ICSV14
Cairns • Australia
9-12 July, 2007

14th
International Congress on
Sound & Vibration

TOWARDS A MUSIC SYNTHESIZER CONTROLLED BY TIMBRAL ADJECTIVES

David M Howard, Alastair C Disley and Andy D Hunt

Audio Lab, Intelligent Systems Research Group, Department of Electronics, University of York,
Heslington, York, YO10 5DD, United Kingdom
dh@ohm.york.ac.uk

Abstract

For many musicians, the control parameters associated with electronics synthesis systems are not conducive to their creative processes. Filter setting, resonance controls, frequency or amplitude modulation settings, different “raw” waveforms and varieties of noise are examples of what have emerged as synthesizer controllers designed by engineers. Musicians describe the timbre of sounds using everyday adjectives such as bright, mellow, brash, warm and fuzzy. This paper will describe the development of a pilot music synthesis system based on Pure Data (PD) which uses timbral adjectives as its controllers. The effect that the controls have on the acoustic output is based on a series of listening tests carried out with musicians to elicit commonalities in how they use adjectives to describe the sounds of existing musical instruments. The listening tests were carried out over the internet. To validate the use of the internet in this way, a control group were asked to carry out the listening test alone and in the presence of an investigator in order to establish whether there was any difference in the data obtained. The average differences were smaller than the step size of the test itself, confirming the validity of the internet as a listening test vehicle. The listening test data have been analysed using multidimensional scaling and principal component analysis to establish which adjectives account for the greatest degree of segregation across the listeners and the sounds under investigation. These adjectives are the controllers for the synthesizer, and a number of sound examples will be played to illustrate the effectiveness of the final prototype synthesis system.

1. INTRODUCTION

For the performing musician or composer who wishes to make use of electronic instruments, a wide variety of commercial instruments and software tools currently exist for sound synthesis. However, the majority of these make use of user parameters that control directly low-level aspects of the synthesis process itself such as (depending on the synthesis method employed): the waveshape; settings of one or more oscillators; cut-off frequencies of filters; fundamental frequency; resonance frequency; or time envelope. The relation between such parameters and an individual’s perception of the resulting sound is often both complex, difficult to pin down and non-linear.

Some synthesis techniques, such as frequency modulation (or FM) synthesis, are especially renowned amongst musicians as having no obvious direct relation between the output sound and the parameters employed: *“there is no straightforward perceptual relationship between the modulation values and the timbre produced, and hence it is difficult to use FM synthesis to create a specific sound that you might require”* [1]. Consequently, it can often be very difficult for someone synthesizing sounds to predict in advance how the output sound will be perceived.

Contemporary technology has developed to a point described in [2] where *“the issue is no longer what sound one can produce, but what sound one chooses to produce. The listener and the listener’s perception become the central criteria for artistic choice”*. Risset also provides examples of the complex relation between physical parameters and perception: *“I have, for instance, produced sounds that seem to go down in pitch when their frequencies are multiplied by two. … Similarly, I have produced beats that seem to slow down when one doubles the speed of the tape recorder on which they are played”* [3].

It can be argued that it is theoretically perfectly possible with digital technology to reproduce any sound given the right sequence of samples, but actually finding this sequence is a formidable task. Knowledge of acoustics and psychoacoustics help us to explore the potential of digital technology in a much more effective way, confirming that an interdisciplinary approach is crucial [4]. Musicians are familiar with the notions of pitch, loudness and timbre which are commonly used in practice. Pitch relates to issues such as: notes on a score, key, melody, harmony, tuning systems, and intonation in performance. Loudness relates to matters such as: musical dynamics (e.g. pianissimo, piano, mezzo piano, forte, fortissimo etc.), or the balance between members of a musical ensemble, whether individual parts, choir and orchestra, or soloist and accompaniment. Timbre relates to sound quality descriptions such as: mellow, rich, covered, open, dull, bright, dark, strident, grating, harsh, shrill, sonorous, sombre, colourless or lacklustre. Timbral descriptors are therefore used to indicate the perceived quality or tonal nature of a sound [5].

