Automatic Real-Time Extraction of Musical Expression

Anders Friberg[1,2], Erwin Schoonderwaldt[1,2], Patrik N. Juslin[2] & Roberto Bresin[1,2]

[1)] Royal Institute of Technology (KTH)
Speech Music and Hearing
Stockholm, Sweden

[2)] Uppsala University
Department of Psychology
Uppsala, Sweden

Abstract

Previous research has identified a set of acoustical cues that are important in communicating different emotions in music performance. We have applied these findings in the development of a system that automatically predicts the expressive intention of the player. First, low-level cues of music performances are extracted from audio. Important cues include average and variability values of sound level, tempo, articulation, attack velocity, and spectral content. Second, linear regression models obtained from listening experiments are used to predict the intended emotion. Third, the prediction data can be visually displayed using, for example, color mappings in accordance with synesthesia research. Preliminary test results indicate that the system accurately predicts the intended emotion and is robust to minor errors in the cue extraction.

Introduction

Expressivity is one of the most important issues in music performance. The acoustic parameters (*cues*) of each tone (e.g., sound level, duration, timbre, vibrato) constitute the basic means conveying not only the musical structure but also the expressive intentions of the performer. Previous research on musical expressivity has revealed that a set of emotional expressions can be accurately predicted using a limited set of cues, even without using score information (for a review, see Juslin, 2001). P. Juslin has developed a method which uses a set of statistical cues in conjunction with Brunswik's "lens model" to describe and quantify the efficiency of communication of emotions (e.g., Juslin, 2000; Juslin & Laukka, 2000). This process can be described both in terms of how a performer uses cues to express emotions, and how listeners use the same cues to recognize expressed emotions. However, this framework has not previously been implemented in an automatic recognition system.

Algorithms for extraction of musical notes from audio recordings are now relatively common, and even polyphonic recordings can be transcribed with reasonable accuracy (see Klapuri, 1997; Marolt, 2001). Most systems are focused on recognizing pitches and durations. The further translation of note durations into music notation is a separate and in some cases rather complicated task, because it essentially requires an understanding of the expressive principles used by music performers.

The current system combines a low-level cue extraction algorithm with a listener model to predict what emotion the performer is trying to convey in his or her performance. 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. This paper presents a real-time version of a more sophisticated system intended primarily to be used as a pedagogical tool. The goal of this tool is to provide performers with automatic cognitive feedback to improve their expressive skills (Juslin, Lindström, Friberg, & Bresin, 2001).

Overview of the system

In Figure 1, the overall layout of the system is shown. A performer is playing or singing into a microphone connected to a computer which runs the algorithm. Low-level cues are extracted mainly from the sound level. The cue values are standardized and fed into a listener model in the form of a regression equation. A calibration procedure is needed for the cue standardization (left side of Figure 1). The result is finally displayed graphically.

