社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-5

Programming by Playing and Approaches for Expressive Robot Performances

Angelica Lim, Takeshi Mizumoto, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University
36-1 Yoshida-Honmachi
Sakyo-ku, Kyoto, 606-8501 JAPAN
{angelica, mizumoto, tall, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

This paper extends our work with a theremin-playing robot accompanist. Here, we consider that a good accompanist should play with "expression": small deviations in volume, pitch and timing. We propose a Programming by Playing approach that allows a human flutist to transfer a performance to a robot theremin-ist, keeping these expressive changes intact. We also examine precisely what makes music robots play more or less "robotically, and survey the eld of musical expression in search of a good model to make robots play more like humans.

1 Introduction

A major challenge in human-robot interaction is the current lack of "humanness" in robot communication. Whereas humans express emotions using vocal inflection, expressive gestures and facial expression, robots have difficulty detecting these implicit emotions. Conversely, robot speech and movements remain dry, flat and unnatural. How can we make robots both detect these inexplicit emotions, and respond in emotionally empathetic, expressive ways? In the field of computer music, adding expression to synthesized music has already been a major goal since the 1980's [Todd, 1985a]. Musical expression is the result of adding variations [Sundberg, 1993] to a neutral ("robotic") performance, giving pleasing, natural renditions, sometimes even evoking emotions from listeners. Furthermore, there is evidence that communication of emotions in music follow the same patterns as speech [Juslin and Laukka, 2003]. Thus, we pursue the possibility that by giving robots musical expression detection and production abilities, we are one step closer to natural human-robot interaction.

We first propose a method called *Programming by Playing*: our anthropomorphic robot [Mizumoto *et al.*, 2009] listens to a flutist's performance with its own microphone, then replays the piece on the theremin with the same timing and dynamics as the human (Fig. 1). In the field of music robots, Solis et al. [Solis *et al.*, 2007] have already achieved an impressive increase in expressiveness by training an artificial neural network (ANN) to reproduce a human flutist's vibrato and note length. However, expression is a multifaceted problem that we can attack from many angles; for example, many musicians are able to play a given piece in a "sad" or "happy" manner on demand [Gabrielsson and Juslin, 1996]. How could we make robots play with emotion, too?

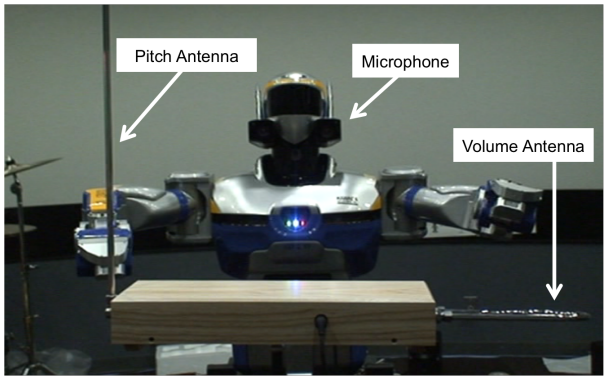


Figure 1: HRP-2 robot listens to a performance with its microphone, then replays it on the theremin by varying pitch and volume.

In the second part of this paper, we survey musical expression research not only from a computational music perspective, but also a psychological perspective. We first review some factors which make a performance expressive or not, then describe a 5-dimensional musical expression model [Juslin, 2003] suggested by music psychologist Juslin for the case of human musicians. We suggest that by extending the Programming by Playing approach to consider such a model, music robots could both perceive human musician's emotional intentions, and produce these emotions in their own playing as well.

2 A programming by playing approach

Let us begin by considering the simplest method for giving robots the appearance of human expressiveness: mimicry. At first sight, translating a human performance to a robot performance seems like a simple problem of music transcription. The naive approach would be to segment the performance into notes (using note onset detection, for example), extract each note's pitch and volume, and create a robot-playable MIDI file that contains each discretized note. This technique has worked well for piano because a piece can be represented simply by 3 parameters for each note: note length, pitch, and key-strike velocity [Raphael, 2009].

We claim that, while MIDI transcription may work well for piano, this note-level representation is an oversimplification for continuous instruments such as flute, voice and violin. Here are some concrete examples:

- *Intra-note volume changes* over the course of a note (e.g. *crescendo* or *diminuendo*) add fullness and expression for many continuous instruments. This is often overlooked because single piano notes cannot change volume in a controlled manner over time.