There is no single subjective rating scale against which timbre judgements can be made; this is completely unlike pitch and loudness which can typically be reliably rated by listeners on scales from *“high”* to *“low”*. The American National Standards Institute [6] formally defines timbre as: “*Timbre is that attribute of auditory sensation in terms of which a listener can judge two sound similarly presented and having the same loudness, pitch and duration as being dissimilar.*” In other words, two sounds that are perceived as being different but which have the same perceived loudness, pitch, and duration are said to differ by virtue of their timbre.

The timbre of a note is the aspect by which a listener recognises the instrument which is playing a note. The definition given by Scholes [7] encompasses some timbral descriptors: “*Timbre means tone quality - coarse or smooth, ringing or more subtly penetrating, “scarlet” like that of a trumpet, “rich brown” like that of a cello, or “silver” like that of the flute ... The one and only factor in sound production which conditions timbre is the presence or absence, or relative strength or weakness, of overtones”*. Timbral descriptions of sounds and their relationship to the acoustic nature of the sounds themselves can be unique to an individual or they might be shared more widely between listeners in terms of the sounds they describe. Miranda et al. [8] propose a new taxonomy for the timbres produced by the *Chaosynth* software synthesiser which is based on cellular automata. The reason they require this is that Chaosynth can produce extremely complex sounds. Examples of defined classes of sounds in their proposed taxonomy include: *chaotic*, *explosive*, *general textures*. They have further observed that: *“potential users have found it very hard to explore its possibilities as there is no clear referential framework to hold on to when designing sounds”*.

A number of commonly used adjectives have been studied in relation to how listeners describe timbre, and some of which have been found to share objectivity amongst a majority of listeners, and the acoustic correlates have been established. Timbre’s inherent multidimensionality is well-known [9, 10], and the application of adjectives to its classification has been previously explored [11-14]. The use of these timbral adjectives by musicians provides clues to the reductive mental system of classification and description they are using. Several studies have examined the relationship between high-level descriptors and timbre for specific instruments, including . Nykänen and Johansson [15] list ten common timbral descriptors used by Swedish saxophone players. Disley and Howard [16] used similar methods to gather English words describing pipe organ ensembles, and refined these in subsequent listening tests [17] to realise an uncorrelated and consistently understood subset of descriptors for use in that context: *thin, flutey, warm, bright* and *clear*.

Other studies have gathered timbral descriptors without musical stimuli. Moravec et al., [18] collected words from Czech musicians and developed a subset based on frequency of occurrence. Many studies have looked for generally applicable auditory cues for individual timbral adjectives in isolation, and these are summarised on pages 76 to 81 of [10]. Von Bismarck [12] summarises timbral scales from many sources and creates a subset of four (*dull-sharp, compact-scattered, full-empty* and *colourless-colourful*) but Kendall and Carterette [13, 14] question both the relevance of these scales to the sounds of real instruments and the wisdom of assuming the opposition of words, going on to demonstrate more success with scales of "*x*" to "*not-x*". The problem with such studies is that their solutions tend to gravitate toward the same few readily measurable auditory phenomena such as the spectral centroid and relative harmonic strengths. This results in multiple theories that are difficult to apply simultaneously. Many of these words do not have a single obvious correlation with spectral features, with some words simultaneously describing both timbral quality and perceived loudness [19, p25]. The attempt to define such general relationships is problematic, as much usage of these words is inherently subjective and dependent on the learning of relationships between timbres and adjectives; an area that has been largely ignored thus far.

It is the success of this work with the sounds of the pipe organ [16, 17] that provided the authors with feasibility evidence and relevant experience for the move from the restricted timbral space occupied by the pipe organ to a wider timbral space encompassing the sounds of other musical instruments. The methodology then and in the current work involved the exploration of timbral descriptors that are commonly employed by musicians alongside the acoustic attributes of the sounds that these timbral descriptors describe. Listening tests were carried out in which musicians were asked to describe a set of sounds using timbral adjectives, and the results were analysed using multidimensional scaling and principal component analysis to establish which adjectives account for the greatest degree of segregation across the listeners and the sounds under investigation. These adjectives provide candidates for use as user controls in a synthesis system, and progress towards the design of such a sound synthesis system is described.

