A Virtual Head Driven by Music Expressivity

Maurizio Mancini, Roberto Bresin, and Catherine Pelachaud

***Abstract*—In this paper, we present a system that visualizes the expressive quality of a music performance using a virtual head. We provide a mapping through several parameter spaces: on the input side, we have elaborated a mapping between values of acoustic cues and emotion as well as expressivity parameters; on the output side, we propose a mapping between these parameters and the behaviors of the virtual head. This mapping ensures a coherency between the acoustic source and the animation of the virtual head. After presenting some background information on behavior expressivity of humans, we introduce our model of expressivity. We explain how we have elaborated the mapping between the acoustic and the behavior cues. Then, we describe the implementation of a working system that controls the behavior of a human-like head that varies depending on the emotional and acoustic characteristics of the musical execution. Finally, we present the tests we conducted to validate our mapping between the emotive content of the music performance and the expressivity parameters.**

***Index Terms*—Acoustic cues, emotion, expressivity, music, virtual agent.**

I. Introduction

WHAT happens when it is a computer listening to the music? In human–computer interaction (HCI) applications, affective communication plays an increasingly important role. Quality of communication between humans and machines would be improved if systems could express what they perceive and communicate it to the humans through visual and acoustic feedbacks.

Listening to music is an everyday experience. But why do we do it? For example, one could do it for tuning her own mood. Research results show that we are not only able to recognize different emotional intentions used by musicians or speakers [1], but that we also feel these emotions. It has been found that when listening to music, people experience a change in biophysical cues (such as blood pressure, etc.). This change may correspond to either the feeling of the emotion arising from listening the music or to the recognition of the emotion evoked by the music [2].

Virtual agents with a human-like appearance and communication capabilities are being used in an increasing number of applications for their ability to convey complex information through verbal and nonverbal behaviors like voice, intonation, gaze, gesture, facial expressions, etc. Their capabilities are useful when being a presenter on the web [3], a pedagogical agent in tutoring systems [4], a talking head helping hearing-impaired people to "listen" to a telephone call by lipreading [5], a companion in interactive setting in public places such as museums [6], [7], or even a character in virtual story-telling systems [8]. The expressivity of behaviors, that is the way behaviors are executed, is also an integral part of the communication process as it can provide information on the state of an agent, such as current emotional state, mood, and personality [9].

In our work, we have implemented a system that provides a visual display of an acoustic source by altering the facial expression and the quality of movement of a virtual head. The system is meant to show the direct connection between head motion and expressivity in music performance [10].

Possible applications of the new system include visual display of expressivity in music collections and visual feedback to students practicing to play music with expressivity. Indeed, one could think of providing a portable music player with a visual feedback in the form of an expressive face changing expression while browsing the music database stored in the player. Another application is a system helping children to learn how to play a musical instrument with expression: the expressive facial expression would provide a visual feedback on their playing style. However, to implement an application of the system proposed here is out of the scope of this paper, which is limited at the description and testing of its basic functionalities and possibilities.

In Section II, we present a state of the art, and we follow by giving some background information on expressivity for human behavior, voice, and music execution. In Section V, we introduce our real-time application for visual display of musical execution. We provide information on the mapping between acoustic cues and animation parameters. In Section VI, we describe the tests we conducted to validate our mapping between music performances and expressivity parameters. Finally, we conclude the paper.

II. State of the Art

Some previous works [11]–[13] have addressed the generation of synthetic human behavior depending on music (or sound) input. The works by Lee *et al.* [14] and by Cardle *et al.* [12] have mainly focused on adapting precalculated animations like walking or dancing to a given music input. These systems analyze the music and extract parameters such as *tempo*. Based on the values of the extracted parameters, the *rhythm* of the animation is changed. In the works by Cornwell *et al.* [15] and by

Manuscript received November 28, 2006; revised March 12, 2007. This work was supported in part by the EU funded Human-Machine Interaction Network on Emotion Network of Excellence (HUMAINE http://www.emotion-research.net/) and COST 287 Action "Gesture CONtrolled Audio Systems" (ConGAS http://www.cost287.org). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rudolf Rabenstein.

M. Mancini and C. Pelachaud are with the IUT de Montreuil-University of Paris 8, 93100 Montreuil, France (e-mail: mancini@iut.univ-paris8.fr; pelachaud@iut.univ-paris8.fr).

R. Bresin is with the School of Computer Science and Communication, Department of Speech, Music, and Hearing, KTH-Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: roberto@kth.se).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.899256

Downie and Lefford [13], interaction between agents are modulated by music and sound. The emotive content of the acoustic source is correlated to the quality of the interaction between agents. For example, a group of agents will tend to collaborate more when listening to a *happy* and *positive* piece of music [15]. [13] also underlines that music can help to *give life* to inanimate objects, increasing their credibility. Taylor *et al.* [16] developed a system that allows a user to adapt the way she plays a music instrument depending on the reaction of a virtual character. The user may vary her execution to make virtual character reacts in some desired way.

Our work is similar to that of DiPaola *et al.* [11], in which they used MIDI-coded expressive music performances and mapped them into facial expression. The authors emphasize that affective information can be delivered through several means (music, facial expression, body movement, etc.) by translating the original message into the language used by each mean. So, if music is the starting mean and facial expression is the output mean, the system elaborates the expressive information coming from music and translates it into facial expressions. Similarly to our system, DiPaola's system provides means to associate head movements to acoustic characteristics of music (for example, rhythmic head movements could be associated to music beats) but it does not contain any implementation of such correspondence, which is left to the final user (the animation designer). In this paper, we propose both a mapping between music and facial expression, and another mapping between music and head movement expressivity. Moreover, both mappings have been validated through perception tests.

III. Expressivity

Human individuals differ not only in their reasoning, their set of beliefs, goals, and their emotive states, but also in their way of expressing such information through the execution of specific behaviors. We refer to these behavioral differences with the term *expressivity*. In Section III-A, we present the definition of expressivity as the quality of movement in human behavior, while in Section III-B, we describe some work in voice and music expressivity.

