Synthesis and decoding of emotionally expressive music performance

Roberto Bresin, Anders Friberg
Department of Speech, Music and Hearing - Royal Institute of Technology
Drottning Kristinas väg 31, 10044 Stockholm, Sweden
{roberto, andersf}@speech.kth.se

ABSTRACT

A recently developed application of Director Musices (DM) is presented. The DM is a rule-based software tool for automatic music performance developed at the Speech Music and Hearing Dept. at the Royal Institute of Technology, Stockholm. It is written in Common Lisp and is available both for Windows and Macintosh.

It will be demonstrated that particular combinations of rules defined in the DM can be used for synthesizing performances that differ in emotional quality. Different performances of two pieces of music were synthesized so as to elicit listeners' associations to six different emotions (fear, anger, happiness, sadness, tenderness, and solemnity). Performance rules and their parameters were selected so as to match previous findings about emotional aspects of music performance. Variations of the performance variables IOI (Inter-Onset Interval), OOI (Offset-Onset Interval) and L (Sound Level) will be presented for each rule-setup. In a forced-choice listening test 20 listeners were asked to classify the performances with respect to emotions. The results showed that the listeners, with very few exceptions, recognized the intended emotions correctly. This shows that a proper selection of rules and rule parameters in DM can indeed produce a wide variety of meaningful, emotional performances, even extending the scope of the original rule definition.

INTRODUCTION

In recent years, an increased effort has been spent in the study of verbal as well as non-verbal communication of emotion. In the music field, in particular, Gabrielsson, Juslin and others at the Department of Psychology of Uppsala University, showed that there are specific parameters in the microstructure of a performance that are manipulated by the performers when they are asked to play the same simple piece of music with different prescribed emotions [1]. This research has concerned so-called *basic emotions* (anger, sadness, happiness, and fear) and also solemnity and tenderness. It has been shown that all these emotions, as conveyed by players, could be clearly recognized by an audience of musically trained and untrained listeners [2], [3]. Other research has shown that it is possible to communicate also more complex emotional states, even though it is not completely clear how to define them in term of adjectives; performers and listeners often use different terms in describing intentions and perceived emotions [4]-[7]. The most recent step in this research trend has been to synthesize performances encoding different emotions by setting the values of certain expressive cues on a commercial sequencer [8], or by using original software [9]. Also under these conditions listeners could recognize and identify the intended emotions. The aim of the present preliminary investigation was to take a further step in this direction; complete automatic performance synthesis was used in an attempt to convey six different emotions.

SYNTHESIS

Gabrielsson [1], [10] and Juslin [2], [8] proposed a list of possible expressive cues that seemed characteristic for each of the emotions fear, anger, happiness, sadness, solemnity, and tenderness. These cues are described in terms of qualitative changes of *tempo*, *sound level*, *articulation* (staccato, legato), *tone onsets and decays*, *timbre*, *deviations of IOI* (Inter-Onset Interval), *vibrato*, *final ritardando.* As a grand piano sound (Kurzweil sound samples of the Pinnacle Turtle Beach soundboard) was used in the present experiment; the cues *tone onset and decay*, *timbre* and *vibrato* could not be used. The remaining cues manipulated here are listed in Table 1.

The Director Musices (DM) program was utilized for synthesizing the performances. This is an expert system for automatic music performance written in Lisp language and containing about twenty rules. These rules attempt to model a performer's rendering of, for example, phrasing, intonation or rhythmic patterns. The rules can affect most of the performance parameters mentioned above. It is an implementation of the "KTH performance rule system" [11]. In order to synthesize performances associated with each of the emotions, the qualitative description of each expressive cue was interpreted by the authors into a quantitative rule description. For each intended emotion, a macro rule was written, each activating a special subset of previously defined rules. In Table 1, the cue profiles for each emotion, as outlined by Gabrielsson and Juslin, are compared with the rule setup utilized for the synthesis. No more than five rules were used: *Duration contrast articulation* rule [12], *Duration contrast* rule [11], [13], *Punctuation* rule [14], *Phrase arch* rule [15], *Final ritardando* rule [16]. The setting of rule parameters was refined with the help of Lars Frydén, expert musician and principal advisor in designing the rules in DM.

