2012 12th IEEE-RAS International Conference on Humanoid Robots
Nov.29-Dec.1, 2012. Business Innovation Center Osaka, Japan

Affective Gesturing with Music Mood Recognition

David K. Grunberg, Alyssa M. Batula, Erik M. Schmidt, and Youngmoo E. Kim

***Abstract*— The recognition of emotions and the generation of appropriate responses is a key component for facilitating more natural human-robot interaction. Music, often called the "language of emotions," is a particularly useful medium for investigating questions involving the expression of emotion. Likewise, movements and gestures, such as dance, can also communicate specific emotions to human observers. We apply an efficient, causal technique for estimating the emotions (mood) from music audio to enable a humanoid to perform gestures reflecting the musical mood. We implement this system using Hubo, an adult-sized humanoid that has been used in several applications of musical robotics. Our preliminary experiments indicate that the system is able to produce dance-like gestures that are judged by human observers to match the perceived emotion of the music.**

I. INTRODUCTION

Humanoid robots are often versatile machines, able to perform tasks in a variety of ways. For example, a humanoid might be able to walk at varying speeds, or with different arm motions. While it may be possible for human users to specify exactly how they would like a particular task to be done, it would be more convenient and efficient if the robot could determine this automatically. Humans naturally use emotional cues to convey some of this information to each other: a 'frantic' mood, for instance, may imply a need for rapid action, while a 'calm' mood could indicate that more time and care could be taken. If robots could understand this information, they could incorporate it into their tasks without requiring a human user to explicitly designate a mood. This would result in simpler and more intuitive control of robots. Similarly, if robots could display such emotions, they could communicate them to humans or each other, again allowing for simpler and more flexible communication than explicitly declaring a certain mood.

Music, often termed the "language of emotions," is one particular area in which we would like robots to be able to identify and communicate moods [1].[1] We wish to enable robots to perform musical tasks, such as dancing or performing in musical ensembles, in ways that resemble human performances. In order for robots to react to music as humans would, they should determine the mood of the music, then incorporate that information into their performances. As different humans may identify the mood of a particular piece of music differently, the robot's mood detection/production system should be robust enough that a variety of people would agree that the emotional content of the robot's motions is similar (or *congruent*) to that of the music. This ability would make the robot more useful as a performance tool or platform for researching musical performance.

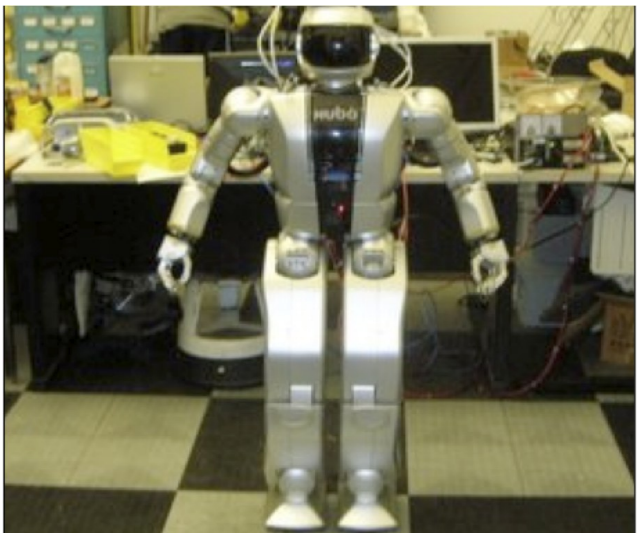

Fig. 1. One of the Drexel Hubos.

We are interested in using humanoids in particular as platforms for performing these musical tasks. Many musical instruments and dance styles are already designed for the human form, so humanoids are more likely than other types of robots to be physically capable of the desired tasks. Similarly, as humans communicate emotions with various gestures and styles of movement, a robot shaped like a human is likely to be able to perform similar gestures or motions to indicate the same moods. We therefore focus on the design of a system that can detect the mood of a segment of music, and then control a humanoid robot to produce a gestural response to convey a congruent emotion.