A. Expressivity in Behavior

Many researchers (Johansson [17], Wallbott and Scherer [9], Gallaher [18], Ball and Breese [19], and Pollick [20]) have investigated human motion characteristics and encoded them into categories. Some authors refer to body motion using dual qualifiers such as slow/fast, small/expansive, weak/energetic, small/large, unpleasant/pleasant. Behavior expressivity has been correlated to energy in communication, to the relation between temporal/spatial characteristics of gestures, and/or to personality/emotion. For Wallbott [21], it is related to the notion of quality of the mental, emotional, and/or physical state and of quantity (somehow linked to the intensity factor of the mental/emotional/physical state). Behaviors encode not only content information (the "What is communicating" through a gesture shape for example) but also expressive information (the "How it is communicating" through the manner of execution of the gesture). There is evidence that some movement qualities are characteristic to emotions. These qualities are the spatial extension of the movement, its energy/power and the activity [21].

In a recent study by Dahl and Friberg [10, (paper IV]), a correlation between body motion and acoustic cues in expressive music performance has been observed. For example, in angry and happy performances, faster body movements of the performer correspond to faster *tempi*, and larger amount of body movement corresponds to louder sound level; in sad performances, more fluent body movements correspond to a more *legato* articulation. In a perceptual test done within the same study, subjects could recognize the intended emotions of the performers by rating muted video clips, each showing one of three different musicians, a marimba player, a saxophone player, and a bassoon player, performing the same score with four emotional intentions (fear, anger, happiness, sadness). In particular, this study highlights the importance of head movements in the communication of the emotional intentions of the player.

B. Expressivity in Voice and Music

Sound is an important mean of communicating emotions. Much of the essence of speech and music concerns the communication of moods and emotions. Sound can be characterized in terms of a number of physical variables (*cues*): onset time, decay time, pitch, loudness, timbre, and tempo as well as the rate of change of these variables. Combinations of these cues can be used for describing the expression of emotion in sound.

In the last 30 years, K. Scherer and coworkers at the Department of Psychology, University of Geneva, have been conducting extensive work in the field of emotions and in particular in the area of perception of emotions in speech. They shed light on the multifaceted problems related to emotions in vocal expression (for an overview see [22] and [23]). Among other findings, they clearly identified how cues are manipulated in the communication of emotions and specially during appraisal processes [24].

In the past decade, research in music communication has focused on the analysis and formalization of expressive communication [25]–[27]. Striking analogies between spoken and musical communication have been revealed, with respect to how emotions and moods are expressed using acoustic cues [1], [28]. It has been noticed that expressive rendering can help in marking more clearly the structure of the message being communicated [28], and several acoustic cues involved in the communication of emotional expression have been identified [1], [29]–[33]. These cues can be combined in various ways for signaling the same emotion, thus affording robustness in communication via redundancy and variation. In a review of 101 papers on vocal expression and 41 on music performance [1], similarities were found between both channels in their use of acoustic cues for the communication of emotions.

IV. Modeling Expressivity

A. Expressivity for Virtual Agents

1) Behavior Expressivity Parameters: In order to increase its credibility and life-likeness, a virtual agent should not only be able to show an emotional state but also to show it with a

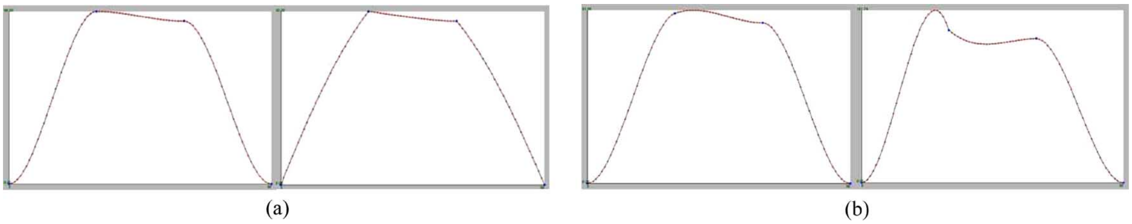


Fig. 1. (a) Fluidity variation. Left diagram represents high fluidity, right diagram represents low fluidity for the same behavior. (b) Power variation. Left diagram represents movement executed with low power, while the right diagram represents the same movement with high power.

certain quality [34]; that is, the agent should be able to alter its way of expressing a given emotion through the application of some *modifications* on the quality of its movements. In our work, we are more interested in what is visually perceived of a given behavior than in the internal reasons (for example, mental state, personality, mood, etc.) that have triggered that behavior. We base our model on perceptual studies. Starting from the results reported in [9] and [18], we have defined the expressivity of behavior over six *dimensions*[34]. In the present work, five of these dimensions influence qualitatively the animation of a virtual agent head. They are as follows.

- *Overall Activity*: Amount of activity (e.g., passive/static or animated/engaged). This parameter influences the number of single behaviors happening during the communication. For example, as this parameter increases, the number of head movements per unit of time will increase. Its value is a floating point number ranging from 0 to 1, where a value of *zero* corresponds to *no activity*, and a value of one corresponds to *maximum activity*.
- *Spatial Extent*: Amplitude of movements (e.g., expanded versus contracted). This parameter determines the amplitude of head rotations. The attribute, like all the following, is a real number defined in the interval $[-1,1]$. A value of *zero* corresponds to a *neutral* behavior, that is the behavior of our virtual agent without any expressivity control; a value of -1 corresponds to the reproduction of very small head rotations, while value of 1 corresponds to very wide rotations.
- *Temporal Extent*: Duration of movements (e.g., quick versus sustained actions). This parameter modifies the speed of execution of movements. Low values produce very fast head rotations while higher values produce slower rotations.
- *Fluidity*: Smoothness and continuity of movement (e.g., smooth, graceful versus sudden, jerky). Higher fluidity allows smooth and continuous execution of movements while lower value creates a discontinuity in the movements. Fig. 1(a) shows the same movement executed with different fluidity values.
- *Power*: dynamic properties of the movement (e.g., weak/relaxed versus strong/tense). Higher (respectively lower) values increase (respectively decrease) the acceleration of the head rotations making movement looking more (respectively less) powerful. Increasing this parameter will also produce movement *overshooting*. Fig. 1(b) shows some examples of curves with different power.