Two clearly different pieces of music were used. One was the melody line of a Swedish nursery tune (Ekorrn satt i granen, henceforth *Ekorrn*, "The squirrel sat on the fir-tree", composed by Alice Tegnér) written in major tonality (Figure 1). The other was a computer generated piece (henceforth *Mazurka*), written in minor tonality in an attempt to portray the musical style of Fréderic Chopin [17]. The phrase structure with regard to the levels of sub-phrase (level 6), phrase (level 5) and piece (level 4) was marked in each score (in Figure 1 the marking for *Ekorrn* is shown). This was needed as an input to the *Phrase arch* rule.

Each of the two scores was performed in seven different versions by Director Musices. The rule setup shown in Table 1 was applied for fear, angry happy, sad, solemn and tender versions, while no rules were used for synthesizing a dead-pan version, henceforth referred to as "no-expression". Each macro rule, starting from the no-expression case, produced deviations in the performance parameters. The original tempo was 187 quarter notes per minute for *Ekorrn* and 96 quarter notes per minute for *Mazurka.* Note that the same rule setup was used for both *Ekorrn* and *Mazurka.*

0-7803-5731-0/99/$10.00 ©1999 IEEE

Table 1 Cue profiles for each emotion, as outlined by Gabrielsson and Juslin, are compared with the rule set-up utilized for the synthesis with Director Musices.

Emotion	Expressive Cue	Gabrielsson and Juslin	Director Musices
Fear	Tempo	Irregular	*Tone IOI* is lengthened by 80%
	Sound level	Low	*Sound level* is decreased by 6 dB
	Articulation	Mostly staccato or non-legato	Duration contrast articulation rule (k = 2)
	Time deviations	▪ Large ▪ Structural reorganizations ▪ Final acceleration (sometimes)	▪ Duration contrast rule (k = 4) ▪ Punctuation rule (k = 2) ▪ Phrase arch rule applied on phrase level (level = 5, k = -1.5, turn. position = 0.2, next = 1.3, amp = –4.0) ▪ Phrase arch rule applied on sub-phrase level (level = 6, k = -1.5, turn. position = 0.2, amp = –4.0, last = 0.2) ▪ Final ritardando (k = 1.0, q = 3)
Anger	Tempo	Very rapid	*Tone IOI* is shortened by 15%
	Sound level	Loud	*Sound level* is increased by 8 dB
	Articulation	Mostly non-legato	Duration contrast articulation rule (k = 1)
	Time deviations	▪ Moderate ▪ Structural reorganizations ▪ Increased contrast between long and short notes	▪ Duration contrast rule (k = 2, amp = 0) ▪ Punctuation rule (k = 2) ▪ Phrase arch rule applied on phrase level (level = 5, k = -0.7, turn. position = 0.5, next = 1.3, amp = 4) ▪ Phrase arch rule applied on sub-phrase level (level = 6, k = -0.7, turn. position = 0.3, amp = 4, last = 1)
Happiness	Tempo	Fast	*Tone IOI* is shortened by 20%
	Sound level	Moderate or loud	▪ *Sound level* is increased by 3 dB ▪ High loud rule (k = 1.5)
	Articulation	Airy	Duration contrast articulation rule (k = 2.5)
	Time deviations	Moderate	▪ Duration contrast rule (k = 2) ▪ Punctuation rule (k = 2) ▪ Final ritardando rule (k = 0.3, q = 4)
Sadness	Tempo	Slow	*Tone IOI* is lengthened by 30%
	Sound level	Moderate or loud	*Sound level* is decreased by 6 dB
	Articulation	Legato	
	Time deviations	Moderate	▪ Duration contrast rule (k = -2) ▪ Phrase arch rule applied on phrase level (level = 5, k = 1.5, turn. position = 0.3, next = 1.3, amp = 2) ▪ Phrase arch rule applied on sub-phrase level (level = 6, k = 1.5, turn. position = 2, amp = 4, last = 0.2)
	Final ritardando	Yes	Obtained from the Phrase rule with the *next* parameter
Solemnity	Tempo	Slow or moderate	*Tone IOI* is lengthened by 30%
	Sound level	Moderate or loud	▪ *Sound level* is increased by 3 dB ▪ High loud rule (k = 1.5)
	Articulation	Mostly legato	Duration contrast articulation rule (k = -1)
	Time deviations	Relatively small	Punctuation rule (k = 1)
	Final ritardando	Yes	Final ritardando rule (k = 0.3, q = 2)
Tenderness	Tempo	Slow	*Tone IOI* is lengthened by 30%
	Sound level	Mostly low	*Sound level* is decreased by 6 dB
	Articulation	Legato	
	Time deviations	Diminished contrast between long and short notes	Duration contrast rule (k = -4, amp = 0)
	Final ritardando	Yes	Final ritardando rule (k = 0.5, q = 1.5)