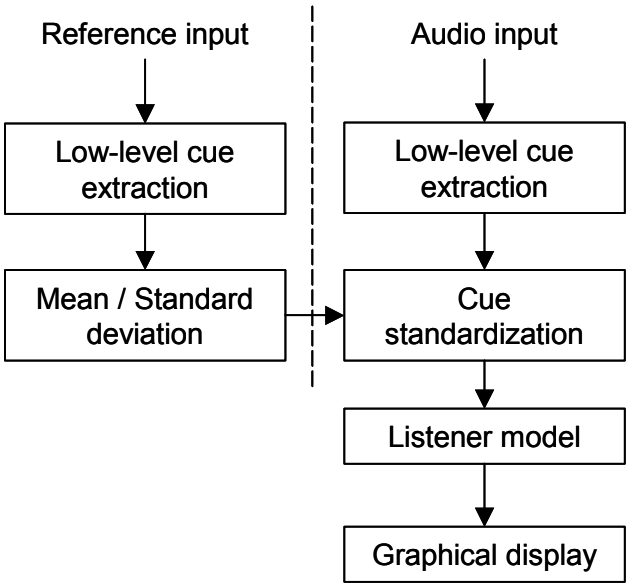

Figure 1. Overview of the system

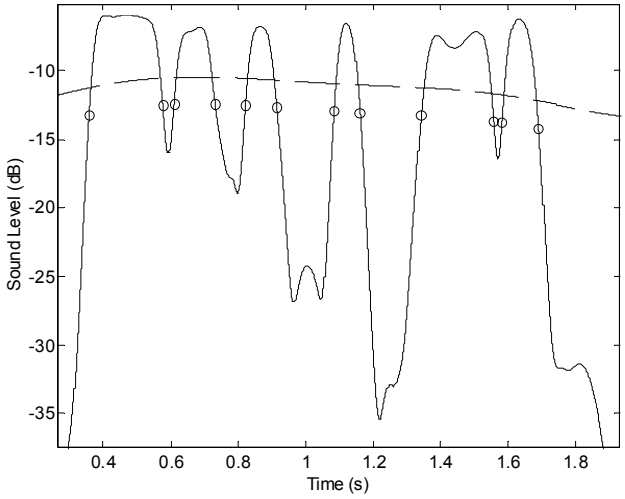

Figure 2. An example of the onset and offset detection applied on a short phrase played on a flute. The solid line is the tone profile, the dashed line is the phrase profile, and the circles indicate the detected onsets and offsets.

Cue extraction

The cues are computed from tone parameters. Thus, the point of departure is an onset-offset detector. Such a unit is needed in many different audio content applications such as beat tracking, automatic transcription, score following etc. There exists a variety of methods for onset-offset detection including the use of spectral content, changes in amplitude, derivatives of amplitude changes, pitch (F_0) extraction, and auditory filterbanks.

The current implementation employs a simple method of tone onset and offset detection based on the sound level of the incoming audio signal. The sound level is filtered using two different lowpass filters, yielding a *tone profile,* using a cutoff frequency of about 30 Hz, and a *phrase profile,* using a cutoff frequency of about 1 Hz. The lowpass filters each consist of four one-pole filters in cascade in order to get zero ripple. The tone profile is delayed, compensating for the slow response of the phrase profile filter. The points were the tone profile and the phrase profile (with an added small negative offset) cross define the onset and offset times (see Figure 2).

Obviously, this method only detects onsets when there is a dip in the sound level between tones. This will not happen when certain instruments are played with *legato* articulation. In such cases, either a pitch-based onset detection method (Monti and Sandler, 2000) or possibly a filterbank method (Klapuri, 1999) is necessary. Such a refinement is currently under construction. However, the use of statistical cues in combination with regression models used in the subsequent processing makes the system very robust to minor errors in the tone detection algorithm.

Using the extracted onset and offset times, the following parameters are computed for each tone: interonset duration (IOI), relative articulation (amount of silence relative to IOI), peak sound level, attack velocity, and spectral ratio. The spectral ratio is simply defined as the difference in sound level below and above 1000 Hz. The acoustic cues used in the subsequent analysis are obtained by computing running averages and standard deviations of the parameters. Thus, for each new tone there is an output from the cue detection.

The selected cues were found to be the most important for discriminating different emotions in previous listening experiments (cf. Juslin, 2001). There is, however, some cue redundancy depending on the particular instrument played. For example, sound level is often correlated with spectral slope. Therefore, it is necessary to select a relevant subset of cues for each instrument family.

From cues to emotions

In order to make the system work for different musical instruments and different types of music, as well as system dependent variables, such as record level and microphone setup, the cues are transformed to their corresponding *z-scores*. This requires a short calibration procedure in which the performer plays a few examples of the relevant music with variation with respect to dynamics, tempo, articulation and timbre. The standardization of the cues is performed by subtraction of the mean values of the respective cues of this reference input and division by the standard deviations.

