EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Towards expressive musical robots: a cross-modal framework for emotional gesture, voice and music

Angelica Lim*, Tetsuya Ogata and Hiroshi G Okuno

Abstract

It has been long speculated that expression of emotions from different modalities have the same underlying 'code', whether it be a dance step, musical phrase, or tone of voice. This is the first attempt to implement this theory across three modalities, inspired by the polyvalence and repeatability of robotics. We propose a unifying framework to generate emotions across voice, gesture, and music, by representing emotional states as a 4-parameter tuple of speed, intensity, regularity, and extent (SIRE). Our results show that a simple 4-tuple can capture four emotions recognizable at greater than chance across gesture and voice, and at least two emotions across all three modalities. An application for multi-modal, expressive music robots is discussed.

Keywords: affective computing, gesture, entertainment robots.

1 Introduction

Music robots have succeeded in entertaining and enthralling audiences around the world with their virtuoso performances. Take Shimon [1], a music robot that has toured Europe and the United States–this robot plays the marimba and interacts harmoniously with human musicians on stage. LEMUR bots, orchestrated teams of robot musicians, play complicated scores for piano and percussion with perfect timing, synchronization and repeatability [2]. In Japan, a flute-playing robot [3] plays Flight of the Bumblebee with speed, precision, and endurance comparable to the world's top human flutists. From a technical standpoint, these performances are not unlike watching an amazing guitarist on stage–they are awe-inspiring and extremely fun to watch.

We propose that the next great challenge is to create music robots that engage listeners in a different way: playing the piece in a way that stirs up emotions and moves the listener. Needless to say, this is an extremely difficult task for robots, as they lack emotions themselves. Neurologist and musician Clynes [4] gives us insight into the power that skilled (human) musicians possess, p. 53:

> "In the house of Pablo Casals in Puerto Rico, the Master was giving cello master classes. On this occasion, an outstanding participant played the theme [...] from the Haydn cello concerto, a graceful and joyful theme. Those of us there could not help admiring the grace with which the young master [...] played. Casals listened intently. "No," he said, and waved his hand with his familiar, definite gesture, "that must be graceful!" And then he played the same few bars–and it was graceful as though one had never heard grace before, so that the cynicism melted in the hearts of the people who sat there and listened. [...] What was the power that did this? A slight difference in the shape between the phrase as played by the young man and by Casals. A slight difference–but an enormous difference in power of communication, evocation, and transformation."

Although achieving Casals' level of expression is still far off, there remains a large gap to be filled between his play and that of current music robots. The problem is known ironically as "playing robotically", stepping from note to note exactly as written, without expression. Casals himself attributed his mastery of expression to a divine talent, saying, "It comes from above" [4]. Trying to algorithmically describe this "divine talent" of score shaping could

* Correspondence: angelica@kuis.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto University, Kyoto, Japan

help us better understand not only the difference between an amateur and expert musician, but perhaps also the difference between a human and a robot.

1.1 Definitions

Expression is the most important aspect of a musician's performance skills, reports a survey of music teachers [5]. But what is expression exactly? It has been often defined as "deviations from the score", but this definition is rather vague. Music psychologist Juslin provides a useful break down of these "deviations" into five components [6]: *generative rules, emotional expression, random fluctuations, motion principles*, and *stylistic unexpectedness* (GERMS). *Generative rules* are similar to speech prosody, varying features such as dynamics, timing, and articulation to clarify piece boundaries, harmonic structure, and so on. *Emotional expression* communicates emotions like joy or tenderness through variations of features like tempo and volume. Adding *random fluctuations* and mimicking human *motion principles* of dynamics have been shown to improve perception of naturalness. *Stylistic unexpectedness* describes deviation from stylistic convention to add tension and unpredictability.

An ultimate goal is to create a robot with all of these expressive dimensions, but as a first step, we focus on *emotional expression*. The majority of responses to an open-ended survey of 135 of expert musicians defined expression in terms of "communicating emotions" or "playing with feeling" [7]. Listeners also report that music communicates emotion; 76% of participants in [5] respond that music expresses emotion "often". On a practical level, emotion-laden music has also been shown to be therapeutic: patients that listen to positive music show increased tolerance to pain by 20-25% [8]. For these reasons, improving expression of emotion for music robots is the main focus of this study.

1.2 Related work

Within music robotics, a few studies address the problem of "deadpan" or "robotic" performances. Ogata et al. [9] developed a violin playing machine which could change timbre according to adjectives such as "moist" or "dry". Solis et al. [10] developed an expressive flute-playing robot, which learned parameters such as note length and vibrato based on a human flutist's performance. Nakano and Goto extracted the vocal [11] and facial expression from a famous singer, and later reproduced the performance on the hyper-realistic android HRP-4C. Lim et al. [12] used a programming by playing approach to transfer expression from a human flute player's performance to a robot thereminist. In the last three cases, the expressive content was learned with respect to a specific score; no expressive generation for a novel score has yet been reported. The power of expressive movement has also been harnessed by some robot researchers. Instead of using movements exclusively to play notes, the Shimon marimba playing robot [1] performs expressive gestures like head-bobbing to add humanness to its performance. Humanoid violin and trumpet robots from Toyota shift weight from one leg to the other, raising and lowering their instruments [13]. To summarize, no experiments have yet been conducted to assess the emotional communication between music robots and humans.

Of course, expressive generation has been long explored in the field of computer music. The annual RenCon competition welcomes researchers to render expressive musical performances given a nominal score, reviewed in [14]. Yet, we find three major advantageous aspects of studying expression with embodied robot musicians over computer-generated music.

Aspect #1. Multiple modalities for expression

We hypothesize that music robots with multi-modal redundancy can convey a desired emotion to an audience with a higher reliability than music alone. For instance, it has been shown that non-musicians rely more heavily on visual rather than audio information to determine emotional content of a performance [15]. Features linked to expressiveness, such as note length, can be perceptually modified with gesture: [16] showed that marimba sounds can be perceived as shorter or longer depending on the percussionist's arm trajectory. Emotionally expressive robots such as Kobian [17] effectively convey emotions using facial features and pose, such as a slouched posture for sadness. By compounding multiple sources of affective information, a music robot may have a better chance at communicating emotional intent to an audience.

Aspect #2. Situatedness

Context is a major factor controlling the emotion, if any, induced by music. For instance, music performed at a memorial service would have a different effect as the same song played at a bustling restaurant. Adapting to the listener is important, too; "slow and soft" may indicate sad music in normal situations, but would the same music amplified for a listener who is hard of hearing still sound sad, and why? The physical distance between a robot and audience can also play a role; leaning closer to an interlocutor can indicate aggression or anger, and farther away imply fear. The fact that a robot is embodied and always situated in a particular location opens the door to a new area of research where these contextual factors of emotion can be explored in a non-contrived manner.