2) Perceptual Tests for Behavior Expressivity: To validate our expressivity model, we performed two perceptual tests [35]. In the first study, we aimed at evaluating the implementation of each expressivity parameter, while in the second study we aimed at understanding if the set of expressivity parameters would allow us to model expressive behaviors. Both tests gave positive results. Subjects could perceive relatively well each expressivity parameter and which movement quality was intended (for more details on the studies see [35]).

B. Automatic Extraction of Expressivity in Music Performance

CUe EXtraction (CUEX) is an algorithm developed at KTH and Uppsala University for extracting acoustical cues from an expressive music performance and mapping them into the corresponding emotional expression [36], [37]. Acoustical cues that can be extracted by CUEX are articulation (*legato* or *staccato*), local tempo (number of events in a given time window), sound level (dB), spectrum energy above 1000 Hz, attack speed (dB/s), musical tone, and *vibrato*. The CUEX algorithm has been validated by testing it on real monophonic expressive performances[1] played with electric guitar, piano, flute, violin, and saxophone. On average, about 90% of tone onsets were correctly detected. CUEX has also been tested with voice and gave similar results. In the future, it will be possible to design algorithms able to extract information about expressiveness even in polyphonic recordings. Indeed, algorithms for the expressive morphing of polyphonic music have been developed [40], [41], and their reverse engineering could be used for the analysis of expression in polyphonic music performance.

Research in music performance has shown that musicians control acoustic cues for communicating emotions when playing [42], [43]. Particular combinations and relative values of the cues correspond to specific emotions. In Table I, we present the use of acoustic cues by musicians when performing with happiness, anger, or sadness. Complete data have been reported by Juslin [43]. CUEX maps the extracted acoustic cues into a two-dimensional space that represents the emotional expressiveness of the performance. The two-dimensional space is defined by the axes pleasure-displeasure (valence) and degree of arousal (activity) as proposed by Russell [44]. In the present work, CUEX maps the extracted acoustic cues onto this two-dimensional activity—valence space using a fuzzy logic approach [45]. For example, if a piece of music is played with *legato* articulation, soft sound level, and slow tempo, it will be classified as “sad”; while it will be classified as “happy” if the performance is characterized by a more *staccato* articulation, louder sound level, and faster tempo. In some situations, it

[1] These performances were collected, and rated in listening tests in previous experiments [38], [39]. Listeners were able to identify the intended emotions in musicians’ performances.

TABLE I
MUSICIANS' USE OF ACOUSTIC CUES WHEN COMMUNICATING EMOTION IN MUSIC PERFORMANCE (FROM [43])

Emotion	Acoustic cues	Emotion	Acoustic cues
Sadness	slow mean tempo	Anger	fast mean tempo
	large timing variations		small tempo variability
	low sound level		high sound level
	legato articulation		staccato articulation
	small articulation variability		spectral noise
	soft duration contrasts		sharp duration contrasts
	dull timbre		sharp timbre
	slow tone attacks		abrupt tone attacks
	flat micro-intonation		accent on unstable notes
	slow vibrato		large vibrato extent
	final ritardando		no ritardando
Happiness	fast mean tempo		rising micro-intonation
	small tempo variability		fast tone attacks
	small timing variations		bright timbre
	high sound level		sharp duration contrasts
	little sound level variability		staccato articulation
	large articulation variability		

can happen that the acoustic cues do not reach their extreme values on the different scales, but values which are in between those characterizing different emotions. In this case, the CUEX algorithm will provide three values which indicate how much the current performance is happy, sad, and angry; that is, CUEX outputs a blend of emotions. CUEX is implemented both in Matlab and PD [46]. The latter is a simplified version of the Matlab version with less precision, but it runs in real-time. In this study, we used the PD implementation.

V. VISUALIZATION OF EXPRESSIVITY: FROM ACOUSTIC CUES TO AN ANIMATED VIRTUAL HEAD

In this section, we turn our attention to an explicit visual representation of expressivity and emotion in music performances. The system, called Music2Greta, has been realized by integrating the CUEX system described in Section IV-B with the Greta virtual agent [47], see Fig. 2. It works as follows. Acoustic cues extracted by CUEX and the deducted emotional content are elaborated in real-time; they drive the behavior of the Greta's head. The module called *Acoustic params to expressivity* applies a mapping (Section V-A) between the acoustic cues and the expressivity parameters of the Greta agent (Section IV-A). At the same time, the *Emotion blending* module generates the facial expression that is displayed by the Greta's face (Section V-B). Finally, the animation of the virtual head is generated by the *Animation generation* module as described in Section V-C.

A. Mapping Acoustic Cues to Expressivity Parameters

The acoustical cues extracted by CUEX (that is, sound level, tempo, articulation) are linearly mapped into the behavior expressivity parameters using a scaling factor to adapt their ranges of variation. The variation of each expressivity parameter is as follows.

- *Sound level*. The current sound level of the music performance is linearly mapped onto the *Spatial Extent* and *Power* expressivity parameters. Thus, it influences the angle of rotation of head movements (*Spatial Extent*) as well as their acceleration and quantity of overshooting (*Power*).

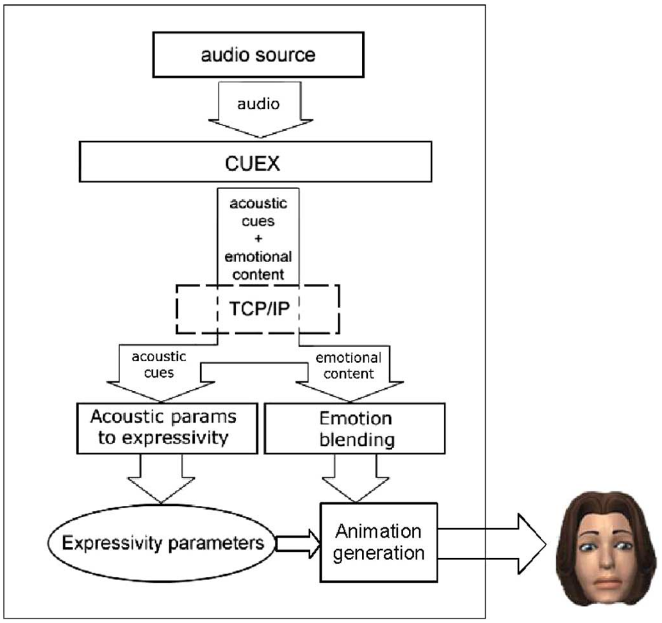


Fig. 2. Music2Greta architecture.

