Emotional Responses to the Perceptual Dimensions of Timbre: A Pilot Study Using Physically Informed Sound Synthesis.

Sylvain Le Groux¹ and Paul F.M.J. Verschure¹,²

¹ The Laboratory for Synthetic, Perceptive, Emotive and Cognitive Systems (SPECS).
Universitat Pompeu Fabra
Roc Boronat, 138 08018 Barcelona, Spain
² Institucio Catalana de Recerca i Estudis Avancats (ICREA)
Barcelona, Spain
{t.sylvain.legroux,paul.verschure}@upf.edu

Abstract. Music is well known for affecting human emotional states, and most people enjoy music because of the emotions it evokes. Yet, the relationship between specific musical parameters and emotional responses is not clear. A key question is how musical parameters can be mapped to emotional states of valence, arousal and dominance. Whereas many studies have focused on emotional responses to performance or structural musical parameters [6–8], little is known about the role of timbre on emotions. Here, we propose a sound synthesis-based approach to the study of sonic emotion using physically-inspired sound models, and investigate the emotional effects of three perceptually salient timbre descriptors, namely log-attack time, spectral centroid and spectral flux. Our results show that an increase in spectral centroid and spectral flux parameters produces an increased emotional response on the arousal scale. Moreover we could observe interaction effects between log-attack time and spectral flux on the dominance scale. This study suggests a rational approach to the design of emotion-driven music systems for multimedia installations, live performance and music therapy.

Key words: Timbre, Perception, Emotion, Physically Inspired Sound Synthesis

1 Perceptual Dimensions of Timbre

Sounds can be described by five main perceptual components, namely pitch, loudness, duration, timbre, and spatialization [25]. To a first degree of approximation, pitch, loudness and duration are unidimensional parameters which relate to fundamental frequency, sound level, and time respectively. On the other hand, timbre is notoriously difficult to define [28]. The American Standards Association (ASA) proposes the following definition [32]: timbre is the perceptual attribute
of sound that allows a listener to distinguish among sounds that are otherwise equivalent in pitch, loudness and subjective duration.

Over the years, several approaches to the study of timbre have been investigated, leading to disparate but related results. One possible approach is to search for a reduced semantic field that would describe sounds accurately. Following this approach, [4] asked subjects to rate sounds on 30 verbal attributes. Based on a multidimensional scaling analysis of the semantic-differential data, he found four statistically significant axes, namely the dull-sharp, compact-scattered, full-empty and colorful-colorless pairs.

Other type of studies have tried to relate the perceptual dimensions of sounds to acoustics descriptors. Those descriptors can be spectral, temporal or spectro-temporal [22], and generate a “timbre space” [35].

The timbre space is determined using multidimensional analysis of data derived from experiments where a listener estimates the dissimilarity between pairs of sounds with different timbral characteristics. Multidimensional analysis techniques allow to extract a set of “principal axes” or dimensions that are commonly assumed to model the criteria used by participants to estimate the dissimilarity. It is then possible to find acoustical descriptors that correlate with the dimensions derived from multidimensional analysis. The three descriptors consistently reported by dissimilarity studies on instrumental sounds are the spectral centroid, the log-attack time and the spectral flux [21, 9, 10].

In this study we propose to investigate the relationship between the timbre of a musical note and the emotional response of the listener. To this aim, we use a physically informed synthesizer that can produce realistic sounds efficiently and intuitively from a three-dimensional timbre space model mapped to a reduced set of physical parameters.

2 Sound Generation: Modal Synthesis

In modal synthesis, a sound is modeled as a combination of modes which oscillate independently (Figure 1). While this kind of modeling is in theory only accurate for sounds produced by linear phenomena, it allows for efficient real-time implementations [5, 1, 33].