We have selected the Hubo as our robot platform (Fig. 1). Hubo is an adult-sized humanoid developed by the Korean Advanced Institute for Science and Technology (KAIST). We have used Hubo in several other tasks involving robotic reactions to audio, such as moving in synchrony with audio beats [2]. Hubo possesses over forty degrees of freedom and can perform smooth and graceful motions such as tai chi, so it is a suitable choice for a robot that will need to move in human-like ways and display emotion.

*This work is supported by NSF Awards DGE-0947936, OISE-0730206, CNS-0960061, and the Graduate Research Fellowship.

**The authors are with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, 19143 {dgrunberg, batulaa, eschmidt, ykim}@drexel.edu

[1]Though 'emotions' and 'mood' have slightly different connotations in some fields, these terms are used interchangeably in this paper

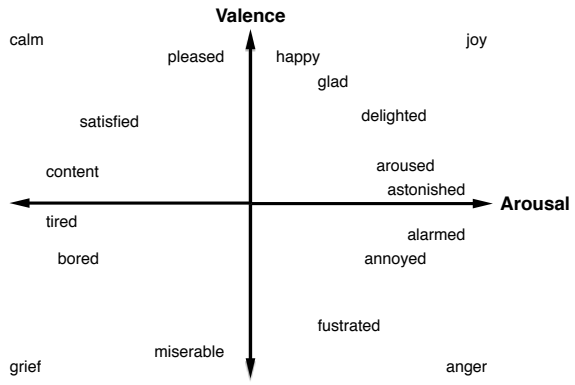

Fig. 2. An A-V plot with several emotions labeled.

II. RELATED WORK

In order to perform musical tasks, a system must often determine the locations of the the beats in the audio. This is particularly the case when operating on live performances, where the beats may not be known in advance. Efficient beat tracking algorithms have been developed by Scheirer and Davies [3], [4]. These trackers respectively calculate the subband energy envelopes or the complex spectral difference and then perform efficient operations (such as autocorrelations) to find periodicities in these features. The periodicities are used to find equally-spaced sets of beats. Other systems have been proposed that use more complex algorithms such as Linear Discriminant Analysis [5]. These systems, though, are often more computationally expensive and slower than the simpler versions.

We have made several developments in automatically determining the emotional content of audio. One important decision was that of representing musical emotion as the two scalar metrics of *arousal* and *valence* [6]. Arousal represents the intensity of an emotion, while valence indicates whether it is positive or negative (Fig. 2). We have also obtained annotations of the arousal and valence (A-V) values in audio from human users, as this ground-truth data is invaluable in training emotion recognition systems [7]. Finally, our group has developed several methods for automatic music emotion recognition. These include computationally efficient methods, such as calculating the spectral contrast and Mel-Frequency Cepstral Coefficients of audio and mapping those features into the A-V space [8]. We also developed and evaluated more elaborate methods, such as using Deep-Belief Networks to learn new features that are more correlated to A-V values [9].

Research by Camurri et al. indicates several important factors in dance motions that determine what mood they will convey [10]. For example, angry gestures tend to take less time than sad gestures, contain more tempo changes, and carry more dynamic tension. Lourens et al. enabled a robot to identify the emotional content of a user based on their gestures [11] by modeling the emotions with Laban notation, a type of notation used specifically to record dances [12]. Nakata independently verified that robots are capable of displaying emotion by following Laban principles [13].

Musically expressive humanoids have been developed by various groups. Asimo, a humanoid robot developed by Honda, is able to step in response to music [14]. Similarly, the humanoid robot HRP-2 can produce human-style dance gestures obtained via motion-capture technology [15]. Neither of these systems, however, consider mood. The RoboCup Junior competition includes a 'dancing' event, in which robots dance to a piece of music[2]. The music is known in advance and played from file, though, so musical features can be marked by hand instead of being extracted from the music. Shiratori studied synthesizing dance performances based on human perceptions of musical mood [16]. This system, though noncausal, calculated emotional psychometrics such as 'intensity' from both a piece of music and a human dancer's response to that music, and used those values to drive the robot.