Aspect #3. A platform for developing holistic models for emotion

Simply playing music in an expressive manner is not enough. As stated in [14], "neurological and physical modeling of performance should go beyond [artificial neural networks] 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." Indeed, many automatic music expression systems emphasize performance, to the detriment of general application and explicative power. Humanoid robots are polyvalent, and thus are perfect platforms for synthesizing the work from specialized, disparate domains. The idea is to develop a general 'emotional intelligence'–an ability to express emotion in music as well as speech, gesture, and other modalities.

In this study, we explore in particular Aspect #3: we develop and test an emotion framework that encompasses voice, gesture, and music for humanoid robots. In particular, we research the "how"–*how* one plays music, how one moves, or *how* one speaks, and test whether this emotional "style" is consistent across modalities. The eventual goal is to convey emotional music in a rich, multi-modal manner (as described in Aspect #1) using one unified emotional framework. If achieved, this could also give evidence to speculations that emotional expressions in music, movement and voice have a common form (e.g., [4,18]).

1.3 Requirements

A framework for emotional music robots should be powerful, flexible, yet simple:

1. **Simple**: Music researchers have found an abundant list of features to analyze and create emotional music. For instance, high tempo is correlated with happiness or anger, and low volume with sadness [14]. How can we most economically account for the differences in emotion? A good approach should reduce this feature set to the most meaningful dimensions.
2. **Powerful**: First, the model should be powerful enough to be used for both analysis and generation of emotional performances. The most popular affect model, the Circumplex model of arousal-valence, is commonly used for generation of emotion-laden media (e.g., Kismet robot [19]). On the other hand, it is not always clear how to map the dimensions for analysis. Features ranging from speed to voice quality to pitch changes have been found to be correlated with pleasantness [20], which is why model-free classification using high-dimensional feature vectors is such a popular approach for emotion recognition (e.g., [21,22]). Second, the representation should be *nuanced*, taking into account that a portrayal may be a mix of emotions, instead of discrete states. As Fellous argues: "Implementing emotions as 'states' fails to capture the way emotions emerge, wax and wane, and subside." [23].
3. **Cross-modal**: The model should allow for emotion expression across pieces, instruments and modalities. Studies have shown that musicians of various instruments can play arbitrary passages, even simple repeated notes, in a given emotional style [24]. From an engineering viewpoint, an instrument-independent emotion approach is useful for robot portability as well. Consider also that piano teachers may show a student how to perform a *furioso* passage by singing, and conductors transmit their expressive intent to orchestra members through gesture. A model that generalizes across pieces, instruments, and modalities would reflect the *essence* of emotion that humans store, which is important from an artificial intelligence standpoint.

Our proposed requirements are well summarized in a recent review of emotion/affect models:

> "Despite the existence of diverse affect models, the search for an (1) optimal *low-dimensional* representation of affect, (2) for *analysis and synthesis*, and (3) for *each modality or cue*, remains open." [25] (our emphasis and numbering).

2 A representation and framework for emotion across modalities

A grouping of common factors across modalities is the basis of our proposed emotion representation. Here, we refer the reader to key reviews for both emotion *recognition* and *generation* in music, speech, and gesture. For instance, Livingstone et al. [26] provide an excellent up-to-date review of research in musical emotion, and emotional speech is reviewed in Cowie et al. [20]. Emotional gesture has been less studied, though Pelachaud's study in animated characters [27] may be the most state-of-the-art. The results of our review and feature groupings are summarized in Table 1.

Our review found that the most salient factors for emotion recognition and generation could be clustered perceptually into *speed, intensity, regularity*, and *extent*. The most robust feature has been called speech rate, velocity of gesture, or tempo; in other words, *speed*. The dichotomy between fast and slow is the simplest way to distinguish between happy and sad voices, music and gestures. However, this simple distinction often confuses happiness with anger or fear, which is why other features are needed. Another major feature is *intensity*, which we define as the perceptual distinction between gradual and abrupt. For instance, anger is often characterized with abrupt movements or attacked words and musical notes [18,28]; sad voices, music, and gestures are correlated with low intensity, gradual changes. *Regularity* is the

Table 1 DESIRE parameters and associated emotional features for modalities of voice, gesture and music

		Modality mappings to relevant emotional features		
Parameter	**Description**	**Voice**	**Gesture**	**Music**
Speed	slow vs. fast	*speech rate* [20], pauses [18]	*velocity* [35], animation [36], quantity of motion [28]	*tempo* [24,26]
Intensity	gradual vs. abrupt	*voice onset rapidity* [18], articulation [20]	*acceleration* [35], power [37]	*note attack* [24], articulation [26]
Regularity	smooth vs. rough	*jitter* [18], voice quality [18,20]	directness [35], *phase shift* [38,39], fluidity [27]	*microstructural irregularity* [26], timbral roughness [24]
Extent	small vs. large	*pitch range* [20], *loudness* [18]	*spatial expansiveness* [36,37], contraction index [35]	*volume* [24,26]

Features in *italics* are used in our experiments.

perception of smooth versus rough. For example, fear can be indicated in a voice with a breathy voice quality [20], quivering (as opposed to smooth) gestures, and music with irregular, sporadically played notes. Finally is the idea of *extent*: for gesture, large, expansive movements could be characteristic of happy or (hot) anger. Smaller movements can indicate depression or making oneself small due to fear.

2.1 DESIRE Framework

We propose a representation of emotion through the parameters of speed, intensity, regularity, and extent, based on the result of our literature review. For short, we call this parameter set *DESIRE: description of emotion through speed, intensity, regularity, and extent*, or simply *SIRE*. The DESIRE framework in Figure 1 illustrates how we extract a SIRE representation, and express it through different modalities.

In short, the DESIRE framework is:

1. *DESIRE representation*, dynamic parameters representing universally accepted perceptual features relevant to emotion (SIRE). We define them as a 4-tuple of numbers $S,I,R,E \in [0, 1]$.
2. *Parameter mappings*, between the dynamic parameters and hardware-specific implementation.

The *parameter mappings* can be divided into two-layers (see Figure 1):

- *Hardware-independent layer*: A mapping from DESIRE to perceptual features. These mappings are those outlined in Table 1.
- *Hardware-specific layer*: A mapping of perceptual features to a hardware-specific implementation (discussed in Section 4.2).

We have implemented the DESIRE framework on three systems representing three modalities:

1. **Voice:** HRP-2 singing with Vocaloid real-time opera synthesizer (used in [29])
2. **Gesture**: NAO (http://www.aldebaran-robotics.com) gesturing robot (reported in [30]).

