Tom Perkins

Fourth Year Project - Interim Report

Machine learning applied to timbral acoustic analysis

Contents

1. Project Specification

Definition of timbre

Acoustical timbre can be described as the qualities characteristic to a sound, allowing it to be recognised and "distinguished from other sounds at the same pitch and loudness" [1].

Acoustical timbre is defined for the purposes of this project as the qualities characteristic to a sound that allow it to be recognised as having been produced by a particular voice or instrument. Thus, timbral characteristics allow sounds of the same pitch and volume to be differentiated when they are produced by a different body.

Topic: Applying signal processing & machine learning to timbral analysis

Timbral analysis consists of automatically extracting information corresponding to the timbre of a source from a digital audio signal, in order to describe the source's unique sonic qualities. Timbre is a perceptual quality of sound – therefore complex and subjective. Nevertheless signal processing techniques allow us to measure these qualities. Many approaches exist to estimating timbre, none individually capturing the entire character of the sound effectively – need a combination.

Precisely describing the timbre specific to sounds such as an individual human voice or instrument is a challenging problem not usually tackled by a single conventional signal processing method such as spectral analysis in isolation. The combination of multiple analysis methods with machine learning constructs such as neural networks could allow for a system to infer the embeddings encapsulating timbral information, enabling differentiation between closely related musical sounds, thus approximating the ability of the trained human ear to hear fine timbral differences.

Goal: Identification/classification of acoustic piano sounds using timbral analysis

Many research works have attempted to differentiate between families of musical instruments by timbre, but few have focused on identifying the subtle variations that exist between different instances of the same type of instrument.

We will focus our efforts on developing a system to detect the timbral differences between different variations of the same type of instrument, as opposed to classifying them by a subjective taxonomy (i.e. targeting a glossary of perceptual descriptions such as "bright", "mellow"). We aim to propose a system focused on identifying variations between the timbre produced by pianos of various types. This type of fine differentiation task present a relatively novel challenge, and can be tricky even for humans, as it requires a deep level of familiarity with the instruments.

2. Background theory

2.1. Signal processing features and theory for characterising timbre

In this section, we present an overview of the most popular signal processing concepts and features frequently applied in the literature to timbral analysis tasks. Understanding the role of these features, how to compute them and their interpretation will be instrumental in selecting a musically relevant and representative set of candidate features to trial as pre-processed input to a timbral classifier. In the following discussion, sources [2] and [3] are referenced as comprehensive summaries containing more detailed definitions of each timbral feature and their computation, a subset of which are presented in the popular MPEG-7 standard for audio descriptors [4].

Digital audio signals being made up of samples recorded at a rate Fs, many features are calculated over frames of length L, where L is the number of samples in a frame. Temporal features are computed on the waveform in the time domain within a given frame or over several frames, while spectral features are drawn from the magnitude spectrum in the frequency domain, which is obtained by applying the discrete Fourier transform of the signal. The most commonly applied version of this is the Short-Term Fourier Transform (STFT) with a Hamming window, which produces the magnitude spectrum of the discrete signal on a per-frame basis, allowing for analysis of the magnitude of the frequency bins within a frame as well as analysis of the time-evolution of the spectrum over consecutive frames (spectro-temporal features).

Phase information is not usually considered for timbral analysis, as it is broadly assumed that the character of a musical sound can be inferred from its waveform amplitude and magnitude spectrum primarily [5]; thus the phase characteristics of the signal are ignored in our discussion.

2.1.1. Temporal features

For extraction of temporal features, which concern the time evolution of the waveform over the course of the sound segment, we assume the signal analysis is applied to a single, isolated tone representative of the sound. For musical instruments, this would correspond to a single note played in isolation, and recorded from its onset to finish.

Temporal envelope *(energy envelope)*

The envelope of a waveform is a smoothed version of the signal indicating the overall amplitude shape that the signal takes on over time. This can be achieved in its simplest form by taking the local average [3] or maximum of the waveform's amplitude over a moving window, as demonstrated in [6].

Envelope attack, sustain and decay

- The ***attack time*** is defined as the period between the start of the sound until its maximum amplitude is reached [3]. Typically, the start of the attack is estimated by finding the time step at which a threshold (e.g. 10%) proportional to the amplitude's maximum value over the considered sound is surpassed [2].
- The ***attack slope*** over the attack period further parametrises the speed of a sound's rise, and is inferred from the average rate of increase of the waveform magnitude over the attack period [2].
- The ***steady-state***, or ***sustain*** period, corresponds to the phase after the attack during which the magnitude remains approximately constant near its maximum, and can be characterised by its length (sustain time).
- The ***decrease*** or ***decay*** is characterised by the decrease slope, which can be calculated by estimating the rate at which the signal decays from the maximum-energy point [2].

The shape of the envelope characterises important timbral information relating to the articulation and form of a musical sound. For instance, a note played with *staccato* ("attacked") articulation typically has a short envelope with a rapid rise (short attack time), as opposed to a note articulated as a swell, which will have a slower rise due to the note amplitude's gradual increase initially. On a finer level, these envelope parameters depend not only on articulation (how the instrument is played), but also on the type of instrument and variations between different models of the same instrument, and have been shown experimentally to play an important role in humans' perceptual ability to identify instruments [7].

Temporal centroid

The temporal centroid of a sound measures the time instant around which the energy of a sound is centred [3]. This is estimated using the time average over the signal's envelope, weighted by the signals energy.

Zero-Crossings

The zero-crossing count is the number of waveform sign changes in a given frame. This is computed after subtracting the DC offset (average amplitude) within each frame from the signal, and can be expressed as a zero-crossing rate per unit of time for each frame by normalising the count by the frame length L [2].

2.1.2. Spectral features

Spectral features characterise the distribution of frequencies across the magnitude spectrum for a given sound, within each STFT frame. The spectrum can be skewed towards higher frequencies, which is perceived as a brighter sound, or conversely towards lower frequencies, which corresponds to darker, muted sounds. Furthermore, the distribution of energy across the spectrum can either be concentrated in isolated peaks for tonal sounds, or have a broadband spread, which is perceived as a noisy, breath-like sound [2].

Spectral envelope

Analogous to the temporal envelope in the frequency domain, the spectral envelope corresponds to the overall shape of the spectrum, and can be computed by smoothing the energy spectrum of the signal. As stated in [8], the spectral envelope can characterise a sound independently of pitch, and therefore its shape is indicative of timbre. The following features seek to express this information more succinctly using a set of spectral metrics.

Spectral moments [2]

- The ***Spectral Centroid*** characterises the "central" frequency around which the signal's energy is concentrated. It is calculated by the magnitude-weighted mean of the spectrum along the frequency axis. This can be interpreted as a broad measure of perceived "brightness" of the sound, in that it quantifies the proportion of high to low frequency energy [3]. But this does not account for the spread of frequencies; therefore this measure of brightness is especially indicative if the signal is distributed within a narrow-band of frequencies.
- ***Spectral Spread*** characterises how broadly or narrowly energy is distributed about the spectral centroid (the mean). It is measured as the standard deviation of the frequency distribution in the spectrum (weighted by the normalised magnitude of each bin). This measure is also equivalently described as the bandwidth relative to the centroid, for instance in [9].
- ***Spectral Skewness*** describes the skew, or asymmetry, of the spectrum about the spectral centroid. Negative values indicate energy concentrated below the centroid frequency, while positive values indicate the energy is concentrated in higher frequencies relative to the centroid.
- ***Spectral Kurtosis*** measures the spectrum's flatness around the centroid. Particular ranges of the kurtosis value indicate different spectral shapes, as detailed in [2]: "[a kurtosis value of] 3 indicates a normal (Gaussian) distribution, < 3 a flatter distribution, and > 3 a peakier distribution". This allows us to describe with a single value the "peakiness" of the sound, which is an important part of characterising how tonal it is.

Spectral slope *(spectral tilt)*

The spectral slope is the gradient of the spectrum, typically computed using a linear regression over the points in the spectrum to find the slope of the spectral magnitude [2] or the log-power spectrum, depending on the definition used. This is another descriptor which, similarly to the spectral centroid and skewness, characterises the overall relative prevalence of high and low frequencies in terms of spectral energy.

Spectral Roll-off frequency

The spectral roll-off attempts to measure the cut-off point of the spectrum, as another descriptor of the spectrum's overall shape. This is computed as the frequency below which a majority of the

energy in the spectrum is condensed [3], for instance in [2], "the frequency $f_c(tm)$ below which 95% of the signal energy is contained" is used. This is particularly relevant in characterising low-pass signals, as the roll-off frequency will yield an estimate of the cut-off or corner frequency of a filtered signal.

