Towards a Theoretical Framework for Sound Synthesis based on Auditory-Visual Associations

Kostas Giannakis, Matt Smith
School of Computing Science
Middlesex University - Bounds Green
London N11 2NQ - United Kingdom
T: (+44) (0)181 362 6727
F: (+44) (0)181 362 6411
k.giannakis@mdx.ac.uk; m.r.smith@mdx.ac.uk

Abstract

In this paper, we provide a critical review of research efforts that attempt to identify the high-level perceptual dimensions of two distinct sensory percepts, musical timbre and visual texture. These dimensions are tested against a number of evaluation criteria in order to define appropriate sets for further empirical investigation of auditory-visual associations.

1 Introduction

Our age is characterised by a growing interest in the application of computers in arts. Computer technology has become an everyday tool for creative expression in almost every area of human activity. As a result, traditional ways of expression have been transformed and most importantly new forms of electronic art have been introduced. Music, both as a form of art and a field of science, has undergone significant changes due to the application of computers in areas such as sound analysis and synthesis, composition, performance, etc.

Computers can generate sounds either for the imitation of acoustic instruments or the creation of new sounds with novel timbral properties. Almost 50 years of research in computer music laboratories worldwide has resulted in the development of a large body of diverse sound synthesis techniques, e.g. additive synthesis, subtractive synthesis, physical modelling, granular synthesis (see Roads, 1996). Usually, a synthesis technique comprises a set of low-level parameters related to Digital Signal Processing (DSP) modules (*unit generators* in computer music jargon) such as oscillators, filters, etc. A common characteristic of synthesis techniques is that a sound is represented as an object consisting of a large number (hundreds or thousands) of short sub-events that can be controlled by numerous time-varying parameters. Inevitably a vast amount of musical data must be defined and modified (ideally in real time) making the process of creating a sound object a very complex, non-musical and tedious task. Therefore, although it is theoretically possible to create almost any sound using one or more techniques, the main problem with current computer sound synthesis is how the composer interacts with the system in order to create and manipulate sound. This problem becomes even more apparent and complicated when we consider the auditory dimension of timbre and its significance as a compositional tool. Traditional music compositional processes have focused mainly on pitch and duration, and treated timbre as a second-order attribute of sound (Wishart, 1996). However, the development of computer-based sound synthesis tools paved the way for a more tractable exploration of musical timbre. Although current sound design tools provide very sophisticated control of the low-level parameters of sound, there is very little investigation on how a perceptual model of timbre can be incorporated in the design process.

In this paper, we argue that recent developments in the understanding of our auditory mechanisms as well as studies related to our sensory experience can provide a useful theoretical framework for the design of sounds in ways that have a strong cognitive basis. Our research focuses on the mapping between perceptual dimensions of auditory and visual percepts. We propose a metaphor for sound design and representation in computers, which is based on *auditory-visual* associations.

First, we set out to find an appropriate set of perceptual dimensions that can be used for the intuitive control and manipulation of timbre. Second, we argue that a similar set of dimensions can be constructed on the basis of results from studies in the perception of visual texture. Finally, we outline the design of an experiment for the empirical investigation of the cognitive associations between timbre and visual texture.

1.1 A Visual language for Sound Synthesis

The significant role of visual communication in computer applications is indisputable. In the case of music it seems that it is very natural for musicians to translate

non-visual ideas into visual codes (see Walters (1997) for examples of graphic scores from J. Cage, K. Stockhausen, I. Xenakis, and others). In the past, associations between auditory and visual elements (e.g. Isaac Newton's colour-pitch associations) inspired a new artistic movement under the title of *visual music* (e.g. Wells, 1980; Goldberg and Schrack, 1986; Peacock, 1988; Whitney, 1980; Pocock-Williams, 1992). In the domain of sound synthesis, modern systems incorporate graphical editors (e.g. for the drawing of waveforms[1]) and/or on-screen interconnections of graphical objects (e.g. oscillators, filters, etc.). However, the utilisation of a visual language for sound synthesis is still based more on low-level acoustic information (i.e. time-domain and frequency domain representations) and less on how composers actually conceive and externalise compositional ideas that are primarily based on high-level perceptual experiences.