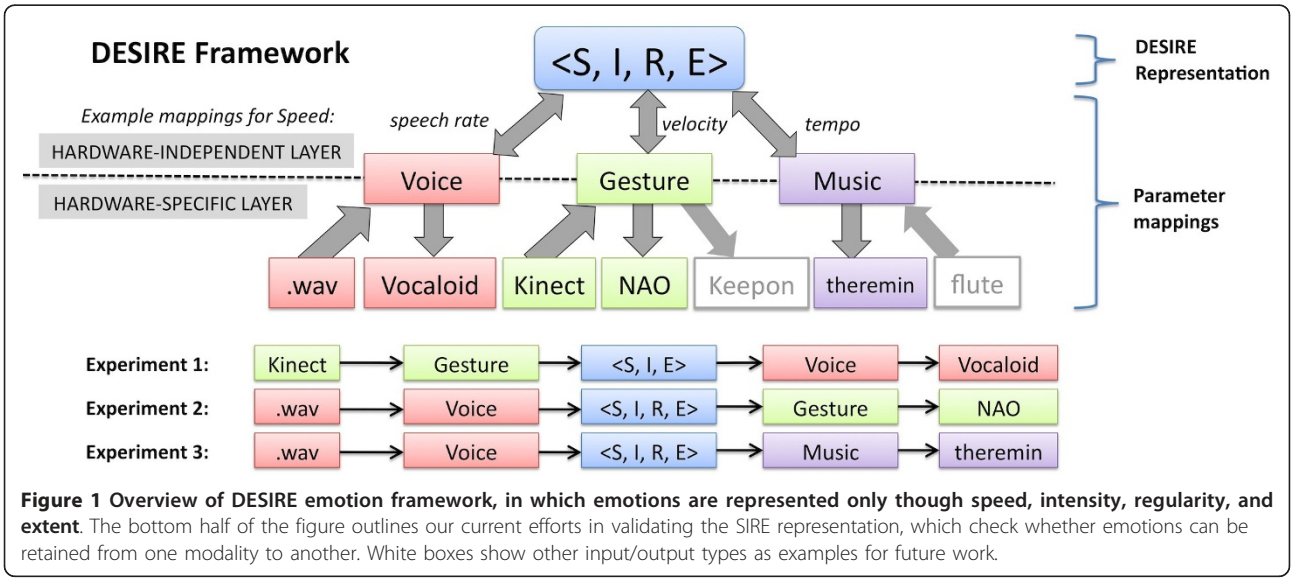


Figure 1 Overview of DESIRE emotion framework, in which emotions are represented only though speed, intensity, regularity, and extent. The bottom half of the figure outlines our current efforts in validating the SIRE representation, which check whether emotions can be retained from one modality to another. White boxes show other input/output types as examples for future work.

3. **Music**: NAO theremin-playing robot (based on [31]).

We use these systems to evaluate the effectiveness of the DESIRE in representing emotion, based on the requirements in Section 1.3.

3 Experiments

In this study, we use the DESIRE model to see whether portrayals of the same emotion in disparate contexts and modalities have the same underlying form. If they do, then we can use DESIRE to control multiple modalities from one source. Therefore, the research questions are as follows: (*Q*1) Does the same emotion in two different modalities have the same underlying form? (*Q*2) Can DESIRE capture that form? and (*Q*3) If so, what are the DESIRE values for each emotion? We first focus on four basic emotions: happiness, sadness, anger, and fear.

Our paradigm to answer these questions is to perform "cross-modality" emotion transfer. In each of three experiments, we *extract* a DESIRE from human portrayals of emotion, and use that DESIRE to *generate* robot portrayals in a different modality. Both the source and generated portrayals are then evaluated by human raters. If both the source and generated portrayals are rated as the same emotion, then we can say that DESIRE is sufficient to represent that emotion across the two modalities.

The experiments, as shown in Figure 1, are as follows:

- **Experiment 1: Gesture to Voice via SIE**. A pilot experiment using only 3 parameters of speed, intensity, extent (SIE) from human gesture to synthesized voice
- **Experiment 2: Voice to Gesture via SIRE**. Testing all 4 SIRE parameters from emotional voice to robot gesture
- **Experiment 3: Voice to Music via SIRE**. Testing all 4 SIRE parameters from emotional voice to theremin-playing robot

3.1 Experiment 1: gesture to voice via SIE

3.1.1 Method

We asked four naive students (3 male and 1 female) from Kyoto University to generate gestural portrayals of happiness, sadness, anger, and fear in front of a 3D sensor. Each emotion was to be acted for around 5-10 s and their anonymized gestures recorded with a standard video camera (as in Figure 2). The participants were not professional actors, but scenarios were provided to help elicit a desired emotion (e.g., "You have just won the lottery. Convey your happiness to the robot").

A Kinect and the OpenNI library were used to detect the position of the participants' hands in 3D (see Figure 2), and the maximum speed, acceleration and extent of the hands were extracted for each performance. Average speed was also informally tested, but did not produce any distinct difference between portrayals, perhaps because speed was diluted over time. Our program converted these values to SIE by linearly scaling them between 0 and 1 based on maximum speed, acceleration, and distance, respectively. The minimum and maximum values were experimentally set prior to the experiment. Future study should explore other mappings: for instance variance relative to a person's average amount of movement would better capture the idea of relatively slow or fast.

As output, the Vocaloid [32] synthesized male opera singer, Tonio was used. We chose a neutral utterance string: "I'm going to the store. Do you want anything?". The phrase was given the hand-made prosody as shown in Figure 3 to add naturalness. Then, the extracted SIE triples were given as input to the voice module as per Table 2. The vocal utterances were recorded as videos with the robot head and shoulders in the frame, as in Figure 4a.

The 16 human gesture videos and corresponding 16 robot voice videos were uploaded to the Internet in the form of an anonymous survey. Rating was performed in a forced-choice manner; according to [33], forced-choice judgments give results similar to free-labeling judgments when evaluators attempt to decode the intended emotional expression. After watching a video of either a human or speaking robot, the evaluator was asked to select the emotion most conveyed, among happiness, anger, sadness, and fear. An average of 9 evaluations for each display of emotion, and 37 for each emotion class were collected for this pilot study.

3.1.2 Results and discussion

The results of the emotion evaluations can be visualized in the confusion matrices in Figure 5. The visualized confusion matrix here can be thought of as a distribution of perceived emotion for a given portrayal class. The intended emotion is shown in the titles, and the average percentage of raters that selected each emotion are given along the dimensional axes. For instance, Figure 5-1a shows that human portrayals of happiness through gesture were recognized on average by raters as happiness by 53% of raters, anger by 22%, sadness by 13% and fear by 13%. In this way, the graphs can also be interpreted as componential representations of emotion portrayals.