- *Tempo*. This parameter represents the local *tempo* of the musical performance. It influences *Temporal Extent* and *Overall Activity* expressivity parameters. That is, it acts on the duration of head movements (*Temporal Extent*), and on the frequency of head movements (*Overall Activity*).
- *Articulation*. It reflects the style and the quantity of the articulation in the music performance, i.e., the amount of *staccato* or *legato*. It varies the *Fluidity* expressivity parameter. This means that it acts on the continuity of consecutive head movements making them less continuous and less coarticulated as the execution of music becomes more and more *staccato*.

We can notice that our mapping establishes a mimicry between sound quality and movement quality: loud sound is matched by large head movement; fast speed tempo by rapid movement; staccato performance by discontinuous movement, etc. These positively correlated relations between acoustic cues

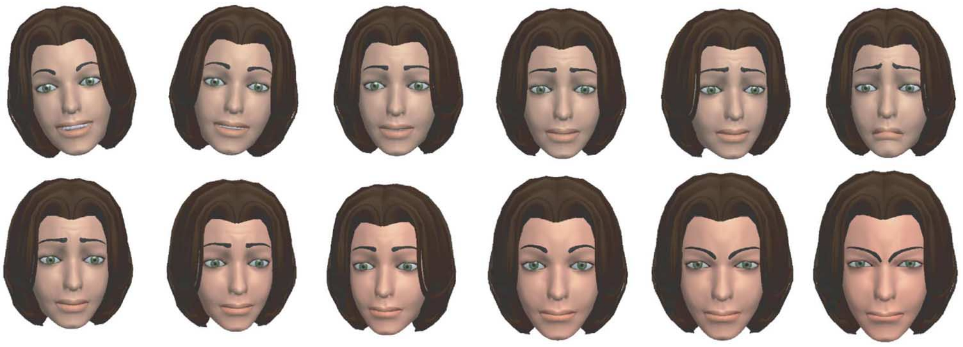


Fig. 3. This sequence shows an example of output of the Music2Greta system. From left to right and top to bottom, we can see the agent tilting her head on the side while displaying happiness. Then this facial expression fades to sadness as the head rotates downward. Finally, its expression changes into anger with the skin reddening and the head leaning towards the user.

variation and behaviors quality variation have been noticed between pitch accent on emphatic words and nonverbal behaviors [48].

B. Mapping Emotional Intention Onto the Agent's Face

The emotional intention recognized in the music performance by the CUEX system is mapped onto the facial expression to be displayed by the agent Greta. As described in Section IV-B, the CUEX system determines the emotional content of a music performance in real-time. The result takes the form of a three-dimensional vector whose coordinates are the amount of happiness, anger, and sadness recognized by the system in the actual performance. As described earlier, it is possible that one, two, or three emotions are present at the same time, but with different strength. In the last two cases, a blending of emotions is computed. The final facial expression is obtained by applying the rules defined by Ekman and Friesen [49] for blend of emotions. Two facial areas (*upper face* (eyes and eyebrows) and *lower face* (cheeks and mouth)) are considered [50], [51]. The expression on the upper face of one expression is combined with the expression on the lower face of the other expression. The combination follows Ekman's research [49] that states that expressions of negative emotions are mainly recognized from the upper face (e.g., frown of anger), while positive emotions are from the lower face (e.g., smile of happiness). By applying these findings, we have elaborated the following rules.

- If *anger* and *sadness* are present: The lower face shows anger (tense lips) and the upper face displays sadness (inner raise eyebrows).
- If *anger* and *happiness* are present: The lower face shows happiness (smile) and the upper face displays anger (frown).
- If *happiness* and *sadness* are present: The lower face shows happiness (smile) and the upper face displays sadness (inner raise eyebrow).
- If *anger*, *sadness*, and *happiness* are all three present: The lower face shows happiness (smile) and the upper face displays sadness (inner raise eyebrow). Anger will be revealed through rapid head movements.

C. Generating Animation

The *Animation generation* module of the main system 2 is an endlessly running process that computes the animation of the virtual agent head in real-time. The animation is obtained as a sequence of *keyframes* where each keyframe is defined as a particular configuration of facial expression and head orientation. Movement is obtained by interpolating between these keyframes using TCB splines [52]. As the module receives the expressivity parameters and a three-dimensional emotion vector from the other modules of the system (see Sections V-B and V-A) it determines the corresponding keyframes and performs the interpolation in real-time.

Finally, the generation module realizes two additional visible effects on the agent face, namely skin color and head direction of movement [49]. For example, when the music performance becomes "angry," the face skin becomes redder while the head leans forward; for sadness emotion, the head tends to look down while face skin becomes paler. Fig. 3 shows an example of output of the Music2Greta system.

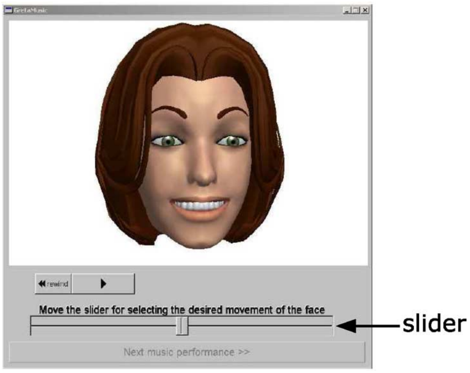
VI. Testing the Mapping Between Emotional Intention and Expressivity Parameters

Our system proposes a mapping through several parameter spaces: from music performance to emotional intention, from emotional intention to facial expression and expressive head movement, and from expressivity parameters to animation. In recent studies, we have already conducted evaluation and validation of the first [31] and third mapping [34] (see Sections IV-A2 and IV-B). In this paper, we present the results of evaluation tests in which the goal was to validate our mapping between the emotional intention (as extracted from the acoustic cues) and the expressivity parameters of the virtual agent head.