As an example, Figure 2 shows the deviations for the synthesized "fear" version of *Ekorrn*. The relative time deviation for each note's IOI varies between -20% and +100% of the nominal values, as seen in the top graph in Figure 2. Most of the deviations are positive indicating a slower tempo than in a neutral performance. Note that the per cent IOI deviation curve is not a straight line, but rather contains quite great and quick oscillations, thus reflecting what Gabrielsson described as "irregular tempo". Larger time deviations are associated with the shorter notes. The graph also shows that a strong *ritardando* appears in the end. The *Duration contrast articulation* rule introduced quite large articulation pauses after all comparatively short notes, thus rendering a "mostly staccato articulation"[2], [10]. The *Punctuation rule* inserts larger articulation pauses at automatically detected structural boundaries. All this

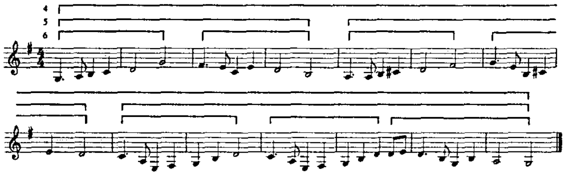

Figure 1 The Ekorrn melody, including phrase (5), sub-phrase (6) and piece (4) markings as used by the phrase arch rules.

corresponds to the positive deviations in the off-time duration curve of Figure 2.

The sound level curve of the *afraid* version of *Ekorrn* (Figure 2, bottom graph) shows negative deviations. This reflects the "lower sound level" referred to by Gabrielsson. An unusual setting of the *Phrase arch rule* produced a sequence of decrescendo-crescendo patterns, with louder notes at phrase and sub-phrase boundaries. This was an attempt to realize the "structural reorganization" proposed by Gabrielsson.

In figure 5, the average IOI deviations are plotted versus the average sound level together with their standard deviations for all six performances of each piece. Negative values of sound level deviation indicate a softer performance than the no-expression one. "Anger" and "happiness" performances are thus played quicker and louder while "tenderness", "fear", and "sadness" performances are slower and softer relative to a no-expression rendering.

Figure 6 shows relative mean and standard deviation for the deviations of IOI for all six performances of each piece. Negative values of IOI deviation imply a tempo faster than the original, and vice versa. The "fear" and "sadness" versions have larger standard deviations obtained mainly by exaggerating the duration contrast and also by applying the phrasing rules.

DECODING TEST

According to Juslin forced-choice judgments and free labeling judgments give similar results in listeners' decoding of a performer's intended emotional expression [3]. Therefore, it was considered sufficient to make a forced-choice listening test to assess the efficiency of the emotional communication. Fourteen performances (7 emotions x 2 examples), originally stored in MIDI files, were recorded as standard sound files and presented to a panel of listeners.