- *Intra-note pitch variation* known as vibrato can vary in speed and depth within a note. In most MIDI representations, vibrato speed and depth are set to constant values, if present at all.

- *Pitch bends*, or purposely playing slightly flat or sharp for expressive effect may be discretized to the nearest semi-tone.

- *Articulation* such as legato, attacked, staccato is produced by musicians using carefully composed note volume envelopes. In MIDI, this is often abstracted into a single average volume per note.

- *Timbre.* For instruments with timbral characteristics, tones can be "bright" or "dull" depending on their spectral composition; this information may be lost, too.

In summary, many critical details that may make a performance expressive can be lost when representing a piece symbolically! Thus, we must take care to represent our score in as rich a way as possible.

2.1 An Intermediate Representation: The Theremin Model

Raphael [Raphael, 2009] has proposed that the essence of an expressive melodic performance can be represented using a simple, but capable "theremin model". The theremin model takes after the electronic instrument of the same name that produces a pure sinusoidal pitch. Players can modulate the theremin's pitch frequency and volume independently, by moving their hands closer or farther from the respective pitch or volume antennas. We therefore represent a performance as a pitch trajectory and volume trajectory that continuously varies over time. Equation 1 represents the discrete sound signal s at time t:

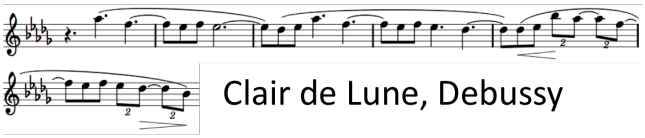


Figure 2: Example piece played by human flutist

$$s(t) = a(t) * \sin(f(t) * 2\pi * t), \quad (1)$$

where:

- $a(t)$ is the amplitude (a.k.a. power)
- $f(t)$ is the fundamental frequency (a.k.a pitch)

With a sufficient number of samples per second, this representation can capture almost all of the subtle information described in the previous section. For example, an attacked note would be equivalent to a sharp increase and quick drop in $a(t)$. Vibrato and note changes are captured in modulations over time in $f(t)$. Unfortunately, timbral characteristics, otherwise known as tone color, are not representable here, as a theremin's sound is characteristically composed of only a pure sine wave. See [Raphael, 2009] for a modified theremin model which adds timbre as a function of amplitude using hand-designed functions.

This simple representation captures the essential details of a performance while allowing for inter-instrument transfer. As noted in [Williamon, 2004], "The communication of emotion in music is generally successful despite individual differences in the use of acoustic features among performers... and different musical instruments." In more concrete terms, we can take as input a recording of a human's performance on flute, and output a performance by our robot thereminist.

2.2 Acoustic Processing

The input to our system is a wave file recording of a piece played by an intermediate flute player. It is recorded using the robot's own microphone, sampled at 44.1 kHz. As an example, consider the excerpt from Clair de Lune as shown in Fig. 2. Processing of the flute recording is composed of three parts: robot noise removal, continuous power extraction, and continuous fundamental frequency extraction.

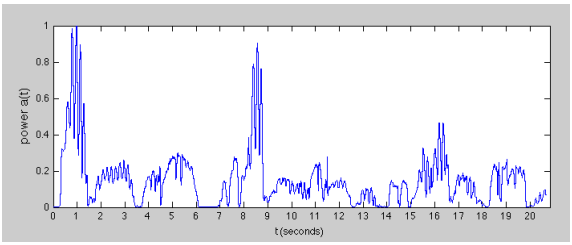
2.2.1 Noise Reduction

To increase robustness in our next steps, we first remove the robot fan noise also captured during recording. We use a filter called a spectral noise gate, which is likened to "background subtraction". By analyzing the frequency spectrum of a "silent" part of the recording (ie. when the flutist is not playing) we can reduce the fan noise by 24 dB from the entire recording (see Fig. 3). An FFT size of 2048 is used, resulting in 1024 frequency bands.

