EXPRESSIVE AUDIO SYNTHESIS: FROM PERFORMANCES TO SOUNDS

Gianluca D'Incà and Luca Mion

Center of Computational Sonology
Department of Information Engineering - DEI
University of Padova - Italy
{gianluca.dinca,luca.mion}@dei.unipd.it

ABSTRACT

In this paper we present preliminary results on expressive synthesis of simple sounds; we control a small set of audio parameters derived from the analysis of simple musical gestures, in order to synthesize sounds with different expressions. Then, from results of listening tests, we show that a control strategy based on the deviations of parameters leads to an effective communication of expressive content, even if using a very simple rendering model.

1. INTRODUCTION

The control of expressiveness is a very important concept in audio communication: expressive performance synthesis is a widely explored field, and several synthesis models have been proposed (e.g Canazza [1], Friberg [6] and Katayose [7]). The idea behind these models is to control the expressiveness of structured musical performances, using the score as reference. This approach led to the implementation of effective synthesis models, and its success is due to the fact that rendering models follow the behavior of human performers. In fact, musicians enrich their performances with expression by acting on their available degrees of freedom [5], and introducing deviations from a mechanical playing of the score. Apart from the interpretations of the score, expression is present in non-structured sounds: for example [9], in 60's science fiction movies synthetic sounds were used to communicate feeling of artificiality and sense of disquietude. Expressive synthesis and control of non-structured sounds have been less studied and explored; nevertheless, it is an interesting field for possible applications in the field of Human Computer Interaction (HCI). Expression can be added to systems for synthesizing and manipulating non-speech sounds like *Auditory Icons*, which refer to everyday listening sounds that can be either simple sounds or patterns [10]. Also, expressive content could be used in *Earcons*, that are defined as abstract musical tones that can be used in structured combinations to create sound messages [11]. Other applications can be found in well-defined problem like the design of alarm enunciators: in this field several studies were conducted on the mapping between sonological parameters and urgency [12], and expressiveness control can be added in order to give to the user the information about different levels of warning. Furthermore, expression can be used in other fields such as artistic productions, medical-therapeutic systems, applications in musicotherapy and so on.

As regards the expressiveness model, the fundamental task is to find the relationship (mapping) between *expressive cathegories* and *model parameters*. To solve this task, the performance paradigm can useful. In order to communicate different expressive intentions, musicians act on their playing by adding expressive content: this action results in deviations of sonologic parameters' from a performance played in a scholastic way, that we call *neutral* [8]. These deviations could be embedded in a synthesis model to add expressive content to a *neutral sound*. In this kind of approach the *mapping* is easily extrapolated from analysis of recordings.

In light of the above considerations, in this work we started from the analysis of simple musical gestures played by professional musicians [4]. From the analyses we derived a small set of audio parameters that we used for synthesize expressive sounds. Then we performed a listening test, both on real and synthetic stimuli. Results of test confirmed that expression can be differentiated and effectively communicated in simple sounds, controlling a simple synthesis model with a small set of parameters. This paper is organized as follows: next section describes the analysis and synthesis of expressive sounds; results of previous works are briefly recalled in Section 2.1, while Section 2.2 describes the synthesis model. In Section 3, listening test and results will be presented. Finally, conclusions and remarks will be discussed in Section 4.

2. EXPRESSIVE ANALYSIS AND SYNTHESIS

In this section we present the synthesis by analysis framework. We started from a previous work on the analysis of expression in musical gestures, which is briefly recalled in the next section.

2.1. Data analysis

In [4], a dimensional approach to conceptualize emotions has been used. This approach consists in placing expressions on a space with a small number of dimensions. Dimensions are used to simplify the study of highly structured concepts, and yields structures more suitable for a general comprehension of the phenomena. We refer to the Kinematics Energy space and to the Valence Arousal space.