Spectral Flatness Measure (SFM)

The SFM aims to measure how close the spectrum approaches white noise, whose spectrum is ideally flat. This is estimated by taking the ratio of the geometric mean to the arithmetic mean of the spectral amplitudes in a given frame [2]. Beyond describing the shape of the spectrum, flatness measures such as SFM and Spectral Kurtosis allow us to place the periodicity of a sound along a scale between tonal and noisy sounds, where on one end we have an ideal single sine tone, and on the other extreme white noise, which can be approached using an infinite sum of sinewaves of different frequencies uniformly distributed across the spectrum. The space between these extremes is occupied by sounds of increasing complexity as more tones are combined; this will be further explored in our discussion of harmonic features, in section 2.1.4.

2.1.3. Spectro-temporal features

Spectrogram

The magnitude spectrogram of a signal is a frequency representation of the signal over time, made up of the magnitude spectrum computed over consecutive time frames. The frequency axis (typically plotted along the y-axis, with time on the x axis) and magnitude range in the spectrogram can be scaled to approximate the way humans perceive pitch and volume; by using a Mel frequency scale and/or a logarithmic magnitude scale to form a log-Mel spectrogram representation [10].

The resulting 2-dimensional signal, in which visualisations usually use a colour intensity scale to show the magnitude of each time-frequency bin, represents a fingerprint of the input signal's frequency distribution over time. For instance, the fundamental and harmonic frequencies (see section 2.1.4.) and their associated intensities are visible in this representation, as well as the evolution of individual frequency components over the signal envelope; therefore the spectrogram gives a fairly complete representation of a signal's timbral profile, although it is not pitch-invariant.

Spectro-temporal envelope

The spectro-temporal envelope characterises the shape of the signal in both the frequency and the time domain, by encapsulating the evolution of the energy contained in each frequency bin over consecutive frames. The result is the shape of the signal over a given time period as a function of both time and frequency, and as stated previously, this can be seen in the spectrogram.

In [3], a feature characterising the spectro-temporal envelope is estimated for each frequency bin, calculated over a window of several frames by taking an average of the magnitude of each spectral component over consecutive frames (which the authors name “global spectral envelope”). The authors use this to derive “Harmonic Spectral Deviation”, which is a measure how much each spectral amplitude component differs from its neighbouring spectral envelope.

Spectral flux *(spectral variation)*

The spectral flux or variation is a measure of the spectrum’s rate of change over time. Two different definitions exist in the literature, both consisting of comparing the spectral distribution at consecutive time frames. Both calculations produce a function of time from the spectra of two successive time frames.

In [2], spectral variation is computed as one minus the correlation between two consecutive spectral amplitudes (normalised by the spectral energy at both time steps). In [3], the sum (over the spectral components) of squared differences between spectral magnitudes at two consecutive time frames is used.

2.1.4. Harmonic features

In the context of harmonic analysis, complex sounds such as those produced by musical instruments are modelled as a sum sinusoidal components of differing frequency and amplitude, which are called partials. Of these, harmonic components of the sound are those located at integer multiples of the fundamental frequency (the pitch of the musical sound).

The fundamental is not directly implicated in timbral analysis, since timbre is defined as being independent of pitch, but this feature can be used for estimation of the expected harmonic frequencies of a signal as described in detail in part II.B.4. of [2], and potentially to normalise pitch-dependent features such as the spectrogram (see section 2.1.3). Notably, the harmonic peaks present in a musical sound may deviate from the theoretical evenly-spaced harmonic distribution along the frequency axis, which the authors of [2] qualify as “inharmonic distortion”. The harmonic modelling context introduces a number of features which are commonly used to analyse the timbre of musical sources by characterising the distribution of harmonics (harmonic centroid, spread and variation) as well as the extent to which the signal conforms to harmonic assumptions (i.e. the tonality of the sound).

Harmonic Centroid [3]

The harmonic centroid frequency, analogous to the first spectral moment, is the amplitude-weighted mean frequency of the harmonic peaks identified in the spectrum.

Harmonic Spread [3]

The harmonic spread, similarly to the second spectral moment, is measured as the amplitude-weighted mean across the detected harmonics of the standard deviation of each of the harmonic peaks. This is expressed normalised by the harmonic centroid frequency.

Inharmonicity and ***Harmonic energy skewness*** [9]

Inharmonicity is defined in [9] as the measure of how much the first 4 partials differ from the corresponding theoretical harmonic frequencies (integer multiples of the fundamental). This is computed as the sum of the distances on the frequency axis between each partial and the corresponding expected harmonic frequency (each distance is normalised by that harmonic frequency). The calculation of harmonic energy skewness is similar to that of inharmonicity, but each distance is scaled by the spectral energy in a neighbourhood of the considered partial, yielding a measure analogous to the third spectral moment (skewness).

Harmonic Variation [3]

The harmonic variation is similar to the spectral flux between consecutive time frames. It is computed by the normalised correlation between the harmonic peak amplitudes between two consecutive frames.

Harmonicity Rate and ***Harmonic Brightness***

These features, as defined in [3], both aim to characterise the prevalence of harmonic content in the sound, giving a measure of how tonal or noise-like (inharmonic) the sound is.

The harmonicity rate aims to measure the proportion of the wider magnitude spectrum that corresponds to the harmonics of the sound. This is estimated by computing the maximum of the normalised autocorrelation of the signal. The harmonic brightness aims to quantify the prevalence of upper harmonics (higher frequency overtones) in the signal by taking the ratio of the sum of index-weighted partials' magnitudes over the sum of the harmonics' magnitudes.

Tristimulus values

Developed as an analogy to the three channels of visual colour, these coefficients aim to characterise the harmonic "colour" of a sound based on the distribution of energy over the harmonic series in the signal. Their computation, detailed in [2], relies on the sums of the amplitudes of the harmonic partials of the signal in a given frame.

Odd-to-Even harmonic energy ratio

Sounds containing mostly even harmonics are perceived as "smoother" than those in which the odd harmonics dominate the share of energy in the spectrum [2]. The odd-to-even ratio is computed by the sum of squares of the odd harmonic amplitudes divided by that of the even harmonics.

2.1.5. Formant analysis and the source-filter model

Formant analysis is a popular approach in speech timbre analysis and synthesis [11], often applied to speaker differentiation and identification [12], and can be compared to harmonic analysis in the musical context (formants in speech processing corresponding to harmonics in musical contexts). The approaches and features involved in formant analysis and source-filter modelling could provide useful results in characterising the timbre of musical instruments via analogy to the human voice.

Source-filter model and ***formants***

The source-filter model, typically applied to modelling the human vocal tract, interprets a sound as resulting from a linear system, by which a source (exciter) being passed through a filter (resonator), as shown in **Figure 1**. The excitation at the model source accounts for the noise-like qualities, while the order and characteristics of the filter account for the resonant (tonal) qualities of the resulting sound.

Formant analysis concerns the study of the resonator, and consists of determining the resonant frequencies and bandwidths of the filter modelled for the signal, which are particular to the shape and nature of the body generating the sound. As discussed in [13], the relationship between these formant frequencies is relatively constant across different pitches played by musical instruments, indicating that the formant frequencies, magnitudes and bandwidths are a relevant set of features for characterising instrument timbre. Furthermore, many of the computations described in section 2.1.4. could be applied to characterising formants and their prominence in the spectrum analogously to harmonics.

Linear Predictive Coding (LPC)

The most common scheme for estimating the frequency and magnitude of formants from a waveform is Linear Predictive Coding, which predicts each value of the signal by linear combination of previous samples, as described in [11]. This corresponds to an auto-regressive filter model, which is computed by using the least squares solution to determine each filter coefficient (linear prediction coefficients) using a pre-determined order for the filter. The resulting LPC filter is an estimate of the filter part of the source-filter model, and the location of the filter's poles in the z-plane yields the formant frequencies (the peaks in the filter's frequency response), as shown in **Figure 1**.

