PITCH-AWARE REAL-TIME TIMBRAL REMAPPING

Dan Stowell
Queen Mary, University of London
Centre for Digital Music

Mark Plumbley
Queen Mary, University of London
Centre for Digital Music

ABSTRACT

We propose an approach to *timbral remapping*, a process which maps the timbre variations of one audio source onto the timbre variations of another, for real-time control of synthesis. Puckette [17] has made a foray into such a concept, but there are two important issues which must be addressed: how best to construct the timbre space for remapping purposes; and how to perform this remapping efficiently in real-time. We review some of the acoustical features used in the literature to represent timbre, and consider how to combine these usefully. We also describe some of our recent work on warping timbre space for effective coverage, and on an optimised real-time database lookup procedure suitable for use in live remapping.

Keywords – timbre, timbral analysis, performance, real-time

1. INTRODUCTION

Timbre is typically defined as the features of a sound which are not loudness and pitch [1][11], a vague definition which must include a wide range of perceptual qualities. In recent decades the perceptual and acoustic dimensions of timbre have been studied in various contexts, whether comparing different instruments [12], comparing different notes played on a single instrument [14], studying voice quality [13, chapter 3], or analysing the timbral similarity of music recordings [2]. Perceptual studies have found that some of the most readily identifiable dimensions of timbre include “brightness” (which correlates well with the *spectral centroid* measure) and attack time [21][12][18].

With recent improvements in timbre analysis and in computer power, the potential arises to analyse the timbre of a signal in real-time, and to use this analysis as a controller for synthesis or for other processes – in particular, the potential to “translate” the timbral variation of one source into the timbral variation of another source. We call this process *timbral remapping*. De Poli and Prandoni [4] made an early attempt at such control, and more recently Puckette [17] took the concept a little further.

In this paper we discuss some of our current work in this area. One of the main issues is the construction of a useful *timbre space* for the purpose of timbral remapping. There are many options as to which acoustic features to derive, and how to transform them, so as to provide a continuous space that provides useable control to the performer. Some features may exhibit interactions with pitch, and the variation of some features may depend strongly upon the type of source.

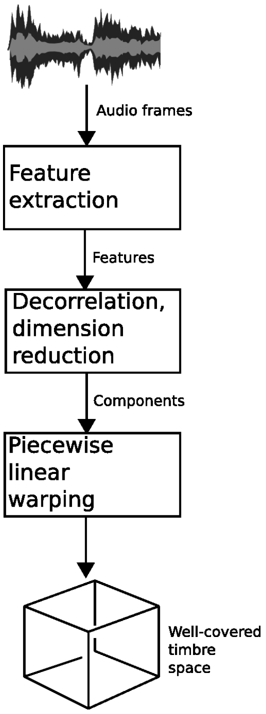

Figure 1. Constructing a timbre space

We also discuss issues connected with the real-time operation of such a system. Real-time operation imposes certain demands and restrictions, and our approach includes specific techniques which make the system useable in real-time. The requirement of extremely fast real-time table lookup in fact has a strong influence on the design of the system.

2. CONSTRUCTING A TIMBRE SPACE

In order to be able to map continuously from the timbre of one source to that of another, we must construct a timbre space in which to perform that mapping. Note that in this discussion we are restricting ourselves to “instantaneous” timbre measurements, applicable to each of a stream of audio frames. This does not necessarily exclude the use of temporal features (such as attack time), but in the present

Email: dan.stowell@elec.qmul.ac.uk

discussion we will not consider temporal features.

Constructing an acoustic timbre space typically consists of deriving one or more acoustic features, and then performing some kind of decorrelation and dimension reduction to transform those features into the "main" timbral components (e.g. [4], [2], [19]). In addition to this, we wish to transform the timbre spaces so that the timbre trajectory of one source can reliably be mapped to the timbre trajectory of another source, so we warp the dimensions of our timbral space in order to improve the likelihood that a good cross-mapping can be found in real time.

We now consider issues relating to each of these steps.

