Assessing the usefulness of neural networks to predict the BPM of music

Newcastle University

William Caine (16026326), supervised by dr. Jaume Bacardit

2020

# 1. Introduction

As a part-time DJ using modern digital tools, I often found myself frustrated with the amount of time that needed to be spent preparing my music on the computer before a performance. One of the most time-consuming parts of this preparation is labelling each track with its beats per minute (BPM) to aid in live mixing during a performance. As such, most DJ software’s come with tools to automatically detect this value, however the accuracy of these tools is not ideal, leaving the DJ with inaccurate labels or requiring them to manually check the results. There are a variety of approaches and algorithms used to accomplish the BPM prediction but none of the common methods use machine learning, although there are several commercial software’s that keep their methodology a closely guarded secret.

With this in mind I set out to assess the potential of using machine learning to accomplish this task and to see if it could be a viable alternative to the most commonly used algorithms. The task involves identifying patterns in the timing of the music, particularly by identifying percussive elements, and so this pattern recognition struck me as something that a machine learning algorithm may be suited to. Additionally, as a DJ myself, I already had access to a fairly large dataset of manually labelled music that could be used as training data in a supervised learning setting.

## Document Outline

### Chapter 1 - Introduction

Introduction to the project covering the motivation and rationale for the research as well as the specific aims and objectives to be completed.

### Chapter 2 - Background

Some background information on neural networks, how they work and what types might be useful for this application.

### Chapter 3 – Developing a Solution

A detailed account of the process I went through in order to develop and produce the neural network.

### Chapter 4 – Results and Evaluation

An evaluation of how successful the neural network was and how it compares to existing solutions.

### Chapter 5 – Conclusion

A final review of the project as a whole, as per the initial aims and objectives as well as proposed future work.

## 1.2 Aims and Objectives

In my original project proposal, I stated that I wished to predict not only the BPM of the music, but also the onset of the first beat. Using these two values in conjunction, you can determine the exact time at which each beat occurs. As the project progressed however, I found predicting these two values would likely require two separate models trained for each task, or a very different approach. As such, I decided to shift my focus to just predicting the BPM alone as this is the most useful value to know and I believe the success of this model would be a good indication of whether further research would be worthwhile. The final aim of this project was therefore as follows:

Assessing the usefulness of neural networks to predict the BPM of music

In order to achieve this aim, I broke it down into the following key objectives:

* Collect and label a large enough dataset of music samples
* Investigate methods of pre-processing the raw audio for optimal feature detection
* Research different types of neural network and their uses
* Train and evaluate my own neural network model
* Compare the accuracy of my model to existing methods

# 2. Background

Before beginning my research, I first needed to understand the problem at hand as well as neural networks in general to ensure I was working towards the best solution I could.

## 2.1 Existing Solutions

### Sound Energy Algorithm

The most basic algorithm for detecting BPM uses “sound energy” to identify beats (Patin, 2003). This works by splitting the audio into blocks, calculating the energy of each block and comparing it to that of a window of preceding blocks. If block *j* were to contain 1024 samples in stereo, then it’s energy could be calculated as follows:

Blocks above a certain threshold of energy are determined to contain a beat.

### Low Pass Filter Algorithm

One potential way to improve this sound energy algorithm, would be to first process the audio with a low pass filter (Sullivan, n.d.). A low pass filter removes the higher frequency elements of the music such as melodic and harmonic parts, leaving the lower frequencies. This works because most music, particularly western music, contains percussive elements in this register that are indicative of a beat, for example a kick drum. These are both quite primitive methods, however.

### Discrete Wavelet Transform

A more advanced algorithm makes use of discrete wavelet transforms (DWT) in order to process the audio (George Tzanetakis, 2001). The audio is first split into small windows of 1-10 seconds. Each window is split into 4 frequency bands using the DWT algorithm and for each band an envelope is computed.

These envelopes are summed and go through a process called “autocorrelation” to produce the final envelope. From this envelope, the maximum value is identified, and its position is used to identify the tempo of the entire window. This process is repeated over each window of audio and median value is taken to be the BPM of the entire track.

### Fast Fourier Transform

TO DO

## 2.2 Machine Learning

Machine learning (ML) is the term used to describe the process of building mathematical models, without explicit instructions, to infer patterns and features in data (Samuel, 1959). It is considered a branch of artificial intelligence and is an increasingly hot topic for research in recent years. Machine learning is quite a broad term to describe a variety of approaches to building such models, but it is becoming a very commonly used technique for tasks such as computer vision and natural language processing where traditional algorithms can struggle to achieve the desired results.

## 2.3 Supervised Learning

Supervised learning is a specific style of machine learning that makes use of an existing dataset of inputs and desired outputs (Russell & Norvig, 2009). This dataset is referred to as “training data” and consists of a set of examples. One or more inputs is recorded against its desired output or “supervisory signal,” usually represented as a matrix.

Using an iterative approach to optimisation of an objective function, the model is tuned to produce a new function that is able to predict the output of previously unseen inputs. An optimal algorithm will successfully be able to predict the outputs of examples that were not a part of the training data. The process of improving the accuracy of this algorithm is what we refer to as machine learning (Mitchell, 1997).

Supervised learning algorithms include classification and regression. Classification algorithms are used to produce an output belonging to a discrete set of limited values. Regression algorithms, on the other hand, are used to produce a continuous output in a given range.

## 2.3 Neural Networks

Artificial neural networks are one approach to implementing a supervised learning algorithm. The structure of these networks is inspired by biological neural networks in animal brains. The network consists of a series of connected nodes called artificial neurons. These neurons are loosely based on biological cells, taking in an input signal, processing it, and potentially signalling other connected neurons.

This “signal” is typically represented by a real number and the processing done by each neuron to this signal is usually some non-linear function. Each neuron typically has a weight to it that increases or decreases the signal strength. This weight is adjusted during learning to tune the algorithm.

Usually these neurons are aggregated into layers to perform various transformations or functions. An input is passed to the input layer of the network, processed through some number of hidden layers, and finally an output is given by the output layer.

A classic example of this type of network might be used for image recognition. For example, a network could be built to identify the presence of a cat in an image. The inputs here would be an image, represented by a matrix of pixel values. The output might consist of 2 neurons, one representing “cat” and the other representing “no cat”. Whichever neuron has the higher output signal would be considered the prediction.

# References

George Tzanetakis, G. E. P. C., 2001. *Audio Analysis using the Discrete Wavelet Transform.* [Online]   
Available at: https://soundlab.cs.princeton.edu/publications/2001\_amta\_aadwt.pdf

Patin, F., 2003. *Beat Detection Algorithms.* [Online]   
Available at: http://archive.gamedev.net/archive/reference/articles/article1952.html

Russell, S. J. & Norvig, P., 2009. *Artificial Intelligence: A Modern Approach.* 3rd ed. s.l.:Patience Hall.

Samuel, A. L., 1959. Some studies in machine learning using the game of Checkers. I*BM JOURNAL OF RESEARCH AND DEVELOPMENT,* pp. 71-105.

Sullivan, J., n.d. *Beat Detection Using JavaScript and the Web Audio API.* [Online]   
Available at: http://joesul.li/van/beat-detection-using-web-audio/  
[Accessed 20 02 2020].