AUDIO ENGINEERING SOCIETY
Italian Section

ANNUAL MEETING 2005

Paper: 05001

Como, 9-12 November
Politecnico di MILANO

COMPUTATIONAL MODELS FOR AUDIO EXPRESSIVE COMMUNICATION

GIOVANNI DE POLI, GIANLUCA D'INCA', LUCA MION[1]

[1] *Centro di Sonologia Computazionale, Dept. of Information Engineering, University of Padova, Italy*
{depoli, gianluca.dinca, luca.mion}@dei.unipd.it

Audio objects have an important role in multimedia communication, and audio expressive content can enrich the Human Computer interaction in multimedia systems. In this paper expressive analysis, modelling and detection paradigms are reviewed and future research efforts are discussed.

1 INTRODUCTION

Expression is an important aspect of music performance. It is the added value of a performance, and is part of the reason that music is interesting to listen to and sounds alive. Moreover, understanding and modelling expressive content communication is important in many engineering applications. In human musical performance, acoustical or perceptual changes in sound are organized in a complex way by the performer in order to communicate musical content to the listener. The same piece of music can be performed trying to convey a specific interpretation of the score, by adding mutable expressive intentions. The analysis of these systematic deviations has led to the formulation of several models that try to describe their structures, and aim at explaining where, how and why a performer modifies, sometime in an unconscious way, what is indicated by the notation of the score. Modelling paradigms and problems are reviewed and issues for future research efforts are discussed.

2 SOUND IN MULTIMEDIA

In multimedia products, textual information is enriched by means of graphical and audio objects. A correct combination of these elements is extremely effective for the communication between author and user. Usually, attention is put on visual rather than sound, which is merely used as a realistic complement to image, or as a musical comment to text and graphics. With increasing interaction, while the visual part has evolved consequently the paradigm of the use of audio has not changed adequately, resulting in a choice among different objects rather than in a continuous transformation on these. An important task is to evolve the use of audio in multimedia, especially regarding at the expressive content: the use of this information can enrich the Human Computer interaction in multimedia systems, leading the user to a deeper fruition of the product.

Sound expression presents multiple facets, having an important role at different levels of sound complexity.

Expressive content is present in non-structured sounds: for example, in 60's science fiction movies synthetic sounds were used to communicate feeling of artificiality and sense of disquietude. At this level, expression can be added to systems for generating and manipulating Auditory Icons (non-speech sounds) like alarm enunciators where the expression can enrich the information about the types/levels of urgency or warning.

On the other side, expression has a key role in the music composition, which is probably the most structured level of music: at this level the composer introduces into

his works his own emotions and sensations to communicate them to the listeners.
We are interested on the expressive content between these two levels, because even if this content is really important in music communication, it has been less studied and explored than the others two levels. In particular, we focus on the expressive content of structured sounds which is not related to the musical score (composition). At this intermediate level the expression is introduced by performer: while playing music, ha acts on the available freedom degrees to give his own interpretation of the musical score.
Thus, at this level expression refers to different qualities of musical gestures, that can be performed following different expressive intentions. Intentions can be related to sensorial or affective characteristics.
Roughly, we can identify studies on three level of gestures: single gestures, simple pattern-based gestures, and structured gestures. In musical context, single gestures are intended as single tones. These single gestures represent the smallest non structured actions which combined together form simple patterns. Thus, single patterns can be represented by scales or repetition of single tones. Highly structured gestures are the performance of scores. Several studies on music performance have demonstrated that it is possible to communicate expressive content at an abstract level, so to change the interpretation of a musical piece. Results of these studies will be presented on the next section.