2.1. Which features to use?

Acoustical timbre analysis is still relatively immature, and so a variety of different measures have been investigated. Puckette's timbral remapping work [17] uses the energies in fixed frequency bands (suitably decorrelated) as acoustic features to represent timbre. One difficulty with such an approach is that because of the fixed filter frequencies, dependence upon pitch is likely, whereas we wish timbre to be pitch-independent.

Many researchers use Mel-frequency cepstral coefficients (MFCCs) (e.g. [4], [2]). MFCCs are well-known and exhibit useful properties such as preserving Euclidean distance and representing general spectral shape compactly.

There have however been criticisms of MFCCs. As they are designed to model spectral features such as vowel formants, they may be poorly suited to modeling noisy frames such as transients [9][2]. They are sensitive to the presence of noise in the signal [5], and may also be affected by pitch [2]. (MFCCs, as with Puckette's subband energies, are based on fixed-frequency filters.) Aucouturier and Pachet [2] find evidence of what they call a "glass ceiling", an upper limit to the accuracy of MFCC-based timbral similarity algorithms. Some researchers have proposed modifications to the MFCC algorithm to ameliorate some of these issues [5][20].

Instead of looking for a single set of features following from a single analysis method (e.g. from a cepstrum), it may be worthwhile to pursue the alternative approach of combining various features which have been found to characterise certain aspects of timbre. Features which have been found useful by musical timbre researchers include spectral centroid [21], spectral flatness [10], and subband spectral flatnesses [8].

A further source of acoustic timbral features is the field of "voice quality", the term generally used by voice researchers to refer to vocal timbre [13, chapter 3]. Measures that have been found to be useful for vocal timbre include the strength of the first interharmonic [6], the strength of the second harmonic [7][3], and the long-term balance of energies in different subbands [16]. Our research has an emphasis on the analysis of voice so these features are under consideration, although their generalisation to other types of signal may or may not be suitable.

Given this abundance of possible features, it becomes important to consider their relative merits, separately or in combination. The question also arises of how best to combine acoustic features which do not form a single orthogonal basis set and may be quite heterogenous. Decorrelation and dimension reduction, used appropriately, can help with this.

2.2. Decorrelation and dimension reduction

Whichever set of features is used, the set typically provides more dimensions than the experimenter wishes to handle, and the features often exhibit correlations, which are unhelpful when we want to represent timbral dimensions parsimoniously. Often these problems are overcome by the use of projection methods such as self-organising maps or Principal Component Analysis (e.g. [4], [19]). Both of these methods can be used to project a high-dimensional space down to a lower number of dimensions, while preserving most of the variation in the data.

In the present case we wish to project the data down to a relatively low-dimensional space so that it becomes feasible to map from input to output rapidly and usefully in real time. However, we do not need to project to as low as 2 or 3 dimensions, as is often done for visualisation purposes, or in timbral research to try and determine the "main" dimensions of timbre (e.g. [4]). In the present study we use Principal Component Analysis as a simple method to reduce the dimensionality of the data, although we are also considering other approaches.

Part of the motivation for our choice of Principal Component Analysis is to enable fast real-time mapping, as will be discussed further in section 3.1. It may be possible in future to optimise self-organising maps and similar systems to use fixed-point arithmetic etc.

2.3. Spanning the space: piecewise warping

We may wish to map between the timbre trajectories of very different signals. A standard way to normalise a feature space uses the minimum and maximum values retrieved, so that (for example) the minimum of each dimension is mapped to 0 and the maximum is mapped to 1. This does not in general guarantee that two datasets will overlap well in the timbre space: for example, the "centre of mass" of different datasets (e.g. the mean or median) may lie in very different positions in the space, which may lead to difficulties when attempting to map points from one dataset to points in the other. We must also consider the constraint that we may wish to use online processing, i.e. performing the mapping procedure before all data points have been received, and therefore without a definitive characterisation of the entire data distribution.

