A Neural Network Approach for Synthesising Timbres from Adjectives

Alex Gounaropoulos and Colin G. Johnson
Computing Laboratory, University of Kent
Canterbury, Kent, England

Abstract—This paper describes a computer sound synthesis system, based on artificial neural networks, that constructs a mapping between adjectives and adverbs that describe timbres, and sounds having those timbres. This is used in two ways: firstly, to recognize the timbral characteristics of sounds supplied to the system, and, secondly, to make changes to sounds based on descriptions of timbral change.

I. INTRODUCTION

The system is used by firstly loading a sample of an instrument from a wave file. The audio is analysed to extract additive synthesis parameters from the audio, consisting of amplitude and detuning envelopes for each harmonic. However, this representation is unsuitable for use with neural networks. Not only is this representation complex, but worse still the number of synthesis parameters required for a sound depends on the duration of the audio sample. A longer sound requires more parameters than a short sound. A different representation is therefore used for compatibility with a neural network which has a constant number of inputs/outputs. This representation has a fixed number of parameters, and also reduces...
Fig. 1. An overview of the system: the relationships between words and sound features are learned indirectly.

Fig. 2. Neural network for recognition of timbral features.

Fig. 3. Neural network for sound synthesis control.
the complexity of the data by keeping only the most perceptually important information.

Once the sound has been analysed and the simplified representation created, the values are processed using the timbre recognition neural network. This network is trained to take a list of sound parameters as its input and map them onto outputs representing a description of the timbre. Each output value corresponds to a certain adjective, and the value ranges from 0 to 1 to indicate how strongly that characteristic is present in the sound. The graphical interface is then updated to display to the user how the timbre has been characterised. It is then possible to specify a change to one feature of the timbre, while keeping other characteristics of the sound unchanged.

Once the user has set the description of the desired sound, the system must modify the parameters of the synthesis algorithm in order to produce the required change in timbre. This involves using the synthesis neural network which has been trained to map a set of timbre description values onto the simplified sound representation. Finally, this data is used to transform the low-level additive synthesis parameters. The user can then use a MIDI keyboard to play the new sound through the synthesis engine.

III. SOUND REPRESENTATION

The primary requirements of the sound representation use in the system are:

- to reduce the dimensionality of the additive synthesis parameters
- to present the information in a form that makes significant patterns easier to identify
- to move from an additive synthesis representation which has a variable number of parameters to a fixed number of parameters which are suitable for use with a neural network.
- to be general enough to allow a useful range of sounds to be represented.

The representation is informed by various studies on timbre (e.g. [3]) that identify perceptually significant features. However, most of this work is concerned with timbre recognition, and many features that are useful for recognition do not work as well for synthesis. For example, the spectral centroid is a common measurement which measures the average frequency, weighted by amplitude, of a spectrum. This is useful for timbre classification tasks since it correlates strongly with brightness. While the spectral centroid can be calculated from a sound spectrum, it is obviously not possible to calculate an inverse to reconstruct the original spectrum from a spectral centroid value. This measure is clearly more suited to timbre categorisation tasks than synthesis.

The information required to represent a sound can broadly be classified as either spectral or temporal. Spectral features measure the relative amplitudes of the various frequency components in a sound. Temporal features are concerned with how a certain aspect of the sound changes over the duration of a note.

A training set is constructed by firstly choosing a selection of adjectives for the timbral characteristics that need to be controlled. A number of instrument samples are needed, each recording consisting of a single note. The sounds in the training set must adequately demonstrate all the timbral characteristics that need to be learnt. For instance, if one of the chosen adjectives is “brightness”, then a range of sounds ranging from very bright to very dull need to be included. Once a set of sounds has been

An overview of the sound representation designed for this system is shown in table I below.

A prominent feature of this system is that spectral information is stored using a list of peak amplitudes for the first 64 harmonics. This is a relatively low-level representation compared to the spectral measures commonly used in the literature, such as the spectral centroid or ratio of odd/even harmonics. This representation retains more of the information required for synthesis, whereas other measures are more suited to categorisation than synthesis tasks. It is a generic representation that does not need to make assumptions about what patterns in the harmonic amplitudes are perceptually significant. Instead, it is left to the neural network to discover patterns in the harmonics during training; this could include the patterns that are traditionally used as input into such algorithms, or it could discover new patterns. This generality comes at the cost of having a relatively high number of parameters to represent a sound, and requiring more work to successfully train the neural network.

The amplitude envelope of the sound is summarised using four measurements: attack time, release time, attack curve, and release curve. Figure 4 shows how these measures describe an amplitude envelope.

The amplitude envelope is the primary source of information for determining the articulation of the instrument sound. Plucked, bowed, and percussion instruments have recognisable envelope shapes. For instance, percussion sounds have short attack and release times, whereas plucked instruments have a short attack with a longer release. The model makes certain simplifying assumptions, namely that is an attack section where the amplitude increases, and this is followed by a release section where the amplitude fades towards silence. A best-fit curve is calculated for each section to approximate its shape.

IV. TRAINING THE NEURAL NETWORKS

A training set is constructed by firstly choosing a selection of adjectives for the timbral characteristics that need to be controlled. A number of instrument samples are needed, each recording consisting of a single note. The sounds in the training set must adequately demonstrate all the timbral characteristics that need to be learnt. For instance, if one of the chosen adjectives is “brightness”, then a range of sounds ranging from very bright to very dull need to be included. Once a set of sounds has been...
Table I

<table>
<thead>
<tr>
<th>Number</th>
<th>Value name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-64</td>
<td>Harmonic amplitudes</td>
<td>Peak amplitude of the first 64 harmonics</td>
</tr>
<tr>
<td>65</td>
<td>Harmonic damping</td>
<td>Rate at which the amplitude of higher harmonics decays compared to lower harmonics (i.e. sound becomes duller over time)</td>
</tr>
<tr>
<td>66</td>
<td>Attack time</td>
<td>Time taken for amplitude to reach maximum amplitude</td>
</tr>
<tr>
<td>67</td>
<td>Release time</td>
<td>Time from maximum amplitude to end of note</td>
</tr>
<tr>
<td>68</td>
<td>Amplitude attack curve</td>
<td>Describes the general shape of the attack section of the amplitude envelope</td>
</tr>
<tr>
<td>68</td>
<td>Amplitude release curve</td>
<td>Curve of the release part of the amplitude envelope</td>
</tr>
</tbody>
</table>

Summary of measures used to represent a sound.

When the training set has been assembled, the separate recognition and synthesis neural networks are trained. The sound samples are analysed and stored using the compact sound representation discussed earlier. The recognition network is then trained to map an input sound onto the timbre description given in the listening tests. The synthesis network is then trained to perform the inverse mapping, taking a timbre description as input and outputting a sound representation.

V. Conclusions

In this paper we have described a machine learning method for associating timbre words with sounds. This includes both networks for timbre recognition, and for modifying synthesis parameters based on a given set of words.

This system has been implemented and a series of listening trials are currently underway to determine the effectiveness of the system. A screenshot showing the interface can be found in figure 5. Results from an earlier version of the system, based on evolutionary computation methods, can be found in [2].

References