3 MUSICAL GESTURES AND EXPRESSIVE INTENTIONS

Music is an important means of communication where three actors participate: the composer, the performer and the listener. The composer instils into his works his own emotions, feelings, sensations and the performer communicates them to the listeners. The performer uses his own musical experience and culture in order to get from the score a performance that may convey the compositor's intention.
Different musicians, even when referring to the same score, can produce very different performances. The score carries information such as the rhythmical and melodic structure of a certain piece, but there is not yet a notation able to describe precisely the temporal and timbre characteristics of the sound. The conventional score is quite inadequate to describe the complexity of a musical performance so that a computer might be able to perform it. Whenever the information of a score (essentially note pitch and duration) is stored in a computer, the performance sounds mechanic and not very pleasant. The performer, in fact, introduces some micro-deviations in the timing of performance, in the dynamics, in the timbre, following a procedure that is correlated to his own experience and common in the instrumental practice. From such measurements it would be possible to deduce general performance features and principles.
However no musician plays the same piece in the same way on every single occasion. Each performance depends on the performer's emotional state at that particular moment as well as his/her hypothetical dialogue with other musicians and subjective artistic choices. Moreover, the same piece of music can be performed trying to convey different interpretations of the score and emotions, according to different "expressive intentions", which can be even in contrast with the usual performance praxis of that particular piece. A textual or musical document, in fact, can assume different meanings and nuances depending on how it is played.
A major problem of analysis-by-measurements method is that a specific deviation on one note could originate from several different principles, so the "true" origin may be impossible to trace. It is difficult to identify a multidimensional structure underlying a surface level, merely by analyzing this surface level.
Some musical performance studies tried to understand how the expressiveness is conveyed in music performance using the analysis-by-synthesis approach. To this purpose, some models or rules systems for generating automatic performance were developed. First, an assumption is hypothesized, then it is realized in terms of a synthetic performance, and finally it is evaluated by listening. If needed, the hypothesized principle is further modified and the process repeated. Eventually, a new rule has been formulated. In other words, the method is to teach the computer how to play more musically. The success of this method is entirely dependent on the formulation of hypotheses and on competent listeners.
For the great variety in the performance of a piece, it is difficult to determine a general system of rules for the execution. An important step in this direction was made by Sundberg (KTH) and co-workers. They determined a group of criteria which, once applied to the generic score, can bring to a musically correct performance [4]. Further on, the performer operates on the microstructure of the musical piece not only to convey the structure of the text written by the composer, but also to communicate his own feeling or expressive intention. Quite a lot of studies have been carried out in order to understand how much the performer's intentions are perceived by the listener, that is to say how far they share a common code. Gabrielsson in particular, studied the importance of emotions in the musical message. An increasing number of studies is concerned with how the musician's intentions affect the performance. In this case the experimental material is obtained by asking the performer either to provide different performances of his own choice and to describe the intentions behind them, or to play with a certain intention in mind (see [5] for a review).

In this context, we started researches in order to understand the way an expressive intention can be communicated to the listener and we realized a model able to explain how it can be possible to modify the performance of a musical piece in such a way that it may convey a certain expressive intention. In the following section the main aspects of analysis, modeling and classification of expressive intentions will be presented with reference to our results.

4 ANALYSIS OF EXPRESSIVE INTENTIONS

4.1 Perceptual analysis

Aim of the perceptual analysis is to see whether the performer's intentions are grasped by the listeners and to determine the judgment categories used by the listeners.

We selected a set of scores from Western Classical and Afro-American music. For each score, different performances (correlated with different expressive intentions) were played by professional musicians. Two different factor analysis were made. Factor analysis on adjectives (e.g. see fig. 2) allowed us to determine a semantic space defined by the adjectives proposed to the listeners. By means of factor scores, it was possible to insert the performances into this space. The comparison between the performance positions and the evaluation adjectives demonstrated a good recognition of the performers intentions by the subjects.

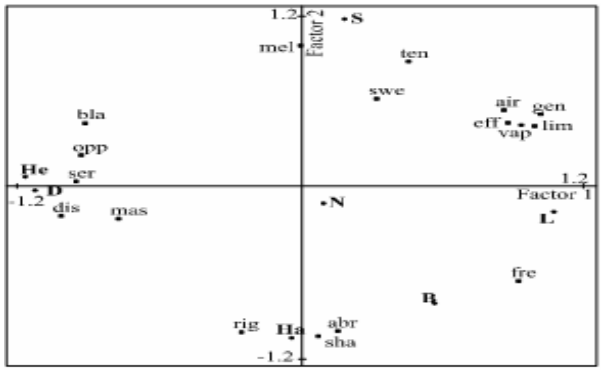

Fig. 2: Factor analysis on adjectives. Evaluation adjectives: black, oppressive, serious, dismal, massive, rigid, mellow, tender, sweet, limpid, airy, gentle, effervescent, vaporous, fresh, abrupt, sharp. Performances: Neutral, Light, Bright, Hard, Dark, Heavy, Soft. It can be noticed a good recognition of the performers intentions by the subjects.