We look in particular for similar distribution shapes in each column–this would indicate a similar perceived emotion for both input gesture and output voice. For instance, the voice generated using the same SIE values as Figure 5-1a was rated as happiness by 58% of raters, anger by 20%, sadness by 16%, and fear by 7%, as shown

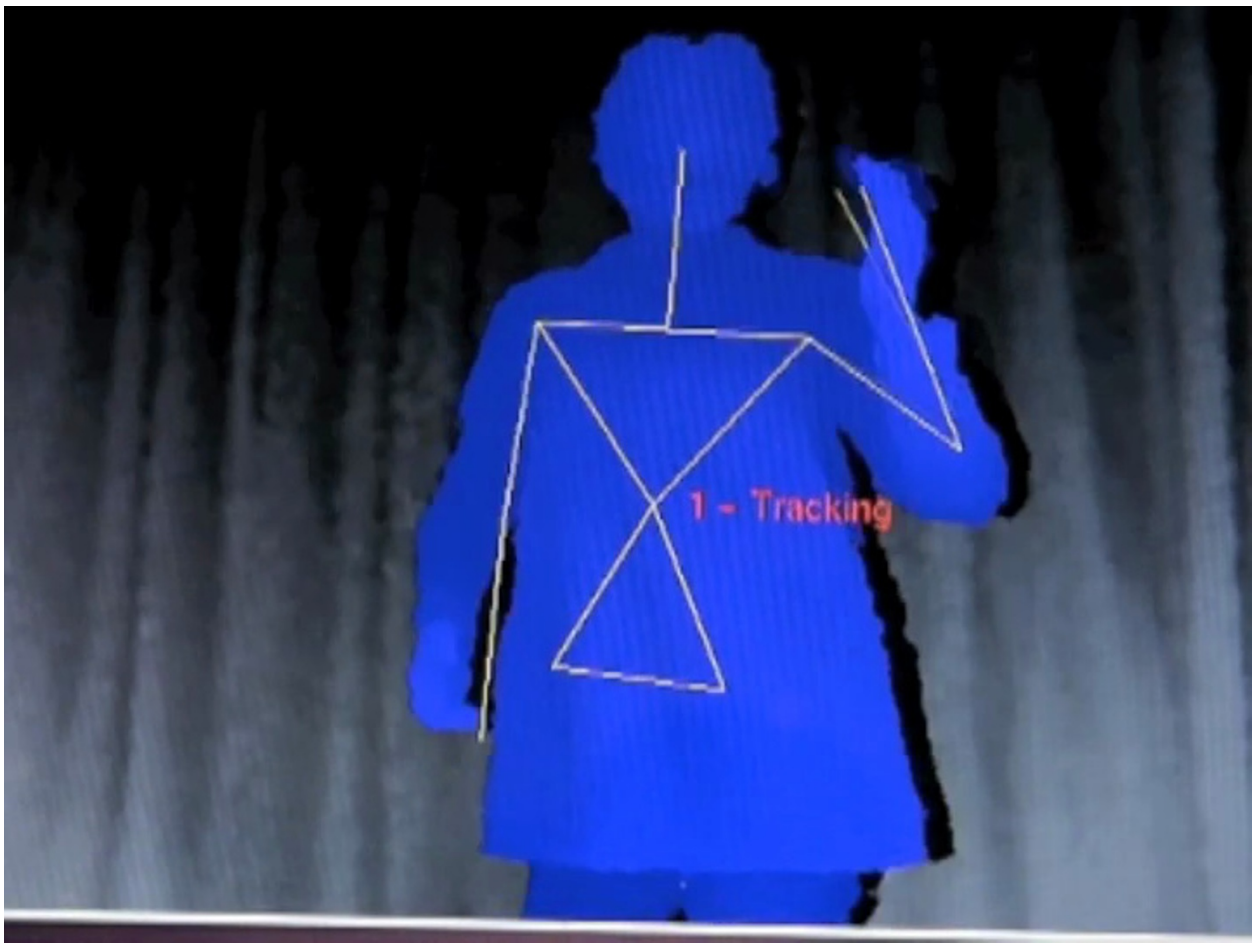


Figure 2 Body pose estimation using the Kinect 3D sensor to extract hand locations.

in Figure 5-1b. This large overlap, plus the result significantly over chance (25%) suggests that SIE indeed was sufficient for transferring happiness from gesture to voice.

The reasons why the acted emotions were not perfectly recognized may be better interpreted upon analyzing the gestural portrayals in a qualitative manner.

Fear was extremely well-recognized for human gesture portrayals, but not for transferred voice. Gestural portrayals included pulling back in fear sporadically, arms clutched to chest or to their sides. Two possibilities are possible: pose may have had a great effect on the understood emotion, which could not be transferred to the voice, or the SIE parameters are not sufficient for transferring this emotion.

Anger was portrayed by two participants in a prototypical manner–balled fists, gestures as if hitting a table, and approaching the camera. Interestingly, these were sometimes confused with sadness, presumably looking similar to outbursts of grief. On the other hand, one participant shook a finger at the camera, recognized by 100% of raters as anger. A stern, "cold anger" pose was also well-recognized.

Sadness portrayals contained prototypical hunched shoulders, hanging arms, and relatively low energy for two participants. Two of the more dramatic participants

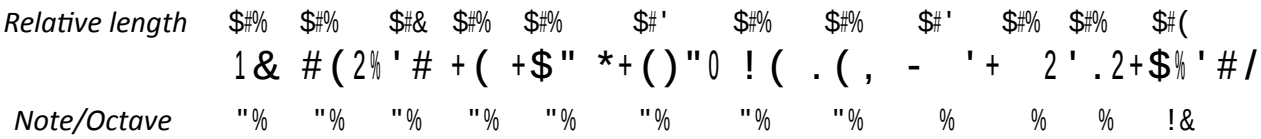


Figure 3 Prosody for the utterance.

Table 2 Experiment 1 parameter mappings

Gesture mapping	Parameter	Voice mapping
Hand Velocity	Speed	Tempo
Hand Acceleration	Intensity	Attack (onset delay)
Inter-hand Distance	Extent	Volume

showed sadness with both hands to head, and bending at the waist in pain.

Happiness was an interesting case, as only one participant demonstrated prototypical "jumping for joy". Another participant danced, another one made gestures to the sky in thankfulness, and the last shook her arms to the side in excitement while running in place. Interestingly, the dancing portrayal was the most well recognized by raters, not "jumping for joy", which was sometimes mistaken for anger. The gestures toward the sky were often perceived as grief or sadness.

This discussion, along with the results from Figure 5, allows us to draw three observations:

1. Happiness, sadness, and anger were transferred at greater than chance, despite the varied gestural interpretations for each emotion.
2. Fear was not well transferred. The irregular, sporadic backwards movements in fear portrayals could not be captured solely through speed, intensity, and range, which is one reason why we add the regularity parameter to Experiments 2 and 3.
3. The impoverished source gestures are not perfectly recognized, underlying the importance of multimodal redundancy. In addition, this suggests that studies should not aim at perceiving one "correct" transferred emotion at high rates, but also focus on the distribution of recognition, as in Figure 5. For instance, if a gesture is rated as 50% angry looking and 50% happy, the vocal output should also be 50% angry and 50% happy.