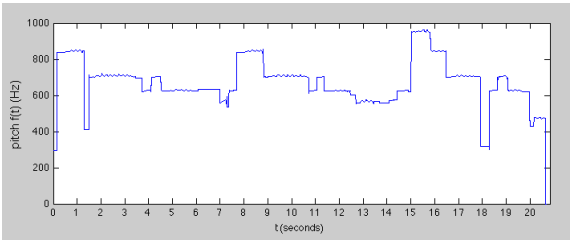
(a)

(b)

Figure 3: Clair de Lune original recording before (a) and after (b) fan noise reduction.



(a)



(b)

Figure 4: Continuous power a(t) and pitch f(t) extracted from flutist's Clair de Lune recording.

2.2.2 Continuous Power Estimation

We now have the filtered recorded signal $x(t)$. To extract the power $a(t)$, we use window sizes of 512 and sum the values of $x(t)^2$ of each bin. We then normalize the result to values between 0 and 1. The resulting power is plotted in Fig. 4(a).

2.2.3 Continuous Fundamental Frequency Estimation

Using the same input signal $x(t)$, we estimate the fundamental frequency at windows of 2048 using multi-comb spectral filtering and a hopsize of 1024. Instead of discretizing to the nearest semi-tone on the melodic scale, we measure to the nearest frequency in Hz. We can visualize the pitch estimation in Fig. 4(b).

2.3 From Representation to Performance

To convert the theremin model representation to a performance, we must first consider two constraints: instrument-related constraints and player (robot) constraints. Finally, we can convert our intermediate representation to a score playable by our robot thereminist.

2.3.1 Instrument-related Constraints

In this step we modify our performance representation depending on our target instrument. Consider that during silent sections of the recording where $a(t)$ is 0, the detected frequency $f(t)$ could have an arbitrary number of possible settings. To relate this situation to other instruments, a marimba player, for example, may return to "home position" during silent rests, and perhaps a flute player may hold the flute neutral with no keys depressed. In the case of our target instrument, the theremin, we assume that a theremin player would anticipate the next note during rests. Concretely, where $a(t)$ is 0, we set $f(t)$ to the next non-zero value of $f(t+k)$ where k is positive. Other possible modifications that may fall under Instrument-related constraints may include changing register (in case the human's instrument is, for example, a bass instrument, and the robot's instrument is soprano).

2.3.2 Player-related Constraints

Beginner and expert musicians have very different capacities. In our case, our player is an HRP-2 robot produced by Kawada Industries. However, in [Mizumoto *et al.*, 2009] Mizumoto et al. showed that the theremin-playing capabilities can be easily transfered to other robots, including a humanoid robot developed by Honda. In tests with another Kawada Industries robot, Hiro, we found that Hiro can change notes faster than HRP-2, due to a difference in arm weight. Thus, we must either modify our representation to be "easy" enough for our particular robot to play, or program these constraints into the motor module directly. For now, we scan our representation for any changes in frequency or volume that would violate the maximum acceleration of our robot arm, and remove them.

2.3.3 Generating a Robot-Playable Score

In this final step, we convert our intermediate representation score to a robot playable score. In preliminary experiments, we found that our system could handle a score with 3 pitch/volume targets per second (i.e., an update rate of 3 Hz) and still play in real-time using feedforward control. Using our Programming by Playing method, we thus update our robot's target note and volume multiple times per note, achieving more subtle tone and volume variations.

2.4 Preliminary Results and Improvements

We implemented Programming by Playing coupled with the theremin volume/pitch model to transfer the performance of "Clair de Lune" by a human flutist to a robot thereminist. In informal listening tests, the resulting performance does indeed sound more natural than our score-based method, but the reader is encouraged to evaluate the performance for themselves at

`http://winnie.kuis.kyoto-u.ac.jp/members/angelica/pbp`.

Although vibrato could be heard slightly, our maximum update rate of 3 Hz may have been too little to fully define vibrato (which previously had been hand defined

at 5-10 Hz). It also remains to be seen whether using the theremin model representation could be applied to instrument pairs other than flute-theremin. In particular, we have not implemented timbre into our performance representation, though this could be implemented with a third continuous parameter containing the extracted spectral centroid of the original recording.