2. LISTENING TESTS

Three listening tests have been performed, and these are summarised below. A pilot listening test was performed first in order to confirm the experimental design and to enable the set of timbral adjectives and stimuli to be reviewed and altered if deemed appropriate.

2.1 Pilot listening test

Sixteen listeners (13 male, 3, female; average age 36yrs., range 21-63yrs.) took part in the pilot experiment (UK: 7; USA: 4; Canada, China, Germany, Italy and Sweden: 1 each). All listeners were musical and involved in an occupation related to music or music technology.

They were asked to listen to a set of ten test stimuli one at a time and rate each one using ten timbral adjectives which were provided on an eleven point slider scale via a graphical user interface.

The timbral adjectives were selected for this purpose by choosing words previously studied [10, 14, 17, 18], and avoiding any with ambiguous or similar meanings (e.g. *brilliant* was considered too similar to *bright* to justify the inclusion of both). The ten adjectives chosen were: *bright, clear, warm, thin, flutey, harsh, dull, nasal, metallic* and *colourful*. These were presented alongside a scale of eleven radio buttons to provide answer positions from *not bright* to *bright*, *not-clear* to *clear,* etc. Listeners were also asked to indicate their level of confidence in setting the sliders for each adjective on a five radio button scale from *not confident* to *confident*. At the outset, they were given a trial example using an eleventh stimulus (*koto*) to demonstrate the test procedure with just one adjective (*bright*) being shown along with its confidence scale. Testing was conducted over the Internet.

The stimuli were taken from a Yamaha XG module and consisted of: *piano, xylophone, bell, taiko drum, flute, oboe, viola, brass, sawtooth* and *sinewave*. The stimuli were captured as 1.5 second mono PCM samples (44.1kHz sampling rate, 16 bit resolution) on the note G3 (MIDI note 67, nominally 392Hz), and each was adjusted to have no vibrato or use of multiple or stereo voices.

The main purpose of the pilot test was to indicate whether the chosen adjectives were of any use in practice for labelling sounds, and to gather other words that listeners cared to suggest [20]. *Flutey* was demonstrated to be a poor discriminator, with most sounds being definitely *not flutey* apart from the flute and, to a much lesser extent, the sinewave. For this reason, *flutey* was discarded in the main experiments. *Colourful* was found to have the least level of listener confidence, and so it too was discarded.

2.2 Main listening test

Pilot test listeners suggested that it would be reasonable to use up to 15 adjectives and 12 stimuli in such a test format. Having discarded *flutey* and *colourful*, additional adjectives were included in the main experiment *(wooden*, *rich*, *gentle*, *ringing*, *pure*, *percussive* and *evolving)* alongside those remaining from the pilot experiment *(bright*, *clear*, *warm*, *thin*, *harsh*, *dull*, *nasal*, *metallic)*.

Listeners also indicated that the use of synthetic stimuli was not ideal, and they indicated a desire to hear the sounds of real acoustic instruments. Thus, twelve instrument samples were selected from the MUMS (McGill University Master Samples) library, three from each of the four categories: strings, brass, woodwind and percussion as follows.

- Strings: Viola Bowed, Viola Pizzicato, Electric Guitar
- Brass: Tenor Trombone, Bach Trumpet, Trumpet Harmon
- Woodwind: Flute Vibrato, Alto Saxophone, Oboe
- Percussion: Hamburg Steinway, Tubular Bells, Xylophone.

A total of 59 listeners (29 female, 30 male) took part in the main test, 23 (18-22Yrs.) took the test 10 times (control group) and 36 (19-27Yrs.) took it once. All listeners used high quality headphones. The purpose of the control group was to enable the use of the Internet for testing to be explored with a view to its validation, and they each took the test five times controlled by an investigator and five times alone with their own computer (uncontrolled). The remaining 36 listeners took the test with their own computers (uncontrolled). All subjects were native English speakers and students or staff in the subjects of Music and Music Technology at a UK University. Subjects were paid for their participation and were free to withdraw at any point. Full experimental details are in [20]. The stimulus order of presentation was varied for the control group and no significant effect was found.