Twenty listeners, 24-50 years old, 5 female and 15 male, volunteered as subjects. None of them was musicians or music students. Eighteen of the subjects played or used to play an instrument on a non-regular basis. The subjects were all working at the Speech Music Hearing Department at the Royal Institute of Technology, Stockholm.

The subjects listened to the examples individually. Each subject was instructed to identify the emotional expression of each example as one out of seven alternative substantives: fear, anger, happiness, sadness, solemnity, tenderness, no-expression. The responses were automatically recorded by means of the Visor software system, specially designed for listening tests [18]. Visor presents the sound files in random order as anonymous boxes on the screen. The listeners were given instructions directly on the computer screen. The subjects listened to the stimuli over headphones (Sennheiser HD435 Manhattan), and the output level of the soundboard was set to the maximum level possible.

Each session contained four sub-tests, each presenting the seven performances of (1) *Ekorrn*, (2) *Mazurka*, (3) *Ekorrn*, and (4) *Mazurka*. The order of the performances within each sub-test was automatically randomized for each individual subject by Visor. The average duration of an experiment session was approximately 11 minutes.

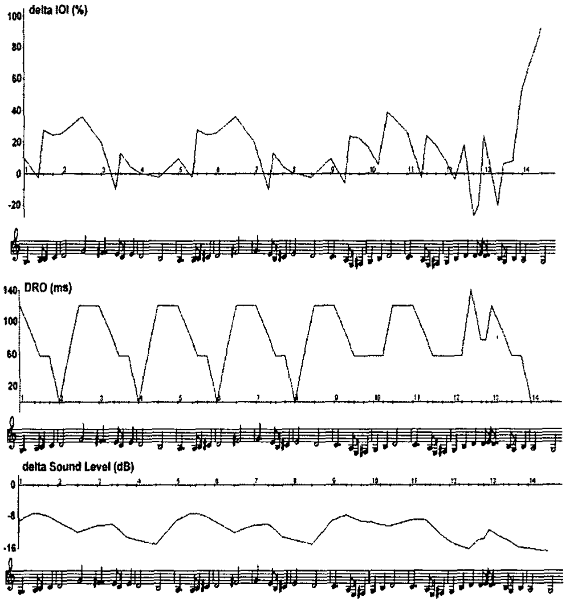

Figure 2 Deviations from no-expression case for the "fear" version of Ekorrn for per cent IOI (top), off-time in milliseconds (middle), and sound level in decibel (bottom).

RESULTS

Four subjects, who gave same answers for repeated stimuli in less than 36% of the cases, were eliminated from the subsequent analysis.

Figure 3 shows the percentage of "correct" responses for the two pieces. In all cases but one, these percentages well exceeded the 14% chance level. The "tenderness" version of *Mazurka* was the most difficult case to identify (13%). For "happiness" the responses differed substantially between *Ekorrn* (97%) and *Mazurka* (69%). The same applied to "no expression" (*Ekorrn* 81% and *Mazurka* 66%). This could partly be due to the fact that *Mazurka* was in a minor tonality, which in the western music tradition is often associated with the moods sadness and anger. Fear, anger, happiness, and sadness received an average percentage of 69% for *Ekorrn* and 59% for *Mazurka.*

To facilitate a more detailed analysis, the subjects' responses for each version of *Ekorrn* and *Mazurka* are presented in Tables 2 and 3. For *Ekorrn,* the subjects mostly chose the "correct" alternative, although for the "fear" version the "tender" alternative was chosen by 34% of the subjects, and for the "tenderness" version the "sadness" alternative was selected by 28%. This may be due to the unavailability of vibrato in piano synthesis; according to Gabrielsson and Juslin vibrato is an important expressive cue in the synthesis of "fear" and "tenderness" performances. Furthermore, according to Juslin (in press) the "tenderness" version could be performed with a higher sound level than the "sadness" version, while the same sound level for both versions was used in this experiment.