III. BEAT, TEMPO, AND MOOD RECOGNITION

In order to make the robot platform as versatile as possible, it should be be able to retrieve various information or features from acoustic signals. We focus specifically on the music information retrieval (MIR) tasks of finding the tempo, beat locations, and A-V values in a piece of music. While a robot could, in certain restricted cases, be capable of knowing these features based on previous performances, this cannot always be assumed to be the case. Live performances, for instance, will always vary slightly from show to show (and may even feature new, improvised music), so a robot that could only use features derived from previous performances of a song would be unlikely to be as congruent with a new performance.

The MIR portion of the system only requires an acoustic waveform as input. The algorithms do not require metadata (such as a score) or a digital description of the audio (such as Musical Instrument Digital Interface, or MIDI, data). The robot can thus operate even when such information is not known to the system.

A. Tempo identification and beat tracking

In order for the robot's motions to be synchronized with the music, the robot must determine the tempo and beat locations. When humans dance, their motions are often spaced according to the tempo and apex on beats, and this information is needed for many musical tasks. Thus, we developed a fast and causal beat-tracker for our musical robots (Figure 3). This system is briefly described below; the interested reader is directed to our prior work [17].

An acoustic signal is divided into short-time frames and split into several subbands. The subband envelopes are smooth and rectified, then autocorrelated. A signal's autocorrelation has large values at lag values that are proportional to the period of the original signal, so by peak-picking from the autocorrelation, the system can obtain an estimate of the period (or tempo) of the original music.

[2]http://rcj.robocup.org/dance.html

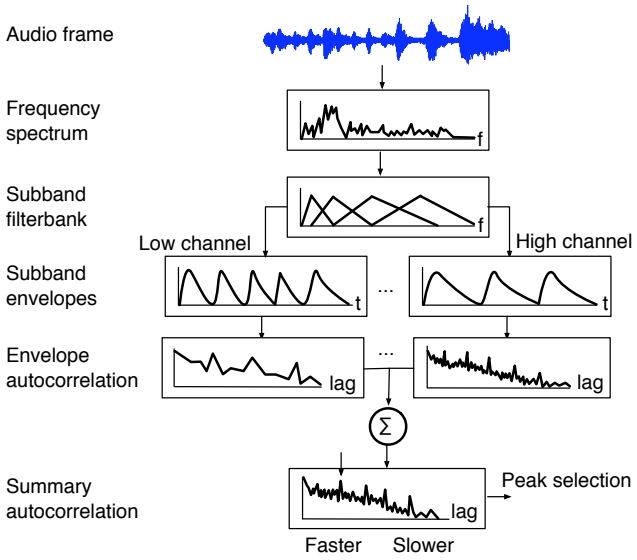

Fig. 3. Flowchart of the beat tracking algorithm.

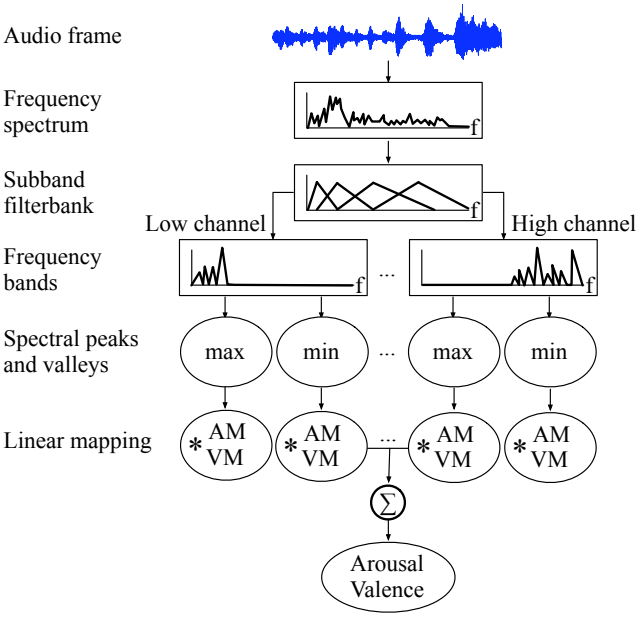

Fig. 4. Flowchart of the mood tracking algorithm. 'AM' and 'VM' represent the arousal and valence mapping values, respectively.