3. LISTENING TEST RESULTS

The results from the control group confirm that the use of the Internet for these listening tests did not have any adverse effect on the results themselves. The average difference in results between the controlled and uncontrolled situations was 0.5% of the rating scale range, and since the scale itself had 11 points (quantised in 10% increments), this is not significant. There was a slight increase in self-evaluated confidence in the uncontrolled situation, but this was found to be non-significant.

ADJ.	INSTRUMENT STIMULI	SPREAD
Bright	Harmon muted trumpet and electric guitar	Good range, nothing very *not bright*
Clear	Electric guitar, Hamburg Steinway and xylophone	Reasonable spread on the positive side, little on the negative
Warm	Flute, <u>not</u> muted trumpet	Good spread, most around middle
Thin	Muted trumpet and electric guitar; <u>not</u> trombone or flute	A reasonable spread
Wooden	Xylophone; <u>not</u> tubular bells, Bach and muted trumpet, and electric guitar	reasonable spread tending negative
Harsh	Muted trumpet; <u>not</u> pizzicato viola or flute	Good spread
Dull	Tenor trombone; <u>not</u> muted trumpet or electric guitar.	Reasonable spread biased negative
Nasal	Muted trumpet and oboe; <u>not</u> xylophone, Steinway, pizzicato viola or tubular bells	Good spread
Metallic	Tubular bells, electric guitar, muted trumpet; <u>not</u> saxophone	Reasonable spread, nothing very negative
Pure	Xylophone, but not by much	Poor spread, around middle
Percussive	Xylophone, tubular bells, pizzicato viola, piano and electric guitar; <u>not</u> all others	Interesting spread, tending bimodal
Rich	Flute and bowed viola	Poor spread, better than *pure*
Gentle	Flute and pizzicato viola; <u>not</u> muted trumpet	Good spread, nothing *very gentle*
Ringing	Electric guitar, tubular bells and the Steinway; <u>not</u> saxophone	Good spread
Evolving	Viola; <u>not</u> saxophone, pizzicato viola or xylophone	Reasonable spread

Table 1: Instrument stimuli with particularly high ratings for each adjective and an indication of the spread each adjective exhibited across all instruments.

Instrument stimuli that had particularly high ratings for each adjective are shown in table 1, along with an indication of the degree of spread across all instruments exhibited by each adjective. Example results averaged across all 59 listeners are shown in Figure 1 for all instrumental stimuli for the adjectives *bright*, *dull*, *pure* and *nasal*. Note that these graphs have no vertical scale, the sample names are spread out vertically to make them readable. The adjectives that had the highest agreement between listeners were: *bright*, *percussive*, *gentle*, *harsh* and *warm*, whilst those with the least were: *nasal*, *ringing*, *metallic*, *wooden* and *evolving*. That which elicited the least confidence was *evolving*, whilst the highest confidence ratings being given to *clear, percussive, ringing* and *bright*.

The objective of this experiment is to attempt to establish a set of timbral adjectives that can be used in a synthesis system to provide a high-level means of its control. However, the use of fifteen such controllers would not really be practical, and therefore some means of reducing the set of fifteen adjectives prior to synthesizer implementation is required. It is highly likely that some adjectives will be describing similar aspects of timbral difference between sounds, and a means of detecting this and evaluating the degree of similarity is needed.

