Visualizing Emotion in Musical Performance Using a Virtual Character

Robyn Taylor[1] and Pierre Boulanger[2] and Daniel Torres[3]

[1] Advanced Man-Machine Interface Laboratory,
Department of Computing Science, University of Alberta
T6G 2E8 Edmonton, Alberta. Canada
robyn@cs.ualberta.ca
[2] pierreb@cs.ualberta.ca
[3] dtorres@cs.ualberta.ca

Abstract. We describe an immersive music visualization application which enables interaction between a live musician and a responsive virtual character. The character reacts to live performance in such a way that it appears to be experiencing an emotional response to the music it 'hears.' We modify an existing tonal music encoding strategy in order to define how the character perceives and organizes musical information. We reference existing research correlating musical structures and composers' emotional intention in order to simulate cognitive processes capable of inferring emotional meaning from music. The ANIMUS framework is used to define a synthetic character who visualizes its perception and cognition of musical input by exhibiting responsive behaviour expressed through animation.

1 Introduction

In his influential 1959 work, "The Language of Music," Deryck Cooke analyses Western tonal music in an attempt to understand how a musical piece conveys emotional content to a listening audience [2].

Cooke uses the term 'content' to represent the sense of exhilaration, despair, or bliss which a listener might report feeling after listening to a powerful piece of music. Although content cannot easily be extracted from the written score in the same way that pitch names, key signatures or harmonic structure may be, Cooke defends his belief that emotional content is inherent to music. He describes content as "not a technical part [of a piece of music], but something more elusive: the interpretation which we put upon the interaction of the technical elements."[4]

> "In other words, the 'content' is inseparable from the music, except as an emotional experience derived from listening to the music. If we use the word to signify 'the emotion contained in the music', we must keep clear in our minds that the emotion is contained, not as a necklace in a box,

[4] Cooke, *The Language of Music*, p.199.

> to be taken out and examined, but as an electric current in the wire: if we touch the wire we shall get a shock, but there is no way whatsoever of making contact with the current without making content with the wire."[5]

Cooke's work references his extensive research in Western tonal music, makes hypotheses concerning the relationships between musical structures and composers' emotional intent, and cites numerous examples from musical literature to support his conjectures.

We have chosen to rely on Cooke's findings to implement an immersive music visualization system which attempts to enrich the experience of observing a live musical performance by echoing the musical work's emotional content – Cooke's "electric current in the wire" – through the behaviours of a life-sized virtual character who simulates an emotional response to the music being performed.

This paper presents a method of visualizing live musical performance through the behavioural responses of a virtual character. Torres and Boulanger's ANIMUS Project [1] enables the creation of animated characters that respond in real-time to the stimuli they encounter in the virtual world [14] [15]. We describe an ANIMUS character with the ability to perceive musical input and to encode it in a way consistent with previously defined organizational models for tonal music. The character is equipped with a cognition mechanism which relates perceived music-theoretical features to emotional states consistent with Cooke's findings, simulating an emotional understanding of the music it 'hears'. Using keyframe animation techniques, the ANIMUS character may express its simulated emotional response through visual animations generated in real-time and displayed on a life-sized stereoscopic screen (see Figure 1). Its expressive behavioural response animations provide a visual accompaniment to the performer's musical performance.

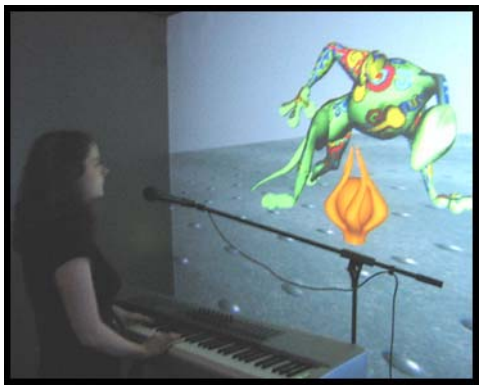


Fig. 1. The User Interacting with the Virtual Character

[5] Cooke, *The Language of Music*, p.199.