It is useful here to distinguish between two modes of conception: *low-level* and *high-level*. In the low-level mode, a composer conceives and plans compositional actions in terms of DSP instruments using one or more sound synthesis techniques. Subsequently, the composer may use either a low-level (e.g. Csound (Vercoe, 1986)) or a combination of low and high-level interfaces (as in the case of ARTIST (Miranda, 1994) where users first design a DSP instrument and then control it using a natural language interface). In the high-level mode of conception, the internal musical idea may be completely abstract (e.g. a velvety sound). Again, this idea may be externalised by using a low-level interface, a high-level interface, or a combination of both. However, there is no direct high-level interface to match the composer's abstract musical idea. In between musical ideas and sound generated by computers lies the control of a synthesis technique. As a result, the focus of composers has shifted from the high-level musical task of sound design to the low-level and cumbersome process of understanding and controlling the sound production mechanism idiomatic to each synthesis technique. In this study, we argue that a visual language that is based on an investigation of auditory-visual associations can provide a direct high-level interface for the control and manipulation of sound.

As an introduction to auditory-visual associations we present research efforts on colour-sound associations. Previous attempts to model sound using colour (e.g. Padgham, 1986; Caivano, 1994) were based on correspondences that may exist between the physical dimensions of sound and colour. For example, in Caivano's approach, hue is associated with pitch since both these dimensions are closely related to the dominant wavelengths in colour and sound spectra respectively. In the same manner, pure (or high-saturated) colours are associated with pure (or narrow bandwidth) tones whereas low-saturated colours (those that involve wider bandwidths of wavelength) are associated with complex tones and noise. Finally, colour lightness is associated with loudness (black and white represent silence and maximum loudness respectively with the greyscale representing intermediate levels of loudness). In further studies to empirically investigate the validity of these associations, Giannakis and Smith (2000) suggested that pitch and loudness can be predicted by colour lightness and saturation respectively (see Figure 1). This latter study was only concerned with pitch and loudness and involved the use of pure tones in order to neutralise the effect of timbral richness.

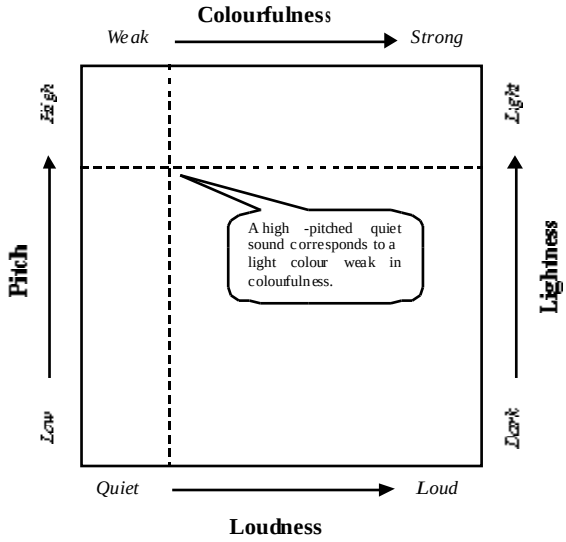

Figure 1: Proposed space for the associations between pitch-lightness and loudness-colourfulness based on Giannakis & Smith (2000).

A natural extension of the above-described studies is the empirical investigation of the associations between dimensions of timbre and dimensions of other visual percepts such as shape, texture, etc. Pitch and loudness are well understood auditory dimensions and both can be ordered on a single scale. In contrast, the perception of timbre is a more complex and multidimensional phenomenon. Recently, visual texture has been proven effective when used in the visualisation of multidimensional data sets (e.g. Ware and Knight, 1992; Healey and Enns, 1998). We have also identified a number of important similarities between timbre and visual texture that suggest further investigation of the potential cognitive associations between these sensory percepts. These similarities will be discussed in more detail in the remaining sections of this paper.