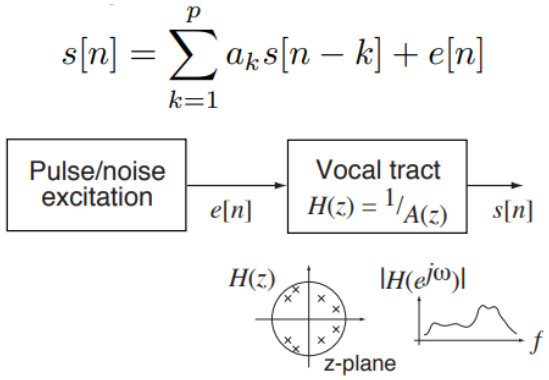

***Figure 1**: Auto-regressive filter equation with coefficients a_k, error e[n] and order p (top). Source-filter model showing z-domain filter transfer function H(z) and representing an illustrative frequency response and pole locations in the z-plane (bottom). Source: From Slide 7 in [14]*

Inverse filtering

Inverse filtering complements formant analysis by attempting to model the excitation, or source, part of the source-filter model through applying an inverse filter to the signal in order to recover the output of the "source" element. The excitation can be estimated by applying Linear Predictive Coding and considering the linear prediction error term e[n] shown in **Figure 1**, which is called the residual [15].

This atonal part of the sound corresponds to unvoiced sounds in speech, and analogously concerns the breathy, inharmonic aspects of the sound produced by musical instruments. Thus, isolating and characterising this excitation may provide interesting results in the way of encapsulating timbre beyond harmonic information.

2.1.6. Cepstrum features & MFCCs

The Cepstrum of a signal is obtained by taking the discrete cosine transform (DCT) of the log-magnitude frequency spectrum. This Cepstral representation shows peaks corresponding to shifted echoes in the original waveform, and therefore reveals a representation of periodic events in a waveform, such as the periodicity corresponding to the fundamental pitch and formants in a complex signal such as speech [16].

Mel-Frequency Cepstrum Coefficients (MFCCs)

The Mel frequency scale is a logarithmic frequency scale based on human perception of pitch relationships. The Mel filter bank is a set of filters whose bandwidths are spaced according to the Mel scale. The Mel-frequency Cepstrum is obtained by mapping the spectrum to the Mel scale using the Mel filter bank and computing the DCT of the logs of the spectral magnitudes across the Mel scale [16]. This Mel Cepstrum has discrete values which form the Mel-Frequency Spectrum Coefficients, a feature set which encapsulates pitch and harmonic information

emulating human perception. These coefficients are therefore compact and powerful descriptors of the perceived harmonic content in a signal over time.

2.2. Machine learning algorithms for identifying timbre

A popular application of timbral analysis of musical audio signals is the automatic classification or grouping of musical instruments from audio recordings; we will guide our attention to the various ways in which musical instruments can be identified from their timbre. In order to draw a classification label from timbral input features, or to cluster samples corresponding to instruments with similar timbres based on an input feature set, many inference algorithms have been applied in the literature. This section gives an overview of the machine learning classification and clustering methods most commonly applied to timbral analysis tasks, especially in the context of musical instrument identification by timbre. We will discuss the most popular methods for timbral classification, while also mentioning clustering methods which are also used in the context of timbral analysis; and then explain the motivation behind our focus on neural network classification.

2.2.1. Non-neural network methods

Traditional machine learning algorithms generally require careful pre-processing of input waveforms into low-dimensional input features to infer the timbre qualities of a signal. These inference models therefore inherently require more structured data as input in order to make informed decisions for classification and clustering, but some are advantaged by their simplicity and their ability to make decision using a smaller amount of data, for instance for methods that do not require training, as opposed to neural network models.

Timbral classifiers (supervised inference)

A supervised classification system seeks to assign one or multiple labels from a pre-defined set to an unseen sample (from the test set), based on the seen samples' known labels (sometimes called the training set if the learning process involves training). This inference results from relating the set of input feature values taken by the unseen example to the set of input features associated with the known labels.

K-nearest-neighbours (KNN) is a simple yet popular scheme for classification of unseen samples given a set of known examples, as described in [17]. The method consists of plotting all samples on the input feature space, such that each known sample forms a point of known label and coordinates in the space resulting from the values taken on by the features for that sample. When presented with an unlabelled sample, the KNN algorithm simply classifies it using the most prevalent label out of the K nearest samples (usually measured by Euclidean distance in the space), where K is a parameter of the method. Variations on this method include distance-weighted voting from the K neighbours. This algorithm performs poorly, however, when using a greater number of input features, as the feature space becomes increasingly sparse according to

the “curse of dimensionality” (a concept coined by Richard Bellman and detailed in [18]). Additionally, its ability to generalise inference to unseen data is limited due to the local nature of the connections established between data points, as noted in [9].

Decision trees (or binary trees) is another straightforward classification scheme described in [17] which builds a tree structure based on the distribution of each feature’s values across a training set. This structure divides the known samples such that each branch groups all the samples taking a particular common range or class for a given feature, by splitting the tree on the point subdividing the feature space with maximum information gain (i.e. entropy reduction). When an unseen sample is input, the tree is traversed from the top down, at each node taking the branch corresponding to the test sample’s input feature value, until a leaf node (where all seen samples grouped by a branch share the same label) is reached and its label is assigned.

Other classification algorithms often cited in timbral analysis work include Support Vector Machines (SVM) and Discriminant Analysis (e.g. Linear, Quadratic, Canonical) as detailed in [9], though these schemes have in recent years fallen out of favour in machine learning research due to their implementation complexity, instead replaced largely by neural network classification.

Despite their limitations, simply-implementable classification schemes such as KNN and decision trees remain useful tools to pre-validate the quality of a choice of input features or data when exploring a classification problem, before moving on to developing a more generalisable classifier such as a neural network model. In particular, the maximum information gain strategy in the construction of a decision tree could help elucidate the most telling features in discriminating between instrument timbres, as noted in [17].

Timbral clustering (unsupervised inference)

Clustering methods differ from classification schemes in that known examples with output labels are usually not supplied; the models are left to relate (or group) samples in a set with one another without supervision, only using their input feature values. This provides a viable alternative to classification for inference when manual ground truth annotations are not available or inconsistent. For timbral analysis, clustering methods such as Gaussian Mixture Models (GMMs) (described in [17]) and Self-Organising Maps (SOMs) are applied in the literature to plotting a low-dimensional timbral space for visualisation and quantification of the relationship between musical instrument sounds [19]; or to provide a system for indexing audio databases by timbral similarity [20].

2.2.2. Neural Network based classification

Definition and variants of Neural Networks applied to timbre classification

In the context of pattern recognition and perceptual tasks such as timbral classification, many of the machine learning algorithms previously discussed have been superseded in recent years by

gradient-based backpropagation learning of multilayer neural networks. As explained in detail in [21], this supervised learning method relies on samples being input to a network of nodes (or artificial neurons), each layer of which is connected to the next layer of nodes with randomly initialised weights followed by a differentiable activation function. When training samples are input to the network, the value for each input feature is fed through the nodes via these weights and activations in a forward pass to produce the activation outputs of the final layer of nodes, which is compared with the ground truth label corresponding to the given known example. The resulting error, which is computed using a given loss function, is then backpropagated through the network in order to update each connection weight using a chosen optimisation algorithm. The general goal of the optimisation method is to take a step towards minimising the error, informed by the error gradient (from which the direction of descent towards a local minimum can be derived).

Classification problems are typically handled by setting the number of output layer nodes to match the number of possible classes, with an output activation function bounded between 0 and 1, and encoding a ground truth label during training with "1" for the output node corresponding to that class and "0" for all the others. For example, in the case of binary classification between two classes, only a single output node is required in the final layer. Then, at test time, the input values of a given test sample are passed forward through the network, and the class corresponding to the output node with the highest activation is selected as the model's prediction.

Despite their inherent advantages over other machine learning methods in terms of their ability for generalisation to unseen data and training set noise insensitivity [21], fully-connected neural network models (in which all the nodes in a layer are individually connected to each node in the following layer with a unique weight) initially provided limited early success when applied to musical instrument classification compared to other classifiers such as KNN [17]. However, Convolutional Neural Networks (CNNs) bring notable performance improvements to perceptual inference tasks such as image recognition, computer vision and audio classification. These models use convolutional connections between layers, where filter kernels with learnable weights are applied to the input multidimensional feature map in each layer in order to produce a modified output map for the next layer. Dimensionality reduction between consecutive layers can be achieved by convolution striding, or pooling (down-sampling) patches of the feature map by average or maximum value.