Finally, we briefly discuss latency issues. The Vocaloid system required around one second to generate high quality voice synthesis used in experiments, or around 200 ms using a lower quality real-time mode. Although not a problem for our manually segmented, out-of-context evaluations, this is an important consideration for any real-time interactive system: a pause between a stimulus and an emotion portrayal could potentially indicate a negative emotion such as sadness or fear.

3.2 Experiment 2: voice to gesture via SIRE

3.2.1 Method

We recruited 20 normal-sighted evaluators from Kyoto University Graduate School of Informatics. The participants were males of Japanese nationality, ranging in age from 21-61 (mean = 27.1, SD = 8.9).

As input, we used 16 audio samples taken from the Berlin Database of Emotional Speech[a], which is a database of emotional speech recorded by professional German actors. Each sample was a normalized wave file at 16 kHz, 1.5 to 3.9 s long, all of the same sentence. Four samples each of happiness, sadness, fear, and anger were used, all with recognition rates of 80% or higher by German evaluators.

Given SIRE values extracted from these audio samples as per Table 3, we generated 16 movement sequences using a simulated NAO shown on a projected screen. A full description of implementation can be found in [30]. Only one type of gesture was shown (an extension of both arms in front of the robot), repeated four times in series for each sequence. After each sequence, the participants chose one of happiness, sadness, anger, or fear in a forced-choice questionnaire.

3.2.2 Results

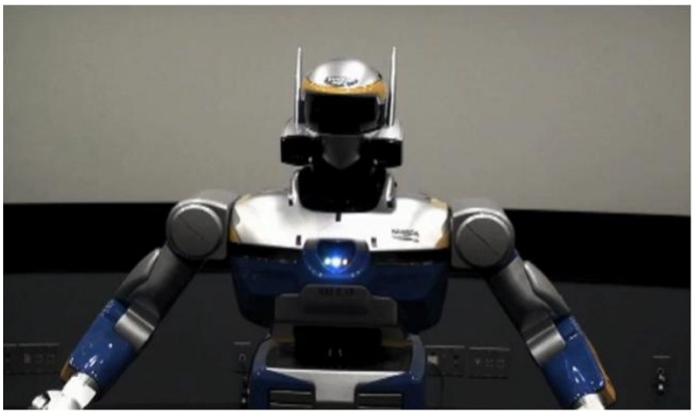
Figure 6 shows the confusion matrices for emotional voice and gesture. Ratings of the German voices is taken from the result of a stationary, speaking robot outlined in [30]. We find that the recognition rates for all emotions are significantly greater than chance (25%), suggesting that the DESIRE framework indeed converts the source vocal emotion to the same emotion in gesture. On the other hand, we can see that happiness (Figure 6-1b) was not clearly distinguished from anger. Further study in [30] suggested interaction with a pose cue: the immobile head of the robot. When compared with portrayals with a moving robot head, the staring, forward-facing head of the robot was significantly rated more often as anger.

3.3 Experiment 3: voice to music via SIRE

3.3.1 Method

Thirty-four participants were recruited over the Internet without respect to cultural or musical background. Average age was 33.2, SD = 12.2. Eight speech files (2 for each emotion) from the set of those used in Experiment 2 were used as input. Self-reported musical experience indicated that 35% of raters had no musical experience, 38% were beginner level, 21% intermediate level, and 6% expert.

The output was generated by the NAO robot playing the theremin with the parameter mappings as shown in Table 4. The robot's right arm was set to control the pitch at 415 Hz. To avoid bias based on song mode (e.g., major or minor), the robot played a simple sequence of quarter notes at the same pitch. This is a standard evaluation method used also in [24]. The left arm of the robot controlled the note volume, which started, shaped and ended the notes. The sounds of the theremin were recorded as sound files and uploaded to the internet in



(a) HRP-2 singing robot

(b) NAO gesturing robot

(c) NAO thereminist

Figure 4 Robot platforms for voice, gesture and music experiments.

the form of anonymous survey. Raters calibrated their headphones or speakers so that they could hear the loudest and quietest samples comfortably, then rated the sounds produced by the NAO thereminist in a forced-choice response.

3.3.2 Results and discussion

The results of the music experiment are shown in Figure 7 in the usual confusion matrix visualization format. We can see that the effectiveness of DESIRE using the theremin modality is limited compared to speech and

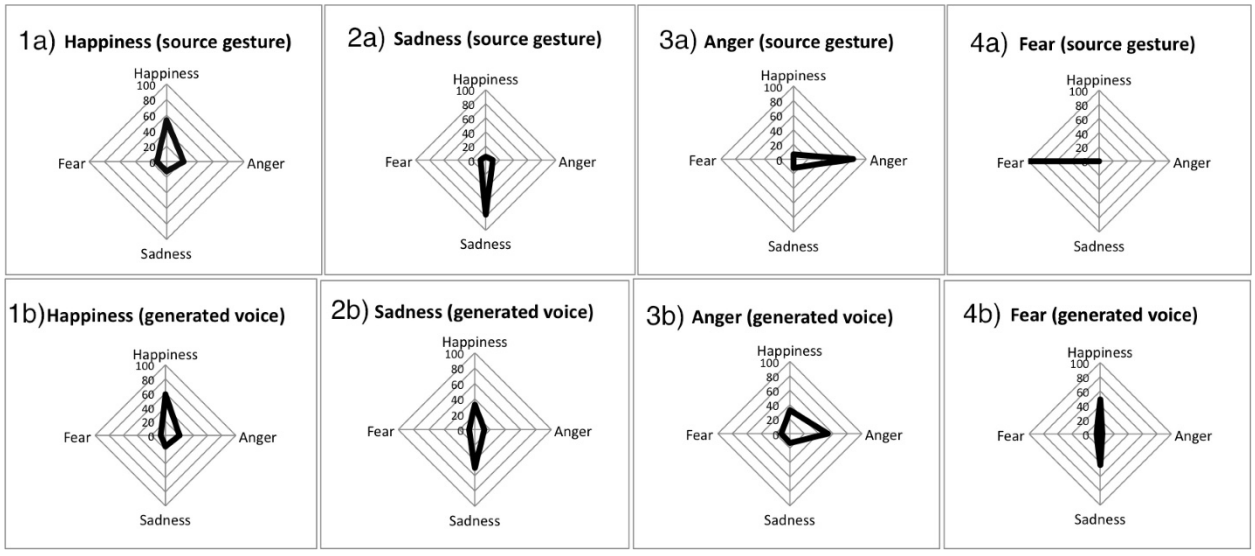


Figure 5 Experiment 1: Visualization of confusion matrices for gesture and voice. Intended emotion is shown in the titles, and the average percentage of raters that selected each emotion are given along the dimensional axes. Pointed triangles indicate that the one emotion was greatly perceived on average. Similar shapes for a given number indicate similar perceived emotion for both input gesture and output voice.

gesture. In particular, happiness and anger could not be reliably expressed. One reason for this may be the theremin sound itself. The theremin is often used for science fiction or horror films due to its eerie timbre, or for romantic, wistful songs such as Rachmaninoff's Vocalise. We find that overall, the evaluations of this modality were skewed towards 34% sadness and 32% fear, whereas only 16% and 19% of all portrayals were perceived as happiness or anger, respectively.