For *Mazurka*, the "fear" version elicited less confusion; it was classified as "tenderness" by only 21% of the subjects. On the other hand, many subjects classified the "tenderness" version as "sadness" (41%) or as "no expression" (31%), and only 13% selected the "tenderness" alternative. The reason for this confusion would be the same as for *Ekorrn,* lack of vibrato and of differentiation of sound level. All other versions of *Mazurka* were mostly classified according to the intended emotion.

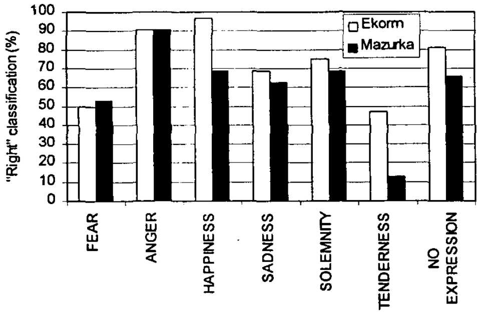

Figure 3 Effect of piece: percentage of "right" classification

The synthesis of each emotional performance was achieved by using a special set-up, including a subset of DM rules, see Table 1. A total of 17 parameters were involved. An attempt was made to reduce the number of dimensions of this space by means of a principal component analysis. Two principal factors emerged, explaining 61% (Factor 1) and 29% (Factor 2) of the total variance. Figure 4 presents the main results in terms of the distribution of the different setups in the two-dimensional space. Factor 1 was closely related to deviations of sound pressure level and tempo: louder and quicker performances had coordinates between *Anger* and *Solemnity*, while softer and slower performances had coordinates between *Sadness* and *Fear*. Factor 2 was closely related to the articulation and phrasing variables. The distribution of setups resembles those presented in previous works and obtained with other methods [6], [7], [9].

The k-values for the *Duration Contrast* rule in the setups, also shown in Figure 4, were highest in the first quadrant, lowest in the second, and intermediate in the third and fourth. The figure also shows an attempt to a qualitative interpretation of variation of this rule in the space.

DISCUSSION

The main result from this exploratory experiment is that the emotions associated with the DM macro rules where classified correctly by the listeners in most cases. This suggests that it is possible to group DM performance rules into macro rules producing performances that can be associated with different emotional states.

An important observation is that the same macro rules could be successfully applied both to *Ekorrn* and *Mazurka*, two completely different compositions. This is not surprising since the used performance rules previously have been modeled so as to work correctly in different musical contexts. On the other hand, it may be more advantageous to use somewhat different versions of the macro rules for scores of differing characters; for example different quantities of *staccato* and phrase marking may be preferable depending on music style.

Using the DM system is not the only possible way for evoking associations with emotions, many other recipes certainly exist. However the findings in the present work also indicate interesting new potentials for the DM system. One of its limitations has been that it produced only one performance of each piece, due to its deterministic structure. The results of the present study indicate that, by complementing the DM with macro rules, performances can be obtained, that significantly differ in emotional quality.

The range of future applications is of course unpredictable. One possibility would be to design an "emotional tool-box" where users can chose different ways of playing the same pieces of music by selecting a button or a combination of buttons associated with different emotions. This could be applied to large MIDI music databases on the Internet.

Table 2 Confusion matrix (%) for the classification test of seven synthesized performances of *Ekorrn.*

Intended Emotion	FEAR	ANGER	HAPPINESS	SADNESS	SOLEMNITY	TENDERNESS	NO EXPRESSION
FEAR	**40**	0	8	3	3	40	8
ANGER	5	**88**	3	0	5	0	0
HAPPINESS	8	5	**85**	0	3	0	0
SADNESS	3	0	0	**63**	3	28	5
SOLEMNITY	3	10	3	5	**65**	3	13
TENDERNESS	5	0	3	33	3	**45**	13
NO EXPRESSION	0	0	18	0	10	5	**68**

Table 3 Confusion matrix (%) for the classification test of seven synthesized performances of *Mazurka.*