2.1 Modal Synthesis

The dynamics of a simple unidimensional mass-spring-damper system (our model of a mode) is given by the second order differential equation:

\[ m\ddot{x} + \mu \dot{x} + kx = 0 \]  

where \( m \) is the mass, \( \mu \) the damping factor, \( k \) the tension and \( x \) the displacement.

The solution is an exponentially decaying oscillator:

\[ x = e^{-\alpha t} A \cos(\omega_0 t + \phi) \]
where $\alpha = \frac{\mu}{2m}$, $\omega_0 = \sqrt{(k/m - (\frac{\mu}{2m})^2)}$, $A$ is the amplitude and $\phi$ the phase.

The modal approximation consists in modeling the time-varying sound $s(t)$ by a linear combination of damped oscillators representing the vibrating properties of a structure. In the general case, the damping factor $\alpha_n$ is frequency-dependent and relates to the physical properties of the material.

$$s(t) = \sum_{n=1}^{N} s_n(t) = \sum_{n=1}^{N} A_n e^{2\pi f_n t} e^{-\alpha_n t}$$

Fig. 1. Within the modal paradigm, a vibrating structure is decomposed into a sum of simple modes.

2.2 Implementation

By discretizing equation 1 with finite difference and using a sampling interval $T = \frac{1}{f}$ we obtain:

$$x[n] = x[n-1] \frac{2m + Tr}{m + Tr + T^2k} - x[n-2] \frac{m}{m + Tr + T^2k}$$

which is the formulation of a standard IIR two-poles resonant filter. Consequently we implemented a real-time synthesizer as a bank of biquad resonant filters in Max5 [11, 37] excited by an impulse (Figure 2).

3 Perceptual Control of The Synthesis Parameters

Previous psychoacoustics studies on timbre have emphasized the perceptual importance of features such as spectral centroid, log-attack time and spectral flux (cf. Section 1). As a means to generate perceptually relevant sounds, we have built a simple and efficient physically-inspired modal synthesizer that produces realistic percussive sounds. We will now show how to map those psychoacoustical features to sound synthesis parameters.
Fig. 2. The sounds are produced by an impulse that injects energy in a bank of exponentially-decaying resonant filters, modulated by a time envelope.

3.1 Background

Few studies have explored the control of audio synthesis from perceptual parameters. We can distinguish four main trends: the machine learning view, where a model of the timbre is learned from audio data, the concatenative view, where the timbre is “constructed” by the juxtaposition of pre-existing sonic grains, the signal processing view, where the transformations on timbre are model-dependent and direct applications of the signal model possibilities, and the physical model view, where the sound generation is modulated by intuitive physical parameters. One of the earliest examples of the machine learning point of view can be found in [36] where an additive synthesis model was driven by pitch, loudness and brightness using artificial neural networks. This paradigm has been more recently generalized and adapted to Support Vector Machine learning [16]. In the case of concatenative synthesis [29], a sound is defined as a combination of pre-existing samples in a database. These samples are already analyzed, classified and can be retrieved by their audio characteristics. For signal models, a set of timbral transformations based on additive modeling is proposed. For instance by explicitly translating and distorting the spectrum of a sound, one can achieve the control over vibrato, tremolo and gender transformation of a voice [30]. In addition, a few studies have investigated the relationship between perceptual and physical attributes of sounding objects [2, 20].

Here we follow this last approach with a simple physical model of impact sound and propose a set of acoustic parameters corresponding to the tridimensional model of timbre proposed by psychoacoustics studies [21, 9, 10] (cf. Section 1).
3.2 Tristimulus / Spectral Centroid

We propose to control the spectral centroid of the sound by using the tristimulus values to specific brightness surfaces [24, 26].