In Section 2 of this paper we discuss previous work in the area of immersive music visualization. Section 3 introduces the ANIMUS Project, the framework upon which our virtual characters are based. Section 4 describes our proposed model for organizing extracted musical feature data using methods derived from an existing music theoretical model. Section 5 discusses Cooke's research into the relationship between music and emotion, and how it can be used to create believable character response behaviour. In Section 6, we describe how character emotion is expressed through interpolated keyframe animation.

2 Previous Work

There exist numerous examples of previous research in musical visualization which correlate audio content and associated imagery. Of particular interest to us are real-time systems which leverage high-end visualization technology to create life-sized, compelling imagery. Imagery of this scope blurs the boundary between the physicality of the performer and the virtual world within which he or she is immersed.

"The Singing Tree" [8], created by Oliver *et al.* at MIT's Media Laboratory, immerses a user inside an artistic space comprised of computer graphics and installed set pieces. The environment is visibly and audibly responsive to the sound of his or her voice. "The Singing Tree" provides audio-visual feedback to participants in order to prompt them towards the goal of holding a prolonged note at a steady pitch.

Jack Ox's visualizations within a three-walled immersive CAVE system explore the harmonic structure of musical input [9]. Her "Color Organ" allows viewers to visualize harmonic relationships in a musical piece by creating three-dimensional structures and landscapes that observers can explore and navigate at will.

An audio-visual performance installation piece, "Messa di Voce", created by Golan Levin and Zachary Lieberman [7], relates the physicality of the performers to the physicality of the virtual space within which they are performing. It visualizes abstract representations of live vocalizations which originate from the locations of the vocalists' mouths. The scale of the imagery makes the performers appear to be a part of the virtual space within which they are performing.

Taylor, Torres, and Boulanger [12] have previously described a system that parameterizes live musical performance in order to extract musical features and trigger simple behaviours in Torres and Boulanger's ANIMUS characters [14][15]. ANIMUS characters can be displayed upon a large stereoscopic projection screen, enabling performers and audience members to perceive them as life-sized and three-dimensional.

We wish to increase the complexity of our music visualization system by creating examples of ANIMUS characters which both perceive live musical input in a 'human-like' fashion and appear to interpret it in such a way as is consistent with Cooke's research into the relationship between musical features and emo-

tional content. It is our hope that this will provide a novel way of visualizing music through human interaction with responsive virtual characters.

3 The ANIMUS Architecture

Torres and Boulanger's ANIMUS Project [1] is a framework which facilitates the creation of 'believable' virtual characters. They describe a believable character as "[appearing] to be alive, giving the illusion of having its own thoughts, emotions, intention and personality"[14].

ANIMUS characters are animated characters which respond to events in their environment in ways that illustrate their particular personalities. 'Shy' characters may cringe as if startled when another character makes a sudden movement, while 'curious' characters may come forward to investigate changes in their environment.

In order to create these types of responsive characters, Torres and Boulanger break the task of information organization and processing into three layers:

- **Perception Layer:** ANIMUS characters must perceive features and events in the world around them. Examples of perceivable events could include user input or actions of other characters in the virtual world. For our purposes, ANIMUS characters must be able to perceive important features in live musical performance. We will describe, in Section 4, how our specialized Musical Perception Filter Layer makes this possible.
- **Cognition Layer:** The characters must analyze the input they have perceived and determine appropriate response behaviours. In this layer, character 'personality' is created by defining how perceived events affect the character's internal state. In Section 5, we discuss how our ANIMUS characters simulate an emotional understanding of perceived musical features.
- **Expression Layer:** When an ANIMUS character has processed perceived data and determines that physical response behaviour is warranted, these behaviours are expressed through animation. The ANIMUS system generates these animations at run-time by using key-frame animation to interpolate between various combinations of pre-defined poses. Animated behaviours vary in mood and intensity in order to illustrate the virtual character's cognitive state.

