DeepITF: Inverting Time Frequency representations using Deep Neural Networks

Nathanaël Perraudin

Student project - Fall 2023

Figure 1: Two representations of the same sound. Left: waveform. Right: STFT magnitude and phase derivatives. From the magnitude, recovering the time signal is often difficult. We believe that reconstructing the phase derivatives in a first step can significantly improve the quality of the recovered waveform or reduce the computation burden.

1 Abstract

Despite impressive recent advances with neural networks [15, 7, 6, 10], generating waveform in audio synthesis remains a challenge in machine learning. As a result, many recent works use specific encoders such as [5] to synthesize sound and avoid problems of generating waveform. Although this approach has been very successful [8, 4] in music generation and speech synthesis, it forces the use of a specific encoder, which can be restrictive in many applications.

In this work, we want to explore a different way to synthesize waveform. Instead of generating the waveform directly, we propose to generate a Short Time Fourier Transform (STFT) representation of the waveform. The STFT magnitude is a much more stable sound representation that does not oscillate in time. However, previous attempts to produce high-quality audio using STFT often failed because of the difficulty to recover the phase from the STFT magnitude. In this work, we take a different approach to overcome this challenge. In addition to the magnitude, we suggest producing the phase gradients with the neural network [1, 3]. Then, the waveform can be recovered using phase integration and STFT inversion [14].
2 Details of the project

Concretely, the main goal of this project will be to build neural networks that invert different STFT and MEL representations to demonstrate the superiority of the magnitude + phase gradient representation. To achieve this, we will first solve the phase recovery problem for any STFT representations, which can be summarized as finding that minimizes a loss of the form

\[ l(x) = \| |Ax| - y \|_2^2, \]  

where \( A \) is the STFT operator and \( y \) is the target magnitude. Existing algorithms \cite{9,14,13} have various degrees of success depending on the representation parameters \cite{11}. We will use a neural network to recover the phase derivatives or to transform the input data into an invertible STFT representation (see Figure 1). Second, we will invert other representations such as MEL spectrograms or Constant-Q Transform \cite{2}.

Our main goal is to show that the phase gradient representation is superior to other representations for audio synthesis. Therefore, we will compare different algorithms and losses for phase recovery and inversion.

3 Student tasks

Stage 1: Setting up data pipeline
- Implement data pipeline for the STFT magnitude and phase gradient
- Test the pipeline to ensure inversion is working

Stage 2: Phase recovery
- Implement Neural Networks for phase recovery
- Compare with traditional algorithms such as \cite{9,14,13} as well as direct inversion using a NN.
- Can we add an adversarial loss to improve the quality of the recovered waveform such as in \cite{5}?

Stage 3: Invert other representations
- Implement Neural Networks for MEL spectrograms and Constant Q Transform
- Compare with other algorithms

Stage 4: Reporting
- Write a report
- Perform a project presentation

4 Additional information

What will you learn? Numerical Analysis, Deep Learning, Generative Modeling, PyTorch, WandB, PyTorch Lightning, Time Frequency analysis, Audio synthesis
Requirements: Machine Learning fundamentals, linear algebra, good Python level, experience with Git

Supervisor: Nathanaël Perraudin, nathanael.perraudin@sdsc.ethz.ch

Is a publication possible? Yes, if the project goes well.

What should I read if I want to know more? You can start with TifGAN [12].

References


    generation of time-frequency features with application in audio synthesis. arXiv preprint

    In 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,

    17–21, 2016.

    Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A