The tristimulus analysis of timbre proposes to quantify timbre in terms of three coordinates (x, y, z) associated with band-loudness values. Inspired from the tristimulus theory of colour perception, it associates high values of x to dominant high-frequencies, high values of y to dominant mid-frequency components and high values of z to dominant fundamental frequency. The coordinates are normalized so that \( x + y + z = 1 \).

\[
T_1 = \frac{a_1}{\sum_{h=1}^{H} a_h}, T_2 = \frac{a_2 + a_3 + a_4}{\sum_{h=1}^{H} a_h}, T_3 = \frac{\sum_{h=1}^{H} a_h}{\sum_{h=1}^{H} a_h}
\]

Fig. 3. An intuitive 2D representation of the tristimulus timbre space is a triangle. The arrow represents a specific time-varying trajectory of a sound in the tristimulus timbre space [24].

For synthesis, the sound is approximated by a sum of weighted damped oscillators at frequencies multiple of the fundamental. We use an additive model and neglect the effect of phase:

\[
s(t) \simeq \sum_{k=0}^{N-1} e^{-\alpha_k t} A_k \cos(2\pi f_k t)
\]
where \( n \) is the time index, \( N \) is the number of harmonics in the synthetic signal, \( A_k \) is the amplitude of the \( k \)th partial, \( f_k \) is the frequency of the \( k \)th partial, with \( f_k = k \cdot f_0 \) where \( f_0 \) is the fundamental frequency (harmonicity hypothesis).

In the synthesis model, the harmonic modes belong to three distinct frequency bands or tristimulus bands: \( f_0 \) belongs to the first low frequency band, frequencies \( f_2 \ldots 4 \) belong to the second mid-frequency band, and the remaining partials \( f_5 \ldots N \) belong to the high-frequency band.

The relative intensities in the three bands can be visualized on a tristimulus triangular diagram where each corner represents a specific frequency band. We use this representation as an intuitive spatial interface for timbral control of the synthesized sounds (Figure 4). The inharmonicity of the sound was set by scaling the values of partials following a piano-like law proposed by\[2\] \( f_k = k \cdot f_0 \sqrt{1 + \beta k^2} \).

![Fig. 4. The synthesizer GUI allows for graphical control over the tristimulus values (left). It automatically updates and displays the amplitudes of the corresponding partials (right).](image)

### 3.3 Damping / Spectral Flux

While the tristimulus is only a static property of the spectrum, the modal synthesis technique described in Section 2 also allows for realistic control of the time-varying attenuation of the spectral components. As a matter of fact, the spectral flux, defined as the mean value of the variation of the spectral components, is known to play an important role in the perception of sounds [21]. We decided, as a first approximation, to indirectly control the spectral flux - or variation of brightness- by modulating the relative damping value \( \alpha_r \) of the frequency-dependent damping parameter \( \alpha \) (Figure 5 and Equation 3) proposed by [2]:

\[
\alpha(\omega) = \exp(\alpha_g + \alpha_r \omega) \tag{7}
\]

### 3.4 Time Envelope / Log Attack Time

The log-attack time (LAT) is the logarithm (decimal base) of the time duration between the time the signal starts to the time it reaches its stable part
Fig. 5. The synthesizer GUI allows for graphical control of the damping parameters $\alpha_g$ and $\alpha_r$.

[22]. Here, we control the log-attack time manually using our synthesizer’s interface (Figure 6). Each time the synthesizer receives a MIDI note-on message, an Attack-Decay-Sustain-Release (ADSR) time envelope corresponding to the desired LAT parameter is triggered (Figure 7).

$$LAT = \log_{10}(t_{stop} - t_{start})$$  \hspace{1cm} (8)

Fig. 6. The synthesizer GUI allows for graphical control over log-attack time parameter via the Attack Decay Sustain Release envelope.

4 From perception to emotion

4.1 Background

Music and Emotion Music and its effect on the listener has long been a subject of fascination and scientific exploration from the Greeks speculating on the acoustic properties of the voice [13] to Musak researcher designing “soothing” elevator music. It has now become an omnipresent part of our day to day life,
whether by choice when played on a personal portable music device, or imposed when diffused in malls during shopping hours for instance.