An immediate use for Programming by Playing is allowing a human ensemble player to program the robot with his own style. That is, it is much easier to synchronize with a duet player that plays with natural timings, pauses, and articulations similar to one's own. Other uses for this version of Programming by Playing could include embodying famous musicians in a music robot based on their music recording.

Up until now, we have taken a relaxed approach to music expressiveness. As previously conjectured, intra-note volume variation, vibrato, pitch bends, articulation, and potentially timbre all contribute to making a performance more expressive. In the next section, we will see why these minute details are so important, and examine how we can exploit them to generate expressive performances "from scratch".

3 Expressive performances

3.1 Definitions

Expression is the most important aspect of a musician's performance skills, reports a nationwide survey of music teachers [Laukka, 2004]. But what is expression exactly? According to the survey, most teachers define expressivity as the communication of the emotional content of a piece, such as joy, sadness, tenderness or anger. Occasionally an expressive performance can even evoke these emotions in the listener ('being moved'), though it is not obligatory for music to be expressive [Davies, 1994]. What else makes human performers sound so different from the "dead-pan" rendition of a piece by a computer?

Another typical definition of expressiveness is "deviation from the score". Although scores may be marked with dynamic markings such as *decrescendo* or *accelerando*, expert performers contribute other expressive changes to the score [Palmer, 1997]. Typical examples include [Kirke and Miranda, 2009]:

- unmarked changes in tempo (such as playing faster in upward progressions of notes)
- loudness (high notes played slightly louder)
- modifications in articulation (staccato or legato)
- changes in intonation (making notes slightly flatter or sharper)
- adding vibrato at varying frequencies
- changing the timbre, if applicable to the instrument

The regularity of these deviations suggest that performances may be either subject to a set of grammar-like rules, or learned to some extent, and has thus spawned a vast number of attempts to reproduce these human-like qualities using computational methods.

3.2 A Need for Psychological and Physical Models

Automated computer systems for expressive music performance (CSEMPs) are programs which take a score as an input and attempt to output expressive, aesthetically pleasing, and/or human-like performances of the score. A recent survey of CSEMPs [Kirke and Miranda, 2009] outlined the various approaches including rule-based, linear regression, artificial neural network, case-based and others. There are too many approaches to outline here, but it is the conclusion of the survey that sparks the most interest.

According to the review, "Neurological and physical modeling of performance should go beyond ANNs and instrument physical modeling. The human/instrument performance process is a complex dynamical system for which there have been some deeper psychological and physical studies. However, attempts to use these hypotheses to develop computer performance systems have been rare." [Kirke and Miranda, 2009] They cite an attempt to virtually model a pianist's physical attributes and constraints [Parncutt, 1997] as one of these rare cases. Thus, in the following sections, we delve deeper into the phenomenon of expression, in order to better understand this challenge.

3.3 Factors

What factors can make a performance expressive or not? Though researchers typically focus on how the *performer* is expressive, the phenomenon can involve environmental factors, too. We briefly overview these factors from [Juslin, 2003], to better understand the variables involved.

3.3.1 The Piece

The musical composition itself may invoke a particular emotion. For example, Sloboda [Sloboda, 1991] found that certain scores consistently produced tears in listeners: scores containing a musical construct called melodic appogiaturas. Shivers were found in participants during points of unprepared harmonies or sudden dynamic change in the score. Score-based emotions have been well-studied, and in a recent review of 102 studies by Livingstone et al. [Livingstone *et al.*, 2010], it was found that happy emotions are most correlated with pieces in major keys, containing simple harmonies, high pitch heights, and fast written tempos. Loud pieces with complex harmonies, in a minor key with fast tempos were considered "angry", and so on. Though we choose not to treat this score-based emotion in the present paper, this is useful to know so we do not confuse emotion evoked by a written score with emotion projected by a performer.

3.3.2 The Listener

The musical background and preferences of the listener may have an effect on the perceived expressiveness of a piece. For example, listeners with less musical education appear to rely more heavily on visual cues (such as gestures or facial expression) rather aural cues when deciding on an affective meaning of a musical perfor-

mance [Thompson *et al.*, 2005]. However, even children at the age of 5 years are able to differentiate happy and sad pieces based on whether the tempo is fast or slow, and six-year-olds can classify additionally based on major versus minor mode [Dalla Bella *et al.*, 2001]. Interestingly, detection of basic emotions such as joy, sadness, and angry even appear to be cross-cultural: Western and Japanese listeners are able to distinguish these emotions in Hindustani ragas [Balkwill and Thompson, 1999]. Thus, though we should take care during evaluations of expressiveness, we should know that detection of emotion in music is not as elusive as it may seem.