1.2 The Perception of Timbre

Timbre has been defined by the American Standards Association (1960) as "that attribute of auditory sensa-

[1] E.g. the UPIC system by Xenakis (1992)

tion in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". This definition has been strongly criticised (e.g. Bregman, 1990; Slawson, 1985) for being too general and ill-defined. In fact, the term 'timbre' is used in a variety of contexts and it is extremely difficult to agree on a single definition. For example, timbre may refer to a class of musical instruments (e.g. string instruments as opposed to brass instruments), a particular instrument in this class (e.g. violin), a particular type of this instrument (e.g. Stradivarius), the various ways of playing this instrument in order to change the resulting timbre, and so on. Therefore, any attempts to investigate the dimension of timbre should clarify which aspect(s) of timbre are addressed. In this study, timbre is defined as that perceptual attribute which pertains to the steady-state portions of sound. Although temporal characteristics are equally important and necessary for a complete description of timbre (see Grey, 1975), they are more related with the identification of sound sources and their intrinsic behaviour rather than the qualitative characteristics of timbre that are hidden in the steady-state spectrum of sounds.

Many studies attempted to identify the prominent dimensions of timbre (e.g. Bismarck, 1974a,b; Grey, 1975; Plomp, 1976; Ehresman and Wessel, 1978; Slawson, 1985; McAdams, 1999). These studies suggest that there is a limited number of dimensions on which every sound can be given a value, and that if two sounds have similar values on some dimension they are alike on that dimension even though they might be dissimilar on others. However, there is no agreement (with the exception of the dimension of *sharpness*) on the dimensions of timbre that these studies proposed. This is mainly due to the different sets of sounds that were used as stimuli in the experiments (e.g. instrument tones as opposed to synthetic tones) and the different time portions of the sounds that were investigated (e.g. attack transients as opposed to steady states). As a result, these findings hold very well for the limited range of sounds that they refer to but they lack generality of application (see also Bregman, 1990). Nevertheless, these studies suggest that timbre depends on certain characteristics of the sound spectra. Based on the above we can suggest that spectral models for sound synthesis can form the basis for a more intuitive approach to sound design. Spectral models are based on perceptual reality, can be controlled by perceptual parameters, and they provide a general model for all sounds in an analysis/synthesis form (Serra, 1997a). However, research in this area is still young and there is no definite set of appropriate perceptual parameters that can be used effectively in sound synthesis systems.

In the remainder of this section, we first describe dimensions[2] of timbre that have been proposed by the above-described studies and then test them against a number of evaluation criteria. Our goal is to define a set of perceptual dimensions that can be incorporated in further investigations. These criteria can be summarised as follows:

- *Empirical support*. This criterion tests whether a proposed dimension is supported by experimental work.

- *Independence*. This criterion tests whether a perceptual dimension is orthogonal (i.e. independent of changes in other dimensions) or it is somehow correlated with other dimensions (in which case, we can talk of composite dimensions).

- *Measurability*. This criterion tests the existence of concrete measurement methods for perceptual dimensions.

- *Synthesisability*. This is related to the criterion of measurability and refers to existing or potential models of synthesis algorithms that control perceptual dimensions.

1.2.1 Sharpness

Sharpness (other terms include auditory brightness, spectral centroid, etc.) is the most prominent dimension of timbre suggested by the above-described studies. For pure tones, sharpness is determined by the fundamental frequency, i.e. the higher the fundamental frequency, the greater the sharpness. In the case of complex tones, the determining factors for sharpness are the upper limiting frequency and the way energy is distributed over the frequency spectrum, i.e. the higher the frequency location of the spectral envelope centroid, the greater the sharpness (Bismarck, 1974b).