Studies on musical emotion have traditionally been influenced by “standard” research on emotion induction and usually assumed that musical emotion relies on general emotional mechanisms such as cognitive appraisal. Nevertheless, while observing that most emotional responses to music do not involve implications for goals in life, a recent review article [12] proposes to challenge the cognitive appraisal perspective, and put forward six additional mechanisms that can be accounted for. Namely, brain stem reflex, evaluative conditioning, emotional contagion, visual imagery, episodic memory and musical expectancy.

In our experiment, we decided to limit our focus to the investigation of the relationship between perceptual acoustic properties and emotional responses. In order to control for mechanisms such as musical expectancy and emotional contagion, the stimuli consisted of one short note. In fact, previous studies have revealed that only a few seconds of music are necessary to elicit an emotional reaction [3, 23]. There was no conditioning task involved, and the sounds generated by our model were synthetic, but realistic. Therefore within the framework proposed in [12], we most probably look at some reflex-like mechanisms reflecting the impact of auditory sensation in the most basic sense.

**Emotional Responses** There exist a number of different methodologies to measure emotional responses. They can be broadly classified into three categories depending on the extent to which they access subjective experience, alterations of behavior or the impact on physiological states [19]. Self-reports of emotional

---

**Fig. 7.** The ADSR time envelope allows for direct control over the LAT each time a note is triggered.
experience measure the participant’s “subjective feeling” of emotions, i.e. “the consciously felt experience of emotions as expressed by the individual” [31]. Self-reports can be subcategorized into verbal [34] and visual self-reports [27]. We decided to use visual reports during experiments since these are more universal than the verbal measures.

We followed the well established three dimensional theory of emotions: valence and intensity of activation or arousal [27] plus dominance using the Self Assessment Manikin (SAM) scale [14] (Figure 8). Emotions can then be placed in a three-dimensional emotional space, where the valence scale ranges from positive to negative, the activation scale extends from calmness to high arousal and the dominance scale goes from dominated to dominant.

4.2 Methods

Fig. 8. The presentation software S-Blast programmed in Max/MSP uses SAM scales (Dominance, Arousal and Valence)[14] to measure emotional responses to sound stimuli from the subjects.
**Participants** A total of ten university students (2 female; mean age 29 years) took part in the pilot experiment. The experiment was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

**Stimuli** Eight sound samples of about 4 s. duration each were used as acoustic stimuli. They were presented to the participants using stereo headphones (AKG K66).

We investigated the effect of three timbre descriptors (two levels each) on arousal, valence and dominance:

- Attack time: low level corresponded to 1ms while high level corresponded to 150 ms.
- Spectral Centroid (controlled by tristimulus): low level corresponded to energy in the first and second tristimulus band while high level corresponded to energy in the third tristimulus band only.
- Spectral-flux (controlled by damping): high level corresponded to high damping $\alpha_r = 0.6$ while low level corresponded to low damping $\alpha_r = -1.6$ while $\alpha_g$ was kept at the constant value 0.23 (cf. Equation 7).
- All the other parameters of the synthesizer were fixed (partial amplitude decay, harmonicity, even/odd ratio, decay, sustain, release), and the sound samples were normalized in amplitude with Peak Pro (BIAS).

**Procedure** We investigated the influence of different sound features on the emotional state of the patients using a fully automated and computer based stimulus presentation and response registration system. In our experiment, each subject was seated in front of a PC computer with a 15.4” LCD screen and interacted with custom-made stimulus delivery and data acquisition software called S-blast (Figure 8) made with the MAX programming language [37]. Sound stimuli were presented through headphones (K-66 from AKG). At the beginning of the experiment, the subject was exposed to a sinusoidal sound generator to calibrate the sound level to a comfortable level. Subsequently, a number of sound samples with specific sonic characteristics were presented together with the different scales (Figure 8).