3.3.3 The Context

The performance environment, acoustics or influence from other individuals present can also affect the expression perceived [Juslin, 2003]. For example, music at a patriotic event may evoke more emotion in that context than in another. Another example is Vocaloid's virtual singer Hatsune Miku, who performs at concerts to a large fanbase despite being a synthetic voice and personality. In these cases, perceived expressiveness may also depend on factors such as visual and cultural context.

3.3.4 The Instrument

Whereas percussion instruments such as piano can only vary timing, pitch and volume, continuously controlled instruments such as flute and violin have many more expressive features. They can change timbre to obtain "bright" versus "dull" tones [Raphael, 2009], have finer control over intensity and pitch, and can produce vibrato. Interestingly, human voice is also in this set of continuously controlled instruments. Since many studies find that timbre, pitch variations and vibrato [Livingstone *et al.*, 2010] can have an effect on the perceived expressiveness, the choice of instrument can limit or extend the ability to convey a particular emotion.

3.3.5 The Performer

Clearly the most important factor of expression lies in the performer, which is why this factor has been so extensively studied. The musician's structural interpretation, mood interpretation, technical skill and motor precision can all affect the perceived expressiveness. We explore the expressive aspects of a performer in detail in the next section.

3.4 A Model for Performer Expressiveness

Up until now, performer expressiveness has been informally described by a large number of performance features, such as playing faster and louder, and with more or less vibrato. Are there any models that can bring order and sense to these empirically derived findings?

Four computational models for expressive music performance were considered in [Widmer and Goebl, 2004]: KTH's rule-based model [Bresin *et al.*, 2002], Todd's model based on score structure [Todd, 1985b], Mazzola's mathematical model [Mazzola, 2003], and Widmer's machine learning model [Widmer and Goebl, 2004]. However, according to the CSEMP review, they are still not sufficient. As the review points out, we should search for a model that adheres to certain requirements: it should take into account psychological and neurological factors, as well as physical studies.

Music psychologist Juslin proposed a 5-faceted model [Juslin, 2003] [Juslin *et al.*, 2002] that separates expressive performance into a manageable, but all-encompassing space: Generative rules, Emotion patterns, Random variance, Motion-inspired patterns, and Stylistic unexpectedness (called GERMS). Details of each element are described shortly. Juslin et al. implemented the first 4 parts of the model in 2002 using synthesis [Juslin *et al.*, 2002], and tested each facet in a factorial manner. Their results, along with evidence that each of these facets corresponds to specific parts of the brain [Juslin and Sloboda, 2010], make this model promising. Even if Juslin's model is not quite correct, we claim that it is still very useful for designing factorized modules for robot expression.

3.4.1 Generative rules for musical structure

Similar to speech prosody, musicians add beauty and order to their playing by adding emphasis to remarkable events [Juslin and Sloboda, 2010]. By adding the following features, the musician makes their structural interpretation of a piece clear:

- Slow at phrase boundaries [Clarke, 1988]
- Play faster and louder in the center of a phrase [Todd, 1985b]
- Micropause after phrase and subphrase boundaries [Friberg, A. And Sundberg, J. And Fryden, 1987]
- Strong beats louder, longer, and more legato [Palmer and Kelly, 1992]

A complete and slightly different ruleset is listed in Juslin's experiments [Juslin *et al.*, 2002]. Listeners rated synthesized pieces with this component as particularly "clear" and "musical".

3.4.2 Emotion

We previously defined musical expression partly as the ability to communication emotion. Particular sets of musical features can evoke emotions, such as happiness, sadness, and anger. Livingstone et al. recently surveyed 46 independent studies and summarized the main acoustic features corresponding to each of 4 basic emotions [Livingstone *et al.*, 2010]. We reproduce here the most notable of each group. Note that the order may matter (i.e., first features characterizing the emotion more strongly). In the case of conflicting reports, we removed the one with less experimental backing.