The Kinematics Energy space has been conceptualized in [2]: professional musicians were told to play performances according to a set of sensorial adjectives (*hard, soft, heavy, light, bright, dark*). Then, a two dimensional space was derived, using measurements of perceptive nature and Multi Dimensional Scaling (MDS). From a musical point of view, the first dimension is mainly related to *tempo* (Kinematics), while the second one is related to *intensity* (Energy). We can distinguish four main categories situated at the opposite sides of the axis: High Energy (HE), Low Energy (LE), High Kinematics (HK), Low Kinematics (LK). We selected the intentions (adjectives) whose projection in this space was closer

VIOLIN	*Valence Arousal Space*				*Kinematics Energy Space*			
	Angry	**Calm**	**Happy**	**Sad**	**Heavy**	**Light**	**Hard**	**Soft**
NPS	+ 0,34	- 0,49	+ 0,65	- 0,58	- 0,35	+ 0,34	+ 0,13	- 0,27
D	- 0,40	+ 1,05	- 0,51	+ 1,49	+ 0,56	- 0,30	- 0,20	+ 0,41
A	- 0,53	+ 0,90	- 0,69	+ 1,07	+ 0,49	- 0,46	- 0,38	+ 0,60
IOI	- 0,29	+ 1,08	- 0,41	+ 1,58	+ 0,65	- 0,27	- 0,10	+ 0,41
PSL	+ 3,29	+ 0,12	+ 0,50	- 0,05	+ 1,84	+ 0,10	+ 1,67	+ 0,34
SLR	+ 2,53	+ 0,00	+ 0,36	- 0,19	+ 1,40	- 0,06	+ 1,32	+ 0,19
R	+ 0,93	- 0,19	+ 0,48	- 0,21	+ 0,24	+ 0,01	+ 0,69	- 0,10
C	- 0,07	- 0,03	- 0,05	- 0,02	- 0,02	- 0,01	- 0,02	- 0,02

Table 1: Relative deviations of values of the features in the Kinematics Energy space and in the Valence Arousal space. Violin performances are taken into account.

to the categories. Thus we have these correspondences: *hard-soft* (HE-LE); *light-heavy* (HK-LK).

The Valence Arousal Space is derived from the "circumplex model of affect" designed by psychologist Russell, who looks at emotions in terms of pleasure and displeasure (valence) and arousal [3]. Russell established that people organize the emotions in a similar way, when they are asked to place emotions on a two-dimensional space where the y-axis is the degree of arousal and the x-axis is the valence. The subjects, for example, placed the adjectives in order to induce the associations "*happy* vs. *sad*" (valence) and "*angry* vs. *calm*" (arousal). Similarly to the case of the KE space, we selected the intentions whose projection in this space was closer to the categories, thus we considered the correspondences: *happy-sad* (High and Low Valence); *angry-calm* (High and Low Arousal). Thus, in our work we considered 8 intentions represented by a set of 4 sensorial adjectives (*hard, soft, heavy, light*) and 4 affective adjectives (*happy, sad, angry, calm*). Musicians were invited to play several executions with different levels of musical structure complexity (excerpts, scales and repeated single notes), and in addition they were asked to play each performance with neutral expression.

From the recordings we extracted a set of 8 relevant features: *Attack* (A), *Note Duration* (D) , *Inter Onset Interval* (IOI), *Note Per Second* (NPS), *Peak Sound Level* (PSL), *Sound Level Range* (SLR), *Roughness* (R) and *Centroid* (C). The *Attack* is defined as the time required to reach the RMS peak, starting from the onset instant. *Note Per Second* is computed by dividing the number of onsets by the window length within the time window. The computations of $PSL = \max_t RMS(t)$ and $SLR = \max_t RMS(t) - \min_t RMS(t)$ derive directly from the RMS profile. The *Roughness* is calculated using the Synchronization Index Model implemented by IPEM Matlab Toolbox [13]. The spectrum centroid is calculated by following formula:

$$C = \frac{\sum_{k=1}^{N/2} f_k |X(k)|}{\sum_{k=1}^{N/2} |X(k)|} \quad (1)$$

where N is the FFT size, X is the FFT of the input signal x and f_k, $k = 1..N$, is the k-th frequency bin. A principal components analysis (PCA) showed that this set of features well discriminate different intentions, with the exception of categories *calm* and *sad*. However, these intentions are not easy to differentiate, and this is confirmed by the fact that recordings sound very similar. Furthermore, *sadness* in music is often related to high level characteristics of the score (e.g. the mode), which are not controllable in this context. More generally, intentions split into three clusters, as shown in Figure 1, reflecting intuitive associations between sensorial and affective categories.