A. Experimental Setup

Two groups of subjects took part to the tests. Group 1 consisted of researchers and doctoral students of music acoustics and speech technology at KTH, three females and four males, aged 25–46 (average 32), who played a musical instrument on average for 14 years. Group 2 was composed of researchers and doctoral students at Paris 8, two females and four males, aged 24–44 (average 32), who played a musical instrument on average for five years. In total, subjects were 13 and of ten different nationalities.

As musical stimuli, we used performances of two melodies, Brahms' first theme of the *poco allegretto* 3rd movement Symphony Op.90 No.3, in C minor, and Haydn's theme from first



Move the slider for selecting the desired movement of the face

Next music performance >>

Fig. 4. Screen-shot of the interface used for the tests. On top of the window there is the agent's head; below there is the slider used by the subjects for controlling the quality of the head movement.

movement of Quartet in F major for strings, Op. 74 No. 2. The two melodies were performed by a professional guitar player with three emotional intentions: anger, sadness, and happiness. The testing set of musical pieces consisted of $3 * 2 = 6$ stimuli (three emotions, two melodies).

Participants were asked to sit in front of a PC where the test application was running. They were also instructed to read the explanation of the test procedure before starting the test.

For the purposes of this test, we realized a slightly different version of the Music2Greta system. The test consisted in listening to the six pieces of music while watching the virtual agent's head moving on the screen and having the possibility of altering its movement quality (e.g., faster movement, lower amplitude) by moving a slider on the screen (Fig. 4 shows the interface used for the tests).

For each musical stimulus, the agent's face displayed the appropriate emotion. Each subject was instructed to change the position of the slider on the screen and see how the agent's head changed its movement quality (the slider did not affect the emotion shown on the agent's face, which was fixed for each piece of music). When the subject found a good match between the musical stimulus and the agent's head movement, she could go ahead to the next stimulus.

The slider value influenced the agent's head movement quality by altering the value of the five expressivity parameters described in Section IV-A. To do so, we had to create a mapping from a one-dimension variable (the slider value) to a five-dimensions space (the expressivity parameters values).

At first, we have defined expressivity values should be considered as the *expected expressivity values* for each of the three emotions in our tests (anger, sadness, and happiness) based on perceptual studies conducted by Wallbott [21] and Gallaher [18]. Then, we have associated this set of *expected expressivity values* to a randomly chosen value of the slider which could be anywhere along the range of variation of the slider. The position of the slider associated to the *expected expressivity values* was unknown to the subjects.

Let us give an example. Fig. 5 shows a graph with the correspondence between the slider position (X axis) and the five expressivity values (Y axis). In this figure, the value $X = -0.3$ (that is $slider = -0.3$) corresponds to the predecided *expected expressivity values* (the dashed box) for the emotion for a given piece of music. Let us see the variation of just one of the expressivity parameters, e.g., Power (PWR). Position $X = 0$ in Fig. 5 corresponds to PWR $= 0.5$. By moving the slider towards the left, i.e., smaller X values, the value of PWR tends to -1, while by moving the slider towards the right PWR tends to 1. Similarly, the other parameters overall activity (OAC), fluidity (FLD), temporal extent (TMP), and spatial extent (SPC) are simultaneously varied when adjusting the slider position.

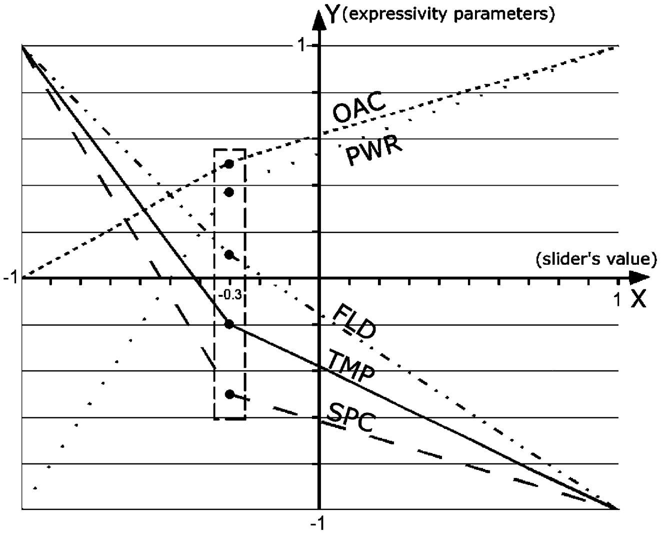


Fig. 5. Example showing the correspondence between the value of the slider (X) and the values of the five expressivity parameters (Y). In the dashed box, we highlighted the *expected expressivity values* for the given example. PWR = Power, OAC = Overall Activity, FLD = Fluidity, TMP = Temporal Extent, and SPC = Spatial Extent.

At the end of the test, the subjects' choice, that is the final position X of the slider, is compared to the expected value ($X = -0.3$ in our example) to check whether our set of *expected expressivity values*, PWR, OAC, FLD, TMP, and SPC is correctly perceived by the subjects.

Subjects could listen to each musical stimulus as many times as they liked to. They could constantly change the position of the slider while watching the corresponding head's behavior on the screen. The order of the six musical stimuli was randomized for each subject. The right and left extremes of the slider were randomly switched between subjects.