Once the tempo is known, the energy in each frame is calculated. The program looks for sequences of frames with high energy (relative to the surrounding frames) that are spaced according to the estimated tempo. Consistent sequences of such frames are marked as likely beat candidates.

We tested our system on twenty pop songs, spanning one hour in duration. We have obtained accuracy of .98 F-Measure for CD-quality audio, and .92 for audio contaminated by robot and room acoustic noise [17].

B. Mood Identification

Our lab previously developed a game called *MoodSwings* which allows users to rate the arousal and valence values of music clips [18]. We selected 240 song clips from the game and examined their acoustic features to identify correlations between the features and the A-V ratings. A feature called *spectral contrast* was found to be strongly correlated with both arousal and valence. Spectral contrast is a measure of the peaks and valleys in a frequency subband, and is calculated with Equations 1 and 2. [19].

$$V_s = \frac{1}{\lceil N_s \alpha \rceil} \sum_{i=1}^{\lceil N_s \alpha \rceil} F(x_i) \quad (1)$$

$$P_s = \frac{1}{\lceil N_s \alpha \rceil} \sum_{i=N_s}^{N_s - \lceil N_s \alpha \rceil} F(x_i) \quad (2)$$

F is the spectrum sorted from smallest to largest value, V and P are the valley and peak values for a subband s, N_s is the number of elements in the subband s, and α is a smoothing parameter, here set experimentally to .02.

This algorithm divides the audio into seven subbands and produces fourteen values, seven peaks and seven valleys, per frame. Values are aggregated over forty frames for one second's worth of data. While using ten seconds of data gives slightly improved results, this also decreases the system's ability to adjust if the audio changes [8]. The resulting values are then linearly mapped to A-V coordinates by multiplying the peak and valley values by a 14x2 matrix to result in a 2x1 A-V value vector. The 14x2 mapping matrix is calculated by using least-squares regression on the 240 clips. While some other methods did produce higher accuracy, such as using Conditional Random Fields, these methods were all either noncausal or too slow to be useful in a realtime application [18].

This system was also found to be accurate in our previous studies. One hour of music was analyzed by human experts to mark A-V values, and then passed into the system [7]. This average error of the system was less than 15% of the total space when taking arousal and valence together, and less than 12% treating them separately.

IV. AFFECTIVE GESTURE GENERATION

The MIR values are used to parameterize the gestures that the robot makes, thereby enabling it to respond appropriately to the music. These parameterizations are based on the research done by the dance community in determining how human dancers convey certain emotions by means of their body language [10], [11].

For this study, we have split the mood space into four overall areas, as dictated by the A-V map (Figure 2) [6]. Starting clockwise from the upper-right, these four areas represent the emotions of 'Joy,' 'Anger,' 'Grief', and 'Calm.' We then designed a basic gesture that could be easily parameterized to represent any of these four emotions. We restricted ourselves to one gesture parameterized four ways instead of different gestures for each mood in order to ensure that the reactions of test subjects would be