1. **Happy**: Tempo fast, Articulation staccato, Loudness medium, Timbre medium bright, Articulation variability large, Note onset fast, Timing variation small, Loudness variability low, Pitch contour up, Microstructure regularity regular, F0 sharp
2. **Angry**: Loudness loud, Tempo fast, Articulation staccato, Note onset fast, Timbre bright, Vibrato

large, Loudness variability high, Microstructural regularity irregular, Articulation Variability large, Duration contrasts sharp

3. **Sad**: Tempo slow, Loudness low, Articulation legato, F0 flat, Note onset slow, Timbre dull, Articulation variability small, Vibrato slow, Vibrato small, Timing variation medium, Pitch variation small, Duration contrasts soft

4. **Tender**: Loudness low, Tempo slow, Articulation legato, Note onset slow, timbre dull, Microstructural regularity regular, Duration contrasts soft

In the evaluation of this factor, happiness versus sadness were implemented by varying tempo, loudness, and articulation. Upon adding emotional cues, listeners judged the piece as "expressive" and "human" by a large factor.

3.4.3 Randomness

Humans, unlike computers, cannot reproduce the exact same performance twice. In studies on finger tapping [Madison, 2000], even professional musicians varied 3-6% (of the inter-onset interval) in tapping precision. It is thus why some software programs such as Sibelius add some random fluctuation to make MIDI playback sound more human [Kirke and Miranda, 2009]. Interestingly, these fluctuations are not completely random; the variation can be simulated by a combination of 1/f noise and white noise [Gilden *et al.*, 1995]. Motor delay noise was simulated in [Juslin *et al.*, 2002] by adding white noise to each note onset time and sound level. Internal timekeeper lag was added by white noise as a function of the note length, filtered to obtain 1/f pink noise.

Although the idea of making robots purposely less precise sounds intriguing, it remains to be seen whether music robots do actually play as perfectly as the computer clocks that control them. Do they achieve perfect timings despite variations in environment such as network lag and motor delay? In computer synthesis tests this randomness factor made performances more "human" over the neutral versions.

3.4.4 Motion constraints

The fourth component refers to two kinds of motion constraints. One pertains to voluntary patterns of human biological motion. Mainly, the final ritardandos of musical performances has been found to follow a function similar to that of runners' decelerations [Friberg and Sundberg, 1999], but more examples can be found in [Juslin *et al.*, 2002]. The other kind of motion constraint is information that specifies that the performer is human. For example, a pianist could not physically play two distant notes faster than two notes side-by-side. This is an involuntary motion constraint.

In terms of robot implementation, safety mechanisms are probably already programmed into lower level motor controls of our music robots. This corresponds to the latter, involuntary constraint. However, similar to the Player-related constraints described in our Programming by Playing approach, it could be possible to add additional motor constraints that mimic natural human movement curves. For example, our pitch or volume trajectories could be smoothed or interpolated with splines. As for the effect of adding the biological motion constraint: listeners rated synthesized pieces more "human".

3.4.5 Stylistic unexpectedness

Despite the systematic discovery of many common expressive features among musicians, humans of course have the freedom to change their style on a whim. For examples, some performers may intentionally play the repeat of a same phrase differently the second time, or a musician may pause longer than usual for dramatic effect. Indeed, in a study on pianists playing the same piece, it was found that graduate students had rather homogenous timing patterns, whereas experts showed more originality and deviations [Repp, 1997].

This element was not included in Juslin's tests due to the difficulty in implementation. Indeed, this could be the crux of what gives originality to a robot's performance. Could we use Programming by Playing to learn the probabilistic tendency of one or many human artists? Could we shape a music robot's "personality" based on this factor (more or less showmanship, or extroversion)? How exactly to approach this module is an open area for research, and perhaps AI in general.