B. Results and Discussion

Since the data collected for the two groups of subjects were not significantly different, they were pooled together in the analysis that follows. The emotional intention of the face and the performances (the independent variable) had a considerable effect on the listeners positioning of the slider, that controls overall activity, spatial extent, temporal extent, fluidity and power (the dependent variables). The main results for the angry, sad, and happy emotional intentions are plotted in Fig. 6. As one can observe, there is an interaction of the tonality of the musical stimuli with the expressive movements chosen by the subjects. Note that tonality is not one of the cues identified by the CUEX algorithm. It is well known that minor tonality is often associated to sadness and major tonality to happiness. This can explain

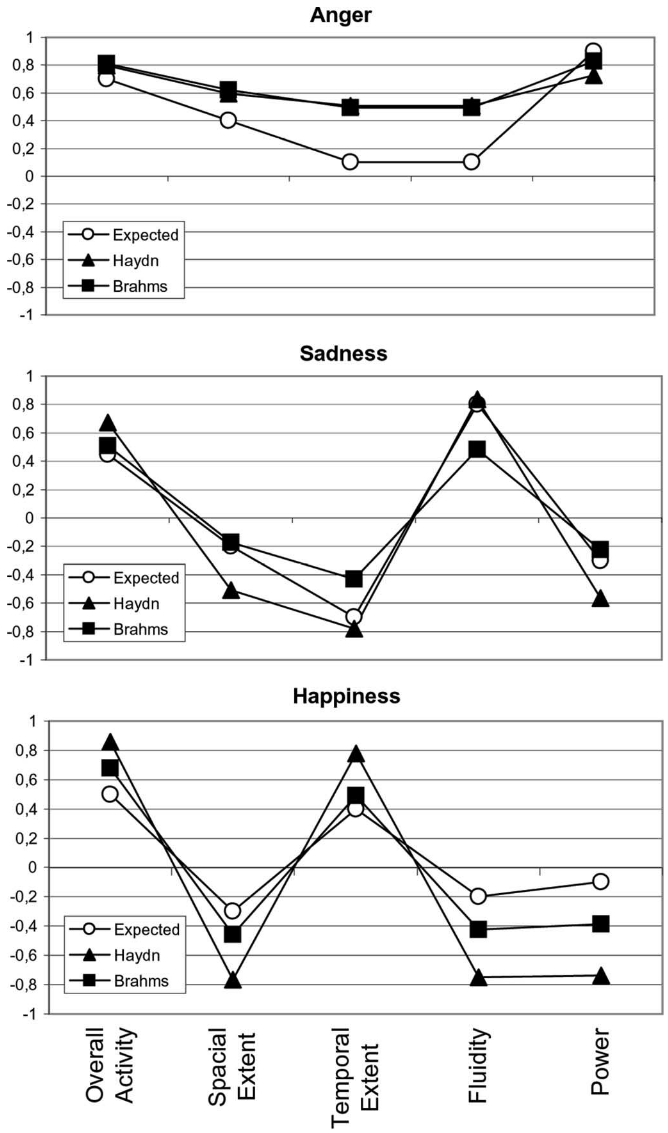


Fig. 6. Mean values for overall activity, spatial extent, temporal extent, fluidity and power as resulted from the test. Empty circles represent the expected values. Full triangles represent the mean values when the major melody was played. Full squares represent the mean values when the minor melody was played.

why in this experiment, for the emotional intentions of sadness and happiness, subjects tend to choose slightly different expressive movements for the same facial expression depending on the musical stimulus. In particular, when a facial expression was presented together with the Haydn's melody, in major tonality, subjects tend to choose expressive facial movements with higher overall activity. Subject did not match the expected angry facial expression, when they were listening to angry versions of the two melodies. Nevertheless, subjects selected the same expressive facial movements for the angry versions of both melodies.

The *expected expressivity values* were originally chosen for facial expressions that were not necessarily associated to music listening. This could partly explain the deviations from expected values observed in this experiment. Greta's expressivity settings, as given by subjects, were also influenced by the tonality, and the overall character of the musical piece. Subjects' choices suggest that the expressive movements showed by Greta could depend on the modality and the context. In our case, it is clear that the expressive movements should be controlled differently when providing display of expressive music than when speaking.

VII. Conclusion

In this paper, we have presented an application in which an animated virtual head is used as a visual display on an expressive music performance. The acoustic parameters in the performance, such as tempo, sound level, and articulation, are extracted and analyzed. Their values are used to identify the emotional intention of the performer. The values of these parameters are also mapped onto the behavior expressivity parameters controlling the movement quality of the virtual head. Finally, the emotional intention and the expressivity parameters are sent to the virtual head that shows facial expressions of emotion, and expressive head movement. Thus, we have a visualization of expressive music performance which can lead to the applications proposed in the introduction of this paper. Future developments of the system will allow analysis and display of any audio stream including voice. A possible application would therefore be a virtual butler whose expressive behavior is driven by the acoustic input from the local, remote, or virtual environment. The butler would give real-time, silent, and informative feedback about the acoustic environment.

Acknowledgment

The authors would like to thank B. Hartmann for developing the gesture expressivity model and to S. Buisine for conducting the evaluation test of this model.

References

[1] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychol. Bull.*, vol. 129, no. 5, pp. 770–814, 2003.

[2] C. L. Krumhansl, "An exploratory study of musical emotions and psychophysiology," *Can. J. Exp. Psychol.*, vol. 51, no. 4, pp. 336–352, 1997.

[3] H. Welbergen, A. Nijholt, D. Reidsma, and J. Zwiers, "Presenting in virtual worlds: Towards an architecture for a 3D presenter explaining 2D-presented information," in *Lecture Notes Computer Science*. Berlin, Germany: Springer-Verlag, 2005, vol. 3814, pp. 203–212.

[4] W. L. Johnson, "Animated pedagogical agents for education training and edutainment," in *Proc. ICALT*, 2001, p. 501.

[5] J. Beskow, I. Karlsson, J. Kewley, and G. Salvi, "Synface—A talking head telephone for the hearing-impaired," in *Proc. Comput. Helping People With Special Needs—ICCHP 2004*, K. Miesenberger, J. Klaus, W. Zagler, and D. Burger, Eds., 2004, pp. 1178–1186.

[6] L. Chittaro, L. Ieronutti, and R. Ranon, "Navigating 3D virtual environments by following embodied agents: A proposal and its informal evaluation on a virtual museum application," *Psychol. J. (Special Issue on Human–Computer Interaction)*, vol. 2, no. 1, pp. 24–42, 2004.

