Speech Time-Scale Modification with GANs

Blog & Code

Eyal Cohen[1], Felix Kreuk[2], Joseph Keshet[1]

[1]Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Israel
[2]Meta AI Research, FAIR
cohen.eyal@campus.technion.ac.il

TL;DR Time Scale Modification (TSM) means speeding up or slowing down a sound without affecting the frequency content, such as the perceived pitch of any tonal component. In this work, we propose a novel unsupervised learning algorithm for TSM of speech called **ScalerGAN**.

Goal

Given a speech utterance, our goal is to **speed up** or **slow down** the speech by a **given rate** $r \in \mathbb{R}$ while keeping the **intelligibility and speaker identity** as much as possible.

Previous work

- Previous works used advanced signal processing techniques such as Time-domain overlap-add [1] and Spectral-domain overlap-add [2], [3].
- All those methods **assume quasi-stationarity** of the input speech; Hence they suffer from perceivable artifacts in the generated waveforms.
- None of them use machine learning.

Our approach

- Generate synthetic speech that fills in the missing speech and maintains the speaker's voice.
- Design a machine learning algorithm that can generate different scaled speech despite not having supervision of matching genuine speech utterances with varying speaking rates.

References

[1] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993.*

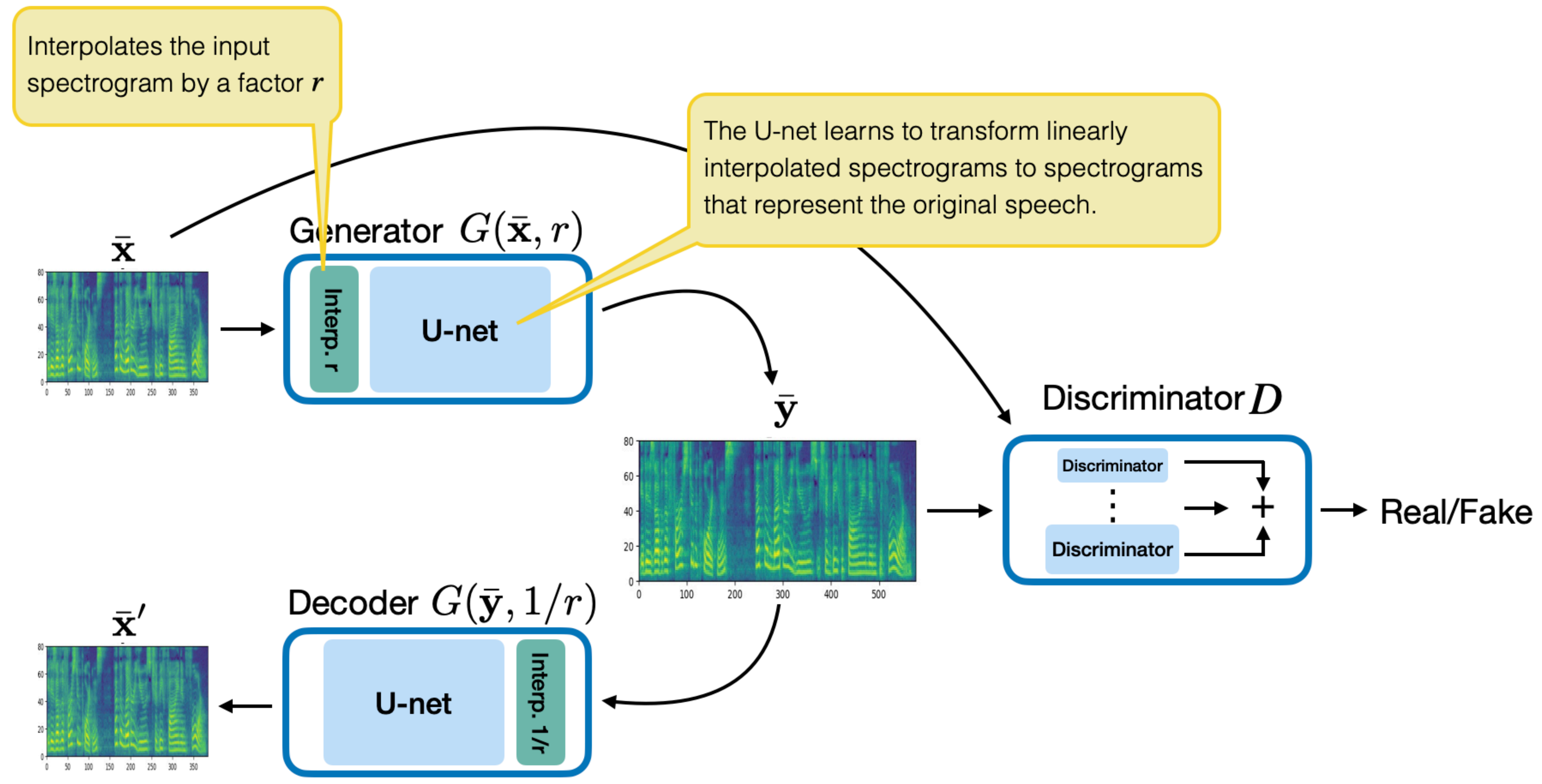
[2] J. Laroche and M. Dolson, "New phase-vocoder techniques for pitch- shifting, harmonizing and other exotic effects," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 1999.*