3.5 Towards an Expressive Music Robot

It seems clear that an expressive music robot should thus have 5 modules:

1. **Prosody controller**: to clarify music structure

2. **Emotion controller**: to store and produce an intended emotion

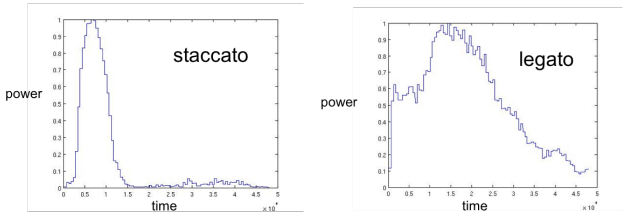
3. **Humanness controller**: to add randomness to imitate human imprecision

4. **Motor smoothness controller**: to mimic human biological movement

5. **Originality controller**: to add unexpected deviations for originality

Although we are still far from implementing this model in full, we have started by implementing the Prosody and Emotion controller. We start with a hand-entered score of the traditional folk song, Greensleeves. Then, it is modified using the generative rules for musical structure mentioned previously. We then address Emotion using Programming by Playing. Focusing on the articulation feature, we record a flutist playing notes in each of the Happy (staccato) and Sad (legato) styles.

We extract volume envelopes for each type as shown in Fig. 5, and apply the volume envelopes to all notes in the continuous volume representation. Our result is two different performances, one to convey sad emotion and the other conveying happiness. It is unclear whether the robot performances effectively convey the emotions as desired, but expressiveness again seems improved over the neutral version. In addition, we have achieved expressiveness without resorting to mimicry.



(a) (b)

Figure 5: Volume envelopes for staccato and legato articulations.

In an ideal version of Programming by Playing, more features (not only articulation) should be extracted. By extracting these acoustic features automatically, perhaps similar to [Mion and De Poli, 2008], we could recognize the emotional content of the human musician.

4 Conclusion and future work

In this paper, we introduced a paradigm called Programming by Playing. We showed how it could be used for expressive robot performance through both mimicry and generation. A key point of the approach was that small details in performance can have a great impact on a performance's expressive content; thus, a good symbolic representation is important.

We also tried to demystify the phenomenon called expression – by applying a 5-facet model to music robot design, we realize that features for structural clarity and emotion are distinct. Another interesting find was that in order to sound more human, we may need to add slight human imprecision. This may be contrary to our current efforts to make "virtuoso" music robots that play faster, but more unrealistically. And finally, the key ingredient missing before music robots will be accepted is a kind of originality or "personality", giving the element of surprise to performances.

All of these factors may be applicable to robot design in general, for example making synthetic voice and movement less "robotic". Yet, what is the goal for music robots? Do we want them to sound more realistic, more human? If that is the case, this complex phenomenon called expression may be the missing ingredient.

Acknowledgments

This work was partially supported by a Grant-in-Aid for Scientific Research (S) No.1910003 and the Global COE Program from JSPS, Japan.

References

[Balkwill and Thompson, 1999] Laura-Lee Balkwill and William Forde Thompson. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception*, 1999.

[Bresin *et al.*, 2002] R. Bresin, A. Friberg, and J. Sundberg. Director musices: The KTH performance rules system. *SIGMUS*, pages 43–48, 2002.

[Clarke, 1988] E.F. Clarke. Generative principles in music performance. *Generative processes in music: The psychology of performance, improvisation, and composition*, pages 1–26, 1988.

[Dalla Bella *et al.*, 2001] S Dalla Bella, I Peretz, L Rousseau, and N Gosselin. A developmental study of the affective value of tempo and mode in music. *Cognition*, 80(3):B1–10, July 2001.

[Davies, 1994] Stephen Davies. *Musical meaning and expression*. Cornell University Press, 1994.

[Friberg, A. And Sundberg, J. And Fryden, 1987] L. Friberg, A. And Sundberg, J. And Fryden. How to terminate a phrase. An analysis-by-synthesis experiment on a perceptual aspect of music performance. *Action and perception in rhythm and music*, 55:49–55, 1987.

[Friberg and Sundberg, 1999] Anders Friberg and Johan Sundberg. Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *The Journal of the Acoustical Society of America*, 105(3):1469, 1999.

[Gabrielsson and Juslin, 1996] Alf Gabrielsson and Patrik N. Juslin. Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. *Psychology of Music*, 24(1):68–91, April 1996.

[Gilden *et al.*, 1995] D. L. Gilden, T. Thornton, and M. W. Mallon. 1/f noise in human cognition. *Science*, 267(5205):1837, 1995.