1.2.2 Compactness

Compactness is a measure of a sound on a scale between complex tone and noise, i.e. the difference between discrete and continuous spectra. However, the formulation of such a scale has been proven difficult (e.g. Bismarck, 1974a). Compactness is also related to the concept of periodicity. An ideal periodic (or harmonic) spectrum contains energy only on exact integer multiples of the tone's fundamental frequency. Malloch (1997) suggested that *cepstrum* analysis as a method to measure the periodicity of a sound could also be used in the measurement of compactness.

[2] Dimensions of timbre that refer to temporal characteristics (e.g.) have been excluded from this discussion for the reasons stated earlier.

1.2.3 Spectral Smoothness

Spectral smoothness is a dimension of timbre discussed in McAdams (1999). It describes the shape of the spectral envelope and it is a function of the degree of amplitude difference between adjacent partials in the spectrum of a complex tone. Therefore, large amplitude differences produce jagged envelopes, whereas smaller differences produce smoother envelopes. A formula for the measurement of spectral smoothness can be found in McAdams (1999).

1.2.4 Roughness

Roughness is related to the phenomenon of *beats*. When two pure tones with very small difference in frequency are sounded together, then a distinct beating occurs that gives rise to a sensation of sensory dissonance (Sethares, 1999). In a series of experiments with pairs of pure tones, Plomp (1976) found that roughness reaches its maximal point at approximately 1/4 of the relative critical bandwidth. For complex tones, roughness can be estimated as the sum of all the dissonances between all pairs of partials (see Sethares, 1999).

1.2.5 Discussion

As far as the criterion of empirical support is concerned, the above-described dimensions are the results of rigorous empirical investigations involving musical and/or non-musical subjects within the limitations of their sound stimuli. Usually, two experimental techniques have been employed: multidimensional scaling (e.g. Grey, 1975; Plomp, 1976; Ehresman and Wessel, 1978; McAdams, 1999) and semantic differential scales (e.g. Bismarck, 1974a). The former is based on subjects' judgements of similarity or dissimilarity between sets of stimuli. These similarities and dissimilarities are then represented in the form of a geometric configuration. In semantic differential techniques, subjects are asked to rate sounds along bipolar scales such as sharp-dull, hard-soft, etc. Based on the above, we can argue that all the presented dimensions of timbre satisfy the criterion of empirical support. However, not all dimensions appear to be independent of each other. For example, roughness has been studied as an individual perceptual dimension and has not been part of a larger set of orthogonal dimensions that is usually produced by multidimensional scaling techniques. The dimension of sharpness has been found orthogonal to compactness in studies by Bismarck (1974a) and orthogonal to spectral smoothness in studies by McAdams (1999). Therefore, further investigation is needed to tell whether compactness and spectral smoothness are orthogonal themselves or correlated in some way. These studies have proposed ways of measuring perceptual dimensions of timbre and the reader is referred to the individual studies for more detailed information on computational issues. Finally, these studies have been mainly concerned with the analysis of sounds and there is no explicit discussion about the synthesisability of these dimensions. However, it seems feasible to create synthesis algorithms that are based on the measurement formuli.

1.3 The Perception of Visual Texture

Even though texture is an intuitive concept, an exact definition of texture either as a surface property or as an image property has never been adequately formulated. In this study, texture is considered as a visual percept.

In vision research, there are two main computational approaches to the analysis of texture: the *stochastic* approach and the *structural* approach. The stochastic approach relies primarily on pre-attentive viewing (i.e. when textures are viewed in a quick glance) and is based on statistics and the theory of probability. In the structural approach, a texture is composed of a primitive pattern that is repeated periodically or quasi-periodically over some area. The relative positioning of the primitives in the pattern are determined by placement rules. In a different attempt, Francos et al. (1991) describe a texture model which unifies the stochastic and structural approaches. This model allows the texture to be decomposed into three orthogonal components: a harmonic component, a global directionality component, and a purely non-deterministic component[3].