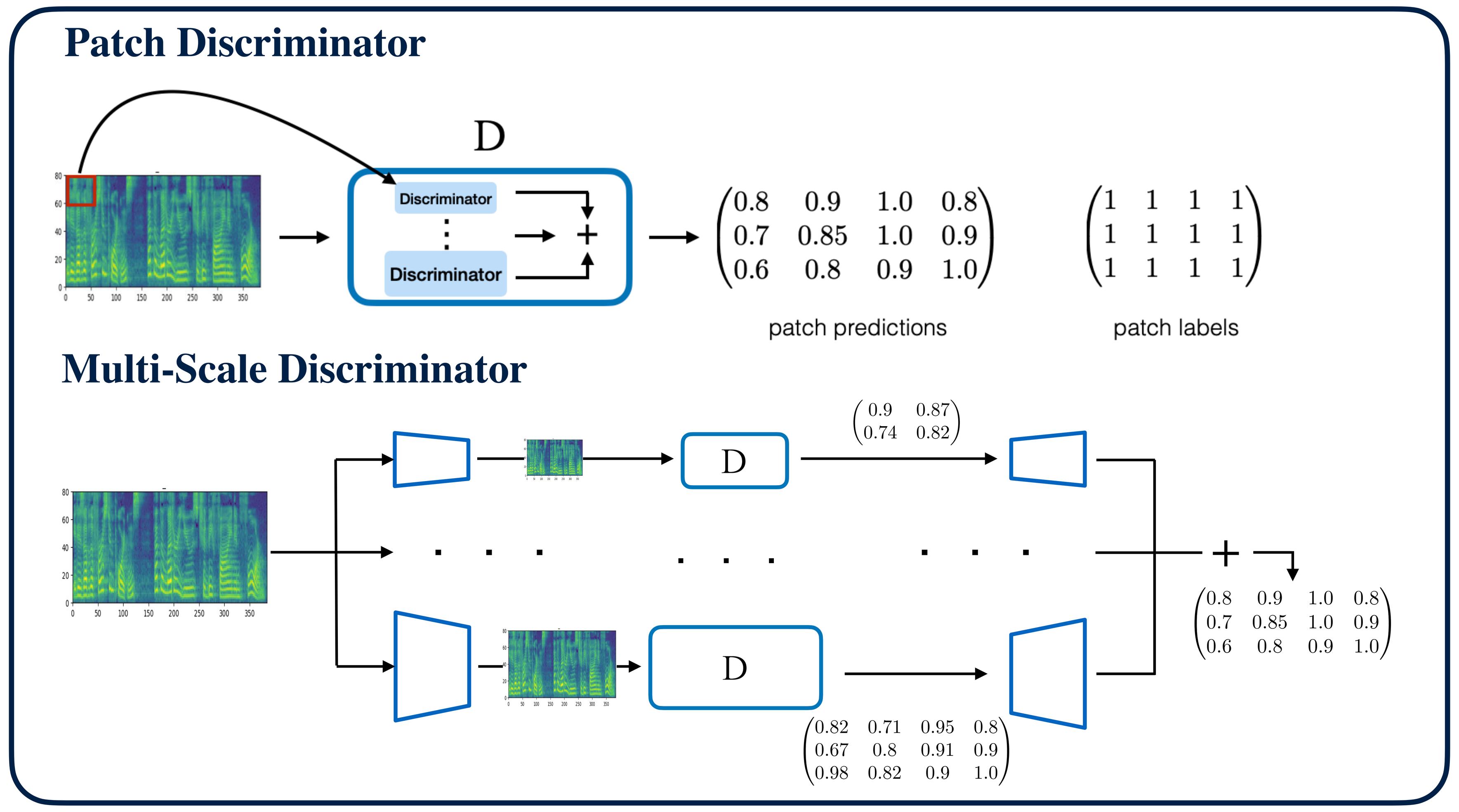
[3] T. Karrer, E. Lee, and J. O. Borchers, "PhaVoRIT: A phase vocoder for real-time interactive time-stretching," in *International Computer Music Conference (ICMC), 2006.*

[4] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/ LJ-Speech-Dataset, 2017.

