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

國立中山大學資訊工程學系

110 學年度大學部專題製作競賽

組員：B073040018 朱劭璿
B072010029 陳居廷

指導教師：陳嘉平教授

Abstract

We proposed a novel approach in the field of time-scale modification on the audio signals. While traditional methods use framing technique and spectral approaches use short-time Fourier transform 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 a two-dimensional image with real values, we apply some existing image resizing techniques on the Neuralgram and decode it using our neural decoder to obtain the time-scaled audio. Our method yields little artifacts and opens a new possibility in the research of modern time-scale modification.

Contents

1 Introduction

With the advance of technologies and digitalization, we can store and reproduce multimedia contents nowadays. We can even manipulate the materials in a way that we couldn't imagine before the digitalization. For example, image resizing and video editing, which changes the dimensionality of the digital pictures spatially and temporally, respectively. Another ubiquitous application regarding audio signals called time-scaled modification (TSM) is used in our daily life. It's also known as playback speed control in the video streaming platform such as YouTube.

With the power of artificial intelligence (AI) and modern computation hardwares, however, we haven't discovered any method using AI to refine TSM algorithm and leverage the quality of the synthetic audio to the next level. Consider we have pragmatic AI tools in similar domains like image super resolution [15] and motion estimation and motion compensation (MEMC) [1], etc.

1.1 Time-scale modification

Time-domain approach The main idea of TSM is that instead of scaling the raw waveforms on the time axis, which leads to pitch shifts due to the changes of wavelengths, we segment the audio into small chunks of fixed length, a.k.a frames or windows to keep the wavelength intact. In order to minimize the boundary breakage after processing, the adjacent frames are overlapped and rearranged to obtain the synthetic audio. As shown in Figure 1. The original distance between the start of each frame is called analysis hopsize. After frame relocation, the distance becomes synthesis hopsize, and the ratio of the analysis hopsize and the synthesis hopsize is the rate at which the audio is speed up or down [3]. In addition, the Hann window [4] is applied to each analysis frame to maintain the amplitude of overlapped areas. The main challenge is the harmonic alignment problem, as shown in Figure 2. With significant periodicities, an unconstrained ratio of the analysis to synthesis hopsize can cause a discrepancy with the original waveform. Specifically, the phases of the same frequency components in the frames do not synchronize properly, which leads to serious interference despite the identical waveforms. Several enhancements have been proposed to solve the synchronization problem [8][18][27]. However, the resultant sound is usually non-natural and contains audible clipping artifacts. This is the negative effect of the framing technique. Moreover, the time-domain TSM only preserve

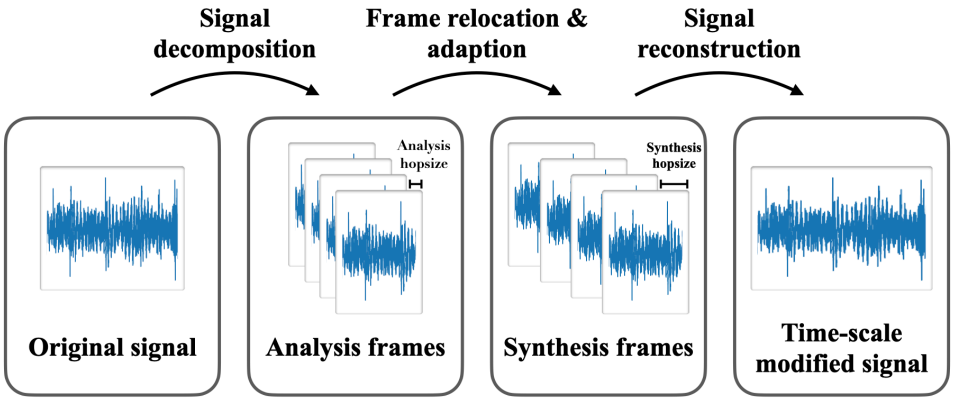

Figure 1: Generic processing pipeline of time-domain time-scale modification (TSM) procedures.