The second factor analysis used performances as variables (e.g. see fig. 3). It showed that the subjects had placed the performances along only two axes. The two dimensional space (Perceptual Parametric Space, PPS) so obtained represents how subjects arranged the pieces in their own minds. The first factor (expressive intention bright vs. expressive intention dark) seem to be closely correlated to the acoustic parameters which regard the kinetics of the music (for instance Tempo). The second factor (expressive intention soft vs. expressive intention hard) is connected to the parameters which concern the energy of the sound (intensity, attack time). Other acoustic parameters (e. g. legato, brightness) are related to the PPS axes and were deduced from acoustic analyses [1].

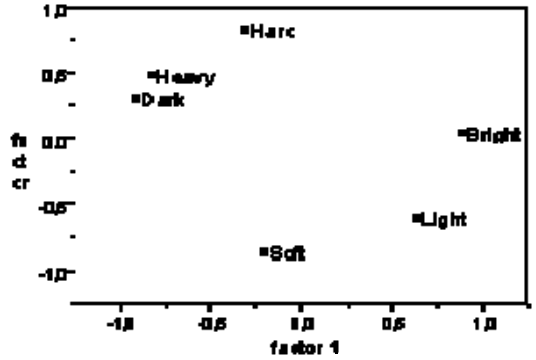

Figure 3: factor analysis using performance as variables. First factor is correlated with kinetics of music; second factor is correlated with energy of the sound.

4.2 Acoustic analysis

Acoustic analysis aims to identify which physical parameters, and how many of them, are subject to modifications when the expressive intention of the performer is varied.

Every musical instrument has its own expressive resources (vibrato in strings, the tongue in wind instruments, etc.), which are used by the musician to communicate his expressive intention. It is inevitable, therefore, that the results of any acoustic measure depend, not only on the score, but also on the characteristics of the instrument used and the strategy adopted by the musician. Consequently, it is necessary to compare the data relative to different scores, musicians and instruments, in order to identify the expressive rules that can be considered valid in a general way and which are specific cases. Several acoustic analyses have been carried out on various musical pieces using different instruments and performers. The recordings are either in MIDI or audio format. A typical relation among expressive intentions and acoustic parameters variation in relation to a neutral (i.e. a performance without any expressive intentions) performance is shown in tab. 1.

	Hard	Soft	Heavy	Light	Bright	Dark
Tempo	+		- -	+ +	+ + +	

Legato		+	+	-	--	
Attack Duration	-	+		+	- -	-
Dynamic		+		-	-	+
Up-beat / Down-beat		-		+		
Envelope Centroid	beginning	center	beginning			center
Brightness	+ +	- -	+	-	+ +	- -
Vibrato		+ +	+			+ +

Table 1: relation among expressive intentions and acoustic parameters variation in relation to a neutral performance in Violin Sonata Op. V by Arcangelo Corelli.

5 MODELING EXPRESSIVE INTENTIONS

According the analysis-by-synthesis method, using results of the analysis and expert's experience, some models for producing performances with different expressive intentions has been developed. Combinations of KTH performance rules and of their parameters were used for synthesizing interpretations that differ in emotional quality.

We developed models suitable to compute the expressive deviations necessary to the rendering step in order to synthesize an expressive performance starting from a neutral one. The rendering step can be done in MIDI or by real time post-processing a recorded human performance. The system was used to generate performances of different musical repertories. Besides the fact that the models were developed mainly for western classical music, they showed a general validity in its architecture, even if some tuning of the parameters is needed. Expressive syntheses of pieces belonging to different musical genres (European classical, European ethnic, Afro-American) verified the generalization of the rules used in the models [1].

