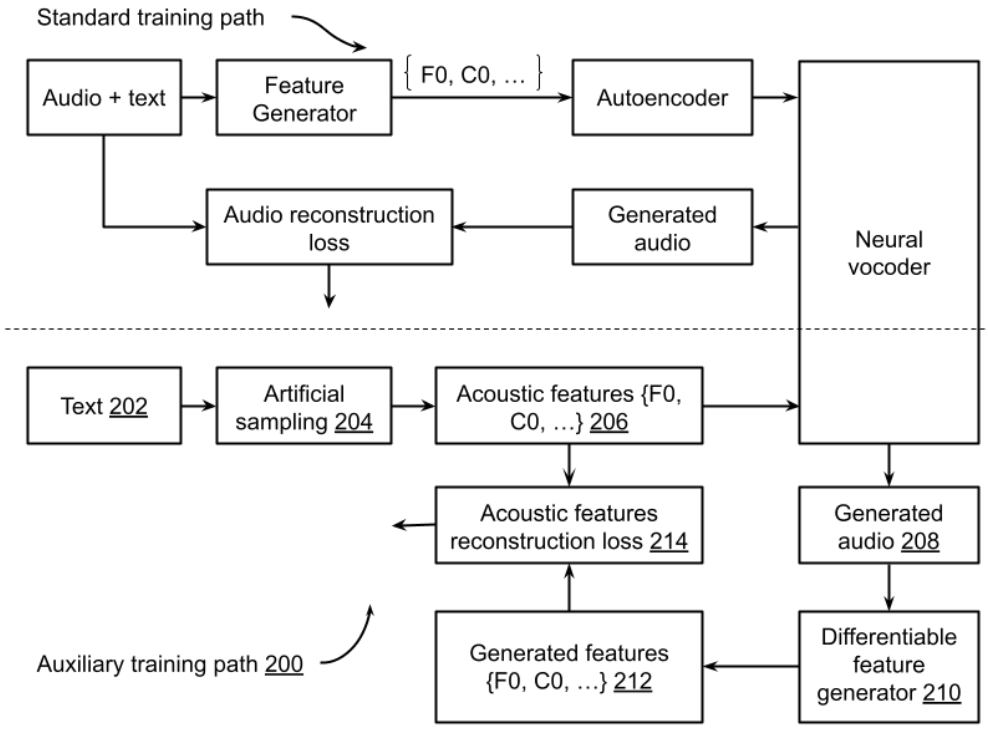

Fig. 2: Auxiliary training path that augments the standard training path of a neural vocoder

Fig. 2 illustrates an auxiliary training path (200), which augments the standard training path of a neural encoder. Unlike the standard training path, which relies on associated pairs of audio and text, the auxiliary training path generates training examples purely out of text (202). In generating examples purely out of text, the auxiliary training path leverages the vastly larger corpus of text as compared to associated pairs of audio and text.

Sampled features (206) are generated by subjecting the text to artificial sampling (204). The sampled features are fed to the neural vocoder (alongside the autoencoder output from the standard training path), which uses them to generate audio (208). A differentiable feature generator (210) recovers acoustic features of the audio. Lacking an audio waveform, the reconstruction loss signal which is a measure of loss in acoustic features (214) of the auxiliary training path is determined by comparing the generated acoustic features (212) of the resulting synthetic audio with the acoustic features (206) obtained from artificial sampling (e.g., using CHiVE-generated data). As with the standard training path, the reconstruction loss signal (214) drives the training of the neural vocoder. However, in contrast to the standard training path, whose reconstruction loss is a function of the difference between audio signals, the reconstruction loss of the auxiliary training path is a function of the difference between acoustic features.

Artificial examples can be built from acoustic features using CHiVE inference with random sampling. For example, for a variational autoencoder (VAE) based system, sampling from a Gaussian distribution may be performed using a large standard deviation. This type of sampling can cover acoustic feature peculiarities that may not be present in a purely audio-based training set.

Synthetically generated acoustic features, when used to train a neural vocoder alongside standard training examples, are not strictly comparable to the standard audio reconstruction loss function, since the synthetically generated acoustic features lack the original audio signal. Therefore, some acoustic features, e.g., F0, C0, etc., are tracked from the audio produced by the neural vocoder and compared with the original acoustic features. This represents a loss function different from the traditional loss function based on audio signals.

The type of the training example - synthetic or natural - plays a role at the level of the cost function. The type is not provided to the neural vocoder, thereby preventing the neural vocoder from developing a behavior or path that is unduly influenced by one or the other type of training example. The neural vocoder trains on natural or synthetic data presented in random order. Advantageously, the described techniques are relatively simple and obviate the collection and tagging of data, since the vocoder is trained using the already trained CHiVE model. The techniques apply to TTS and related products.

In contrast to data augmentation techniques, the synthetic training examples described herein are of a nature different from natural training examples. The synthetic training examples described herein do not have corresponding groundtruth audio. If generated, the synthetic audio may possibly sound like noise to a human. The non-requirement of the intermediate synthetic audio is tied to the non-use of a difference-of-audio loss function in the auxiliary training path.

CONCLUSION

This disclosure describes techniques to improve neural vocoder performance, e.g., reduce audio artifacts, make the vocoder more robust to unusual acoustic feature variations, generally be more forgiving of errors made by the feature generator, etc. The techniques entail the use of an

auxiliary training path that is driven by synthetic training examples generated by CHiVE inference with some random sampling far enough from the mean (zero).

REFERENCES

[1] Huang, Wen-Chin, Yi-Chiao Wu, Hsin-Te Hwang, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, Tomoki Toda, Yu Tsao, and Hsin-Min Wang. "Refined WaveNet vocoder for variational autoencoder based voice conversion." In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1-5. IEEE, 2019.

[2] Zhao, Yi, Shinji Takaki, Hieu-Thi Luong, Junichi Yamagishi, Daisuke Saito, and Nobuaki Minematsu. “Wasserstein GAN and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a WaveNet vocoder.” *IEEE Access* 6 (2018): 60478-60488.

[3] Iyer, Rakesh, and Vincent Wan. “Predicting parametric vocoder parameters from prosodic features.” U.S. Patent Application 17/647,246, filed Jan. 6, 2022.

[4] Vincent Wan; Chun-an Chan; Kenter, Tom; Vit, Jakub; and Clark, Rob. “CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network.” *Proceedings of the Thirty-sixth International Conference on Machine Learning*, pp. 3331-3340. PMLR, 2019.

[5] “Data augmentation.” Available online https://en.wikipedia.org/wiki/Data_augmentation accessed May 10, 2023.

[6] Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang et al. "Tacotron: Towards End-to-End Speech Synthesis." *Proc. Interspeech 2017* (2017): 4006-4010.

[7] Jia Ye, and Julie Cattiau “Recreating Natural Voices for People with Speech Impairments” available online at https://ai.googleblog.com/2021/08/recreating-natural-voices-for-people.html accessed June 14, 2023.