Intended Emotion	FEAR	ANGER	HAPPINESS	SADNESS	SOLEMNITY	TENDERNESS	NO EXPRESSION
FEAR	**53**	0	0	9	3	22	13
ANGER	6	**91**	0	3	0	0	0
HAPPINESS	0	28	**69**	0	3	0	0
SADNESS	3	0	0	**63**	0	34	0
SOLEMNITY	0	16	0	13	**69**	0	3
TENDERNESS	13	0	0	41	3	**13**	31
NO EXPRESSION	3	0	6	0	16	9	**66**

Another possibility would be to use the new macro rules as a tool for objective analysis of emotional aspects of performances. This possibility may be interesting from a musicological point of view. The macro rules could be used in reverse in order to analyze the emotional content of a performance. For example, the parameters of the rules can be automatically fitted to a performance (c.f. Friberg, 1995b). In this way, it may be possible to classify deviations in various performance parameters according to intended emotion. This possibility seems tempting to explore in the future.

The DM rules are all triggered by the structure of the music, i.e., the combination of note values, intervals etc. Hence, the rules can only reflect the structure. Nevertheless, by varying the selection of rules and their quantities, performance were generated that could be readily interpreted in emotional terms. This indicates that an emotionally expressive performance can be directly derived from the musical structure.

ACKNOWLEDGEMENTS

This work was supported by the Swedish National Council for Research in the Humanities and Social Sciences, and by the Bank of Sweden Tercentenary Foundation. The authors are grateful to Johan Sundberg, who assisted in editing the manuscript, to Patrik Juslin for valuable discussions, and to the twenty anonymous subjects who participated in the listening test.

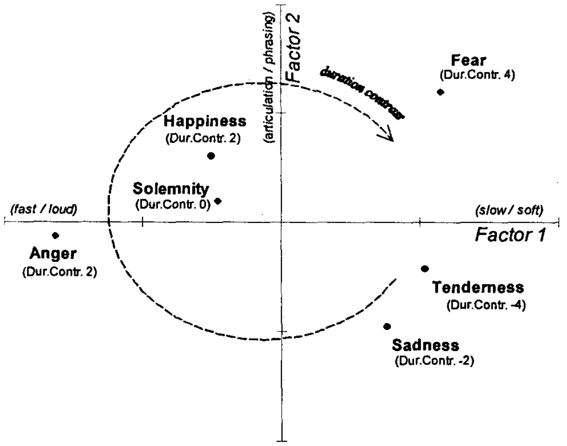

Figure 4 Two-dimensional space of the substantives derived from principal component analysis of the different emotional rule setups.

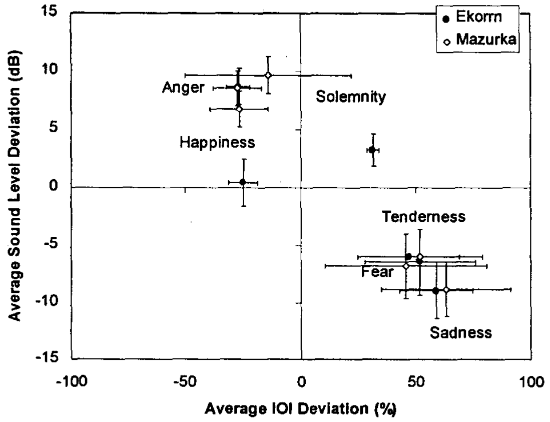

Figure 5 Average IOI deviations versus average sound level deviations for all six performances of each piece. The bars represent the standard deviations.

REFERENCES

[1] Gabrielsson, A. Intention and emotional expression in music performance. In A. Friberg, J. Iwarsson, E. Jansson & J. Sundberg (Eds.) *Proceedings of the Stockholm Music Acoustics Conference 1993*, Stockholm: Royal Swedish Academy of Music, 1994, 108-111.

[2] Juslin, P.N. Emotional communication in music performance: a functionalist perspective and some data. *Music Perception,* 1997a, ***14*** *(4)*, 383-418.