6 AUTOMATIC DETECTION OF EXPRESSIVE INTENTIONS

Some attempts have been made to identify an expressive model that could render different expressive intentions of a human performer [6]. Moreover, automatic detection of expressive performances is quite recent. An important work on machine learning musical style recognition has been done by Dannenberg et al. in [7]. They showed that high-level understanding of musical performance like style recognition is highly beneficial from a machine learning approach. Another study on automatic analysis of expressiveness by Friberg [8] shows a system able to predict what emotion the performer is trying to convey. One or several types of "listener panels" can be stored as models which are used to simulate judgments of new performances based on results from previous listening experiments. In [9] Bayesian Networks have been employed for the recognition of expressive content in piano improvisations, and the following expressive intentions were recognized: slanted, heavy, hopping, vacuous, bold, hollow, fluid, tender. The intentions are derived from the Laban's basic effort theory of expressive movement. In [10], performances on various instruments were investigated; the expressive intentions were recognized with reference to the Kinematics Energy expressive space and the expressive content was classified using machine learning techniques. Also, several experiments on analysis of expression on simple pattern-based musical gestures have been previously carried out. In [11] short sequences of repeated notes recorded with MIDI piano were investigated. In [12], an experiment on expression detection has been done on audio data using professional recordings of violin and flute single repeated notes and short scales. In this work several features were extracted implementing automatic extraction algorithms. Audio cues were extracted over overlapping windows with 4s duration. After sliding over the audio file, mean and variance values for each feature were calculated. The relevance of the features was investigated by applying ANOVA tests over performances. After this test 7 relevant features were selected: Attack, Note Duration, Inter Onset Interval, Note Per Second, Peak Sound Level, Sound Level Range, Roughness and Spectral Centroid.

Violin and flute performances were recorded with reference to both Kinematics Energy Space and Valence Arousal Space categories. A PCA analysis showed a well separated clustering of intentions, as shown in Fig.4. This suggested that selected features provide a good differentiation between performances played with different intentions, although Calm and Sad clusters were close for both instruments.

Automatic detection of expressive intention was implemented using Naïve Bayesian classifier. For example, Table 2 shows the confusion matrix for Violin performances in the Valence Arousal Space.

	HAPPY	*SAD*	*ANGRY*	*CALM*
HAPPY	**77,46**	0,00	16,55	5,99
SAD	0,00	**56,21**	1,59	42,20
ANGRY	4,87	0,32	**89,29**	5,52
CALM	1,26	41,55	1,90	**55,29**

Table 2: Confusion matrix for violin in the Valence Arousal space. Baseline accuracy = 25%.

These studies on automatic detection of expressive content are interesting for both scientific and application field: from a scientific point of view, these results are important for understand performers' strategies for discriminating different intentions, and audio cues that are relevant for this discrimination.
In the application field, these studies are important for realistic modelling and synthesis of expressive multimedia systems, where expression can be used to enhance human computer communication. For example, automatic detection systems will permit to develop systems able to understand the emotional content of a human voice. Also, these system will be able to render expressive sounds to communicate with users, in a bidirectional way. From a larger point of view, expression can be used for audio enhancement of interfaces in several fields, from artistic productions to medical-therapeutic systems, mixed-reality environments for extreme-gaming and entertainment industry in general.

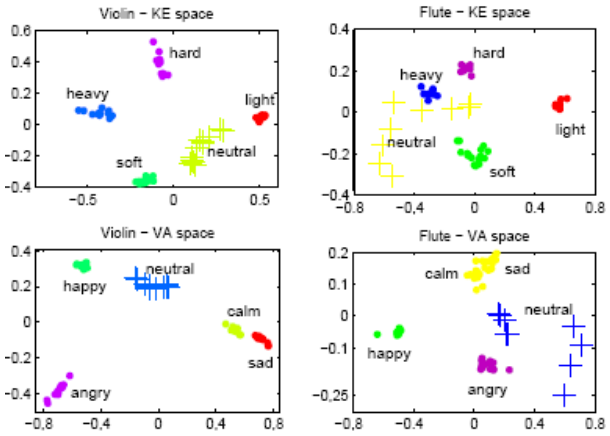

Figure 4: Principal Component Analysis on violin (left) and flute (right) performances in the KE space (top) and in the VA space (bottom).