As initially demonstrated on visual character recognition [22], CNNs are more naturally suited by their convolution kernel weight-sharing structure to higher dimensional inputs such as images or spectrogram inputs (both of which are 2-dimensional vectors), compared to fully-connected neural networks. The combination of sizes of filter kernels that can be used even within a single CNN layer also contributes to their success, as analysis and transformations can be applied at multiple scales and levels of abstraction in a single model [23]. CNNs also have the added benefit of having fewer optimisable parameters between layers since the convolutional weights are

applied like a filter over a small window that moves across a layer's feature map; therefore reducing the cost of each optimisation step. For these reasons, CNNs are adapted to dimensionality reduction and to fusion of high dimensional features, for instance in the context of fusing timbral features [24]. For these tasks, the trend in recent years for solving complex inference tasks is towards deep networks (DNNs) which have a large number of sequential convolutional layers, each using a varied set of filters [23].

RNNs, LSTMs allow for a more dynamic response for time-sequential inputs (more so than CNNs) to be learnt, taking as input consecutive frames of features, nowadays popular in speech recognition and Natural Language Processing (NLP) tasks since they are designed to model the temporal relationship between consecutive inputs [23]. Therefore, these models would also be appropriate for analysing musical audio, and could provide improvements over systems that do not take into account the temporal evolution of timbre.

Advent and popularity of end-to-end learning, letting the network learn feature extraction itself. This type of approach will not be considered since we want to make the most of the mature conventional signal processing timbral feature extraction methods described in section 2.1.

Design, training and data considerations

In general, on top of choosing which type of neural network architecture to use for a classifier problem based on the nature of the given problem and input features, designing a neural network involves the choice of a large number of architectural hyperparameters, including the number of nodes (dimension) and activation functions in each layer, and the total number of layers in the network. In the case of CNNs, we choose the number of convolutional filters to apply between layers, as well as their dimensions. Added to these are training hyperparameters, such as the number of times random batches of training samples are passed through the network, the size of each batch of training samples, as well as the loss function, the optimisation algorithm and its parameters (e.g. optimisation step size). Describe typical training, hyperparameter tuning, cross-validation, testing procedure to be applied to this project.

Deep CNNs require a large amount of training data in order to learn complex patterns [23]. Therefore, a number of considerations are required to properly handle data, especially to prevent overfitting a model to a particular dataset and to avoid bias. Strategies to mitigate these pitfalls include ensuring the employed dataset is class-balanced and representative of a wider population, employing data normalisation, data augmentation, transfer learning [25], early stopping, and regularisation measures such as early stopping and drop-out [26].

3. Literature review of timbral analysis methods

A wide variety of approaches have been applied in the literature to accurately analysing and classifying the timbre specific to a sound, such as that of a particular instrument. We subdivide

these experiments into those focusing mainly on timbral feature extraction via traditional signal processing methods, and methods which employ neural networks to learn the timbre of musical instruments for classification; although we note that these machine learning methods also inherently rely on signal processing theory and constructs, and usually depend on pre-processing of input features. In this section, we will first give an overview of research on musical instrument timbral analysis performed with an emphasis on signal description by applying methods described in section 2.1.. We will then present a survey of more recent works which apply neural networks, specifically CNNs, to the problem of classifying the timbre of musical instruments, research on related tasks, before turning our intention to available datasets which could be used for machine learning of musical instrument timbres.

3.1. Instrument identification by timbre via conventional signal processing

Many authors have tackled identifying or differentiating instruments by their timbre using signal processing descriptors, and the selection of an optimal set of these features for this purpose is often discussed in the literature.

3.1.1. Multi-instrument type classification

Firstly, we consider research aiming to discriminate between different instruments, as opposed to identifying the more subtle timbral variations between models, types or instances of the same instrument. One of the earlier such works on timbral instrument classification focuses on the steady-state part of individual isolated musical notes [27]. This approach narrows the analysis to the harmonic, or tonal, qualities of the sound as opposed to analysing the time evolution of the envelope and transient qualities over the course of a note. As a result, the author finds that discrimination between the considered orchestral instruments is achieved most effectively using spectral and harmonic features such as the spectral moments and the first two harmonic frequencies. The performance of the proposed system, which is a KNN-based classifier, also varies with the considered instrument, reporting much higher accuracy at identifying typically longer-sustained sounds, such as wind instruments, than shorter, more transient sounds such as plucked (*pizzicato*) violin. This shortcoming could be explained by the fact that the shorter sounds tend to be sonically recognisable primarily through their temporal envelope, specifically their harsh attack, as opposed to more sustained instrument timbres, which tend to contain richer tonal and harmonic information.

Another work which focuses on spectral and harmonic feature-based identification of instruments is [9], in which various classification methods (SVM, decision trees & Discriminant Analysis) are applied to identifying classical orchestra instruments from single-note recordings. Similarly to [27], the most representative features for classification are found to be the spectral centroid and the first partial's energy; as well as inharmonicity, which as explained in section 2.1.4. quantifies the prominence of harmonics – thus expressing the extent to which the sound is tonal or noise-like. As a result of incorporating this additional information along with purely tonal descriptors, the authors report reasonable success in identifying both sustained and

transient sounds such as *pizzicati*, although once again the highest scores are reported for identification of wind instruments. In the research presented in [28], many of the same spectral and harmonic features are also applied to classifying a broader range of instruments, including a large corpus of non-western sounds. Beyond the tonal features, the authors also find the attack slope, which characterises the speed of the transient, to be among the most indicative features; along with descriptors of the spectral envelope. These findings indicate that the accurate timbral classification of a wide range of sounds benefits from a combination of spectral/harmonic and time-envelope analyses.

Other research works integrate cepstral features as well as spectral and temporal descriptors in order to classify musical instruments by timbre. For instance in [29], cepstral and time-autocorrelation coefficients are found to be instrumental in accurately differentiating between the timbres of four woodwind instruments using a GMM. Notably, the input to this model consists of extracts from single-instrument performances as opposed to single-note samples. The input information is therefore more complex in nature, but may contain more timbral cues than a single isolated note. This indicates that cepstral features, paired with temporal descriptors, are well-adapted to capturing complex information efficiently; especially since only a small number (10) of cepstral coefficients is used. Similarly, the method presented in [30] applies MFCCs computed at different time scales as well as the time-derivatives of MFCCs to identify instruments using an SVM classifier from excerpts of single-instrument performances. This approach aims to capture timbral features at different time resolutions, both considering the finer details within a single note's envelope as well as on a larger scale for short monophonic musical phrases. The temporal scale of analysis for MFCCs in the context of musical instrument identification was previously examined in [31], a work which also takes on the task of classifying clips of real single-instrument performances using a large number of temporal, spectral, wavelet, and cepstral features.

3.1.2. Intra-instrument classification

A small number of research works focus on differentiating between the timbre of different instances of the same instrument type using signal processing features; either aiming to identify playing techniques or variation in the sonic qualities of the instruments themselves. For instance, in [32], the author examines the timbre of Irish traditional flute playing in order to detect the articulation, phrasing and the model of flute captured in a given recording. The analysis of playing style corresponds to higher-level longer timescale timbral characteristics such as trills, while the single-note study of timbral variation between flutes made from different materials is most relevant to our purposes. Many pre-processing steps are applied for feature extraction and note onset detection: fine variations in timbre between instruments are quantified using spectral and harmonic peak analysis on the harmonically-stable (steady-state) portions of single note recordings, which tends to be effective for timbral characterisation of wind instruments as we have seen in [27].

In the same vein, [19] attempts a similar task by examining the timbre of violins from two different eras. A set of spectral, temporal, harmonic and envelope features are selected to help make this differentiation on sets of single-note and musical scale recordings. Dimensionality reduction (via a clustering method called t-distributed Stochastic Neighbour Embedding (t-SNE)) of the feature space allows for a projected 3-dimensional visualisation to be used to compare the timbral similarity between the instruments. This paper finds that analysing the steady state and decay phases of the violin note sounds allows for differentiation between the two classes of contemporary and historical violins. This differentiation is made on the strength of certain timbral features, determined using feature selection and ranking tests; in particular are selected the spectral distribution features (the spectral moments, roll-off and flatness measures), as well as the Spectral Flux and MFCCs. The relevance of the selected candidate timbral features is verified by inputting them to a SVM for classification between the two types of violins considered: "contemporary" and "historical". Low classification error rates are reported especially when identifying single-note recordings of open strings, which are considered to have more complex harmonic content in the sustain and decay phases of the note envelope, allowing for finer differentiation between violins of different quality.