4 Identifying Musical Features Through A Specialized Perception Layer

ANIMUS characters receive and share information about the world around them using a 'blackboard' system. Information about events occurring within the virtual world is entered on the blackboard, and characters monitor this blackboard in order to perceive these events.

Our ANIMUS characters must be able to 'listen' to a live musical performance and perceive meaningful data within that performance. This is necessary in

order for them to assign cognitive meaning to aspects of the music they have 'heard.' The stream of incoming live music must be parsed and organized in a perceptually relevant way.

For this purpose, we have created a specialized perception layer called the Musical Perception Filter Layer (see Figure 2). The live musician interfaces with this layer by singing into a microphone and playing a digital piano. The Musical Perception Filter Layer is implemented in a distributed fashion, leveraging the capabilities of a dedicated machine to handle the audio analysis tasks in a real-time manner.

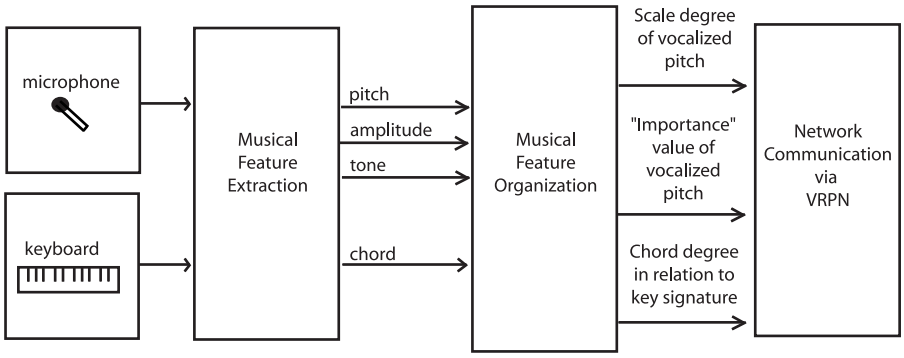


Fig. 2. Musical Perception Filter Layer

Our tasks within the Musical Perception Filter Layer are twofold. First we must parse a complex stream of incoming musical data in order to extract meaningful features upon which further analysis may be made. Second, in order to undertake the task of simulating human-like emotional responses to musical input, we must organize these features in a way that is consistent with previous research in the area of human musical perception.

4.1 Extracting Musical Feature Data from Real-Time Audio and MIDI Signals

We use a Macintosh G5 system to extract important features from a stream of live musical input, and then communicate these extracted features across a network to the PC running the ANIMUS engine.

In order to parse the stream of live music, our system uses functionality provided by Cycling '74's Max/MSP [3] development environment. Max/MSP provides users with a graphical environment within which they may create audio applications. There exists a large community of Max/MSP users. Often, developers freely share their 'patches' and custom made 'objects' with other community members.

We have created a special Max patch combining existing user-created Max objects with our own custom made objects in order to handle the task of extracting features from live musical input.

The features we are interested in extracting are the pitch and amplitude of sung vocals, data describing the singer's vocal tone quality, and chord information obtained when the user plays upon the digital keyboard.

Feature extraction from sung vocal input is done using the `fiddle~` object created by Puckette *et al.*[11]. The `fiddle~` object extracts information about the singer's pitch and amplitude. Additionally, `fiddle~` produces raw peak data describing the harmonic spectra of the user's singing voice.

Upon examination of this harmonic spectra, we define a numerical descriptor indicating whether the user's vocal tone amplitude is mainly concentrated at the fundamental frequency, or whether there is also significant tone amplitude distributed amongst higher partials. This allows us to numerically describe an aspect of the singer's vocal timbre. This analysis could be further expanded in the future in order to produce a more detailed measure of vocal timbre, but currently produces a simple parameter that a vocalist can control by modifying the vocal tone he or she employs.

Our own sub-patch monitors MIDI events in order to determine what chords are being played on the keyboard.