For each feature and intention, we calculated the relative deviation from the *neutral* performance, using equation 2:

$$D_{F,i} = \frac{F_i - F_n}{F_n} \quad (2)$$

where F_i indicates the mean value of feature F of intention i and F_n indicates the mean value of F in the neutral performance. Table 1 summarizes values of relative deviations for violin performances.

In this work we used the results of the analysis to implement a prototype of an expressive tone generator, and then we synthesized different sound stimuli for the listening test. The sound synthesis system will be described in the next section.

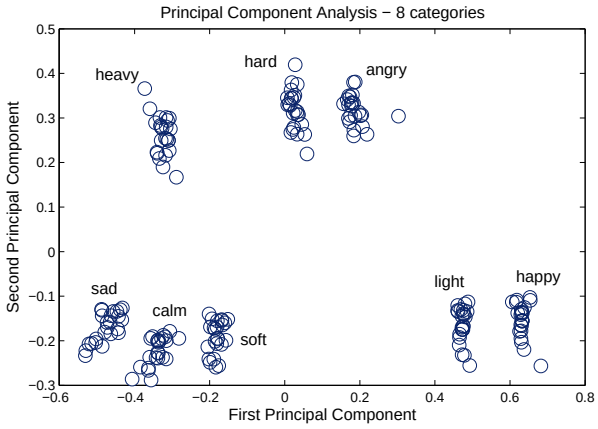

Figure 1: Clustering of the 8 categories: the positions reflect intuitive associations: *light* with *happy*, *sad* with *calm* and *soft*, and *angry* close to *hard* and *heavy*.

2.2. Sound synthesis

We implemented the expressive tone generator using the real-time synthesis environment *pd* (Pure Data). In the first version of the application, sound synthesis is very simple: we control the ADSR envelope of an additive synthesizer with 5 harmonic partials, varying three different kind of parameters: Tempo related features (*Attack, Duration, Note Per Second*), Intensity related features (*Peak*

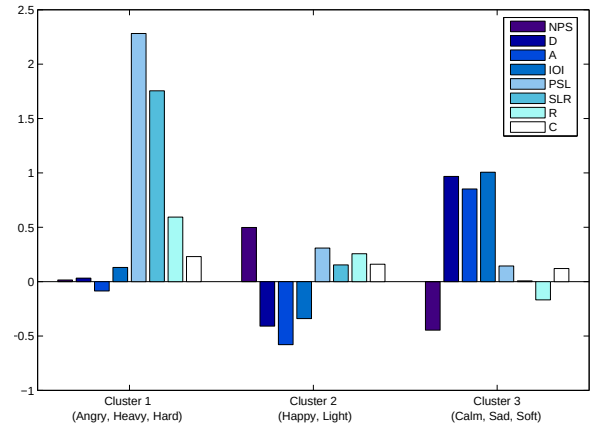

Figure 2: Diagrams of the relative deviations of features values from the neutral performance within the three clusters (single repeated notes).

Sound Level and *Sound Level Range*) and perception related features (*Roughness* and *Centroid*). Intensity and Tempo parameters are controlled using the ADSR envelope values, while *Roughness* and *Centroid* are modified by changing frequency and amplitude of the partials.