3.2. Instrument identification by timbral analysis using neural networks

We now turn our attention to research presented on the subject of identifying musical instrument timbre using neural networks; in particular, we find that CNNs are very popular in recent works in the literature when it comes to analysis and classification of instrument timbre. These CNNs are commonly input Mel or log-Mel spectrograms covering 1s-long time windows, since they encapsulate a variety of spectral, harmonic and envelope information, a feature which CNNs are able to handle and learn from efficiently by their propensity to process large amounts of data in 2-dimensional feature maps (see section 2.2.2.) such as time-frequency maps.

One such system [33] is used to identify the predominant instruments in recordings of multi-instrument mixtures, using a deep CNN made up exclusively of small 3x3 filters in each convolutional layer, separated periodically by max-pooling layers for dimensionality reduction from the input spectrograms via abstracted feature maps to the low-dimensional fully-connected output layers. Broadly speaking, this sort of deep and narrow CNN architecture has been commonly applied across the literature in recent years. For instance, [34] improves upon this system by adding source separation of the instrument mixture as a pre-processing step, achieving improved instrument classification results on identifying jazz instruments with closely related timbres. Contributing to this method's success, the authors also cite transfer learning as a good way of getting around limited training data. This is achieved by using a model pre-trained on a different, larger dataset as a starting point for the CNN to then learn more application-specific timbral mappings from the small targeted dataset of jazz instruments.

Another piece of research considering classification the most prominent instrument from recordings of pieces played by multi-instrument mixtures is [35], which also uses log-Mel spectrograms as the input feature map. This work focuses on how to design CNN architectures to effectively capture timbral information, using musically-informed intuitions such as the fact that timbre should be inferred independently from pitch, duration and volume. Given that the network's layers operate on the spectro-temporal domain, the optimal choice of dimensions in time (number of frames) and frequency (number of bins) of the convolutional layers is discussed. The experiments on sung phoneme classification and instrument classification lead the authors to conclude that the first layer of the CNN benefits from using a diversity of filter dimensions to capture different scales of time-frequency feature mappings. Additionally, max-pooling layers over the frequency dimension are used in order to reduce the effects of pitch, as it can be shown that max-pooling (as opposed to average pooling) mitigates the effect of shifting the input of a CNN layer on its output feature maps.

Other CNN-based methods using Mel spectrograms are applied to less challenging data in the form of isolated recordings of individual instruments, often restricted to playing a single note at a time. In [36], the authors apply this class of method to recognising classical instrument families using a variety of recording types, comparing the performance of CNNs trained on Mel spectrograms drawn from isolated notes, monophonic melodies, and polyphonic pieces. It is found that CNNs trained on one sort of data do not generalise well to classifying the same instruments from another type of recording: for instance, high accuracy is reported for a model trained and tested on single-note recordings, but this performance does not carry over to testing the same model on recordings of pieces played on the same set of instruments. This implies that in the paradigm of CNNs using Mel Spectrograms as input, models are sensitive to polyphony, and that the embeddings that allow them to differentiate between single note samples of instruments differ from those characterising timbre in recordings of a piece of music being played on those same instruments.

3.3. Research on related topics

Besides musical instrument classification, a great deal of recent research has concerned applications of and tasks related to timbral analysis. These include:

- *Music Information Retrieval* (MIR), including musical genre identification, for the automatic indexing of audio metadata. Genre or style identification is usually achieved by similar feature-based means to instrument classification, but applied to learning the texture or other style cues of a full mix of music as opposed to that of an instrument. A popular reference in the literature for this task is the approach presented in [37], where many of the same timbral features we describe in section 2.1. are applied alongside features related to pitch and rhythm description for genre classification.

- *Source separation*, which aims to separate out audio corresponding to the different instruments in a mix. Notably, this task is tackled in [38] and [39], where the problem of overlapping harmonic partials (formants) between 2 instruments in a mix is mitigated using other timbral descriptors. In [38], the estimated spectral envelope of each instrument in the mixture is used to help separate the instruments from one another, while [39] aims to separate the sources by exploiting differences in their amplitude and frequency modulation characteristics (i.e. vibrato and tremolo playing effects respectively in musical terms).
- *Musical synthesis* and instrument *timbre transfer*: [40] uses a generative system (variational auto-encoder) to map musical instrument audio to a latent timbral space, a type of approach which the authors recently applied [41] to synthesising new sounds by selecting points in the space to transfer the timbre of one instrument to a different instrumental performance (e.g. between orchestral instruments and voice).
- *Speech recognition & diarization* (differentiation of who is speaking when): similarly to identifying variations between instruments, the timbral quality of the human voice is often used to help identify a speaker, which can be applied to diarization for dialogue transcription and voice authentication. Recent work [42] applies CNNs using Mel-spectrograms as input (as seen in section 3.2.) to the timbral classification of different voices on the basis of gender and age labels.

Another promising recent development relevant to timbral analysis is the integration of conventional signal processing elements, such as those studied in section 2.1 for extraction of timbral features, into an end-to-end neural network architecture [43]. This allows signal processing functions to be used within a deep-learning framework, as opposed to being limited to use as pre-processing steps. Notably, qualitatively promising results for this system are demonstrated on timbre transfer from voice to violin, as well as on the decomposition of musical instrument sounds into noise-like and tonal components.

3.4. Survey of available datasets

Table 1 presents the main attributes of musical instrument sample databases available online, many of which are used in works reviewed in sections 3.1 and 3.2. These datasets were initially destined for various applications including signal processing research, music information research, and digital music creation. We place a special emphasis on databases containing samples of a range of acoustic pianos, since this will be our target instrument; and using a larger amount and variety of training data for our classifier will improve its ability to generalise timbral inference.

We have also sought out datasets in which timbral information on the articulation (the playing technique used, e.g. vibrato), type, or model of instrument is annotated, since these are the sort of timbral labels we plan to predict with our classifier. For piano sounds, these labels could

include different body shapes and sizes, such as upright and grand, different mechanisms (e.g. use of the felt pedal on an upright piano) affecting the timbre, or dynamics (e.g. pianissimo, forte, etc.). Many databases destined for music creation, called sample libraries or virtual instruments, capture a given instrument in high-quality note-by-note audio recordings, sometimes with several passes over the range of notes available on the instrument, at multiple dynamic levels (called velocity layers). These detailed recordings could be particularly useful for our purposes, since they include a large number of carefully parametrised and labelled samples for training and testing a neural network.

Dataset name	**Affiliation / Reference**	**Intended purpose**	**No. / type of instruments**	**Timbral Annotations**	**No. / type of pianos**	**No. of samples**
SOL	IRCAM [44]	Research	16 wind + string	Articulation	None	25000
MUMS Revised	Eerola et al. [45]	Research	100+ varied	Model Articulation	3 Upright, Grand	N/A (large)
SHARC	G. Sandell [46] Derived version of MUMS containing only steady-state portions	Research	39 orchestra	Articulation	None	1338
conTimbre	T. Hummel [47]	Various	150 orchestra	Articulation	1 N/A	4073
MIS	University of Iowa [48]	Research	30+ orchestra	Model Dynamics	1 Grand	N/A (large)
MAPS	Telecom ParisTech [49]	Research	9 pianos	Articulation Type & Model Dynamics Conditions	9 Grand, Upright, Hybrid	N/A
RWC	Real World Computing Partnership [50]	Research	50 varied	Articulation Dynamics	5 Acoustic, Electric	2000+
Concert Piano	N. Plath [51]	Research	1 piano	Model Dynamics Conditions	1 Grand, before and after concert use	600+
BiVib	Papetti et al. [52]	Research	2 pianos	Type & Model Dynamics	2 Grand, Upright	1000+
Piano Pedalling	L. Beici [53]	Research	1 piano	Articulation (pedal) Type & Model Dynamics	1 Grand	500+
Pianobook	C. Henson [54]	Music Creation	450+ varied	Articulation (pedals) Type/Model Dynamics	100+ Grand, Upright, Electric	N/A (large)

***Table 1**: Comparison of single-note musical instrument sample databases*

A combination of these datasets could be collated to provide a greater diversity of types, recording conditions and sources of instruments, as well as a larger number of samples to support training and testing of a classifier. Some of this merging work has already been done by the authors of the Single-Note Database (SNDB) [55], which combines RWC, MUMS, MIS and further samples from the Vienna Symphonic Library (destined or music creation). The resulting database is assembled specifically for the purposes of training machine learning systems, thus featuring uniformly-formatted annotations. This could form the starting point for assembling a larger dataset using the additional databases listed in **Table 1** sharing timbral annotation fields with SNDB.