[5] F. Fang, J. Yamagishi, I. Echizen, M. Sahidullah, and T. Kinnunen, "Transforming acoustic characteristics to deceive playback spoofing countermeasures of speaker verification systems," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

Training



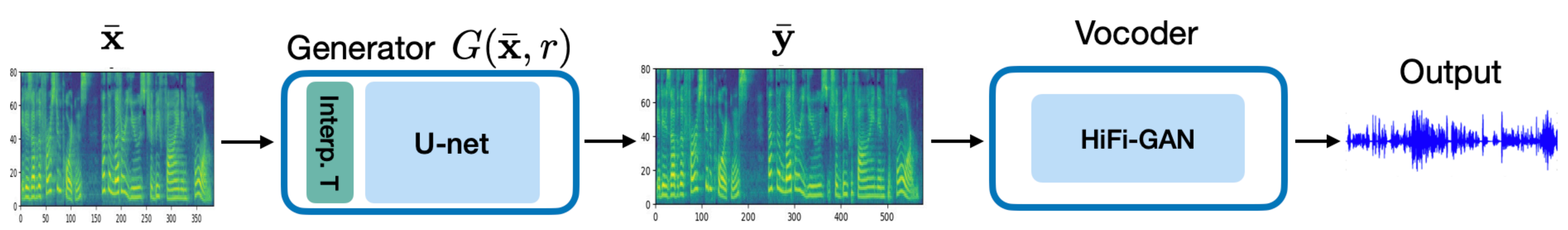


Loss functions:

$$\mathcal{L}_R(G) = \|G(G(\bar{\mathbf{x}},\ r),\ 1/r) - \bar{\mathbf{x}}\|_1$$

$$\mathcal{L}_{LS}(G,D) = \mathbb{E}_{\bar{\mathbf{x}}\sim p(\bar{\mathbf{x}})}\left[(D(\bar{\mathbf{x}}) - J)^2\right] + \mathbb{E}_{\bar{\mathbf{x}}\sim p(\bar{\mathbf{x}})}\left[D(G(\bar{\mathbf{x}}, r))^2\right]$$

Inference



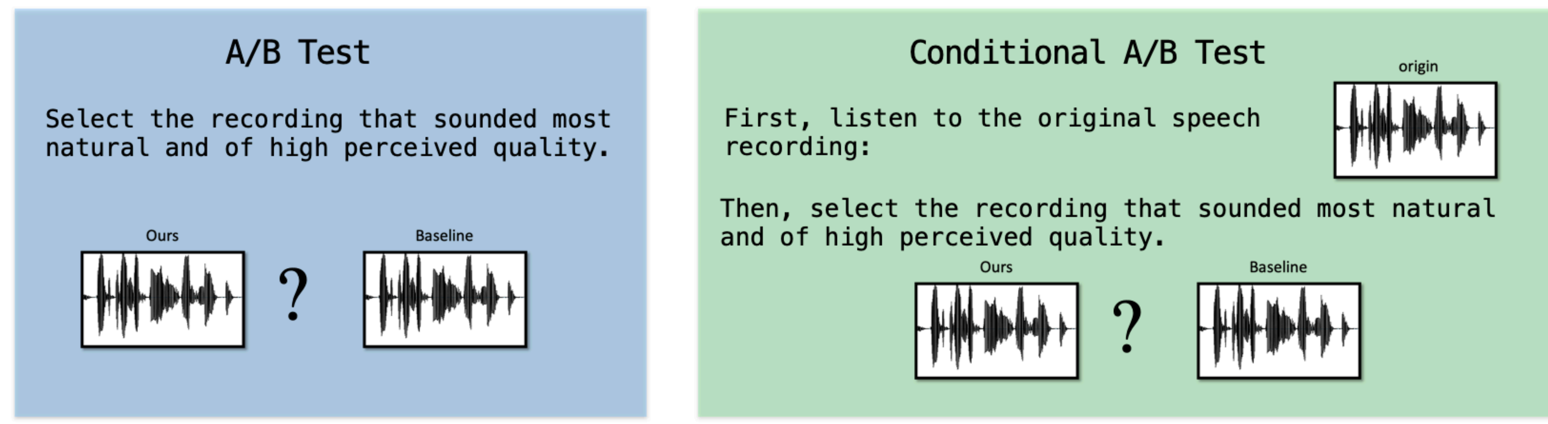
Empirical evaluation

Datasets:

- **LJSpeech** [4] - A dataset consists of 13,100 short audio clips of a single female speaker reading passages from 7 non-fiction books with a total length of approx. 24 hours.
- **DR-VCTK** [5] - A subset of the VCTK dataset: 28 speakers, 14 males and 14 females for training, and 1 male and 1 female for testing.

Evaluation methods:

- Single speaker dataset and multiple speaker dataset.
- Comparison between our method and 11 SOTA methods.
- Amazon MTurk platform with Native American English raters.
- 6 different time-scaled versions for every utterance.



ScalerGAN A/B Test preference rates when compared to other methods and across various rates. Values greater than 50% indicate ScalerGAN **was preferred** over the specified method.