Another reason may be the maximum speed of the theremin–unlike instruments such as piano or flute, the theremin cannot change notes quickly without the sounds becoming indistinct. The results are thus inconclusive as to whether the emotions were maintained across voice to music, as the modality itself may have an overwhelmingly large influence. On the other hand, if our main purpose is a music robot that can play slow, emotion-laden music as opposed to a virtuoso performance, we find that there are SIRE parameters which 62% of raters recognize as sadness: S = 0.12, I = 0.44, R = 0.72, and E = 0.42. In addition, the SIRE parameters of S = 0.95, I = 1.0, R = 0.13, E = 0.37 produced a performance recognized as fear by 53% of evaluators. In experiment 2, these same SIRE parameters produced gestures that were recognized as sadness at 76% and fear at 65%. These results, coupled with the fact that the source of these parameters were sad and fear voices, suggest that emotions can be captured through SIRE across three modalities. Further experiments with a more versatile musical instrument such as piano are needed to confirm the effectiveness for happiness and anger.

Use of the theremin robot also highlighted the difficulties of robot platforms over computer-based music generation. For instance, for portrayals where intensity was high, the robot swayed to absorb the impact of a fast attack. Though not easily perceptible by ear, post-experiment analysis of the theremin sound showed a slight difference in the sound's spectral envelope. In physical robot gesture experiments in [30], the movement of motors induced sounds that some raters commented sounded "sad". Strong movements also caused residual vibration, which could look like a shivering or fearful effect. Additionally, the maximum speed of the robot motors limit the fastest movements it can make; the maximum arm speed of the robot was limited such

Table 3 Experiment 2 parameter mappings

Voice mapping	Parameter	Gesture mapping
Syllable rate	Speed	Arm velocity
Voice onset rapidity	Intensity	Arm acceleration
Jitter	Regularity	Inter-arm phase shift
Pitch range	Extent	Gesture extent

Table 4 Experiment 3 parameter mappings

Voice mapping	Parameter	Music mapping
Syllable rate	Speed	Tempo
Voice onset rapidity	Intensity	Note onset rapidity
Jitter	Regularity	Note timing offset
Pitch range	Extent	Maximum volume

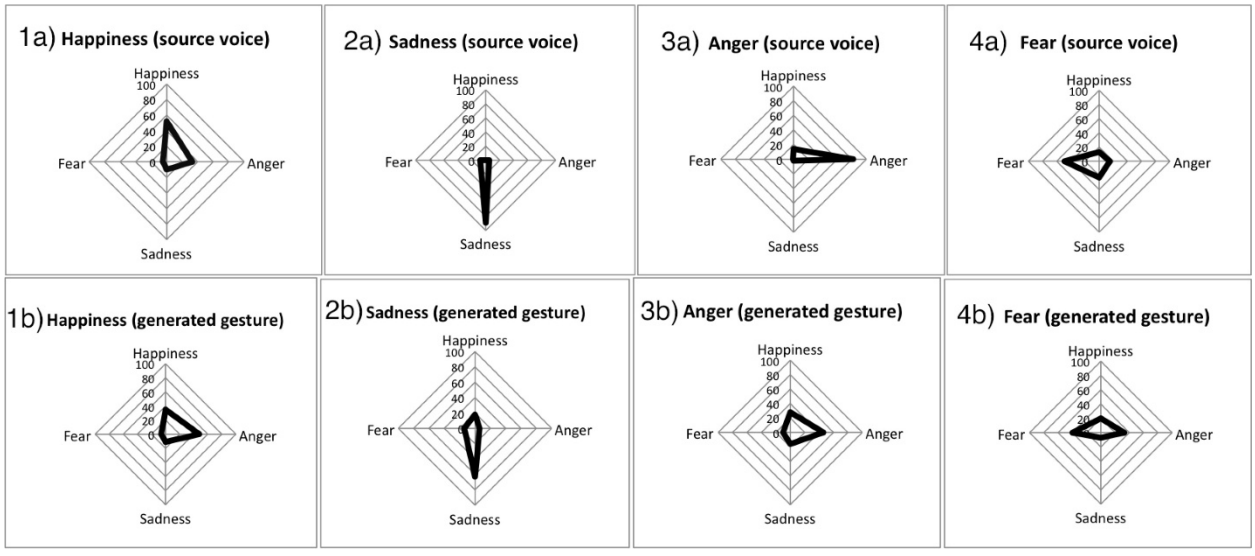


Figure 6 Experiment 2: Visualization of confusion matrices for voice and gesture. Similar shapes for a column indicate similar perceived emotion for both input voice and output gesture.

that it would not fall over after an intense gesture. Although all efforts were made to reduce the effects of using a real robot on the perceptual experiments, we outline them here as precautions for future experiments.

4 Outlook and conclusions

The results of our three experiments give a promising outlook for the DESIRE framework for both analyzing and generating recognizable emotions. We can respond to the research questions stated in Section 3.

- *(Q1) Does the same emotion in two different modalities have the same underlying form?* Yes, our results suggest that voice and motion have the same underlying form for happiness, sadness, anger, and fear. Voice and music have a similar form for at least sadness and fear with our theremin experiments.
- *(Q2) Can DESIRE capture that form?* Yes, but not to the same extent for every portrayal; some values are more cross-modally recognized than others.