4. Implementation and Evaluation Plans

4.1. Software standards and toolkits

We plan to develop the classifier primarily in Python, mainly as a result of prior experience with Python, as well as its popularity in the machine learning community and the availability of various open-source neural network development libraries. MATLAB and libraries for MATLAB are also included in the list of software tools used, since they are useful both for initial experimentation and for feature extraction, potentially for integration into the Python environment [56]. A key aspect of the software libraries used is the fact that they are open-source, allowing for examination of the underlying source codebases for understanding, debugging and modification.

Signal processing feature extraction toolkits

- VOICEBOX for MATLAB [57]: includes a wide range of standard audio Digital Signal Processing functions, including timbral analysis, for voice processing, many of which are appropriate also for processing musical audio.
- Rhythm & Timbre Feature Extraction from Music library for Python, MATLAB & Java [58]: includes spectral, temporal and modulation descriptors (e.g. Statistical Spectrum Descriptor, Modulation Frequency Variance Descriptor) for timbre characterisation. These features are described in the accompanying paper [59].
- The Librosa [60] library for Python: implementation of MIR signal processing algorithms, including spectral, harmonic, statistical, and temporal analysis, and extraction of timbral features described in section 2.1.
- The Essentia [61] library for Python, C++ & Java: another popular library for analysis of music signals, including many relevant functions for timbral feature extraction and signal description.
- The Timbre Toolbox [2] implemented for MATLAB [62]: a set of temporal, harmonic and spectral descriptors for timbral characterisation of signals, many of which are those described and referenced in section 2.1.

Machine learning development libraries for Python

- NumPy for standard mathematics functions and data manipulation.
- PyTorch for designing, training and testing neural network architectures.
- TensorFlow may be used as it is a popular alternative in the literature to PyTorch, although the preference is for PyTorch due to personal familiarity from prior experience.
- Scikit-learn for additional machine learning and data science functions, including evaluation tools.

4.2. Implementation & Evaluation Plans

4.2.1. Completed Tasks (September 2020 – January 2021)

- *Specification*: Initial project proposal conceived and approved.
- *Research*: Study of background theory of signal processing for timbral feature extraction, and machine learning methods for timbral classification (see section 2).
- *Experiment*: Formant extraction and analysis using Linear Predictive Coding coefficients. Wrote a MATLAB script to extract formants from single-note recordings of a flute played at different pitches. LPC functions provided by the VOICEBOX toolkit were used to estimate LPC coefficients from the waveforms, which were then translated to estimated formant frequencies and bandwidths. The relative frequencies of the first few formants were plotted across the different pitches in the range of the instrument in order to confirm the pitch-invariance of the ratio between the formant frequencies of a given instrument, which is one of the reasons for which formants are considered as descriptors of timbre (see section 2.1.5.)
- *Research*: Literature review of signal processing feature extraction and machine learning applied to timbral analysis. Categorisation of papers into principally conventional signal processing methods and CNN-based approaches, as well as related tasks and applications of timbral analysis.
- *Research*: Collection of datasets for timbral analysis of musical instruments.
- *Specification*: Refining the project to classification of single-note piano sounds using neural networks, on the basis of research on existing methods in the literature and the available databases.
- *Experiment*: set-up and debugging of off-the-shelf code for one of the research works on CNN-based timbral classification [35] discussed in section 3.2, from the code repository published online by the author [63]. At the time of writing, the author, Rong Gong, has been contacted for additional guidance in reproducing the reported results using the supplied pre-trained models.
- *Deliverable*: Completed the interim report.

4.2.2. Planned future milestones

This section presents the tasks which are currently planned for the project, including any foreseen risks and contingencies considered for each task. For a Gantt chart projecting the

planned chronology for these tasks, see **Figure 2**. Each bullet point in the implementation and evaluation plans corresponds to a row in the Gantt chart.

Testing methods from the literature

- *Experiment*: Continuing on from the above, reproduce the results achieved by the CNN presented in [35] for the demo experiments described in the code repository [63] (sung phoneme classification and identification of the predominant instrument in a mixture).
- *Experiment*: Apply the off-the-shelf CNN method [35] to classifying single-note isolated piano sounds, using one of the datasets presented in **Table 1**. This architecture could be used as a proof-of-concept and baseline for future experiments, or if initial results are promising, it could serve as a starting point for the proposed system specialised to our target application. Performing this task depends on having debugged and successfully reproduced the musical instrument classification results reported in the paper (see prior experiment task).
  Risk: The experiment in the paper involves identifying the predominant instrument in excerpts of pieces played on a mixture of instruments. Since the data is fundamentally different, models will most likely need to be re-trained to perform well on single-note samples, since CNNs are not known to generalise well as discussed in [36] (see section 3.2.). Adapting the proposed architecture to single-note piano sounds may prove difficult or impossible for this reason.
- *Experiment / Fallback plan*: if no success is reported with the above considered CNN, we may attempt reproducing and training a different CNN from the literature (section 3.2.) by using an online implementation, or, in the absence of public code repositories, by following the architecture specification outlined in one of the papers.

Feature extraction

- *Specification*: Select which timbral features to extract and pre-processing steps to apply for input to the musical instrument timbral classification system. This task will depend on having fully researched the considered timbral features, how they can be combined, and how they have been applied in the literature to instrument classification.
  Risk: while this choice will be informed by the survey of existing methods in the literature, it is impossible to predict which features will provide the optimal performance for the system we will propose before it has been developed – it is therefore possible that we later re-examine and modify this selection based on performance of the classifier.
- *Implementation*: Develop timbral feature extraction code, using one or more of the software toolkits cited in section 4.1. This depends directly on the previous selection task, as we will seek to integrate only the method(s) we have chosen as input pre-processing for the classifier.

Development of a Neural Network architecture

- *Specification*: Select a Neural Network architecture type (Fully-connected, CNN, RNN, or LSTM). This is currently planned to be a deep CNN architecture, given this sort of design's prevalence and success in the literature on timbral classification of musical instruments.
- *Specification*: Select the technical characteristic (e.g. type, model, dynamics or articulation) of pianos to target as timbre label for the classifier. This co-depends directly on which data sources are used (see the following task), since our selection of a characteristic to predict using the classifier is limited by which ground truth annotations are available in the chosen databases, in order for our system to be trained and tested on the dataset.
- *Implementation*: Prepare a training & testing dataset of piano sounds for the task, pulling from multiple database sources. Design and implement a scheme for splitting the dataset into training and testing subsets, without introducing artificial bias to any of the subsets.
  Considerations & risks: The dataset will need to be consistently and accurately labelled in a common format, which will require some effort to merge databases originating from multiple sources. We also need to cover as wide a range of variations as possible, and ensure the dataset is balanced as possible to ensure each class is sufficiently represented to train an unbiased classifier. If we later find that the size or bias of the dataset is a limiting factor for the classifier's performance, we may explore the possibility of using data augmentation to extend or re-balance the dataset, for instance by using slight pitch shifting of samples.
- *Implementation*: Draft a neural network architecture that takes in the selected timbral features and predicts the target label relating to timbre. Train the classifier on the chosen dataset of labelled piano sounds.
- *Implementation*: Tune the architecture and learning hyperparameters of the model, for instance using a cross-validation scheme (within the training set), in order to optimise performance without overfitting the architecture to any one subset of the data. Use these experiments to finalise the design for the neural network architecture.
  Risks & uncertainties: no clear stopping condition here, simply the feeling that the network is "good enough". Uncertainties concerning whether the proposed architecture is any good should nevertheless be determined in the evaluation phase of the project.

Evaluation plan for the proposed system

- *Implementation*: Choose whether the system's predictions are evaluated on a sample-by-sample basis, or allow the classifier to vote for the most likely match by inputting multiple samples at a time (i.e. a set of samples corresponding to the same instance of an instrument).
- *Specification*: Research, select and implement a set of scoring metrics such as accuracy, precision, recall, F1 and AUC (area under the curve) scores to assess the proposed model's classification performance.