4.2 A Perceptual Model for Musical Feature Organization

Western tonal music, while at the lowest level consisting of pitches, durations and amplitudes, is constrained by rules of harmonic structure. Experiments in psychoacoustics by Koelsch *et al.* [6] have shown that perceivable deviations from these harmonic structural rules produce noticeable event-related brain potentials (ERPs). These occur even in the brains of non-musicians, indicating that humans exposed to Western tonal music internalize at some level its harmonic structure. Patel *et al.*[10] report that a trained musician's P600 ERP component responds to harmonic incongruity in the same way as it responds to linguistic incongruity, suggesting similarities between linguistic and harmonic syntactic processing.

Since we are intending to use Western tonal music as input to this system, we have chosen to implement our ANIMUS character's musical perception skills in such a way as is consistent with tonal music theory. We give our ANIMUS character the ability to identify the harmonic context of the vocalized pitches sung by the live musician. This understanding of pitch within a harmonic context is vital to the cognitive processes described in Section 5 of this paper. The rules upon which we base our cognition system assume that the musical input is tonal, so our organizational scheme will facilitate the cognitive processes necessary to infer emotional meaning.

Currently, we are focusing our efforts on the perception of sung vocal melody, as it is the feature that we wish to highlight most prominently through the interaction between the performer and the virtual character.

To organize vocal input within a harmonic context, our system incorporates aspects of an existing music-theoretical melody encoding system devised

by Deutsch and Feroe [4]. Their encoding system must, of course, be adapted for our needs. The real-time nature of our system makes our encoding task different from one based on a traditional harmonic analysis. We do not have the ability to assess an entire musical work at once, but rather must operate in a linear fashion as the musical input arrives.

Deutsch and Feroe describe the representation of absolute pitch values (which we extract from vocalization and keyboard input using Max/MSP patches) as the lowest level of musical information representation. They propose a higher-level system which integrates the raw pitch information into the tonal context within which it is contained. Their system assesses a musical phrase to identify an appropriate pitch *alphabet* (chromatic scale, diatonic scale, etc...) which contains each element of the phrase to be described. They choose a dominant event in the phrase to use as a *reference element*. Each note in the phrase is then described in terms of its position in the specified alphabet with relation to the reference element. Additionally, their model allows complex or repetitive phrases to be described in a hierarchical fashion.

We have complied with their notion of an alphabet, a reference element, and the specification of all other elements in terms of their relationship to the reference element. Our system does not encompass all the features of Deutsch and Feroe's model, as it is merely a subset of their extensive music theoretical model. Instead, it takes a more basic approach with the intent that it could be expanded in the future. One way in which our approach must necessarily deviate from Deutsch and Feroe's is that, while their assessment of a dominant element in a melodic phrase benefits from the ability to make this determination after-the-fact, our system must address and encode melodic information as it is performed live, with no previous knowledge of the score. For this reason, we have chosen to use the tonic note of the key signature of the piece as the reference note. All further notes perceived are encoded as they are related to the tonic note of the scale.

– **Example:** Within the key of C, the tonic note, **C**, would be the *reference note*. If the chosen *alphabet* was the chromatic scale (a scale which contains 12 semitones), **D** (a major second above C) would be represented as being *two steps in the alphabet above the tonic reference note*, while **Eb** (a minor third above C) would be *three steps in the alphabet above the tonic reference note*.

As will be discussed further in Section 5, encoding vocal input within its harmonic context will help to simplify the process of relating associated cognitive meaning to intervallic relationships within the melody line.

Although we are not identifying a dominant musical event within each phrase in order to aid in melody encoding, we are still interested in trying to assess the importance of each sung note in the phrase. In order to determine which notes are emphasized by the performer, we choose to characterize emphasized notes as those which can be differentiated from notes in their surroundings due to

increased volume (volume is described by Cooke [2] as a "vitalizing agent"[6] which implies emphasis) or a sudden shift in register (notes significantly higher or lower than their surrounding notes). Our system calculates a suggested 'importance' value for each sung note.