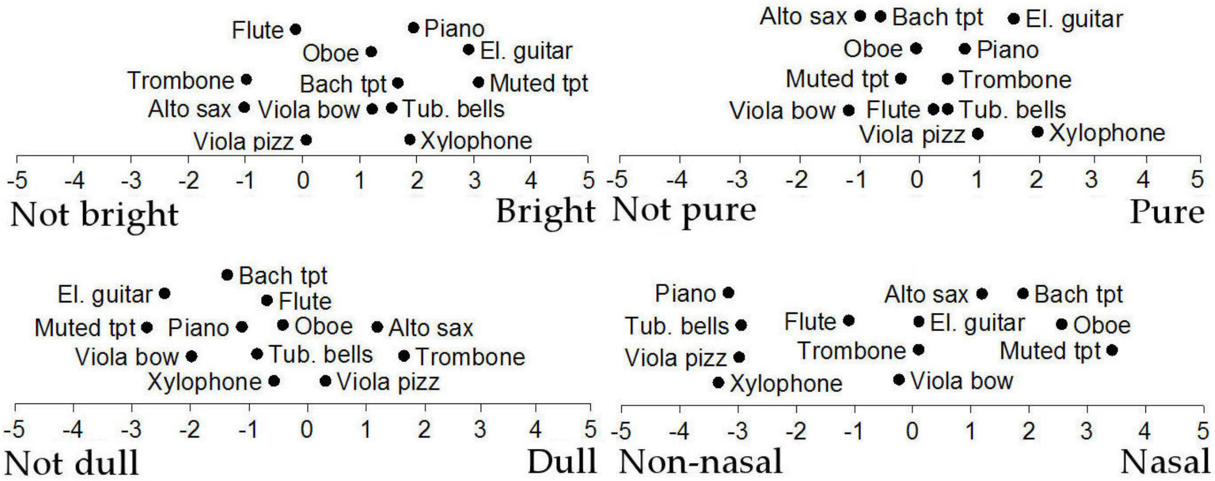

Figure 1: Stimulus ordering for 4 adjectives averaged across all 59 listeners.

Cluster analysis (CA) and principal component analysis (PCA) has been applied to all fifteen adjectives [21]. The CA results suggest that *bright* and *clear*, and *warm* and *gentle* are indicating very similar aspects of the sounds in the context of the particular set of samples used in the listening experiment. The PCA (PCA) indicates how many underlying dimensions there are to the data, and for these data, PCA suggests that 91.7% of the variance can be explained by four principal components. In terms of the timbral adjectives themselves, the first scale has *bright, thin* and *harsh* at one end and *dull*, *warm* and *gentle* at the other. The second has *pure* and *percussive* at one end and *nasal* at the other. The third is mainly accounted for by *metallic* to *wooden*, and the fourth by *evolving*. These data suggest that *pure* and *rich* are not good discriminators for this data set.

As a further consideration, the degree of agreement between listeners in the choice of particular adjectives was considered by reference to the standard deviations obtained in the listening tests. The highest standard deviations were found for: *nasal*, *ringing*, *metallic*, *wooden* and *evolving*, suggesting that these are not being used consistently, and that they therefore would not be particularly suitable as universal timbral descriptors.

4. SYNTHESIZER DESIGN CONSIDERATIONS

Following the analysis of the timbral adjectives, the following remain from the initial fifteen adjectives as the final set to be used as synthesizer controllers: *bright*, *clear*, *warm*, *thin*, *harsh*, *dull*, *percussive* and *gentle*. In order to understand the nature of the acoustic changes between the stimuli and how these relate to the adjectives being applied to the sound by the listeners, a further listening test in which listeners compared pairs of sounds was carried out, using a design based on the classic experiment by Gray [22]. A multidimensional scaling analysis of the results produced two dimensions. The stimuli were analysed in terms of the following set of acoustic features: *spectral centroid, spectral slope, spectral smoothing, attack time, decay time,* and *the ratio of the energy in the first harmonic (fundamental) to that in harmonics 5 to 8*, and these results were compared with the two dimensions. Figure 2 shows the results with respect to the adjectives and acoustic properties respectively. It can be seen that dimension 1 relates to: *spectral centroid, spectral slope, spectral smoothness, attack time* and *ratio f0:harmonics 5-8*. The main adjective for dimension 1 is *percussive.* Dimension 2 relates to some degree to: *ratio f0:harmonics 5-8, spectral centroid* and *decay time*. The main adjectives for dimension 2 are: *bright, thin* and *harsh* to *warm, dull* and *gentle.*