- *Evaluation*: Metric-based evaluation of the implemented classifier and any previously established baselines using the targeted testing dataset. Confusion matrices may also be used to present the classification results.
- *Evaluation*: Metric-based evaluation of the proposed neural network architecture on publicly available datasets, for benchmarking by comparison to systems from the literature which were evaluated on the same datasets.
  Risk: a possible issue for this task is the lack of works in the literature focusing on the specific problem of intra-instrument type timbre differentiation between different models of pianos; we will have to adapt our system to perform a related task, such as inter-instrument type classification, in order for it to be compared to a method in the literature. This will require additional implementation time and may not provide results representative of our proposed system's performance on the task we actually targeted.
- *Evaluation*: Perform dataset size evaluation by plotting the prediction accuracy of models trained on subsets of various sizes of the actual training data, in order to determine whether the amount of data used is a limiting factor in the performance of the classifier. For instance, if we find that the performance gains brought by increasing the portion of training data used taper off as we approach full training set utilisation, we can rule out the hypothesis of the dataset being too small for a given architecture.
- *Evaluation*: Devise and perform a subjective testing experiment to compare the proposed system's ability to differentiate piano sounds to that of humans, preferably surveying a cohort of respondents with relevant musical experience for analysing the sound of pianos. This will require researching methods for performing an effective and representative tests of human perception.
  Risk: this evaluation step may require a sizeable amount of time and attention in order to produce representative results; therefore, depending on whether we judge the evaluation to be sufficiently thorough without comparison to human perception, this task may not be undertaken given the time-frame of the final deliverables of the project.

Final Deliverables

- *Deliverable*: Draft and complete the final report, iterating upon the interim report as more experiments are attempted and completed.
- *Deliverable*: Prepare the final presentation as a summary of content in the final report.

TASK	START DATE	END DATE
Testing methods from the literature		
Reproduce results reported in [35]	1/2	12/2
Apply [35] to piano classification	13/2	28/2
Attempt reproducing a different CNN	20/2	10/3
Feature extraction		
Select timbral features	8/3	10/3
Implement feature extraction	10/3	26/3
Developing a neural network architecture		
Select a neural network type	27/3	29/3
Select the targeted timbre labels	29/3	1/4
Prepare the dataset	29/3	19/4
Draft & train the NN architecture	12/4	26/4
Tune the NN architecture	26/4	10/5
Evaluation		
Choose classifier's prediction scheme	3/5	12/5
Select & implement scoring metrics	10/5	17/5
Evaluation on the targeted dataset	14/5	20/5
Evaluation on public dataset	17/5	31/5
Training data size evaluation	17/5	19/5
Subjective testing	17/5	7/6
Deliverables		
Draft & complete final report	20/5	16/6
Prepare final presentation	1/6	21/6

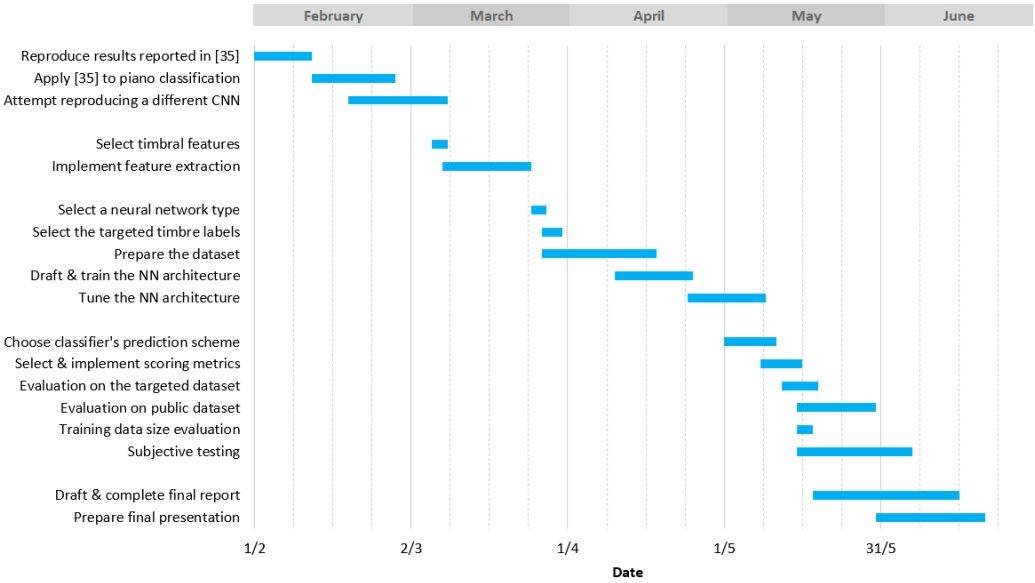

***Figure 2**: Gantt Chart presenting the timeline of planned tasks with start and end dates for each.*

4.3. Legal & ethical considerations

4.3.1 Use of data and licensing of software tools

For training and evaluation, a large amount of data is required to be collected and processed by the neural network. This data will be sourced from public databases and, if not free, acquired via legal means, giving full credit to the original authors and any intermediaries who have contributed to making the data available. Additionally, the licensing information associated to any data sources used will be verified to ensure that our use of the data falls within its intended or allowed uses.

4.3.2. Ethical considerations for subjective testing

Before any survey on humans is undertaken, research into the best ethical practices for subjective testing and proper use of personal data will be performed. Any private data collected will be handled as such, following data protection laws.

References

[1] N. M. McLachlan, “Timbre, Pitch, and Music,” Oxford Handbooks, 2016. [Online]. Available:

https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199935345.001.0001/oxfordhb-9780199935345-e-44. [Accessed January 2021].

[2] G. Peeters, B. Giordano, P. Susini, N. Misdariis and S. McAdams, “The Timbre Toolbox: Extracting audio descriptors from musical signals,” *The Journal of the Acoustical Society of America,* vol. 130, pp. 2902-2916, 2011.

[3] X. Zhang and Z. W. Ras, “Analysis of Sound Features for Music Timbre Recognition,” in *International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, Seoul, 2007.

[4] ISO/IEC, *MPEG-7: Information Technology – Multimedia Content Description Interface - Part 4: Audio (ISO/IEC FDIS 15938-4:2002),* 2002.

[5] R. Plomp and J. M. Steeneken, “Effect of Phase on the Timbre of Complex Tones,” *The Journal of the Acoustical Society of America,* vol. 46, pp. 409-421, 1969.

[6] M. Müller, “Fundamentals of Music Processing: Timbre,” International Audio Laboratories Erlangen, 2015. [Online]. Available: https://www.audiolabs-erlangen.de/resources/MIR/FMP/C1/C1S3_Timbre.html. [Accessed January 2021].

[7] C. Elliot, “Attacks and Releases as Factors in Instrument Identification,” *Journal of Research in Music Education,* vol. 23, no. 1, pp. 35-40, 1975.

[8] D. Shwarz, “Spectral Envelopes in Sound, Chapter 3.3,” Institut fur Informatik, Universitat Stuttgart, Stuttgart, 1998.

[9] G. Agostini, M. Longari and E. Pollastri, “Musical Instrument Timbres Classification with Spectral Features,” *EURASIP Journal on Advances in Signal Processing,* 2003.

[10] J. Pons, O. Slizovskaia, R. Gong, E. Gómez and X. Serra, “Timbre Analysis of Music Audio Signals with Convolutional Neural Networks,” Music Technology Group, Universitat Pompeu Fabra, Barcelona, 2017.

[11] B. Atal and S. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” *The journal of the acoustical society of America,* vol. 50, no. 2B , pp. 637-655, 1971.

[12] F. Nolan and C. Grigoras, “A case for formant analysis in forensic speaker identification,” *International Journal of Speech, Language and the Law,* vol. 12, no. 2, pp. 143-173, 2005.

[13] J. McCarty, “Timbral Analysis: Formant Analysis,” CCRMA Stanford Center for Computer Research in Music and Acoustics, 2003. [Online]. Available: https://ccrma.stanford.edu/~jmccarty/formant.htm. [Accessed January 2021].

[14] D. Ellis, “Linear Prediction (LPC),” Dept. Electrical Engineering, Columbia University, 2013. [Online]. Available: https://www.ee.columbia.edu/~dpwe/e4896/lectures/E4896-L06.pdf. [Accessed January 2021].

[15] J. Smith, “Physical Audio Signal Processing for Virtual Musical Instruments and Audio Effects: Inverse Filtering,” Center for Computer Research in Music and Acoustics, Stanford, 2010. [Online]. Available: https://ccrma.stanford.edu/~jos/pasp/Inverse_Filtering.html. [Accessed January 2021].

[16] A. Oppenheim and R. Shafer, “From Frequency to Quefrency: A History of the Cepstrum,” *IEEE Signal Processing Magazine,* vol. 21, no. 5, pp. 95-106, 2004.

[17] P. Herrera-Boyer, G. Peeters and S. Dubnov, “Automatic Classification of Musical Instrument Sounds,” *Journal of New Music Research,* vol. 32, no. 1, pp. 3-21, 2003.

