國立中山大學 工學院 聯合專題競賽與展示

資訊工程學系
Department of Computer Science and Engineering

第 17 組：朱劭璿、陳居廷　　指導老師：陳嘉平教授

TSM-Net: 以對抗式時序壓縮自編碼器為基礎的音訊變速演算法
TSM-Net: Temporal Compressing Autoencoder with Adversarial Losses for Time-Scale Modification on Audio Signals

Introduction

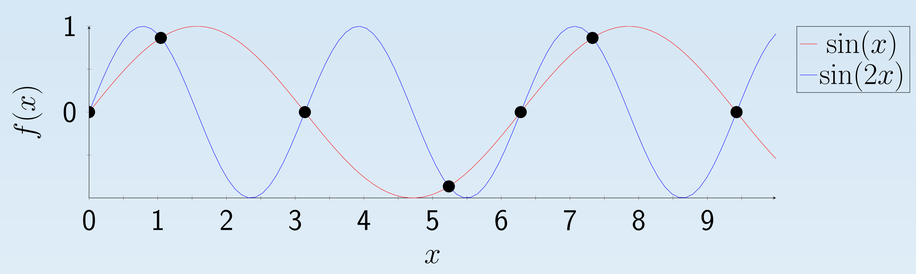
With the advance of technologies, we can manipulate multimedia contents nowadays. An ubiquitous application regarding audio signals called time-scaled modification (TSM) is used in our daily life. It's also known as playback speed control. Interestingly, we haven't discovered any method using (artificial intelligence) AI to refine TSM algorithm and leverage the quality of the synthetic audio.

We proposed a novel TSM approach. While traditional/spectral methods use framing technique/STFT to get high-level units. TSM-Net, our neural-network model encodes the raw audio into a high-level latent representation called Neuralgram. Since the resulting Neuralgram is an image-like data, we apply some existing image resizing techniques and decode it using our neural decoder to obtain the time-scaled audio.

Related Works

Modeling audio is not a trivial task. Models that directly generate raw audio waveform are known as vocoder which can be conditioned on some high-level abstract features. In applications like TTS pipeline, the network often predicts the speech spectrogram of given texts, then uses a vocoder to get the audio. Modern neural-enabled vocoders bring the synthetic quality to a next level.

Decreasing the sampling rate to simplify the dimensionality is another option. However, the Nyquist-Shannon sampling theorem suggests that the low sampling rate would lead to serious aliasing. The figure below shows two signals with different frequency components, which are the aliases for each other in the discrete domain, represented as black dots.

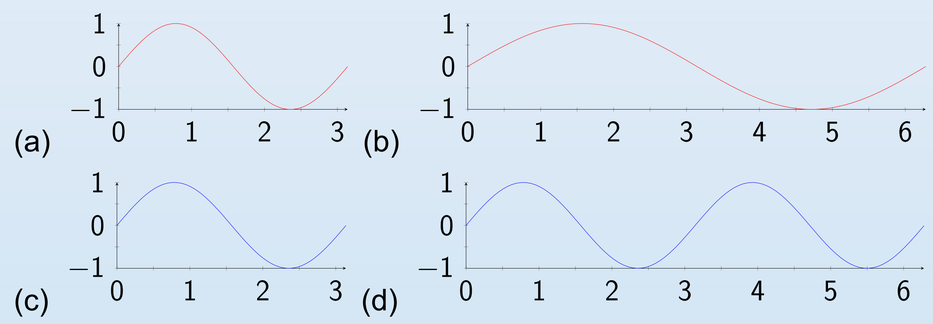


Methodology

Instead of directly scaling on the raw waveform, which leads to the pitch shifting, we encode the raw waveform as a real-valued Neuralgram and scale it. A Neuralgram is applicable on TSM only when the following exists.

- An encoder-decoder pair that is capable of fairly reconstructing the raw waveform.
- A compression ratio that is high enough to put an entire sinusoid of the lowest frequency present into one sample in the Neuralgram

Figure (b) illustrates the result of directly stretching the audio waveform and (d) illustrates the result of stretching the Neuralgram which captures the entire sinusoid.



The encoder-decoder pair mentioned above is a convolution autoencoder trained by GAN architecture, where a discriminator model tries to distinguish the generated audio from the real one. The optimizing loss is defined as below.

$$\min_{D_k} \sum_{k=1}^{3} \mathbb{E}_x \left[\min\left(0, 1 - D_k\left(x\right)\right) + \min\left(0, 1 + D_k\left(A\left(x\right)\right)\right) \right]$$

And the autoencoder tries to fool the discriminator by generating realistic audio. The loss for it is defined as below.

$$\min_{A} \left(\mathbb{E}_x \left[-\sum_{k=1}^{3} D_k\left(A\left(x\right)\right) \right] + \lambda \sum_{k=1}^{3} \mathcal{L}_{FM}\left(A, D_k\right) \right)$$

$$\mathcal{L}_{FM}\left(A, D_k\right) = \mathbb{E}_x \left[\sum_{i=1}^{T} \frac{1}{N_i} \| D_k^{(i)}(x) - D_k^{(i)}(A(s)) \|_1 \right]$$

Experiment

We train our model with Tesla P100 on four Datasets. For the details of the training techniques and sample outputs. Please checkout our paper and visit our demonstration page.

- `https://ernestchu.github.io/tsm-net-paper/tsm-net.pdf`
- `https://ernestchu.github.io/tsm-net-demo/`

Conclusion and future work

The proposed method demonstrated a simple and efficient approach to manipulate the audio. It mitigates the issues found in the traditional TSM. It can also be incorporated with the advancements in other domains such as image interpolation to leverage the audio quality.