There are a small number of studies that attempt to identify the perceptual dimensions of visual texture. An early study by Tamura et al. (1978) suggested coarseness, contrast, and directionality as the most prominent perceptual dimensions of texture. The same study also constructed mathematical models for the above dimensions based on extensive psychometric studies. However, the proposed dimensions were based on the authors' subjective views and therefore the question of whether humans use these dimensions in texture judgements was not adequately answered. In another study, Ware and Knight (1992) proposed a visualisation method based on a set of texture dimensions comprising orientation, size, and contrast. Again, the selected dimensions were not empirically derived. In order to address this problem Rao and Lohse (1996) performed a series of experiments that tried to identify the high-level dimensions of texture perception by humans using a variety of experimental designs and statistical methods (e.g. multidimensional scaling, hierarchical clustering, principal components analysis) to analyse and support their results. This latter study confirmed some of the dimensions proposed by earlier

[3] Note the similarity of this approach with a sound synthesis model proposed by Serra (1997b) based on a deterministic plus stochastic model.

studies as being prominent in texture perception. The perceptual space proposed by Rao and Lohse (1996) comprises the following three orthogonal dimensions:

- *Repetitiveness.* This dimension refers to the way primitive elements are placed and repeated over a texture image. The degree of repetition (e.g. periodic, quasi-periodic, random) can be specified and controlled by placement rules.

- *Contrast and Directionality*. This is a composite dimension due to a high correlation coefficient. Contrast is related with the degree of local brightness variations between adjacent pixels in an image (i.e. sharp vs. diffuse edges). The directionality of a texture is a function of the dominant local orientation within each region of texture.

- *Granularity, Coarseness, and Complexity*. These dimensions are very similar to each other (this is supported by high correlation coefficients) and refer to the size (small vs. large) and structure (fine vs. coarse) of the texture grains.

Based on the above discussion and using the same criteria as for the perceptual dimensions of timbre we can suggest that the model proposed by Rao and Lohse (1996) satisfies all the criteria and therefore consists a suitable set of dimensions for the description of visual texture.

1.4 Conclusions and Further Work

In this paper, working towards a theoretical framework of auditory-visual associations, we attempted to combine the findings of various studies in the perception of timbre and visual texture. Timbre and visual texture have been shown to share some very important characteristics. First, timbre and visual texture are both multidimensional perceptual phenomena that can be described by a small set of prominent dimensions. Second, studies in both fields are based on a similar research methodology, i.e. rigorous empirical investigations of how humans perceive and describe sensory percepts. The findings of these studies were evaluated using a number of important criteria. Table 1 summarises the two sets of perceptual dimensions we propose as suitable for further investigation.

Table 1: Identified perceptual dimensions of musical timbre and visual texture.

Timbre	Texture
Sharpness	Repetitiveness
Compactness	Contrast, Directionality
Spectral Smoothness	Granularity, Coarseness, and Complexity
Roughness	

The next step in our research is to design and conduct an experiment based on these sets of dimensions for timbre and visual texture. The objective of this experiment is the identification and investigation of associations (if any) between these auditory-visual dimensions. Based on our discussion, a number of initial hypotheses for such associations can be made. For example, sharpness may be related with contrast, periodicity with repetitiveness, roughness with granularity and coarseness, and finally compactness with complexity. In this experiment, the sound stimuli will consist of steady-state timbres varying systematically in one or more dimensions. Similarly, a collection of textures with varying dimensions will be used for a timbre-texture association task. These textures can be either generated by computer-based texture synthesis algorithms or they can be selected from larger texture databases that are extensively used in texture perception research (e.g. Brodatz textures). The results will provide significant evidence on the analogies between the dimension sets proposed in this paper and will contribute towards the development of a cognitively useful visual language for sound synthesis.

Acknowledgements

Thanks to Prof. Ann Blandford for comments and suggestions on initial drafts of this paper. This research is supported by a Middlesex University research studentship.