The subject had to rate each sound sample in terms of their emotional content (valence, arousal, dominance) by clicking on the SAM manikin representing her feelings [14]. The subject was given the possibility to repeat the playback of the samples. The SAM graphical scale was converted into an ordinal scale (from 0 to 5) where 0 corresponds to the most dominated, aroused and positive and 5 to the most dominant, calm and negative (Figure 8). The data was automatically stored into a SQLite database composed of a table for demographics and a table containing the emotional ratings. SPSS (IBM) statistical software suite was used

---

3 Available from http://www.dtic.upf.edu/~slegroux/confs/CMMR10

4 http://www.bias-inc.com/

5 http://www.sqlite.org/
to assess the significance of the influence of sound parameters on the affective responses of the subjects.

4.3 Results

We verified that, for each condition, the data passed the Kolmogorov-Smirnov test of normality. We then conducted a factorial (three-way) repeated measure ANOVA to look at the influence of perceptual parameters of timbre on the emotional responses of the participants as measured by a five points SAM scales.

Arousal  Mauchly’s test indicated that the assumption of sphericity had been violated for effects of attack, centroid and damping ($p<0.05$). Therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity.

There was a significant main effect of centroid ($F(1,9)=11.663$, $p<0.05$) as well damping ($F(1,9)=5.857$, $p<0.05$) on arousal. Post-hoc pairwise comparisons with Bonferroni corrections showed a significant mean difference of -0.725 between high spectral centroid (M=1.825) and low spectral centroid (M=2.55) showing higher values produce higher ratings along the arousal scale (i.e. lower value on the scale). Similarly, a significant mean difference of -1.075 was observed between low damping (M=1.65) and high damping (M=2.725), showing low damping was more arousing than high damping (Figure 9).

Valence  No sound property significantly affected ratings along the valence scale.

Dominance  Mauchly’s test indicated that the assumption of sphericity had been violated ($p<0.05$). Therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity.

There was a significant interaction effect between attack and damping ($F(1,9)=7.23$, $p<0.05$) on the dominance scale. A contrast analysis for repeated measures revealed the increased dominance found when using low damping (compared to high damping) is actually reversed when high log-attack time is used instead of low log-attack time. In other words, the increase in dominance due to low damping is significantly greater when used in conjunction with a short log-attack time (low value).

5 Conclusion

In this study we investigated the role of timbre on emotional responses as measured by the SAM scale. We proposed a physically-inspired model to synthesize sound samples from a set of well-defined and perceptually-grounded sound parameters, and showed their correspondence to arousal and dominance scales. In particular, spectral centroid and damping were significantly related to arousal.

6 http://www.spss.com/
Fig. 9. Boxplot of the different conditions on the arousal scale (from 0 to 5 with 0 maximum arousal). Notation: a 3 character string codes for condition (Attack, Centroid and Flux in that order) and levels (High, Low). For example, HHL corresponds to High Attack-time, High Spectral Centroid and Low Spectral flux.

Fig. 10. Interaction graph between log-attack time and spectral flux on the dominance scale (from 0 to 5 where 5 represents highest dominance). Estimated Marginal Means are obtained by taking the average of the means for a given condition.
subjective ratings, while interaction effects between log-attack time and damping were found, showing that short attack with low damping maximized dominant responses.

The study of the relationship between sonic features and emotional responses combined with techniques for real-time sound synthesis paves the way for the design of affective interactive music systems. We plan to use this type of system for affective state diagnosis, affective state modulation (e.g. involving Alzheimer patients and autistic children [15]) and neurofeedback studies [17].

In future experiments, we want to investigate in more details the potential of timbre and higher level musical parameters to induce specific affective states. We wish to expand our analysis to time-varying parameters and to co-varying parameters. Our results, however, already demonstrate that a rational approach towards the definition of affective interactive music systems is possible. We will integrate this knowledge into SiMS, the biomimetic interactive music system we have developed [18], and explore adaptive emotional mappings of musical parameters using algorithms such as reinforcement learning with a response evaluation loop.

References