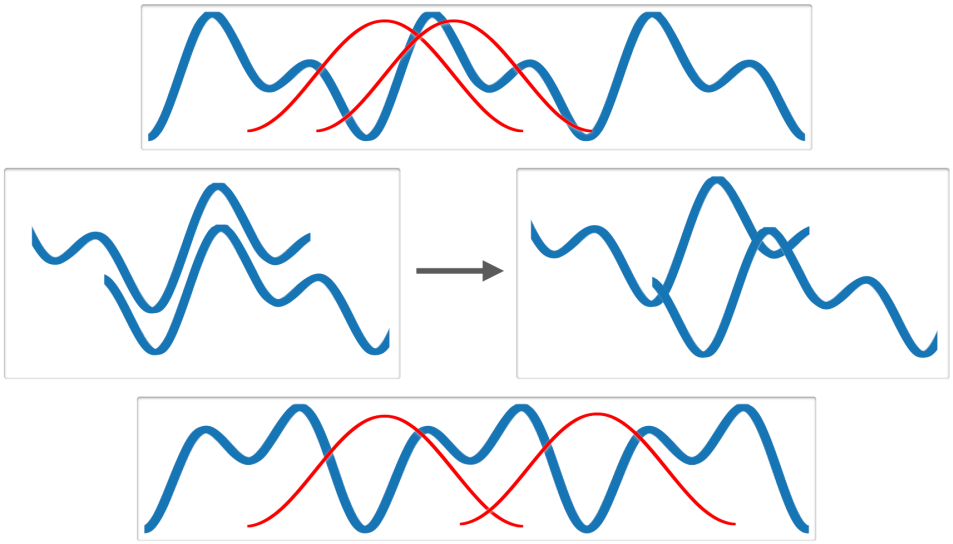

Figure 2: An illustration for harmonic alignment problem. The red Hann windows indicate the rearrangement of the frames. An unconstrained scale ratio would leads to serious interference.

the most prominent periodicity. For the audio with a wide range of frequencies composition, like pop music, symphony and orchestra, the less prominent sound are often erased in the process.

Spectral-domain approach Another approach tries to manipulate the audio in the spectral space, using short-time Fourier transform (STFT) to convert the frequency informations in the raw waveform to a more semantic representation with complex numbers [14]. The magnitude and phase parts can be further derived. Unfortunately, unlike the magnitudes, which gives constructive and straightforward audio features, the phases is relatively complicated and hard to

model. Moreover, due to the heavily correlation between each phase bins, we have to use phase vocoder [5] to estimate phases and the instantaneous frequencies after carefully relocate STFT bins to avoid a peculiar artifacts, a.k.a. phasiness. Despite some refined methods [11][16][19], which improve both the vertical and horizontal phase coherence, the spectral representation is essentially not directly scalable, and the iterative phase propagation process in the phase vocoder is an inevitable overhead.

1.2 Harnessing the power of neural networks

As we've mentioned above, the task requires a highly temporally compressed representation of the audio signals. The neural networks naturally come into our minds. The neural networks are composed of a sequence of linear transformation joined by nonlinear activation functions. With numerous configurable parameters, also known as weights, they are capable of approximating almost any function we desire, including the compression function that transform the raw audio waveforms into low-dimensional latent vectors and back. The parameters are initially random noises and can be gradually inferred using a technique called gradient descent, in which we specify a meaningful loss function such as the difference between the networks' output and the target value, then the parameters can be configured based on the gradients of the loss function so the output would slowly move toward the target value.

TSM-Net To transform back and forth in two domains, we can think of the neural networks as an encoder and a decoder, in the jargon of machine learning, this kind of architecture is called autoencoder [12], which encodes the data into a high-level (and typically low-dimensional) latent vectors and tries its best to decode it back to the original data. In our neural network model, the dimension of the latent vectors is 1024 times smaller than the original one, which means one sample in the latent vector can represent more than an entire wave in the raw audio waveform. Since the latent vector is an image-like multi-dimension vector, we can apply the existing image resizing techniques to scale it. Finally, we decode the resized latent vector to obtain the time-scaled audio waveform.