[7] S. Kopp, L. Gesellensetter, N. Krämer, and I. Wachsmuth, "A conversational agent as museum guide—Design and evaluation of a real-world application," in *Intelligent Virtual Agents*. Berlin, Germany: Springer-Verlag, 2005, pp. 329–343.

[8] E. Figa and P. Tarau, "The VISTA project: An agent architecture for virtual interactive storytelling," in *Proc. TIDSE*, N. Braun and U. Spierling, Eds., Darmstadt, Germany, 2003, pp. 106–116.

[9] H. G. Wallbott and K. R. Scherer, "Cues and channels in emotion recognition," *J. Personality Social Psychol.*, vol. 51, no. 4, pp. 690–699, 1986.

[10] S. Dahl, "On the beat: Human movement and timing in the production and perception of music," Ph.D. dissertation, Speech, Music and Hearing, KTH, Royal Inst. Technol., Stockholm, Sweden, 2005.

[11] S. DiPaola and A. Arya, "Affective communication remapping in musicface system," in *Proc. 10th Eur. Conf. Electron. Imag. Visual Arts (EVA'04)*, London, U.K., Jul. 2004.

[12] M. Cardle, L. Barthe, S. Brooks, and P. Robinson, "Music-driven motion editing: Local motion transformations guided by music," in *Proc. EGUK Eurographics UK Conf.*, Jun. 2002, pp. 38–44.

[13] M. Downie and N. Lefford, "Underscoring Characters," Mass. Inst. Technol. Cambridge, May 1999.

[14] H.-C. Lee and I.-K. Lee, "Automatic synchronization of background music and motion in computer animation," *Comput. Graph. Forum*, vol. 24, no. 3, pp. 353–361, 2005.

[15] J. Cornwell and B. Silverman, "A demonstration of the pmf-extraction approach: Modeling the effects of sound on crowd behavior," in *Proc. 11th BRIMS, SISO*, May 2002, pp. 107–113.

[16] R. Taylor, D. Torres, and P. Boulanger, "Using music to interact with a virtual character," in *Proc. Int. Conf. New Interfaces for Musical Expression*, 2005, pp. 220–223.

[17] G. Johansson, "Visual perception of biological motion adn a model for its analysis," *Percept. Psychophys.*, vol. 14, pp. 201–211, 1973.

[18] P. E. Gallaher, "Individual differences in nonverbal behavior: Dimensions of style," *J. Personality Social Psychol.*, vol. 63, no. 1, pp. 133–145, 1992.

[19] G. Ball and J. Breese, "Emotion and personality in a conversational agent," in *Embodied Conversational Characters*, S. P. J. Cassell, J. Sullivan, and E. Churchill, Eds. Cambridge, MA: MIT Press, 2000.

[20] F. E. Pollick, "The features people use to recognize human movement style," in *Gesture-Based Communication in Human-Computer Interaction—GW 2003*, ser. Lecture Notes in Artificial Intelligence, A. Camurri and G. Volpe, Eds. New York: Springer-Verlag, 2004, no. 2915, pp. 10–19.

[21] H. G. Wallbott, "Bodily expression of emotion," *Eur. J. Social Psychol.*, vol. 28, pp. 879–896, 1998.

[22] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, pp. 227–256, 2003.

[23] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," in *Handbook of the Affective Sciences*, R. J. Davidson, H. Goldsmith, and K. R. Scherer, Eds. New York: Oxford Univ. Press, 2003, pp. 433–456.

[24] K. R. Scherer, A. Schorr, and T. Johnstone, Eds., *Appraisal Processes in Emotion: Theory, Methods, Research*, ser. Affective Science. New York: Oxford Univ. Press, 2001.

[25] A. Gabrielsson, "The performance of music," in *The Psychology of Music*, D. Deutsch, Ed. San Diego, CA: Academic, 1999, pp. 501–602.

[26] A. Gabrielsson, "Music performance research at the millennium," *Psychol. Music*, vol. 31, no. 3, pp. 221–272, 2003.

[27] A. Friberg and G. U. Battel, "Structural communication," in *The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning*, R. Parncutt and G. E. McPherson, Eds. New York: Oxford Univ. Press, 2002, pp. 199–218.

[28] J. Sundberg, "Emotive transforms," *Phonetica*, vol. 57, pp. 95–112, 2000.

[29] P. N. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," *Emotion*, vol. 1, no. 4, pp. 381–412, 2001.

[30] P. N. Juslin and J. A. Sloboda, Eds., *Music and Emotion: Theory and Research*. New York: Oxford Univ. Press, 2001.

[31] P. Laukka, P. N. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cogn. Emotion*, vol. 19, no. 5, pp. 633–653, 2005.

[32] K. R. Scherer, "Emotion expression in speech and music," in *Music, Language, Speech, Brain*, J. Sundberg, L. Nord, and R. Carlson, Eds. London, U.K.: Macmillan, 1991, pp. 146–156.

[33] K. R. Scherer, "Expression of emotion in voice and music," *J. Voice*, vol. 9, no. 3, pp. 235–248, 1995.

[34] B. Hartmann, M. Mancini, and C. Pelachaud, "Implementing expressive gesture synthesis for embodied conversational agents," in *Proc. 6th Int. Workshop Gesture Human–Comput. Interaction Simulation*, VALORIA, Univ. Bretagne Sud, France, 2005, pp. 188–199.

[35] B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud, "Design and evaluation of expressive gesture synthesis for embodied conversational agents," in *Proc. 3rd Int. Joint Conf. Autonomous Agents Multi-Agent Syst. (AAMAS)*, Utretch, Jul. 2005.

[36] A. Friberg, E. Schoonderwaldt, P. N. Juslin, and R. Bresin, "Automatic real-time extraction of musical expression," in *Proc. Int. Comput. Music Conf. (ICMC 2002)*, San Francisco, CA, 2002, pp. 365–367, Int. Computer Music Association.