References

American Standards Association. *Acoustical Terminology SI, 1-1960*. American Standards Association, 1960

G. von Bismarck. Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes. *Acustica,* 30:146-159, 1974a

G. von Bismarck. Sharpness as an Attribute of the Timbre of Steady Sounds. *Acustica,* 30:159-172, 1974b

A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT Press, 1990

J. L. Caivano. Colour and Sound: Physical and Psychophysical relations. *Colour Research and Applications,* 19(2):126-132, 1994

D. Ehresman and D. Wessel. Perception of Timbral Analogies. *IRCAM* Reports, 1978

J. Francos, A. Meiri, B. Porat. Modelling of the Texture Structural Components Using 2-D Determi-

nistic Random Fields. *Visual Communications and Image Processing*, Vol. SPIE 1666:554-565, 1991

K. Giannakis, M. Smith. *Imaging Soundscapes*. In press, 2000

T. Goldberg, G. Schrack. Computer-Aided Correlation of Musical and Visual Structures. *Leonardo,* 19(1):11-17. Pergamon Press, 1986

J. M. Grey. *Exploration of Musical Timbre*. PhD Dissertation. Report No. STAN-M-2. CCRMA. Stanford University, 1975

C. Healey, J. Enns. Building Perceptual Textures to Visualize Multidimensional Datasets. In Proceedings *IEEE Visualization 1998*

S. N. Malloch. *Timbre and Technology: An Analytical Partnership*. PhD Dissertation. University of Edinburgh, 1997

S. McAdams. Perspectives on the Contribution of Timbre to Musical Structure. *Computer Music Journal*, 23(3):85-102. MIT, 1999

E. R. Miranda. *An Artificial Intelligence Approach to Sound Design.* PhD Dissertation. University of Edinburgh, 1994

C. Padgham, C. The Scaling of the Timbre of the Piping Organ. *Acustica* 60: 189-204, 1986

K. Peacock. Instruments to Perform Colour-Music: Two Centuries of Technological Experimentation. *Leonardo,* 21(4):397-406. Pergamon Press, 1988

R. Plomp. *Aspects of Tone Sensation*. Academic Press, 1976

L. Pocock-Williams. Toward the Automatic Generation of Visual Music. *Leonardo,* 25(1):445-452. Pergamon Press, 1992

A. R. Rao, G. L. Lohse. Towards a Texture Naming System: Identifying Relevant Dimensions of Texture. *Vision Research*, 36(11): 1649-1669. Pergamon Press, 1996

C. Roads. *The Computer Music Tutorial*. MIT Press, 1996

X. Serra. Current Perspectives in the Digital Synthesis of Musical Sounds. *Formats 1.* Pompeu Fabra University, 1997

X. Serra. Musical Sound Modelling with Sinusoids plus Noise. In C. Roads et al. (eds), *Music Signal Processing.* Swets & Zeitlinger, 1997

W. A. Sethares. *Tuning, Timbre, Spectrum, Scale*. Springer-Verlag, 1999

W. Slawson. *Sound Color*. University of California Press, 1985

H. Tamura, S. Mori, T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man, and Cybernetics,* 8:460-473, 1978.

B. Vercoe. *Csound*. Computer Music Application. MIT, 1993

J. L. Walters. Sound, Code, Image. *EYE* magazine, 7(26):24-35. Quantum Publishing, 1997

C. Ware, W. Knight. Orderable Dimensions of Visual Texture for Data Display: Orientation, Size, and Contrast. *CHI - ACM Conference on Human Factors in Computing Systems*, 1992

A. Wells. Music and Visual Colour: A proposed correlation. *Leonardo,* 13(1):101-107. Pergamon Press, 1980

J. H. Whitney. *Digital Harmony*. Byte Books, 1980

T. Wishart. *On Sonic Art.* Revised Edition. Harwood Academic Publishers, 1996

I. Xenakis. *Formalized Music.* Revised edition. Pendragon Press, 1992