Distribution modeling In order to make the model generalizes on the unseen data, we have to model the data distribution instead of directly measuring the L2 distance on the existing

dataset, which leads to blurry output upon unseen data. Therefore, we employ a discriminative neural network to help us train the autoencoder [6]. The discriminative network computes the adversarial loss, which measures how well it distinguishes the real and generated data. We will go into detail about the training objective and its effectiveness later.

2 Related Works

Modeling audio is not a trivial task for neural networks. To illustrate it, we can take image generation for example. DCGAN [22] is a generative neural network which synthesizes realistic images with dimension of $3 \times 64 \times 64$. There are 12288 pixels that the model needs to estimate. A 5 seconds stereo audio clips with sampling rate of 22050 Hz has $2 \times 5 \times 22050 = 220500$ samples. Not to mention each pixel is stored in 8 bits while each audio sample is stored in 16 bits and has 256 times possible values than a image pixel. 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.

2.1 Neural vocoder

Models that directly generate raw audio waveform are known as vocoder. A vocoder can be conditioned on some high-level abstract features such as linguistic features or spectrograms. The spectrogram is the magnitude part from the output of STFT. It expresses the frequency composition clearly and being easy to model thanks to its smooth variations over time. As mentioned above, the phases are relative hard to estimate, therefore in applications like text-to-speech (TTS) [25] pipeline, the network often predicts the speech spectrogram of given texts, then uses a vocoder to get the raw audio. The early vocoders include Grinffin-Lim [7], WORLD [17], etc. The modern neural-based vocoders start with WaveNet [26], which predicts the distribution for each audio sample conditioned on all previous ones. However, the autoregressive model runs too slow to apply on real-time applications. FloWaveNet [10] and WaveGlow [21] are neural vocoders based on bipartite transforms. They present a faster inference speed and high quality synthetic audio but require larger models and more parameters to be as expressive as the autoregressive models, and thus harder to trained. WaveGAN [2], MelGAN [13] and VocGAN [28] employ the generative adversarial network [6] training architecture in which a

discriminator is used to measure the divergence of synthetic audio and the real audio and try to learn the generator optimal weights to make the synthetic audio as realistic as possible. The discriminator usually works in multiple scales in order to take care of different frequency bands in the audio data. This kind of approach allows small models to generate high fidelity audio samples.

2.2 Nyquist-Shannon sampling theorem

The sampling rate is the number of samples per seconds that a signal is recorded, and the bit depth is the amplitude precision in which each sample is stored. The analog continuous signal has to be discretized and quantized in order to be stored digitally in the computer. Specifically, discretization and quantization transfer continuous signals into discrete counterparts along the time and amplitude axis, repectively. Common combinations include 44100Hz/24bit, 22050Hz/16bit, 16000Hz/16bit, etc. A lower sampling rate results in lower data dimension, which is beneficial for reducing modeling complexity. However, a lower sampling rate is easier to produce aliasing. The aliasing is a phenomenon where two different signals alias one another in the discrete domain as illustrated in Figure 3. As pointed out and proven by Nyquist [20] and Shannon [23], a signal containing no frequency higher or equal than ωHz, is completely determined by sampling at 2ωHz, i.e., to resolve all frequencies in a signal, it has to be sampled at strictly greater than twice the highest frequency present. Otherwise, the discrete representation cannot determine some high-frequency components which have low-frequency aliases. The aliasing sometimes happens in the real world where some human sensory organs have preceptive limitations. For examples, human eyes can only see about 30 to 60 frames per seconds. For cars driven in a very high speed, the wheels spin too fast to be correctly perceived by human eyes, as if they're spinning backwards. This is known as the wagon-wheel effect. Another example is the moiré pattern. It can be usually observed on the photographs of digital monitors. The resolution of the camera is not high enough to capture the pixels on the monitors and leads to artificial patterns.