Piecewise linear warping offers a simple way to distort a dataset to cover a space in the desired manner. For example, in addition to linearly mapping so that the minimum becomes 0 and the maximum becomes 1, we could also move the mean or median to 0.5, ensuring that the center of mass of the distribution is near the centre of the space. In order to do that we would linearly map all points below

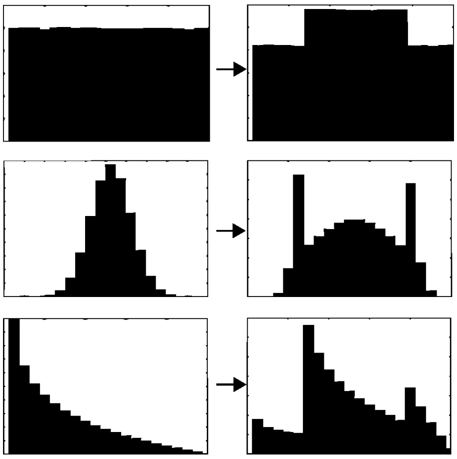

Figure 2. Examples showing the linear piecewise warping of some strongly differing distributions. The distributions become more similar in the way they span the space.

the median or mean to the 0–0.5 range, and separately linearly map the remaining points to the 0.5–1 range.

This piecewise linear remapping can be performed independently with each dimension, using as many pieces as desired. One interesting approach could be to use the positions of the ten-percentile, twenty-percentile, etc, to divide each dimension into ten regions which could then be uniformly mapped into the timbral space axis. Unfortunately percentile measurements are problematic to derive in a real-time online system which can take an unbounded number of input points, because they require the maintenance of sorted lists of values. Sorting algorithms typically take $O(n \log n)$ time to run, which becomes impractical as the number of items n grows large.

Instead we perform the warping using the statistics of minimum, maximum, mean and standard deviation, all of which can easily be calculated online for an unbounded number of inputs. A typical warping might remap the minimum to 0, the mean minus one standard deviation to 0.25, the mean to 0.5, the mean plus one standard deviation to 0.75, and the maximum to 1. This is illustrated in figure 2 for three simulated distribution types.

This warping is required because Principal Components Analysis is merely a linear transformation, and as such does not guarantee what type of distribution will result (e.g. whether it will have “long” or “short” tails along any particular axis). Methods such as self-organising maps can produce well-covered spaces but may lack the efficiency required, as described in the next section.

3. REAL-TIME IMPLEMENTATION

3.1. Fast table lookup

Once a large set of points have been mapped in the timbral space for the “target” synthesiser, and their corresponding synthesis parameters, we wish to be able to translate points from the timbral space “source” signal as fast as they are supplied, which may typically be on the order of 100 or 200 per second. While modern computers can easily perform acoustical analyses at this rate, database search algorithms require a large number of comparisons and data accesses and are thus relatively slow. We therefore require a very efficient database lookup to find a best matching parameter set for each of these points.

Puckette [17] uses minimum Euclidean distance as the criterion for selecting a point from the database. He suggests that this limits the number of analysis points to around 10,000; “we could thus estimate four synthesis parameters simultaneously to 10% resolution, for example”.

In order to facilitate fast search on an ordinary desktop computer, we implement a fixed-point (i.e. integer-based) lookup table with a highly efficient binary masking comparison search. The search proceeds by finding the smallest “cell” (cubic region) of the timbre space which contains both the input point and at least one output point. The dimension warping process helps increase the probability that any given cell (up to a moderate resolution) will contain an output data point, meaning that the output point will typically have closely-matching timbral coordinates.

If there are multiple output points in the cell, we have a number of options for selecting which to use, including Euclidean distance, L_1 (“city block”) distance, random choice, and tending to re-use a recently used point. The choice of which method to use may depend on aesthetic or performance considerations.

This lookup process is highly efficient because it uses simple fixed-point operations. Compare this against the self-organising map, which, although it is relatively efficient once trained, requires the calculation of N Euclidean distances for each input point (where N is the number of map neurons).

3.2. Implementation

Our implementation is being developed using SuperCollider 3 [15]. We currently use GNU Octave to calculate the principal components of datasets (in an offline fashion), which are then passed to SuperCollider which processes audio data in real time. SuperCollider handles feature analysis and remapping, thanks in part to custom C++ plugins written by the first author which enable the fixed-point table lookup.

