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

Timbre is one of the most complicated characteristics of music. By contrast to other characteristics (e.g. rhythm, pitch, volume) there is no clear way to map the space of timbres and the dimensions within that space, nor is it easy to agree on a way of notating timbre [1]. Nonetheless, acoustic instrumentalists communicate about timbre using a wide variety of adjectives and adverbs.

This lack of a systematic mapping between natural language and timbral features of sound makes it difficult for computer music systems to deal effectively with timbre. The aim of this paper is to describe a system that uses neural networks to learn the two core mappings involved in this relationship: the mapping from sounds to timbre-word descriptions, and the mapping from timbrewords to sounds and changes to sounds via a synthesis algorithm.

There are two basic ways in which the idea of timbre has been used, as discussed in our earlier paper [2]. The first of these is referred to as gross timbre, and refers to the placement of sounds into a number of large, discrete categories. The canonical example of this is the use of the word timbre to mean the sound of a particular instrument: "a violin timbre", for example. McAdams et al. [3] have developed a timbre-space for gross timbre, and work such as that by Kostek [4] has applied machine learning techniques to identify gross timbre. The main aim of this kind of work is recognition, i.e. mapping sounds to words; this is used, for example, in automatically tagging sound files with metadata.

The second main use of the term timbre is adjectival timbre [2]. This is the use of words that describe how a sound is different from a neutral sound, how sound changes over time, or the sound produced by a particular technique on an instrument. For example, words such as "metallic", "reedy", "harsh" and "bright" fit into this category. Typically, these are words that: (1) describe a material that produces sound of that type, (2) provide an analogy with some visual or textural feature of objects, or (3) describe some emotional or perceptual quality of the sound.

A small amount of work has been carried out on recognizing adjectival timbre: for example, the work by Disley and Howard [5] is aimed at recognising the features of pipe organ stops. Work by Etherington&Punch [6], Miranda [7] and Vertegaal&Bonis [8] is aimed at synthesising sounds given a particular timbre word or set of words—the work described in this paper can be seen as a continuation of that idea. In particular, we aim for a more general system whose implementation is not tied to a specific set of timbral characteristics as in [6] and [8], and can instead be trained by example to recognise and synthesise arbitrary characteristics.

The work in the remainder of this paper is focused on adjectival timbre, and describes the use of neural networks for timbre recognition and the synthesis of sounds with a particular timbre. The overall concept is illustrated in figure 1. The timbre classifier algorithm is trained on some listener-classified sound samples; this classifier algorithm is then used to train a second algorithm which represents the changes that need to be made to synthesis parameters in order to effect a particular change in the sound.

The remainder of the paper divides into three parts. Firstly, we give an overview of the system. Secondly, we discuss how sounds are represented in the system. Finally, we discuss how the neural networks are trained.

II. System Overview

The system uses a pair of neural networks at its core for recognition and synthesis of timbre (figures 2 and 3). Additive synthesis is used to produce sound, however the complex set of parameters for this synthesis algorithm is hidden from the user. Instead, the user interface provides simple controls labelled with timbral adjectives which are used to describe the desired timbre, and the system automatically adjusts the internal synthesis parameters appropriately.

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

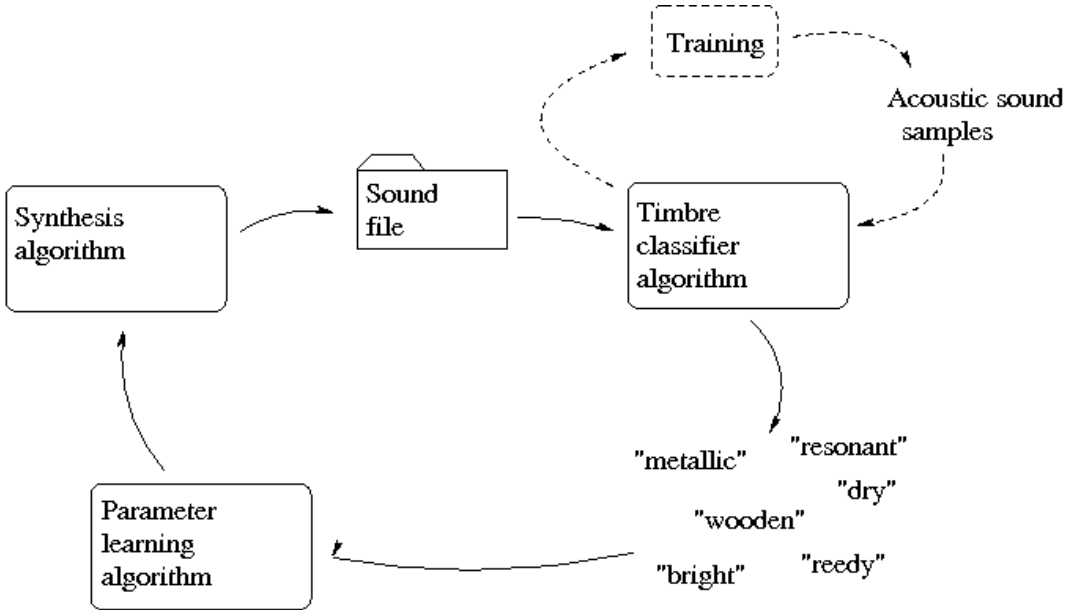

Fig. 1. An overview of the system: the relationships between words and sound features are learned indirectly.