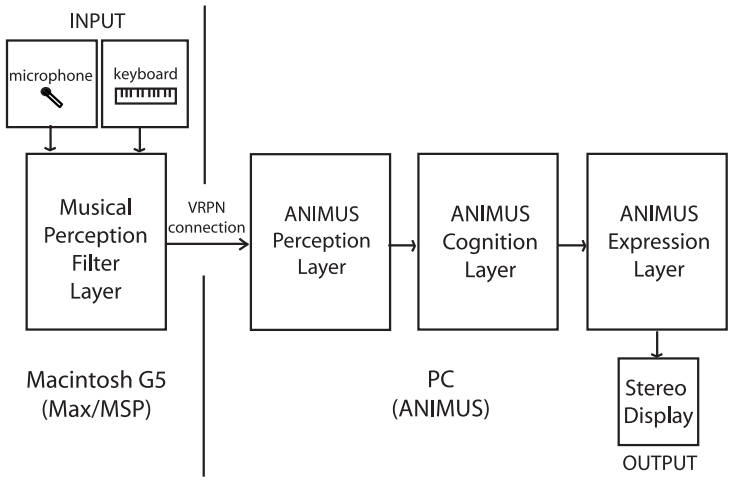


Fig. 3. System Architecture

We must then communicate these extracted events to the PC running the ANIMUS engine via a network connection (see Figure 3 for a system diagram). We have integrated pieces of the Virtual Reality Peripheral Network (VRPN) [13] into the Max environment by custom-creating our own Max external object, `vrpnserver`. This object takes the extracted features as input, and runs a VRPN server on the Macintosh. The PC running the ANIMUS engine's perception layer can connect to `vrpnserver` as a client in order to access the extracted musical feature data. This data is then entered on the ANIMUS blackboard. ANIMUS characters monitor this blackboard to obtain information about the musical performance in order to carry out the cognitive and expressive tasks required to simulate responsive behaviour.

5 Simulating an Emotional Response in the Cognition Layer

In the cognition layer, the information sent by the perception layer must be received and analyzed in order to simulate the internal emotional state of the virtual character. The ANIMUS blackboard contains the information parsed from the live musical performance (sung intervals, data regarding the importance

[6] Cooke, *The Language of Music*, p.94.

of each sung note, information about the singer's vocal timbre, and chord data describing what the user is playing).

We must then simulate a cognitive awareness of the perceived data in order for the ANIMUS character to have an interesting internal state which it may then express through animated behaviours.

Cognitive awareness of musical performance could take many forms. Characters could assign feelings of happiness to a particular melody that they find pleasing, or they could enter a fearful state after perceiving a certain series of chords which they find threatening. Characters could dislike the piercing soprano of the Queen of the Night's coloratura, or express admiration for the rich mezzo tones of Carmen's "Habañera."

In order to create an interesting and flexible cognitive layer for our ANIMUS character, we have chosen to implement aspects of Deryck Cooke's research as described in "The Language of Music" [2]. Cooke's study of a widespread assortment of classical works provides insight into a more generalized relationship between music and emotion. Instead of implementing a character which enjoys one specific melody and dislikes another, we are interested in creating a character which is flexible enough to simulate an emotional response to music-theoretical features within a melody line. Cooke has discerned certain features to be salient features within a large number of musical pieces. He theorizes that certain features used in Western tonal music represent particular emotional concepts in a relatively universal way.

Cooke identifies "the basic expressive functions of all twelve notes of our scale."[7] If a melody contains many instances of the minor third, Cooke's theory states that the proper interpretation of the passage would be "stoic acceptance" or "tragedy".[8] He bases this inference upon many cited examples of musical passages expressive of tragic emotion which contain the minor third, such as Violetta's deathbed scene from Verdi's "La Traviata". Conversely, Cooke cites the American folk song "Polly-wolly-doodle" to exemplify how a major third often signifies "concord" or "joy"[9].

By applying his rules to sung vocal melodies, we can assign cognitive meaning to elements of the live musical performance. As noted in Section 4, our Musical Perception Filter Layer describes all sung melody notes by their tonal context within the existing key signature. Knowing this information, we can easily link each sung pitch to the emotional context extracted by Cooke. This emotional context can then serve to modify the ANIMUS character's internal state and trigger responsive behaviour.