ACOUSTIC FEATURE	Dim. 1	Dim. 2	ADJECTIVE	Dim. 1	Dim. 2
Spectral centroid	0.67	0.43	bright	-0.21	0.68
Spectral slope	0.77	-0.11	clear	-0.38	0.34
Spectral smoothness	0.77	0.28	warm	0.07	-0.89
Attack time	0.63	0.05	thin	-0.19	0.65
Decay time	-0.47	0.40	harsh	0.39	0.82
Ratio of attack to decay	0.59	-0.20	dull	-0.04	-0.57
Ratio f0:harm.2-4	-0.12	0.02	percussive	-0.97	0.18
Ratio f0:harm.5-8	0.61	0.53	gentle	-0.14	-0.83

Figure 2: Acoustic features and timbral adjectives contributions to each of the two dimensions established from multidimensional scaling analysis of the comparison listening test.

This leads to the basic specification for the synthesis system itself. Its high-level controllers will be the eight adjectives, and these will be mapped to varying degrees to the acoustic characteristics listed for the two dimensions, and the output from this mapping process will inform the synthesis engine itself. Given the nature of the acoustic analysis, a harmonic additive synthesis system will be implemented using Pure Data or PD (www.pure-data.org) in which every harmonic can be specified in terms of its own individual amplitude envelope. MIDI (the Musical Instrument Digital Interface) protocol will be used to control the PD implementation and enable musical notes to be specified. In block diagram form, the synthesizer will take the form shown in figure 3.

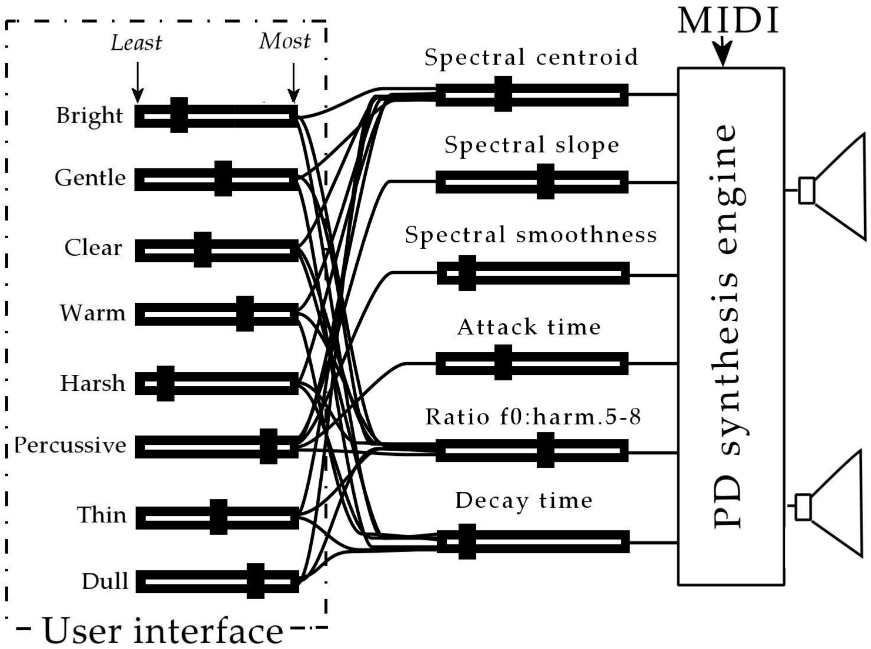

Figure 3: Outline diagram of the proposed synthesizer control system.

5. CONCLUSIONS

A new synthesizer control protocol has been specified that is controlled by high-level timbral adjectives. Listening tests have been carried out to explore how such adjectives are used my musicians, and these have been correlated with the acoustic changes apparent in a set of musical sounds. The synthesizer will be realised in PD and beta tested by practising musicians to establish whether or not such a control system is useful in practice. Music synthesis is a creative activity that should not be hampered by the lack of knowledge in either acoustics or physics. An instrument that uses high-level adjectives to control sound synthesis has the potential to offer music synthesis to all who wish to create music electronically.

6. ACKNOWLEDGEMENTS

The authors thank the subjects who took part in the listening tests. This project was supported by grant number GR/T06384/01 from the Engineering and Physical Sciences Research Council (EPSRC), UK.