In particular, for *Roughness* we used the model proposed by Vassilakis [14]: for a pair of sinusoidal tones with frequencies f_1 and f_2 and amplitudes A_1 and A_2, with $f_{min} = min(f_1, f_2)$, $f_{max} = max(f_1, f_2)$, $A_{min} = min(A_1, A2)$, and $A_{max} = max(A_1, A_2)$, the *Roughness* is calculated using:

$$R_{synth} = X^{0.1} \cdot 0.5 \cdot (Y^{3.11}) \cdot Z \qquad (3)$$

where:

$$X = A_{min} \cdot A_{max}$$

$$Y = \frac{2 \cdot A_{min}}{A_{min} + A_{max}}$$

$$Z = e^{-b_1 s(f_{max} - f_{min})} - e^{-b_2 s(f_{max} - f_{min})}$$

and $b_1 = 3.5$, $b_2 = 5.75$, $s = 0.24/(s_1 \cdot f_{min} + s_2)$, $s_1 = 0.0207$, $s_2 = 18.96$.

The term $X^{0.1}$ represents the dependence of the *Roughness* from the intensity (related to the amplitude of the added sines); the term $Y^{3.11}$ represents the dependence from amplitude fluctuation degree (related to the amplitude difference of the added sines) and the term Z represents the dependence from amplitude fluctuation rate (frequency difference of the added sines) and register (frequency of the lower sine). The overall *Roughness* is calculated by adding the roughness of the individual sine-pairs. The *Centroid* of synthesized sounds is given by the following formula:

$$C_{synth} = \frac{\sum_{k=1}^{5} f_k a_k}{\sum_{k=1}^{5} a_k} \qquad (4)$$

where k is the partial index, f_k and a_k are the frequency and the amplitude of the partial k respectively. Note that in this case we consider sounds with only five harmonics. Figure 3 shows the *pd* patch that allows the control of the frequency and the amplitude of the harmonics in the synthesis process. The tone generator performs a repeated single note with the values of the parameters derived from previous analyses. We defined the *neutral sound* by using the values of parameters from violin neutral performances. Then, we applied the transformations (deviations) resulting from the analysis to synthesize the *expressive sounds*, according to the 8 categories. Synthesized sounds were used in a listening test described in the next section.

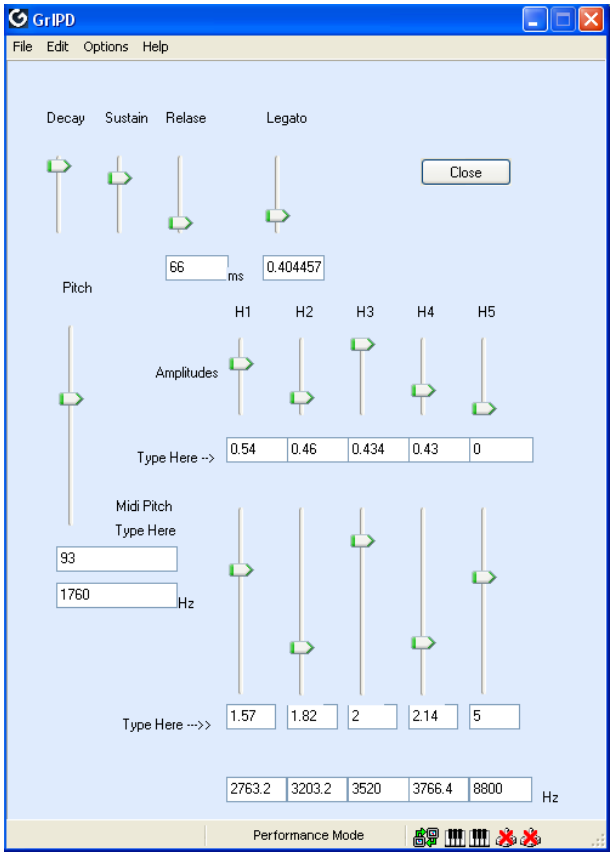

Figure 3: Patch for the control of frequency and amplitude of the harmonics.

3. LISTENING TEST AND RESULTS

For the preliminary listening test we used the forced-choice identification method, asking 26 subjects (age: 18-30) to evaluate the expression of a 16 stimuli series: 8 sounds were both violin and synthetic, and referred to sensorial categories (*hard, soft, heavy, light*); the other 8 sounds referred to affective categories (*happy, sad, angry, calm*). Also in this case, stimuli were both violin and synthetic. The test has two goals: regarding violin stimuli, we want to test the actual possibility to differentiate and communicate intentions in non-structured sounds (in this case, a single repeated note). The second aim is to comprehend the amount of expressive content that a simple synthesis model is able to convey, with respect to the real sounds. Results on violin and synthetic stimuli will be presented in the next two sections.