The ANIMUS system uses a 'driver system' to control character behaviour. Familiar to many computer game users, due to its use in "The Sims" [5], a driver system operates on the principle that once the level of a specific character feature reaches a target maximum or minimum, behaviour is triggered. Sims fall asleep on their feet when their energy driver reaches its minimum, but get up

[7] Cooke, *The Language of Music*, p.89.
[8] Cooke, *The Language of Music*, p.90.
[9] Cooke, *The Language of Music*, p.90.

out of their beds when their energy driver is at its maximum value. Similarly, our ANIMUS characters can use driver systems to express their emotional state.

According to Cooke's theories, a minor third signifies tragedy, while a major third signifies joy. Our ANIMUS character may have a 'happiness driver', the level of which is increased if the singer sings a melody which contains many instances of major thirds, and decreases if a minor third is sung. Since our perception layer also assigns an 'importance' value to each sung note, the amount by which the 'happiness driver' is increased or decreased may also be dependent upon the importance value of the sung notes.

The other extracted musical features (chords, vocal timbre, etc...) can also be linked to the ANIMUS character drivers in order to influence character response.

The cognitive processing of musical input allows the ANIMUS character to maintain a fluctuating emotional state during the course of a live musical performance. This emotional state is then conveyed to the audience via the ANIMUS expression layer, which uses animations to visualize character behaviour.

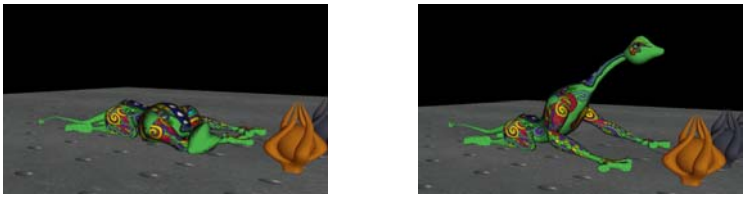
6 Visualizing Character Emotions Through Movement

An ANIMUS character is animated using keyframe-based interpolation. An ANIMUS character is a three-dimensional model that has a controllable skeleton. A variety of endpoint skeletal poses are defined using three-dimensional modelling software. All intermediate poses are generated at run-time in order to generate fluid transitions between these endpoint poses. This allows the animations to be dynamic, a key feature of the ANIMUS engine.

The ANIMUS expression engine allows the designer to define character animations both in terms of which keyframe poses are used to create a motion, and the speed of the transition between these specified poses.

These design decisions can then be linked to the ANIMUS character's cognitive state. For example, Taylor, Torres, and Boulanger describe an ANIMUS character who is attentive to vocalizations within his environment [12]. When the user of the system is silent, the character's cognition layer registers a high value in his 'boredom' driver. His expression layer translates this internal state by displaying him slumped in a 'bored' pose (see Figure 4a). When information about sung vocalizations reaches his perception layer, his cognition layer decreases its 'boredom' driver. His expression layer responds by rapidly transitioning him from his 'bored' pose to his 'attentive' pose, adjusting the position of his head so that he looks towards the perceived source of the sound (see Figure 4b).

We intend to enhance these simple animations by using the extracted emotive properties from a live musical performance as parameters which modify the portrayed 'mood' of the character's actions. Using Cooke's theories to infer an emotional context from a musical performance, extracted emotional indicators can be used to assist in selecting appropriate keyframe poses and transition rates when creating the character's animations.



(a) The 'bored' pose (b) The 'attentive' pose

Fig. 4. A posed ANIMUS character

As an example, consider the case of a 'walking' animation. If the emotional context of a musical passage is inferred to be 'joyful,' keyframe poses can be selected in which the character appears to stand tall and energetic, and transitions between keyframe poses can occur in a smooth and rapid fashion. The character appears to walk with a brisk and confident stride. If the emotional context of the musician's performance begins to reflect 'despair', the keyframe poses can be selected to slump the character's shoulders. Transitions between the poses can slow down in order to reflect a sad and halting step.