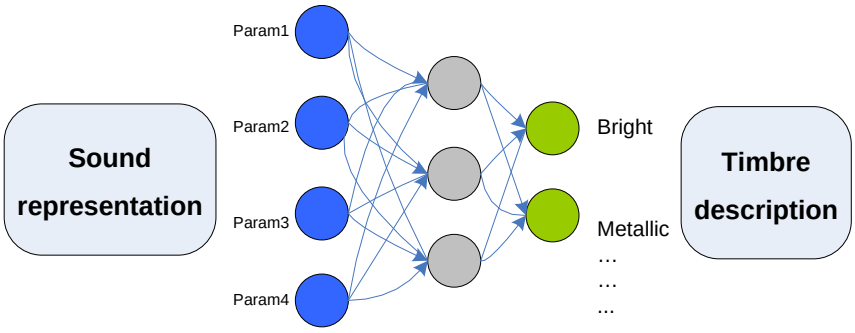

Fig. 2. Neural network for recognition of timbral features.

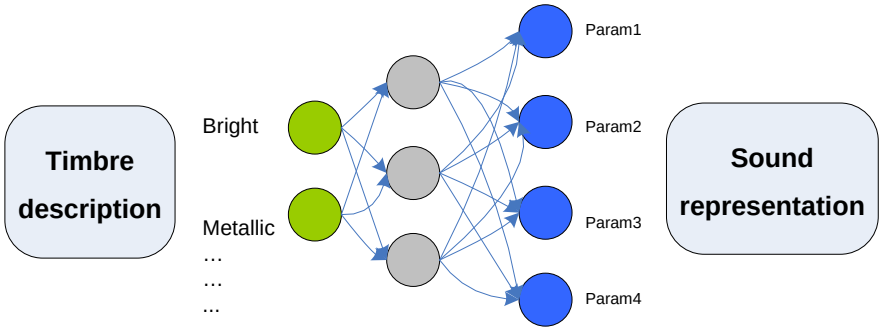

Fig. 3. Neural network for sound synthesis control.

reduces 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.

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.

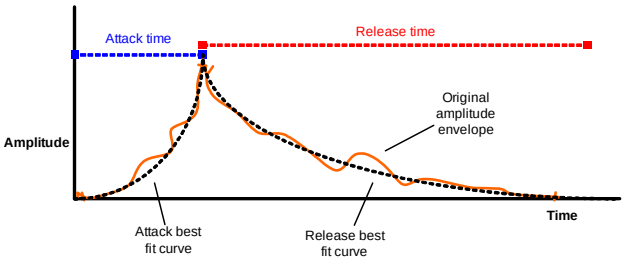

Fig. 4. Features of the amplitude envelope.

The amplitude envelope of the sound is summarized 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 collected, the training set is completed by performing listening tests to assign each sound a rating for each adjective. Ratings are in the range 0-1, with zero indicating the adjective does not apply to the sound, whereas a value of one is used when the characteristic

TABLE I.

SUMMARY OF MEASURES USED TO REPRESENT A SOUND

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

Fig 5. A screenshot of the application

is strongly present in the sound. For many of our experiments, we have used data from an experiment by Darke [9]. This consists of recordings of 15 orchestral instruments, along with listener ratings of the sounds for 12 adjectives (clear, brassy, bright, full, hard, harsh, metallic, muted, nasal, reedy, thin and wooden).

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

[1] Trevor Wishart. *On Sonic Art*. Harwood Academic Publishers, 1996. second edition, revised by Simon Emmerson; first edition 1985.

[2] Alex Gounaropoulos and Colin Johnson. Synthesising timbres and timbre-changes from adjectives/adverbs. In F. Rothlauf et al., editor, *Applications of Evolutionary Computing*. Springer, 2006.

[3] S. McAdams, S. Winsberg, S. Donnadieu, G de Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58:177–192, 1995.

[4] Bo´zena Kostek. *Soft Computing in Acoustics*. Physica-Verlag, 1999.

[5] A.C. Disley and D.M. Howard. Spectral correlates of timbral semantics relating to the pipe organ. *Speech, Music and Hearing*, 46, 2004.

[6] Russ Etherington and Bill Punch. SeaWave: A system for musical timbre description. *Computer Music Journal*, 18(1):30–39, 1994.

[7] Eduardo Reck Miranda. An artificial intelligence approach to sound design. *Computer Music Journal*, 19(2):59–75, 1995.

[8] Roel Vertegaal and Ernst Bonis. ISEE: An intuitive sound editing environment. *Computer Music Journal* 18(2): 21-29, 1994

[9] Graham Darke. Assessment of timbre using verbal attributes. In *Proceedings of the 2005 Conference on Interdisciplinary Musicology*, 2005. http://www.oicm.umontreal.ca/cim05/.