VIOLIN					SYNTHETIC				
	HARD	*SOFT*	*HEAVY*	*LIGHT*		*HARD*	*SOFT*	*HEAVY*	*LIGHT*
HARD	**73,08**	7,69	11,54	7,69	*HARD*	**38,46**	19,23	38,46	3,85
SOFT	11,54	**61,54**	11,54	15,38	*SOFT*	3,85	**73,08**	7,69	15,38
HEAVY	15,38	11,54	**69,23**	3,85	*HEAVY*	11,54	19,23	**65,38**	3,85
LIGHT	3,85	7,69	0,00	**88,46**	*LIGHT*	30,77	15,38	7,69	**46,15**
	HAPPY	*SAD*	*ANGRY*	*CALM*		*HAPPY*	*SAD*	*ANGRY*	*CALM*
HAPPY	**76,92**	0,00	23,08	0,00	*HAPPY*	**46,15**	15,38	34,62	3,85
SAD	0,00	**80,77**	0,00	19,23	*SAD*	0,00	**34,62**	0,00	65,38
ANGRY	19,23	0,00	**80,77**	0,00	*ANGRY*	7,69	3,85	**84,62**	3,85
CALM	3,85	30,77	0,00	**65,38**	*CALM*	0,00	23,08	3,85	**73,08**

Table 2: Confusion matrices for violin and synthetic stimuli in the Kinematics Energy space and in the Valence Arousal space. Baseline accuracy = 25%.

VIOLIN				SYNTHETIC		
	Cluster 1 *Angry, Heavy, Hard*	**Cluster 2** *Happy, Light*	**Cluster 3** *Calm, Sad, Soft*	**Cluster 1** *Angry, Heavy, Hard*	**Cluster 2** *Happy, Light*	**Cluster 3** *Calm, Sad, Soft*
Cluster 1	**83,33**	10,26	6,41	**79,49**	5,13	15,38
Cluster 2	13,46	**82,69**	3,85	36,54	**46,15**	17,31
Cluster 3	7,69	6,41	**85,90**	5,13	5,13	**89,74**

Table 3: Confusion matrices for violin and synthetic stimuli considering the three clusters.

3.1. Violin stimuli

Table 2 shows the confusion matrix of listeners' responses. On the top left the violin stimuli Kinematics Energy Space matrix is presented, related to the sensorial categories, while on the bottom left we have the Valence Arousal Space results (affective categories). We have good results in both the two spaces (baseline accuracy is 25%): affective categories lead to better accuracy of detection, but some confusion between *calm* and *sad* is confirmed. In general, opposite adjectives (e.g. *happy* and *sad*) are well discriminated, while confusion arises when trying to distinguish between other categories. This is confirmed by the remark of some listeners: they tended to make two mental categories (e.g. *hard* with *heavy*, and *light* with *soft*) and then they tried to discriminate inside the categories, having more difficulty.

3.2. Synthetic stimuli

Results from synthetic stimuli are shown on the top and on the bottom right of Table 2. We still exceed the baseline accuracy, but high confusion arises between similar categories, confirming observations by some listeners. On the Kinematics Energy Space, *heavy* is often confused with *hard*; in the Valence Arousal space, *happy* is confused with *angry* and the difficulty to discriminate *sad* and *calm* is also confirmed. We gathered answers from listeners with reference to the 3 clusters mentioned in Section 2.1: results are shown in Table 3. As one may expect, violin stimuli results are good, while some difficulty arises in synthetic stimuli. In particular, *Cluster 2* is often (36% of cases) confused with *Cluster 1*: this fact derives from the confusion between *happy* and *angry*, but also from the not-effective rendering of intention *light* which is often perceived as *hard*.