7 CONCLUSIONS

Music performance researches have been presented, with reference to both analysis, modelling and automatic detection results.
A more intensive use of expressive intention control in multimedia systems will allow to interactively adapt music to different situations. This models will be used in different scenarios, from artistic creation to interactive applications where audio feedback react to user's actions, leading to a deeper fruition of the multimedia product and enhancing the Human Computer interaction.

REFERENCES

[1] Canazza S., De Poli G., Di Sanzo G., Vidolin A. (1998). "A model to add expressiveness to automatic musical performance". In Proc. of 1998 International Computer Music Conference. Ann Arbour. Pp. 163-169.

[2] Canazza S. De Poli G., Drioli C., Rodà A., Vidolin A. (2000). "Audio morphing different expressive intentions for Multimedia Systems". IEEE Multimedia, July-September, 7(3), pp. 79-83.

[3] De Poli G., Roda' A. and Vidolin A. (1998). "Note by note analysis of the influence of expressive intentions and musical structure in violin performance". Journal of New Music Research, 27(3), pp . 293-321.

[4] Friberg, A., Frydén, L., Bodin, L.-G., and Sundberg, J. (1991) Performance Rules for Computer-Controlled Contemporary Keyboard Music, Computer Music Journal, 15-2, pp. 49-55.

[5] Gabrielsson, A. (1999). The performance of music. In D. Deutsch (Ed.), The psychology of music (2nd ed., pp. 501-602). San Diego: Academic Press.

[6] De Poli, G. (2003) "Expressiveness in music performance: analysis and modeling", in Proceedings of the SMAC03 Stockholm Music Acoustics Conference, Stockholm, Sweden, pp. 17-20.

[7] Dannenberg, R., Thom, B., Watson, D. (1997) "A Machine Learning Approach to Musical Style Recognition", in Proceedings of the International Computer Music Conference, San Francisco, USA, pp. 344-347.

[8] Friberg, A., Schoonderwaldt, E., Juslin, P., Bresin, R. (2002) "Automatic Real-Time Extraction of Musical Expression", in Proceedings of the International Computer Music Conference, Göteborg, Sweden, pp. 365-367.

[9] Mion, L. (2003) "Application of Bayesian Networks to automatic recognition of expressive content of piano improvisations", in Proceedings of the SMAC03 Stockholm Music Acoustics Conference, Stockholm, Sweden, pp. 557-560.

[10] Mion, L., De Poli, G. (2004) "Expressiveness detection of music performances in the

Kinematics Energy Space", Proc. Sound and Music Computing Conference (JIM/CIM 04)}, October 20-22, Paris, France, pp. 257-261.

[11] Bonini, F., Rodà, A. (2001) "Expressive content analysis of musical gesture: an experiment on piano improvisation", Workshop on Current Research Directions in Computer Music, Barcelona.

[12] Mion, L., D'Incà, G. (2005) "An investigation over violin and flute expressive performances in the affective and sensorial domains", Sound and Music Computing Conference (SMC 05), Salerno, Italy.

Giovanni De Poli is professor of Computer Science at the Department of Information Engineering of the University of Padova. His main research interests are in algorithms for sound synthesis and analysis, models for expressiveness in music, multimedia systems and human-computer interaction, preservation and restoration of audio documents. He in involved in european research projects: Enactive Network of Excellence and Sound to Sense - Sense to Sound (S2S2) Coordination action.

Gianluca D'Incà received the Laurea degree on 2004 in Computer Engineering from the University of Padua. In 2005 he began the PhD studies in Computer Science, working at the Center of Computational Sonology on models for analysis and automatic detection of the expressive content in music performances.

Luca Mion received the Laurea degree on 2002 in Electronic Engineering from the University of Padua. In 2003 he began the PhD studies in Computer Science, working at the Center of Computational Sonology on models for analysis and synthesis of expressiveness in music, multimedia systems and human-computer interfaces.