[3] Juslin, P.N. Can results from studies of perceived expression in musical performances be generalized across response formats?, *Psychomusicology*, in press

[4] Canazza, S., De Poli, G., Rinaldin, S. & Vidolin, A. Sonological analysis of clarinet expressivity. In M. Leman (Ed.) *Music, gestalt, and computing: studies in cognitive and systematic musicology.* Berlin, Heidelberg, New York: Springer Verlag, 1997, 431-440.

[5] Battel, G.U., & Fimbianti, R. How communicate expressive intentions in piano performance. In A. Argentini & C. Mirolo (Eds.) *Proceedings of the XII Colloqium on Musical Informatics,* Udine: AIMI, 1998, 67-70.

[6] De Poli, G., Rodà, A. & Vidolin, A. A model of dynamic profile variation, depending on expressive intention, in piano performance of classical music. In A. Argentini & C. Mirolo (Eds.) *Proceedings of the XII Colloqium on Musical Informatics*, Udine: AIMI; 1998, 79-82.

[7] Orio, N., & Canazza, S. How are expressive deviations related to musical instruments? Analysis of tenor sax and piano performances of "How High the Moon" theme. In Argentini A & Mirolo C (Eds.) *Proceedings of the XII Colloqium on Musical Informatics*, Udine: AIMI, 1998, 75-78.

[8] Juslin, P.N. Perceived emotional expression in synthesized performances of a short melody: capturing the listener's judgment policy. *Musicae Scientiae,* 1997b, *1 (2)*, 225-256.

[9] Canazza, S., De Poli, G., Di Sanzo, G. & Vidolin, A. Adding expressiveness to automatic musical performance. In A. Argentini & C. Mirolo (Eds.) *Proceedings of the XII Colloqium on Musical Informatics,* Udine: AIMI, 1998, 71-74.

[10] Gabrielsson, A. Expressive intention and performance. In R. Steinberg (Ed.) *Music and the Mind Machine: the Psychophysiology and the Psychopathology of the Sense of Music.* Berlin, Heidelberg, New York: Springer Verlag, 1995, 35-47.

[11] Friberg, A. *A Quantitative Rule System for Musical Expression.* Doctoral dissertation, Stockholm: Royal Institute of Technology, 1995a.

[12] Bresin, R., Friberg, A. Emotional expression in music performance: synthesis and decoding. *TMH-QPSR, Speech Music and Hearing Quarterly Progress and Status Report,* 4/1998, Stockholm, pp. 85-94

[13] Friberg, A. Generative for music performance: a formal description of a rule system. *Computer Music Journal,* 1991, ***15*** *(2)*, 56-71.

[14] Friberg, A., Bresin, R., Frydén, L. & Sundberg, J. Musical punctuation on the microlevel: Automatic identification and performance of small melodic units. *Journal of New Music Research,* 1998, *27 (3)*, 271-292.

[15] Friberg, A. Matching the rule parameters of Phrase arch to performances of "Träumerei": A preliminary study. In A. Friberg & J. Sundberg (Eds.) *Proceedings of the KTH Symposium on Grammars for Music Performance,* Stockholm: Speech Music and Hearing Department, 1995b, 37-44.

[16] Friberg, A., & Sundberg, J. Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America,* 1999, ***105*** *(3),* 1469-1484.

[17] Cope, D. Computer modeling of musical intelligence in experiments in musical intelligence. *Computer Music J*ournal 1992, ***16*** *(2),* 69-83.

[18] Granqvist, S. Enhancements to the Visual Analogue Scale, VAS, for listening tests. *Quarterly Progress and Status Report,* Stockholm: Royal Institute of Technology - Speech Music and Hearing Department, 1996, **4**, 61-65.