4. CONCLUSIONS

Preliminary results on expressive synthesis of simple sounds have been presented. From sonological analysis of musical performances a small set of parameters has been extracted and used for synthesis. Listening tests have been conducted both on real and synthetic sounds; results of listening test confirmed the possibility to communicate different intentions with simple sounds. Furthermore, results on synthetic stimuli evidenced, with some limit, the effectiveness of a very simple rendering model. Current research is focusing on the study of sounds that are not related to musical performance (e.g. physical based synthesized sounds), with the goal of deriving effective mappings between expressive control space and rendering model parameters. This is an ambitious task: without having the explicit control strategies of performers, perceptual tests are needed to find the relationship between model parameters and control space dimensions (e.g. Kinematics/Energy or Valence/Arousal). This work is a further step on expressive control of non structured sounds: more work is needed to increase the knowledge of this interesting but barely explored field.

5. ACKNOWLEDGEMENTS

This research was supported by the European Network of Excellence "Enactive Interfaces" and the Coordination Action "Sound to Sense, Sense to Sound ($S2S^2$)". We thank David Pirrò for developing the sound synthesis prototype.

6. REFERENCES

[1] Canazza, S., De Poli, G., Drioli, C., Rodà, A., Vidolin, A. "Audio morphing different expressive intentions for Multi-

media Systems", *IEEE Multimedia*, July-September, 7(3), pp. 79-83, 2000.

[2] Canazza, S., De Poli, G., Rodà, A., Vidolin, A. "An abstract control space for communication of sensory expressive intentions in music performance", *Journal of the New Music Research*, 32(3), pp. 281-294, 2003.

[3] Russell, J. A. "A circumplex model of affect", *Journal of Personality and Social Psychology*, 39, 1161-1178, 1980.

[4] Mion, L., D'Incà, G. "Analysis of expression in simple musical gestures to enhance audio in interfaces", *Special Issue of Virtual Reality*, In A. Camurri, A. Frisoli (guest eds.) Multisensory interaction in virtual environments, Berlin: Springer Verlag, 2006. *(in press)*

[5] De Poli, G., Rodà, A., Vidolin, A. "A Note-by-note analysis of the influence of expressive intentions and musical structure in violin performance", *Journal of the New Music Research*, vol. 27, no. 3, pp. 293–321, 1998.

[6] Friberg, A., Colombo, V., Frydén, L., Sundberg, J. "Generating Musical Performances with Director Musices", *Computer Music Journal*, vol. 24, no. 3, pp. 23- 29, 2000.

[7] R. Hiraga, H. Katayose, H. Koike, T. Suzuki, K. Noike, and T. Hoshishiba "Performance rendering 2000: Demonstration and panel discussion" *2000-MUS-35*, pages 67-70. IPSJ, 2000.

[8] De Poli, G., Canazza, S., Drioli, C., Rodà, A., Vidolin, A., Zanon, P. "Analysis and modeling of expressive intentions in music performance", *Proc. of International Workshop on Human Supervision and Control in Engineering and Music*, Kassel, Germany, September 21-24, 2001.

[9] De Poli, G., D'incà, G., Mion, L. "Computational models for audio expressive communication", *Proc. Audio Engineering Society Annual Meeting*, November 9-12, Como, Italy, 2005.

[10] Gaver, W. "Auditory icons: using sound in computer interfaces." *Human Computer Interaction*, 2(2), pp. 167-177, 1986.

[11] Brewster, S. A., Grease, M. G. "Correcting menu usability problems with sound", *Behaviour and Information Technology*, 18(3), pp. 165-177, 1999.

[12] Stanton, N. "Human Factors in alarm designs", *London, UK: Taylor & Francis Ltd*, 1994.

[13] Leman, M. "Visualization and calculation of roughness of acoustical musical signals using the synchronization index model (SIM)", *Proceedings of the of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, Verona, Italy, pp. 125–130, 2000.

[14] Vassilakis, P. N. "Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance", *Doctoral Dissertation*, University of California, Los Angeles. Systematic Musicology, 2001.