This implementation is easily capable of lookup in a dataset containing hundreds of thousands of points, hundreds of times per second, on an ordinary desktop computer.

The first author has composed a musical piece based on this implementation, in which the voice is remapped onto the sounds of an 8-bit computer system, for performance in Summer 2007.

4. CONCLUSIONS AND FURTHER WORK

We have proposed an approach to real-time *timbre remapping* which aims to improve on previous work by providing a method for acoustic timbral features (after decorrelation and dimension reduction) to be mapped and retrieved in a well-covered timbral space, which can be queried extremely rapidly in real-time.

We have also considered which acoustic features to use. MFCCs are well known and commonly used but have some potential weaknesses. Timbre research and voice quality research offer a selection of other features that can be useful in characterising timbre. Future work will aim to elucidate the utility of various of these features, including their resilience to pitch variation.

5. ACKNOWLEDGEMENTS

Dan Stowell is supported by the EPSRC under a Doctoral Training Account studentship.

6. REFERENCES

[1] ANSI. *Acoustical terminology*. Number S1.1-1960. American National Standards Institute, New York, 1960.

[2] J.-J. Aucouturier and P. F. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[3] B. Blankenship. The time course of breathinesss and laryngealization in vowels. Master's thesis, Department of Linguistics, UCLA, 1997.

[4] G. De Poli and P. Prandoni. Sonological models for timbre characterization. *Journal of New Music Research*, 26(2):170–197, 1997.

[5] B. Gajic and K. K. Paliwal. Robust feature extraction using subband spectral centroid histograms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, volume 1, pages 85–88, 2001.

[6] B. R. Gerratt and J. Kreiman. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, 29(5):365–381, October 2001.

[7] H. M. Hanson. *Glottal characteristics of female speakers*. PhD thesis, Division of Applied Sciences, Harvard University, 1995.

[8] J. Herre, E. Allamanche, and O. Hellmuth. Robust matching of audio signals using spectral flatness features. In *Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA-2001)*, pages 127–130, 2001.

[9] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai. Music type classification by spectral contrast feature. In *Proceedings of the International Conference on Multimedia and Expo (ICME '02)*, volume 1, pages 113–116, 2002.

[10] J. D. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2):314–323, 1988.

[11] J. Kreiman, D. Vanlancker-Sidtis, and B. R. Gerratt. Defining and measuring voice quality. In *Proceedings of From Sound To Sense: 50+ Years of Discoveries in Speech Communication*, pages 115–120. MIT, June 2004.

[12] S. Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics*, 62(7):1426–39, 2000.

[13] J. Laver. *The Phonetic Description of Voice Quality*. Cambridge Studies in Linguistics. Cambridge University Press, 1980.

[14] M. Loureiro, H. de Paula, and H. Yehia. Timbre classification of a single musical instrument. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR 2004)*, pages 546–549, 2004.

[15] J. McCartney. Rethinking the computer music language: SuperCollider. *Computer Music Journal*, 26(4):61–68, 2002.

[16] M. Nordenberg and J. Sundberg. Effect on LTAS of vocal loudness variation. Technical Report 45, KTH Voice Research Centre, Department of Speech Music and Hearing, KTH, 2003.

[17] M. Puckette. Low-dimensional parameter mapping using spectral envelopes. In *Proceedings of the International Computer Music Conference (ICMC'04)*, pages 406–408, 2004.

[18] E. Schubert, J. Wolfe, and A. Tarnopolsky. Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC 04)*, pages 654–657, North Western University, Illinois, 2004.

[19] S. Shirashi. A real-time timbre tracking model based on similarity. Master's thesis, Institute of Sonology, The Hague, June 2006.

[20] V. Tyagi and C. Wellekens. On desensitizing the Mel-cepstrum to spurious spectral components for robust speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, volume 1, pages 529–532, 2005.

[21] D. Wessel. Timbre space as a musical control structure. *Computer Music Journal*, 3(2):45–52, 1979.