LINKS

Sound and MIDI files of the seven versions of Ekorrn and Mazurka:
http://www.speech.kth.se/~roberto/emotion

KTH performance rules description and software:
http://www.speech.kth.se/music/performance/

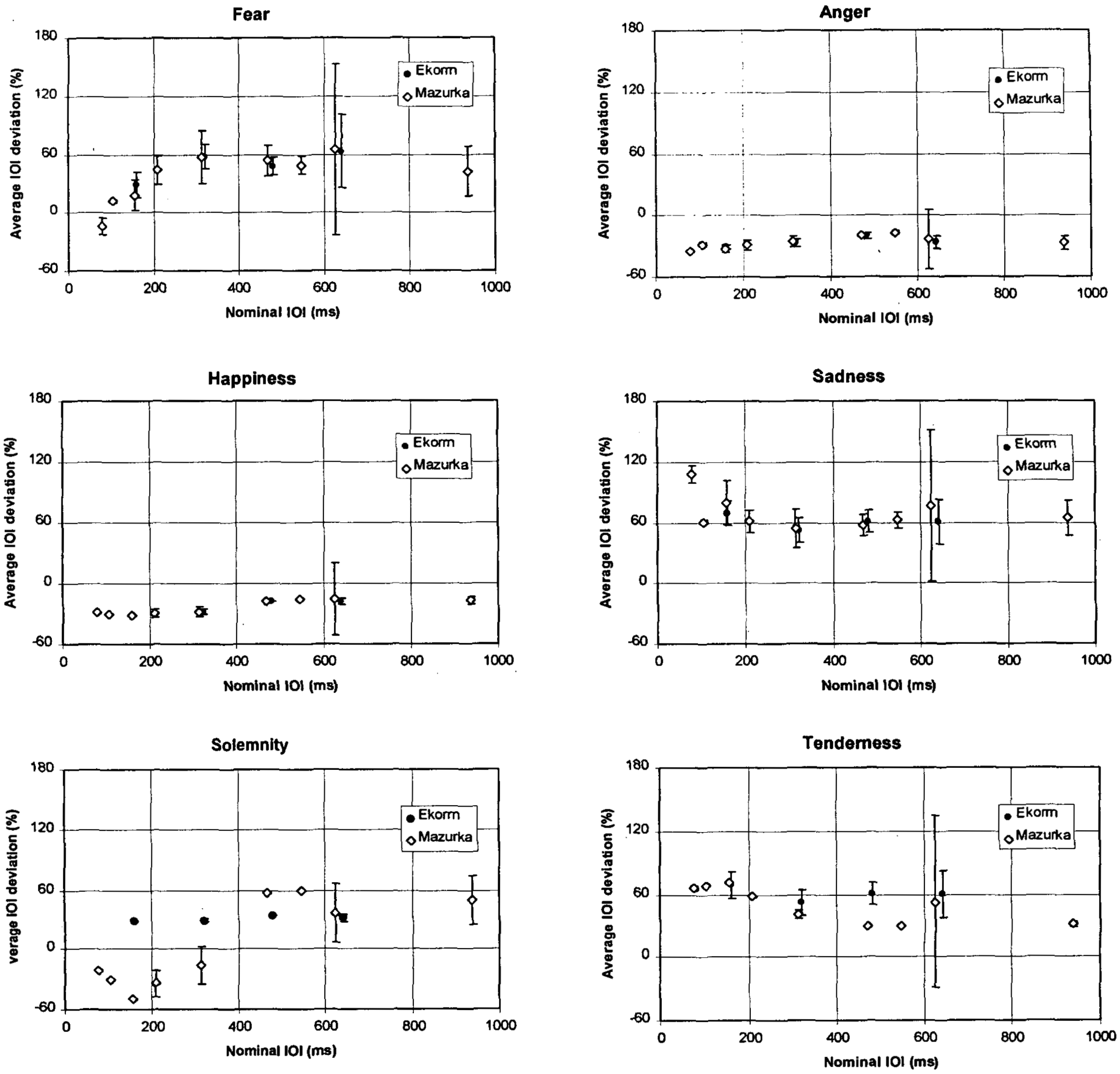

Figure 6 Relative deviations of IOI and for all six performances of each piece. Negative values imply a tempo faster than the non expressive versions, while positive deviations indicate lengthening of tone duration, and thus a slower tempo. The bars show the standard deviations.