The extraction of musical instruments from bachata songs using semantic segmentation for audio source separation

Tamas Illes, Department of Computer Science, Durham University

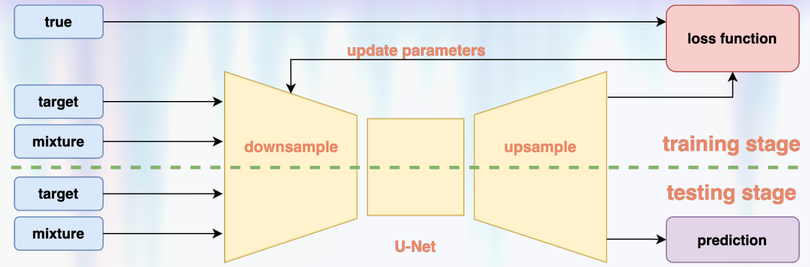
BACKGROUND

In the field of deep learning, **audio source separation via semantic segmentation** is the task of filtering a source signal from a mixture of signals based on learnt knowledge about the signals present in the mixture. While there have been many methods proposed for speech and vocal separation, little work has been done in this area involving **music**. This project explores the possibilities of **extracting instruments from bachata songs** using a U-Net[2] style architecture for semantic segmentation and square-shaped kernels in the convolutional layers of the neural network.

AIMS

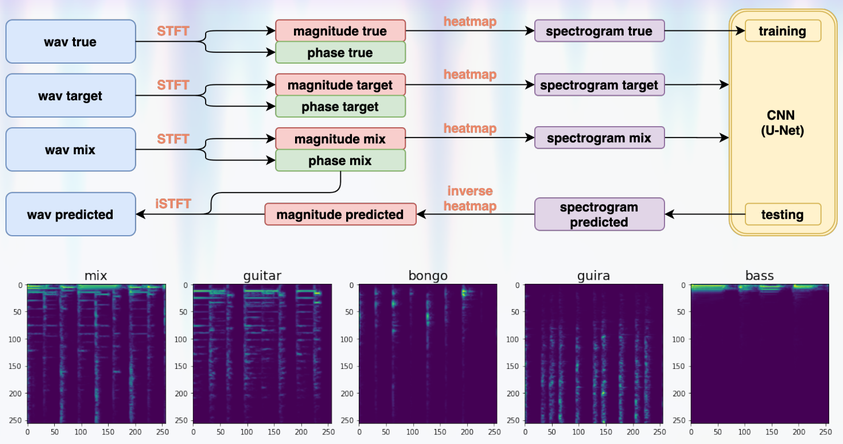
1. Filter a number from a mixture of numbers
2. Separate instruments from audio clips
3. Extract instruments from bachata songs

1. MNIST DATASET



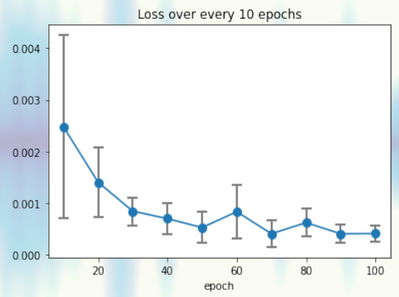
The figure above demonstrates the general flowchart we implemented throughout the project. In the case of the MNIST dataset, the **mixture** was a cluttered image of numbers, while the **ground truth** and **target** were images of the same integer written in two different ways, such that the former one was present in the mixture while the target was not. All images were resized to 32x32 before the 5-level deep U-Net network attempted to filter out the ground truth image from the mixture based on the target image it was given, using MSE as its loss function. The evolution of predictions and the loss behaviour are presented below, using a clutter comprised of 4 distinct integers. The increase in accuracy over time is obvious in both qualitative and quantitative measurements.

2. AUDIO DATASET

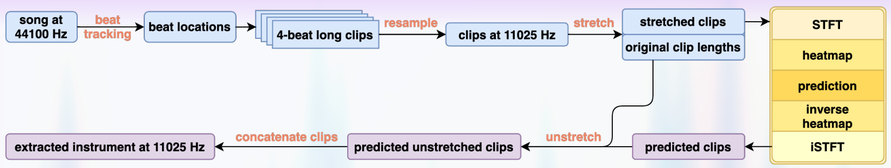


Our audio dataset comprised of **half-bar** (4 beats) long monophonic clips of individual instruments commonly used in bachata music: bongo, guira, bass and guitar. The choice of **bachata**, which is a genre of latin music, was based on its universal structure and the fact that the sound of its instruments has very similar pitch across bachata songs, making them easily detectable [2].

To apply our model design on this dataset, the audio clips had to be converted into images. A useful visual representation of an audio clip is its **spectrogram**, which as obtained by using the **Short-Time Fourier Transform (STFT)** and a **heatmap** function [3]. In this case the ground truth and target spectrograms corresponded to two clips of the same instrument playing different patterns, while the mixture comprised of the four bachata instruments. Our model output the spectrogram prediction of the chosen instrument, from which the filtered audio was constructed using **inverse heatmap** and **inverse STFT** calculations. We obtained **near-perfect** quality predictions, with an overall decreasing tendency in loss, as expected. Qualitative audio results are available in [4].



3. APPLICATION ON SONGS



The application of the trained model on a song required us to **split the song into clips** like those used during training.

Using **beat tracking**, we split the song into 4-beat-long clips which were then **resampled** and **stretched** to match the sample rate and sample length of the training clips prior to feeding them to our neural network. The predicted audio clips were then unstretched and attached to one another to form the **full-length audio of the extracted instrument** at the same tempo as the original song. Our qualitative findings showed that the quality of the guitar and bass extractions always outperformed those of the bongo and guira, however most predictions still contained parts of the **original song in the background**. We also detected a **stronger starting beat** in the clips of the full prediction. This revealed that the 4-beat-long clips produced from songs whose introduction was not comprised of multiples of 4 beats got shifted. Consequently, the input spectrogram did not always correspond to the expected clip of the original song with beat count 1-2-3-4, but for instance to an overlap between two clips with beat count 3-4-1-2. Although this "**intro-issue**" excessively reduced the quality of the extractions, downbeat estimation instead of standard beat tracking could resolve the problem. Qualitative audio samples of the extractions are accessible through [5].

CONCLUSION

Using semantic segmentation with square-shaped kernels in the convolutional layers for audio source separation proved to be successful. Although our design achieved remarkable results on the MNIST and bachata audio dataset, further research is necessary in order to produce good quality instrument extractions from songs.

REFERENCES

[1] Ronneberger, O., Fischer, P., and Brox, T.: *"U-Net: Convolutional Networks for Biomedical Image Segmentation"*, 2015.
[2] Pacini Hernandez, D. : *"Bachata: A Social History of a Dominican Popular Music"*, *1995.*
[3] Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A.: *"The Sound of Pixels"*, 2018.
[4] https://www.dropbox.com/s/3qy3gjo6r4wixbn/sample_clips.zip?dl=0
[5] https://www.dropbox.com/s/u479sggdw25nljo/song_examples.zip?dl=0