[37] A. Friberg, E. Schoonderwaldt, and P. N. Juslin, "CUEX: An algorithm for extracting expressive tone variables from audio recordings," *Acoustica United With Acta Acoustica*, to be published.

[38] A. Gabrielsson and P. N. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychol. Music*, vol. 24, pp. 68–91, 1996.

[39] P. N. Juslin and E. Lindström, "Musical expression of emotions: Modeling composed and performed features," in *Abstracts 5th ESCOM Conf.*, 2003.

[40] F. Gouyon, L. Fabig, and J. Bonada, "Rhytmic expressiveness transformation of audio recordings: Swing modifications," in *Proc. 6th Int. Conf. Digital Audio Effects (DAFX-03)*, London, U.K., 2003, pp. 94–99.

[41] M. Vinyes, J. Bonada, and A. Loscos, "Demixing commercial music productions via human-assisted time-frequency masking," in *Proc. AES 120th Convention*, Paris, France, 2006.

[42] A. Gabrielsson and P. N. Juslin, "Emotional expression in music," in *Handbook of Affective Sciences*, H. H. Goldsmith, R. J. Davidson, and K. R. Scherer, Eds. New York: Oxford Univ. Press, 2003, pp. 503–534.

[43] P. N. Juslin, "Communicating emotion in music performance: A review and a theoretical framework," in *Music and Emotion: Theory and Research*, P. N. Juslin and J. A. Sloboda, Eds. New York: Oxford Univ. Press, 2001, pp. 305–333.

[44] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[45] A. Friberg, "A fuzzy analyzer of emotional expression in music performance and body motion," in *Music and Music Science 2004*, J. Sundberg and W. Brunson, Eds., Stockholm, Sweden, 2005.

[46] M. Puckette, "Pure data," in *Proc. Int. Comput. Music Conf. (ICMC 1996)*, San Francisco, CA, 1996, pp. 269–272.

[47] C. Pelachaud and M. Bilvi, "Computational model of believable conversational agents," in *Communication in Multiagent Systems*, ser. Lecture Notes in Computer Science, M.-P. Huget, Ed. New York: Springer-Verlag, 2003, vol. 2650, pp. 300–317.

[48] D. Bolinger*, Intonation and Its Part*. Stanford, CA: Stanford Univ. Press, 1986.

[49] P. Ekman and W. Friesen*, Unmasking the Face: A Guide to Recognizing Emotions From Facial Clues*. Englewood Cliffs, NJ: Prentice-Hall., 1975.

[50] T. D. Bui, D. Heylen, M. Poel, and A. Nijholt, "Generation of facial expressions from emotion using a fuzzy rule based system," in *Proc. 14th Australian Joint Conf. Artif. Intell. (AI 2001)*, D. C. M. Stumptner and M. Brooks, Eds., Adelaide, Australia, 2003, pp. 83–94.

[51] M. Ochs, R. Niewiadomski, C. Pelachaud, and D. Sadek, "Intelligent expressions of emotions," in *Affective Computing and Intelligent Interaction, 1st Int. Conf.*, ser. Lecture Notes in Computer Science, J. Tao, T. Tan, and R. W. Picard, Eds. New York: Springer, 2005, vol. 3784, pp. 707–714.

[52] D. H. U. Kochanek and R. H. Bartels, "Interpolating splines with local tension, continuity, and bias control," in *Proc. Comput. Graphics (SIGGRAPH '84)*, H. Christiansen, Ed., 1984, vol. 18, pp. 33–41.

Maurizio Mancini received the M.S. degree in computer science at the University of Rome "La Sapienza," Rome, Italy, in 2002, with a M.S. thesis entitled "Analysis and synthesis of communicative gestures for embodied conversational agents." He is currently pursuing the Ph.D. degree under the supervision of Prof. C. Pelachaud at the University of Paris 8, Montreuil, France. His subject is the generation of multimodal behaviors for ECAs, with a particular emphasis in the elaboration of a synchronous scheme to manage the behaviors coming from different modalities (face, gaze, arm/hand gesture, body movement).

In 2003, he participated in the EU-project MagiCster, whose main goal was the creation of believable embodied conversational agents. He has been one of the main developers of the ECA system Greta while a Researcher at the University of Rome "La Sapienza," as well as at the University of Paris 8. He is participating to the HUMAINE Network of Excellence, working on a model of expressive behaviors.

Roberto Bresin received the Ph.D. degree in music acoustics, with a thesis entitled "Virtual virtuosity—Studies in automatic music performance."

He has been Researcher at the School of Computer Science and Communication, Department of Speech, Music, and Hearing, KTH-Royal Institute of Technology, Stockholm, Sweden, since August 1996. His main research interest is expressive music performance. He is currently involved in the EU-financed S2S^2 Coordination Action, BrainTuning, Project in the framework of NEST Measuring the Impossible, HUMAINE Network of Excellence, and in the COST 287 Action ConGAS. He has previously worked on three European projects, The Sounding Object (SOb), A GNU/Linux Audio distribution (AGNULA), and Multisensory Expressive Gesture Applications (MEGA), and in a national project, Feedback Learning in Musical Expression (FEEL-ME). Before coming to Department of Speech, Music, and Hearing, he was a Research Engineer at Centro di Sonologia Computazionale, Padova University, Padova, Italy, for five years.

Catherine Pelachaud received the Ph.D. degree in computer graphics from the University of Pennsylvania, Philadelphia, PA, in 1991.

She is a Professor at the IUT de Montreuil, University of Paris 8, Montreuil, France. She leads the Research Laboratory of Informatics and Communication (LINC). Her research interest includes representation language for agent, embodied conversational agent, nonverbal communication (face, gaze, arm/hand gesture), expressive behaviors, and multimodal interfaces. She has been involved in several European projects related to multimodal communication (EAGLES, IST-ISLE) and to believable embodied conversational agents (IST-MagiCster). She is a member of the Steering Committee of the HUMAINE Network of Excellence and coordinates the workpackage entitled "emotion in interaction." She is currently participating to the IP CALLAS project.