7 Conclusion and Future Work

We have described a method of music visualization which provides a visual interpretation of a musical performance's emotional content by modifying the behaviour of a virtual character. We have chosen to base the character's cognitive understanding of emotional content upon the theories of Deryck Cooke.

Currently, we are working on completing our implementation of the character's cognitive layer in order to create an interesting and 'believable' virtual character. We would like to begin collaborating with a visual artist to create a sophisticated virtual character with a wide library of poses, so that we may have greater flexibility within the ANIMUS expression layer to create dynamic animations, evocative of a wide range of character emotions.

When our development of the cognitive and expression layers is complete, we hope that this system could be used in a live performance setting. The visual dialogue between the live performer and the virtual character enriches the musical experience. The virtual character's responses add a visual component that illustrates the emotive capacity of music. We would like to explore the interaction between the human musician and the animated character through an audiovisual piece composed in tandem by a musician and a visual artist.

We believe that this system represents a novel way to illustrate a 'human-like' perceptual and cognitive understanding of the emotive capacity of music.

8 Acknowledgments

The use of the VRPN library was made possible by the NIH National Research Resource in Molecular Graphics and Microscopy at the University of North Carolina at Chapel Hill.

References

1. The ANIMUS Project: A Framework for the Creation of Emotional and Rational Social Agents. `http://www.cs.ualberta.ca/~dtorres/projects/animus` Computing Science Department, University of Alberta.
2. Cooke, D. The Language of Music. New York: Oxford University Press, 1959.
3. Cycling '74. Max/MSP.
4. Deutsch, D. and Feroe, J. The Internal Representation of Pitch Sequences in Tonal Music. Psychological Review, 88, pages 503-522, 1981.
5. Electronic Arts. The Sims.
6. Koelsch, S., Gunter, T., Friederici, A.D. and Schroger, J. Brain indices of music processing: "Nonmusicians" are Musical. Cognitive Neuroscience, 12, pages 520-541, 2000.
7. Levin, G. and Lieberman, Z. In-situ Speech Visualization in Real-Time Interactive Installation and Performance. In Proceedings of the 3rd International Symposium on Non-Photorealistic Animation and Rendering, pages 7-14. ACM Press, 2004.
8. Oliver, W., Yu, J. and Metois, E. The Singing Tree: Design of an Interactive Musical Interface. In DIS'97: Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods and Techniques, pages 261-264. ACM Press, 1997.
9. Ox, J. Two Performances in the 21st Century Virtual Color Organ. In Proceedings of the Fourth Conference on Creativity and Cognition, pages 20-24. ACM Press, 2002.
10. Patel, A.D., Gibson, E., Ratner, J., Besson, M. and Holcomb, P.J. Processing Syntactic Relations in Language and Music: An Event-Related Potential Study. Journal of Cognitive Neuroscience, 10, pages 717-733, 1998.
11. Puckette, M., Apel, T. and Zicarelli, D. Real-Time Audio Analysis Tools for Pd and MSP. In Proceedings of the International Computer Music Conference, pages 109-112. International Computer Music Association, 1998.
12. Taylor, R., Torres, D. and Boulanger, P. Using Music to Interact with a Virtual Character. In Proceedings of New Interfaces for Musical Expression, pages 220-223, 2005.
13. Taylor II, R.M., Hudson, T.C., Seeger, A., Weber, H., Juliano, J. and Helser, A.T. VRPN: A Device-Independent, Network Transparent VR Peripheral System. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology, pages 55-61. ACM Press, 2001.
14. Torres, D. and Boulanger, P. A Perception and Selective Attention System for Synthetic Creatures. In Proceedings of the Third International Symposium On Smart Graphics, pages 141-150, 2003.
15. Torres, D. and Boulanger, P. The ANIMUS Project: A Framework for the Creation of Interactive Creatures in Immersed Environments. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology, pages 91-99. ACM Press, 2003.