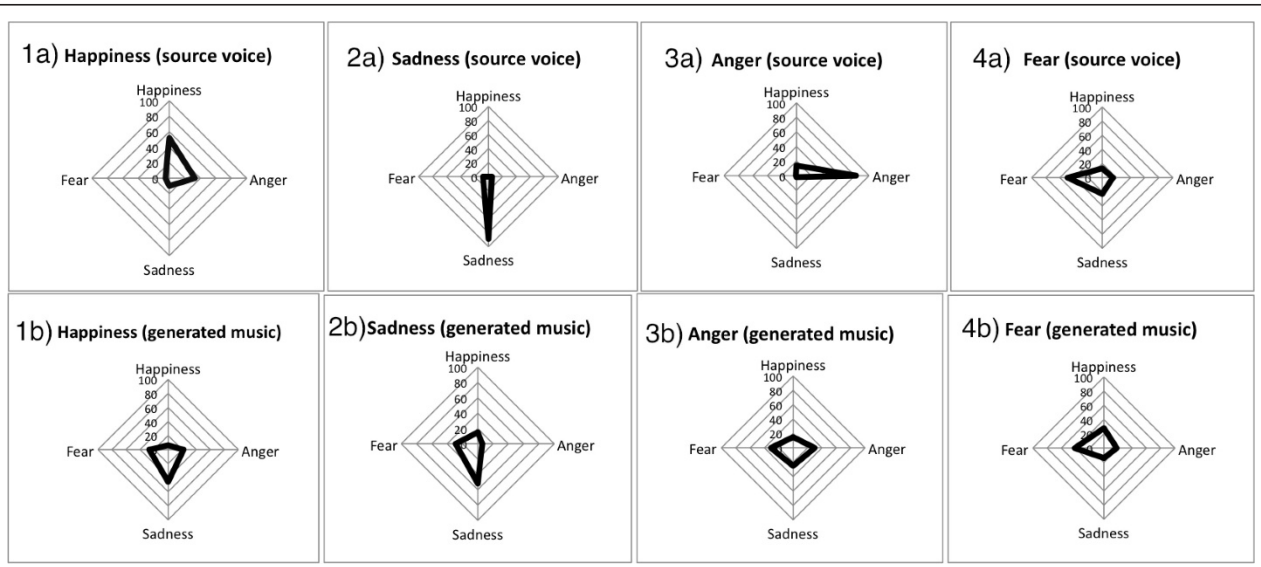


Figure 7 Experiment 3: Visualization of confusion matrices for voice and music. Similar shapes for a column indicate similar perceived emotion for both input voice and output music.

- *(Q3) If so, what are the DESIRE values for each emotion?* Table 5 gives the SIRE values that, at rates better than chance, underlie voice and gesture portrayals of happiness, sadness, anger, and fear.

4.1 Applications and extensions to other modalities

Other modality mappings are certainly a source for future study. For instance, a robot-controlled paintbrush could be mapped to SIRE: a slow versus fast stroke rate, a gradual versus abrupt start, a jagged versus straight stroke, and a small versus large diameter. A virtual fish could be given "emotions" by swimming faster or slower, darting across abruptly or accelerating gradually, making a direct line across a pool versus an irregular path, and by occupying small versus large space. Whether these mappings along with the DESIRE framework produce emotional expressions are open questions.

Perhaps the most obvious source for emotional communication is facial expression. Although facial expressions differ from speech and music in that it can be captured in one still-frame, it still may be possible to map the SIRE dimensions to facial features. For instance, the smile could be mapped to these parameters: **S** could map to slow versus fast smiles, **I** to a gradual versus abrupt smile, **R** to a symmetric or irregular smile, and **E** to a small versus large smile. The difference between a slow versus gradual smile is not clear here though, which is why it is difficult to claim that DESIRE would be useful for facial expression without further investigation.

Finally, the most useful extension would be to a music robot that can move while playing its instrument, such as a piano playing robot, violinist robot, or singing robot. In these cases, instead of mapping the movements to arm gestures, SIRE could be mapped to torso or head motions of varying speed, intensity, regularity, and extent. It should be noted that the theremin-playing robot is not suited to making expressive motion while playing, as any movement in the theremin's electromagnetic space it would affect its sound.

4.2 Parameter mapping design

We briefly discuss how DESIRE mappings can be extended for other modalities. The general principle is to design each dimension such that the majority of raters would perceive S = 0 portrayals as "slower" than S = 1, I = 0 portrayals as less "intense" than I = 1, and so on. In [30], we performed this experiment for gesture: we held each dimension constant and modulated only one parameter. The relative correctness for each parameter was confirmed, although ideal absolute settings remain to be explored. In our case, we designed the maximum mappings such that the robot would still operate safely. The minimum settings were set informally by the experimenter such that, for example, slow speed would be just slightly higher than motionless, but should be automatically set in future work. Indeed, this kind of context-dependent adaptation illustrates the challenge that situated robotics opens up for research, as described in Section 1.2, Aspect #2.

4.3 Limitations of this work

In this study, we have followed the conventional procedure of asking actors to generate expressive displays of emotion (such as those in the German emotional database). Although a standard research tool in emotion studies [4,34], it has a well-known flaw: these portrayals are unrealistic because they induce extreme–rather than natural–portrayals of each emotion. On the other hand, this scheme is beneficial to consistently evaluate the DESIRE model across multiple modalities. Music is fundamentally an on-demand performance modality, similar to dance; therefore, we choose to evaluate with acted portrayals as a first step.

The use of DESIRE for emotion is promising, but far from complete. The reader should note that this study is broad and exploratory in nature, to give intuition into the effect of DESIRE parameters in emotion across modalities. Further experiments involving more subjects are necessary to give more support to this model. Future study will also require realistic, continuous data to evaluate whether DESIRE can handle emotion trajectories as hypothesized, and whether other emotions such as love, pride, or disgust can be generated. In addition, the influence of other cues–such as instrument timbre or body pose–has been observed to be important, but how they interact and the extent of their effects still need to be formally explored. Further, the results were processed offline, nullifying the need for real-time extraction–how to determine the ideal frame size to extract an

Table 5 Gestural sequences with agreement among evaluators and their corresponding SIRE values

Emotion	Human voice (%)	Robot gesture (%)	Robot music (%)	S	I	R	E
Happiness	43	62	6	0.72	0.2	0.22	0.73
Sadness	95	76	76	0.12	0.44	0.72	0.42
Anger	95	86	27	0.71	0.46	0.04	0.73
Fear	33	43	53	0.95	1	0.13	0.37

Low scores for happiness and anger in music may be explained by the difficulty of the musical instrument (theremin) to express these emotions in general.

emotional portrayal is an on-going problem in emotion research in general [25].

Endnotes

[a]http://pascal.kgw.tu-berlin.de/emodb

Competing interests

The authors declare that they have no competing interests.