REFERENCES

[1] A.D. Hunt, and P.R. Kirk, *Digital sound processing for music and multimedia*, Oxford, Focal Press, 1999.

[2] J.C. Risset, Sculpting Sounds with Computers: Music, Science, Technology, *Leonardo*, 27 (3), 257-261 (1994).

[3] J.C. Risset, Pitch and rhythm paradoxes, *Journal of the Acoustical Society of America* **80**, 961-962 (1986).

[4] J.M. Chowning, Digital sound synthesis, acoustics and perception: a rich intersection, *Proc. COST G-6 Conf. on Digital Audio Effects (DAFX-00),* Verona, Italy, December 7-9 (2000).

[5] D.M. Howard, and J.A.S. Angus, *Acoustics and Psychoacoustics*, 3rd Ed., Oxford, Focal Press (2006).

[6] ANSI, *USA Standard Acoustical Terminology (including Mechanical Shock and Vibration)*, Technical Report S1.1-1960, American National Standards Institute, New York (1960, revised 1976)

[7] P.A.. Scholes, *The Oxford companion to* music, London: Oxford University Press, 1970.

[8] E.R. Miranda, J. Correa, and J. Wright, Categorising complex dynamic sounds, *Organised Sound*, **5**, (2), 95-102, (2000).

[9] J.M. Grey, *An Explanation of Musical Timbre*, PhD thesis, Stanford University, 1975.

[10] D.P. Creasey, *An exploration of sound timbre using perceptual and time-varying frequency spectrum techniques*. DPhil thesis, The University of York, 1998.

[11] K.W. Berger, Some factors in the recognition of timbre, *Journal of the Acoustical Society of America* **36**, (2), 1888-1891 (1965).

[12] G. von Bismarck, G., Timbre of steady sounds: a factorial investigation of its verbal attributes, *Acustica*, **30**, (3), 149-159 (1974).

[13] R.A. Kendall, and E.C. Carterette, Verbal Attributes of Simultaneous Wind Instrument Timbres: I. von Bismarck's Adjectives, *Music Perception*, **10**, (4), 445-468 (1993a).

[14] R.A. Kendall, and E.C. Carterette, Verbal Attributes of Simultaneous Wind Instrument Timbres: II. Adjectives Induced from Piston's "Orchestration"*, Music Perception*, **10**, (4), 469-502 (1993b).

[15] A. Nykänen, and Ö. Johansson, Development of a language for specifying saxophone timbre, *Proceedings of the Stockholm Music Acoustics Conference 2003*, (SMAC03), Stockholm, Sweden, **2**, 647-650.

[16] A.C. Disley, and D.M. Howard, Timbral semantics and the pipe organ, *Proceedings of the Stockholm Music Acoustics Conference 2003*, (SMAC03), Stockholm, Sweden, **2**, 607-610.

[17] A.C. Disley, and D.M. Howard, Spectral correlates of timbral semantics relating to the pipe organ, *Proceedings of the Baltic-Nordic Acoustics Meeting*, HUT, Helsinki, (2004).

[18] Moravec, Ondrej and J. Štěpánek, Verbal description of musical sound timbre in Czech language, *Proceedings of the Stockholm Music Acoustics Conference 2003*, (SMAC03), Stockholm, Sweden, **2**, 643-645.

[19] G.J. Sandell, *Aconcurrent Timbres in Orchestration: A Perceptual Study of Factors Determining "Blend"*, PhD dissertation, School of Music, Northwestern University, 1991.

[20] A. Disley, D.M. Howard, and A.D. Hunt, Musicians' use of timbral adjectives, *Proceedings of the Institute of Acoustics*, **28**, (1), 670-679 (2006).

[21] A. Disley, D.M. Howard, and A.D. Hunt. Timbral descriptors of musical instruments, *proceedings of the 9th International Conference on Music Perception and Cognition*, University of Bologna, 22-26 August 2006, 61-68.

[22] J. Grey, Timbre discrimination in musical patterns, *Journal of the Acoustical Society of America* **64**, 467-472 (1977).