[Juslin and Laukka, 2003] PN Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code?. *Psychological Bulletin*, 129(5):770–814, 2003.

[Juslin and Sloboda, 2010] Patrik N. Juslin and John Sloboda. *Handbook of Music and Emotion*. Oxford University Press, USA, 1 edition, February 2010.

[Juslin *et al.*, 2002] PN Juslin, A Friberg, and R Bresin. Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae*, 6(1; SPI):63–122, 2002.

[Juslin, 2003] PN Juslin. Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music*, 31(3), 2003.

[Kirke and Miranda, 2009] A Kirke and ER Miranda. A Survey of Computer Systems for Expressive Music Performance. *ACM Computing Surveys*, 2009.

[Laukka, 2004] P Laukka. Instrumental music teachers' views on expressivity: a report from music conservatoires. *Music Education Research*, 2004.

[Livingstone *et al.*, 2010] Steven R Livingstone, Andrew R Brown, Ralf Muhlberger, and William F Thompson. Modifying Score and Performance Changing Musical Emotion : A Computational Rule System for Modifying Score and Performance. *Computer Music Journal*, 34(1):41–65, 2010.

[Madison, 2000] G Madison. Properties of Expressive Variability Patterns in Music Performances. *Journal of New Music Research*, 2000.

[Mazzola, 2003] Guerino Mazzola. *The Topos of Music: Geometric Logic of Concepts, Theory, and Performance*. Birkhäuser Basel, 1 edition, January 2003.

[Mion and De Poli, 2008] Luca Mion and Giovanni De Poli. Score-Independent Audio Features for Description of Music Expression. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):458–466, 2008.

[Mizumoto *et al.*, 2009] Takeshi Mizumoto, Hiroshi Tsujino, Toru Takahashi, Tetsuya Ogata, and Hiroshi G Okuno. Thereminist robot : development of a robot theremin player with feedforward and feedback arm control based on a theremin's pitch model. In *IROS*, pages 2297–2302, 2009.

[Palmer and Kelly, 1992] C Palmer and MH Kelly. Linguistic Prosody and Musical Meter in Song. *Journal of Memory and Language*, pages 525–542, 1992.

[Palmer, 1997] C. Palmer. Music performance. *Annual Review of Psychology*, 48(1):115–138, 1997.

[Parncutt, 1997] R. Parncutt. Modeling piano performance: Physics and cognition of a virtual pianist. In *ICMC*, pages 15–18, 1997.

[Raphael, 2009] Christopher Raphael. Symbolic and Structrual Representation of Melodic Expression. In *ISMIR*, pages 555–560, 2009.

[Repp, 1997] BH Repp. The aesthetic quality of a quantitatively average music performance: Two preliminary experiments. *Music Perception*, 1997.

[Sloboda, 1991] JA Sloboda. Music Structure and Emotional Response: Some Empirical Findings. *Psychology of music*, 1991.

[Solis *et al.*, 2007] Jorge Solis, Kei Suefuji, Koichi Taniguchi, Takeshi Ninomiya, and Maki Maeda. Implementation of Expressive Performance Rules on the WF-4RIII by modeling a professional flutist performance using NN. In *ICRA*, pages 2552–2557, 2007.

[Sundberg, 1993] J. Sundberg. How can music be expressive? *Speech communication*, 13(1-2):239–253, 1993.

[Thompson *et al.*, 2005] W.F. Thompson, Paul Graham, and F.A. Russo. Seeing music performance: Visual influences on perception and experience. *Semiotica*, 156(1/4):203–227, 2005.

[Todd, 1985a] Neil Todd. A model of expressive timing in tonal music. *Music Perception*, 3(1):33–57, 1985.

[Todd, 1985b] Neil Todd. A Model of Expressive Timing in Tonal Music. *Music Perception: An Interdisciplinary Journal*, 3(1):33–57, 1985.

[Widmer and Goebl, 2004] Gerhard Widmer and Werner Goebl. Computational Models of Expressive Music Performance: The State of the Art. *Journal of New Music Research*, 33(3):203–216, September 2004.

[Williamon, 2004] Aaron Williamon. *Musical excellence: strategies and techniques to enhance performance*. Oxford University Press, 2004.