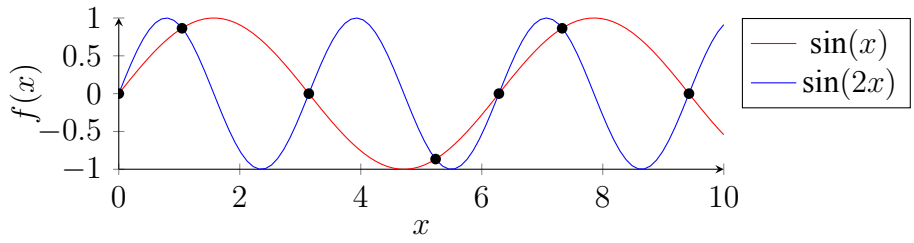

Figure 3: An illustration for aliasing. Suppose the signals are sampled non-uniformly for better illustration, two signals with different frequency components are the aliases for each other in the discrete domain, represented as black dots.

3 Methodology

3.1 Latent representation

We proposed a new representation for audio called Neuralgram to provide a novel approach to TSM. The Neuralgram is a temporally compressed feature map extracted from the middle of an neural autoencoder. A Neuralgram is applicable on TSM only when the following exists.

1. An encoder-decoder pair that is capable of fairly reconstructing the raw waveform.

2. A compression ratio that is high enough to put an entire wave of the lowest frequency present into one sample in the Neuralgram

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 the Neuralgram. Finally, we decode the scaled Neuralgram to get time-scaled audio without pitch shifting, as illustated in Figure 4. Because one sample in neurlgram encodes more than an entire sinusoid for each frequency components, resizing the Neuralgram using Bilinear [24] or Bicubic [9] interpolation can repeat the entire sinusoids in the reconstructed waveform rather than changing their wavelengths and frequencies.

Neuralgram vs. Spectrogram In the literature, most of the works related to neural vocoder use the spectrogram family as the modeling condition. Why do we bother to opt for a new representation? These traditional representations encode different frequency informations into the same amount of samples in the latent space. This leads to different upsample scaling ratios