An estimation of the strength of each modeled intended emotion is obtained from a regression equation taking the standardized cue values as input variables. Thus, continuous rating values are obtained for each emotion. For each output emotion we have a formula of the form:

$$\text{emotion strength} = \beta_1 * CUE_1 + \beta_2 CUE_2 + \beta_3 CUE_3 + \ldots,$$

where the beta coefficients (β_1, β_2 …) of cues of each regression equation were obtained from listener ratings of emotional expression in previous listening experiments. In this way, a "real" listening panel is included in the system. An optional decision-based method can be used to select the current dominant emotion. The fact that communication of emotion in music performance involves a number of partly redundant cues (i.e., cues that are intercorrelated and hence convey similar information) has important consequences for the present system. In particular, the redundancy makes the system relatively robust. For example, lack of information in one cue can be partly compensated for by another cue.

Graphical display

The extracted cues are mapped into colors and shapes on a computer monitor in order to provide a visual display corresponding to the intended emotions used by the player. The mapping of colors has been in accordance with current research in sonification and synesthesia (Barras, 1996: Poast, 2000). A visual display can be used both as a pedagogical tool and for artistic performances. It can provide the music student with immediate feedback of the musical expression without interfering with the sound in any way.

Results

Preliminary testing indicates that the system accurately predicts the intended emotion. Its response is also robust for several types of errors. This reflects inherent properties of the regression method in combination with the correlations between different cues that typically occur in music. Thus, errors in the cue extraction have only a small influence on the resulting prediction values of the different expressions. Further, a listener model from a previous experiment using other music examples and other instruments still makes a reasonable prediction about the intended emotions in new samples. The general impression is that, contrary to what was a priori expected, the prediction of different emotional expressions such as happiness or sadness is easier than, say, to make an automatic transcription of the music.

Implementation

Currently, preliminary versions are implemented both in pd and in EyesWeb (Camurri et al., 2000)

The system will be demonstrated live using a laptop computer. Possible input will be singing or playing a simple keyboard into a microphone. The audience will be invited to try the system.

Acknowledgments

This work has been supported by the Bank of Sweden Tercentenary Foundation and the EU-IST project MEGA. Roberto Dillon, University of Genova, Italy implemented a cue extraction prototype in EyesWeb.

References

Barrass, S. (1996). Sculpting a sound space with information properties, *Organized Sound*, 1(2):125-136

Camurri, A., Coletta, P., Peri, M., Ricchetti, M., Ricci, A., Trocca, R., and Volpe, G. (2000). A real-time platform for interactive dance and music systems. In I. Zannos (Ed.) *Proceedings of the International Computer Music Conference 2000, San Francisco: International Computer Music Association*, 262-265.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1797-1813.

Juslin, P. N. (2001). Communication of emotion in music performance: A review and a theoretical framework. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 309-337). New York: Oxford University Press.

Juslin, P. N., & Laukka, P. (2000). Improving emotional communication in music performance through cognitive feedback. *Musicae Scientiae, 4,* 151-183.

Juslin, P. N., Lindström, E., Friberg, A., & Bresin, R. (2001). *Play it again with feeling: Feedback-learning of musical expressivity.* Paper presented at the Meeting of the Society for Music Perception and Cognition, Kingston, Canada, August 2001.

Klapuri, A. (1997). *Automatic transcription of music*. M.Sc. Thesis, Tampere University of Technology, Finland.

Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1999*.

Monti, G. and Sandler, M. (2000). Monophonic transciption with autocorrelation. In *Proceeding of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy*, 257-260

Marolt, M. (2001). SONIC: Transcription of polyphonic piano music with neural networks. In *Proceedings of the Workshop on Current Research Directions in Computer Music. Barcelona, Spain*, 217-224

Poast, M. (2000). Color Music: Visual Color Notation for Musical Expression, *Leonardo*, 33(3): 215-221