Received: 16 April 2011 Accepted: 17 January 2012
Published: 17 January 2012

References

1. G Weinberg, A Raman, T Mallikarjuna, Interactive jamming with Shimon: a social robotic musician, in *Proceedings of HRI*, San Diego, 233–234 (2009)
2. E Singer, J Feddersen, C Redmon, B Bowen, LEMUR's Musical Robots, in *Proceedings of NIME*, Hamamatsu, 181–184 (2004)
3. J Solis, K Chida, K Taniguchi, SM Hashimoto, K Suefuji, A Takanishi, The Waseda flutist robot WF-4RII in comparison with a professional flutist. Comput Music J. **30**(4), 12–27 (2006). doi:10.1162/comj.2006.30.4.12
4. M Clynes, *Sentics: The Touch of the Emotions* (Prism Press, UK, 1989)
5. P Laukka, Instrumental music teachers' views on expressivity: a report from music conservatoires. Music Educ Res. **6**(1), 45–56 (2004). doi:10.1080/1461380032000182821
6. PN Juslin, J Sloboda, *Handbook of Music and Emotion*, 1st edn. (Oxford University Press, USA, 2010)
7. E Lindstrom, PN Juslin, R Bresin, A Williamon, Expressivity comes from within your soul: a questionnaire study of music students' perspectives on expressivity. Res Stud Music Educ. **20**(1), 23–47 (2003)
8. N Silvestrini, V Piguet, C Cedraschi, MR Zentner, Music and auditory distraction reduce pain: emotional or attentional effects? Music Med. **3**, 264–270 (2011). doi:10.1177/1943862111414433
9. T Ogata, A Shimura, K Shibuya, S Sugano, A violin playing algorithm considering the change of phrase impression, in *Proceedings of SMC*, vol. 2. Nashville, 1342–1347 (2000)
10. J Solis, K Suefuji, K Taniguchi, T Ninomiya, M Maeda, A Takanishi, Implementation of expressive performance rules on the WF-4RIII by modeling a professional flutist performance using NN, in *Proceedings of ICRA*, Rome, 2552–2557 (2007)
11. T Nakano, M Goto, VocaListener: a singing-to-singing synthesis system based on iterative parameter estimation, in *Proceedings of SMC*, Porto, 343–348 (2009)
12. A Lim, T Mizumoto, T Takahashi, T Ogata, HG Okuno, Programming by playing and approaches for expressive robot performances, in *IROS Workshop on Robots and Musical Expressions*, Taipei (2010)
13. Y Kusuda, Toyota's violin-playing robot. Ind Robot: Int J. **35**(6), 504–506 (2008). doi:10.1108/01439910810909493
14. A Kirke, ER Miranda, A survey of computer systems for expressive music performance. ACM Comput Surv. **42**(1), 1–41 (2009)
15. WF Thompson, P Graham, FA Russo, Seeing music performance: visual influences on perception and experience. Semiotica. **156**(1), 203–227 (2005)
16. M Schutz, M Kubovy, Seeing music: The percussionists' strategic use of gesture in live performance, in *Proceedings of Music and Gesture*, Manchester, 148–150 (2006)
17. M Zecca, Y Mizoguchi, K Endo, F Iida, Y Kawabata, Whole body Emotion Expressions for KOBIAN Humanoid Robot–preliminary experiments with different emotional patterns, in *Proceedings of ROMAN*, Toyama, 381–386 (2009)
18. PN Juslin, P Laukka, Communication of emotions in vocal expression and music performance: different channels, same code? Psychol Bull. **129**(5), 770–814 (2003)
19. CL Breazeal, *Designing sociable robots*, 1st edn. (The MIT Press, Cambridge, MA, 2004)
20. R Cowie, E Douglas-Cowie, N Tsapatsoulis, G Votsis, S Kollias, W Fellenz, JG Taylor, Emotion Recognition in Human Computer Interaction. IEEE Signal Process Mag. **22**(1), 33–80 (2005)
21. R Fernandez, RW Picard, Classical and novel discriminant features for affect recognition from speech, in *Proceedings of Interspeech*, Lisbon, 4–8 (2005)
22. Z Zeng, M Pantic, A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell. **31**(1), 39–58 (2009)
23. JM Fellous, From human emotions to robot emotions, in *AAAI Spring Symposium on Architectures for Modeling Emotion: Cross-Disciplinary Foundations*, Palo Alto, 39–46 (2004)
24. L Mion, G De Poli, Score-independent audio features for description of music expression. IEEE Trans Audio Speech Lang Process. **16**(2), 458–466 (2008)
25. H Gunes, B Schuller, M Pantic, R Cowie, Emotion representation, analysis and synthesis in continuous space: a survey, in *Proceedings of FG*, Santa Barbara, 827–834 (2011)
26. SR Livingstone, AR Brown, R Muhlberger, WF Thompson, Modifying score and performance changing musical emotion: a computational rule system for modifying score and performance. Comput Music J. **34**(1), 41–65 (2010). doi:10.1162/comj.2010.34.1.41
27. C Pelachaud, Studies on gesture expressivity for a virtual agent. Speech Commun. **51**(7), 630–639 (2009). doi:10.1016/j.specom.2008.04.009
28. A Camurri, G Volpe, Communicating expressiveness and affect in multimodal interactive systems. Multimedia. **12**(1), 43–53 (2005). doi:10.1109/MMUL.2005.2
29. A Lim, T Mizumoto, LK Cahier, T Otsuka, T Takahashi, K Komatani, T Ogata, HG Okuno, Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist, in *Proceedings of IROS*, Taipei, 1964–1969 (2010)
30. A Lim, T Ogata, HG Okuno, Converting emotional voice to motion for robot telepresence, in *Humanoids*, Bled, 472–479 (2011)
31. T Mizumoto, H Tsujino, T Takahashi, T Ogata, HG Okuno, Thereminist robot: development of a robot theremin player with feedforward and feedback arm control based on a Theremin's pitch model, in *Proceedings of ICRA*, Kobe, 2297–2302 (2009)
32. H Kenmochi, H Ohshita, Vocaloid-commercial singing synthesizer based on sample concatenation, in *Proceedings of Interspeech*, Antwerp, 4011–4010 (2007)
33. PN Juslin, Can results from studies of perceived expression in musical performances be generalized across response formats?. Psychomusicology **16**, 77–101 (1999)
34. L Kessous, G Castellano, G Caridakis, Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. J Multimodal User Interfaces **3**(1–2), 33–48 (2009)
35. M Mancini, G Castellano, Real-time analysis and synthesis of emotional gesture expressivity, in *Proc of the Doctoral Consortium of Intl Conf on Affective Computing and Intelligent Interaction*, Lisbon (2007)
36. PE Gallaher, Individual differences in nonverbal behavior: dimensions of style. J Personal Soc Psychol. **63**(1), 133–145 (1992)
37. HG Wallbott, Bodily expression of emotion. Europ J Soc Psychol. **28**(6), 879–896 (1998). doi:10.1002/(SICI)1099-0992(1998110)28:63.0.CO;2-W
38. K Amaya, A Bruderlin, T Calvert, Emotion from motion. Graph Interface **1999** 222–229 (1996).
39. FE Pollick, HM Paterson, A Bruderlin, AJ Sanford, Perceiving affect from arm movement. J Personal. **82**, 51–61 (2001)

doi:10.1186/1687-4722-2012-3
Cite this article as: Lim *et al.*: **Towards expressive musical robots: a cross-modal framework for emotional gesture, voice and music.** *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:3.