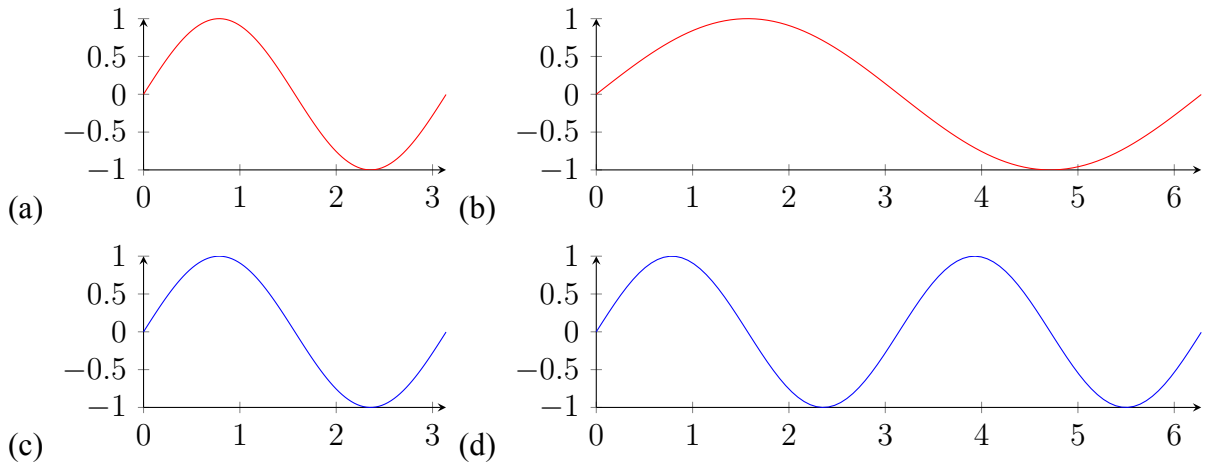

Figure 4: An illustration for the desire TSM. The original signal contains the sinusoid for a single frequency (a,c). The erroneous siganl (b) is the result of directly scaling on the raw waveform, which changes the wavelength of the sinusoid and produces pitch shifting. The desired behavior (d) can be achieved by scaling on the Neuralgram, which compresses the entire sinusoid into one sample.

during the decoding process depend on the frequency since each frequency has different wavelength. Transformation function with variable upsampling ratio is harder to approximate with convolutional generative models. On the other hand, our Neuralgram encodes different frequency components proportionally. This is intuitive for convolutional networks and the model is easier to train.

References

[1] Bao, W., Lai, W.-S., Zhang, X., Gao, Z., and Yang, M.-H. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 3 (2021), 933–948.

[2] Donahue, C., McAuley, J. J., and Puckette, M. S. Synthesizing audio with generative adversarial networks. *CoRR abs/1802.04208* (2018).

[3] Driedger, J., and Müller, M. A review of time-scale modification of music signals. *Applied Sciences 6*, 2 (2016).

[4] Essenwanger, O. *Elements of Statistical Analysis*. General climatology. Elsevier, 1986.

[5] Flanagan, J. L., and Golden, R. M. Phase vocoder. *Bell System Technical Journal 45*, 9 (1966), 1493–1509.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems 27* (2014).

[7] Griffin, D., and Lim, J. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing 32*, 2 (1984), 236–243.

[8] Hejna, D., and Musicus, B. R. The solafs time-scale modification algorithm. *Bolt, Beranek and Newman (BBN) Technical Report* (1991).

[9] Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing 29*, 6 (1981), 1153–1160.

[10] Kim, S., Lee, S., Song, J., and Yoon, S. Flowavenet : A generative flow for raw audio. *CoRR abs/1811.02155* (2018).

[11] Kraft, S., Holters, M., von dem Knesebeck, A., and Zölzer, U. Improved pvsola time-stretching and pitch-shifting for polyphonic audio. In *Proceedings of the International Conference on Digital Audio Effects (DAFx), York, UK* (2012), pp. 17–21.

[12] Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal 37*, 2 (1991), 233–243.

[13] Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebisson, A., Bengio, Y., and Courville, A. Melgan: Generative adversarial networks for conditional waveform synthesis, 2019.

[14] Laroche, J., and Dolson, M. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing 7*, 3 (1999), 323–332.

[15] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[16] Moinet, A., and Dutoit, T. Pvsola: A phase vocoder with synchronized overlap-add. In *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11), Paris, France* (2011).

[17] Morise, M., Yokomori, F., and Ozawa, K. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems 99*, 7 (2016), 1877–1884.

[18] Moulines, E., and Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication 9*, 5 (1990), 453–467. Neuropeech '89.

[19] Nagel, F., and Walther, A. A novel transient handling scheme for time stretching algorithms. In *Audio Engineering Society Convention 127* (Oct 2009).

[20] Nyquist, H. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers 47*, 2 (1928), 617–644.

[21] Prenger, R., Valle, R., and Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 3617–3621.

[22] Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[23] Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal 27*, 3 (1948), 379–423.

[24] Smith, P. Bilinear interpolation of digital images. *Ultramicroscopy 6*, 1 (1981), 201–204.

[25] Tan, X., Qin, T., Soong, F., and Liu, T.-Y. A survey on neural speech synthesis, 2021.

[26] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499* (2016).

[27] Verhelst, W., and Roelands, M. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing* (1993), vol. 2, pp. 554–557 vol.2.

[28] Yang, J., Lee, J., Kim, Y., Cho, H., and Kim, I. Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network, 2020.