[18] E. Keogh and A. Mueen, “Curse of Dimensionality,” in *Encyclopedia of Machine Learning and Data Mining*, Boston, MA, Springer US, 2017, pp. 314-315.

[19] F. Setragno, M. Zanoni, A. Sarti and F. Antonacci, “Feature-based Characterization of Violin Timbre,” in *25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 2017.

[20] A. Eigenfeldt and P. Pasquier, “Real-time timbral organisation: Selecting samples based upon similarity,” *Organised Sound,* vol. 15, no. 2, pp. 159-66, 2010.

[21] S. J. Russell and P. Norvig, “19. Learning in Neural and Belief Networks,” in *Artificial Intelligence: A Modern Approach*, Englewood Cliffs, NJ, Prentice-Hall, 1995, pp. 567-580.

[22] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[23] A. Khan, A. Sohail, U. Zahoora and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review,* vol. 53, no. 8, pp. 5455-5516, 2020.

[24] T. Park and T. Lee, “Musical instrument sound classification with deep convolutional neural network using feature fusion approach,” *arXiv eprint,* vol. arXiv:1512.07370, 2015.

[25] S. J. Pan and Q. Yang, “A Survey on Transfer Learning,” *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, p. 2010, 1345-1359.

[26] N. a. H. G. Srivastava, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research,* vol. 15, no. 1, pp. 1929-1958, 2014.

[27] I. Fujinaga, “Machine Recognition,” in *Proceedings of International Computer Music Conference*, Ann Arbor, MI, 1998.

[28] D. Fourer, J. L. Rouas, P. Hanna and M. Robine, “Automatic Timbre Classification of Ethnomusicological Audio Recordings,” in *International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, 2014.

[29] J. C. Brown, O. Houix and S. McAdams, “Feature dependence in the automatic identification of musical woodwind instruments,” *The Journal of the Acoustical Society of Americ,* vol. 109, no. 3, pp. 1064-1072, 2000.

[30] B. L. Sturm, M. Morvidone and L. Daudet, “Musical instrument identification using multiscale mel-frequency cepstral coefficients,” in *18th European Signal Processing Conference, IEEE*, Aalborg, Denmark, 2010.

[31] C. Joder, S. Essid and G. Richard, “Temporal Integration for Audio Classification With Application to Musical Instrument Classification,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, pp. 174-186, 2009.

[32] I. Ali-MacLachlan, “Computational analysis of style in Irish traditional flute playing,” PhD thesis, Birmingham City University, Birmingham, UK, 2019.

[33] Y. Han, J. Kim and K. Lee, “Deep convolutional neural networks for predominant instrument recognition in polyphonic music,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 1, p. 208–221, 2017.

[34] J. Gómez, J. Abeßer and E. Cano, “Jazz solo instrument classification with convolutional neural networks, source separation, and transfer learning,” in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.

[35] J. Pons, O. Slizovskaia, R. Gong, E. Gomez and X. Serra, “Timbre Analysis of Music Audio Signals with Convolutional Neural Networks,” in *25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 2017.

[36] M. Taenzer, J. Abeßer, S. I. Mimilakis, C. Weiß, M. Müller and H. Lukashevich, “Investigating CNN-Based Instrument Family Recognition for Western Classical Music

Recordings," in *Proceedings of the 20th International Societyfor Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[37] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing,* vol. 10, no. 5, pp. 293-302, 2002.

[38] Y. Lin, W.-C. Chang, T.-M. Wang, A. W. Su and W.-H. Liao, "Timbre-constrained Recursive Time-Varying Analysis for Musical Note Separation," in *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, 2013.

[39] F. Stoter, S. Bayer and B. Edler, "Unison Source Separation," in *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany, 2014.

[40] P. Esling, A. Chemla-Romeu-Santos and A. Bitton, "Generative timbre spaces with variational audio synthesis," *CoRR,* vol. abs/1805.08501, 2018.

[41] A. Bitton, P. Esling and T. Harada, "Vector-Quantized Timbre Representation," 2020.

[42] Y. Lyu, C. Liu, H. Tan, R. Xie, K. Tang and Z. Gu, "Convolutional Neural Network based Timbre Classification," in *CIAT 2020: Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies*, Held online, 2020.

[43] J. Engel, L. Hantrakul, C. Gu and A. Roberts, " DDSP: Differentiable digital signal processing," *arXiv preprint,* vol. arXiv:2001.04643, 2020.

[44] G. Beller, "FullSOL (Studio Online Dataset)," IRCAM, 2020. [Online]. Available: https://forum.ircam.fr/projects/detail/fullsol/. [Accessed January 2021].

[45] T. Eerola, R. Ferrer Flores and V. Alluri, "MUMS Revised," University of Jyväskylä, 2017. [Online]. Available: https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/projects2/past-projects/coe/materials/mums/MUMS. [Accessed January 2021].

[46] G. Sandell, "Sandell Harmonic Archive (SHARC)," Northwestern University, IL, USA, 1991. [Online]. Available: http://gregsandell.com/j/pageSharc.php. [Accessed January 2021].

[47] conTimbre, "conTimbre - Infos," [Online]. Available: https://www.contimbre.com/en/infos. [Accessed January 2021].

[48] L. Fritts, "Musical Instrument Samples (MIS)," University of Iowa Electronic Music Studios, [Online]. Available: http://theremin.music.uiowa.edu/MIS.html. [Accessed January 2021].

[49] V. Emiya, N. Bertin, B. David and R. Badeau, “Midi-Aligned Piano Sounds (MAPS) - A piano database for multipitch estimation and automatic transcription of music,” Telecom ParisTech, Département Traitement du Signal et des Images, Paris, France, 2010.

[50] M. Goto, “RWC Music Database: Musical Instrument Sound Collection,” Real World Computing Partnership (RWCP) of Japan, 2003. [Online]. Available: https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-i.html. [Accessed January 2021].

[51] N. Plath, “Influence of Playing on the Tonal Characteristics of a Concert Piano - An Observational Study,” in *Proceedings of the International Symposium on Music Acoustics (ISMA) 2019*, Detmold, Germany, 2019.

[52] S. Papetti, F. Avanzini and F. Fontana, “Design and Application of the BiVib Audio-Tactile Piano Sample Library,” *Applied Sciences,* vol. 9, no. 5, 2019.

[53] B. Liang, “Dataset for Analysing Effects of Piano Pedalling Techniques,” Zenodo, 2017. [Online]. Available: https://zenodo.org/record/3242149. [Accessed January 2021].

[54] C. Henson, “Pianobook,” 2020. [Online]. Available: https://www.pianobook.co.uk/. [Accessed January 2021].

[55] E. F. Feichtner and B. Edler, “Description of the Single Note Database SNDB,” in *145th Audio Engineering Society Convention*, New York, NY, USA, 2018.

[56] MathWorks, Inc., “Using MATLAB with Python,” [Online]. Available: https://uk.mathworks.com/products/matlab/matlab-and-python.html. [Accessed January 2021].

[57] M. Brookes, “VOICEBOX: Speech Processing Toolbox for MATLAB,” Imperial College London Department of Electrical & Electronic Engineering, [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. [Accessed November 2020].

[58] Vienna University of Technology, MIR group, “Rhythm & Timbre Feature Extraction from Music,” 2015. [Online]. Available: http://ifs.tuwien.ac.at/mir/musicbricks/#RPextract. [Accessed January 2021].

[59] T. Lidy and A. Rauber, “Evaluation of Feature Extractors and Psycho-Acoustic Transformation for Music Genre Classification,” in *6th International Conference on Music Information Retrieval (ISMIR)*, London, UK, 2005.

[60] Librosa development team, “Librosa documentation,” [Online]. Available: https://librosa.org/doc/latest/index.html. [Accessed January 2021].

[61] Music Technology Group, Universitat Pompeu Fabra, “Essentia: Open-source library and tools for audio and music analysis, description and synthesis,” [Online]. Available: https://essentia.upf.edu/. [Accessed January 2021].

[62] V. Perrault, “Timbre Toolbox,” Github, 2018. [Online]. Available: https://github.com/VincentPerreault0/timbretoolbox. [Accessed January 2021].

[63] R. Gong, “Code Repository for 'Timbre Analysis of Music Audio Signals with Convolutional Neural Networks',” GitHub, 2017. [Online]. Available: https://github.com/ronggong/EUSIPCO2017. [Accessed January 2021].