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 [6] 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 frame a sequence of samples, typically larger than the wavelength of the lowest-frequency, and relocate these frames in an overlapping fashion [3][8][10]. However, the resultant sound is usually non-natural and contains audible clipping artifacts. This is the negative effect of the framing technique. Moreover, only the most prominent periodicity can be preserved. 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 from the raw waveform to a more semantic representation with complex numbers [5]. 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 [2] to estimate phases after carefully relocate STFT bins to avoid a peculiar

artifacts, a.k.a. phasiness. Despite some refined methods [4][7][9], the spectral representation is essentially not directly scalable, and the iterative phase propagation 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.

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, 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. 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.

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

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

[4] 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.

[5] 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.

[6] 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).

[7] 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).

[8] 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.

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

[10] 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.