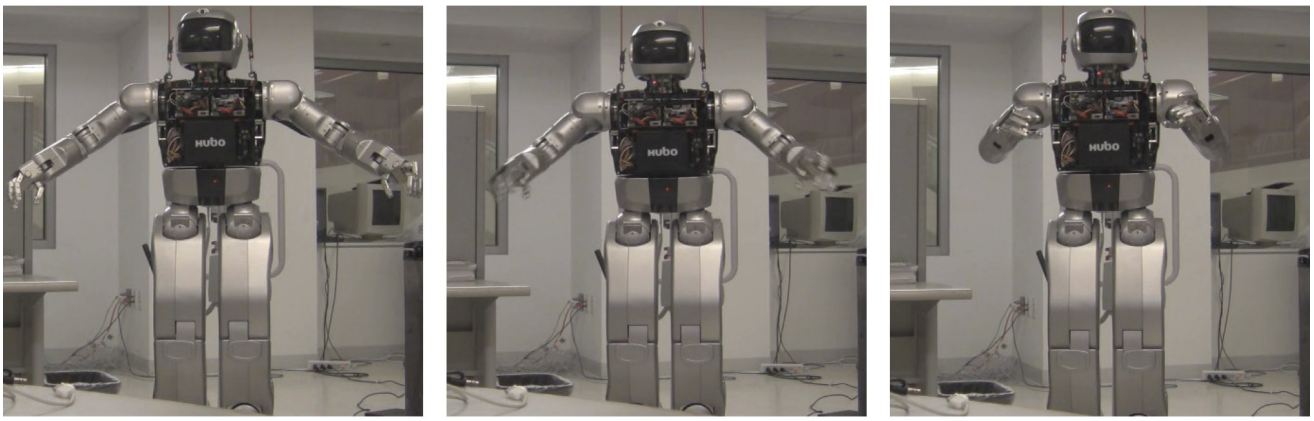

Fig. 5. Hubo moving through the gesture that indicates 'Joy'. Its arms start far apart and pointing down, and they are raised and brought closer to the body as the gesture continues. The result is a sweeping and expansive upwards motion.

TABLE I
DETAILED PARAMETERIZATION OF ROBOT MOTION FOR FOUR EMOTIONS.

	Joy	Calm	Grief	Anger
Motion start time	Immediately	Immediately	Immediately	Delayed ($\frac{1}{3}$ through beat)
Vertical arm motion	$-3"$ to $5"$	$-2"$ to $3"$	$-2"$ to $1"$	$-5"$ to $2"$
Horizontal arm motion	$18"$ to $5"$	$13"$ to $8"$	$11"$ to $5"$	$15"$ to $3"$
Head position	$15°$	$15°$	-15 °	-15 °

solely influenced by the parameterization, instead of the 'fundamental' nature of the gesture itself. In order to avoid destabilizing the robot, we excluded leg motion from the base gesture. The robot's arms and head moved as follows:

- Both arms begin extended away from the robot.
- The arms move horizontally inwards and upwards, towards the robot's head.
- As this is being done, the head tilts to the left or right.
- Once the arms reach their final position, they stop.
- On the next gesture, the arms and head reverse.

This gesture was selected due to its relative simplicity, which would allow a test subject to quickly see and analyze it, as well as its ease of parameterization. The parameter sets we made are as follows:

- Joy: the arms are raised high and extended far from the body (Figure 5). As a result, the gesture is wide, expansive, and relatively fast compared to the other emotions, since there is more space to cover in the same amount of time. The head is raised up as well.
- Calm: the arms are raised to a similar height as in the 'Joy' emotion, but are closer to the body. This results in a slower and more constrained motion. The head is at the same position as with the 'Joy' emotion.
- Grief: the arms are at a similar distance from the body as in the 'Calm' gesture, but are lowered to be closer to the ground. The head is lowered as as well. This gives the robot a downcast and somber appearance.
- Anger: the arms begin low and far from the robot's body. Unlike the other gestures, the robot only moves during two-thirds of the beat; the robot's motions are therefore more 'agitated' for this gesture.

The exact specifications of the motions were set with the above points as a guide, and then tuned by hand to fit the constraints of the robot (such as the maximum allowable speed). Table I more precisely specifies the parameterization of our gesture. Vertical arm motion is measured from the waist, and horizontal arm motion from the sides of the robot.

V. EXPERIMENTAL SETUP

Both the beat-tracking and the A-V value prediction are performed simultaneously in a single MIR system. The MIR information is then transmitted via Universal Datagram Packet (UDP) to the robot. UDP is used to minimize latency, as even a small delay between a motion and a beat location can look visibly incongruent to a human observer. While UDP does carry a risk of dropping packets, this was not found to be a problem during testing.

Twelve clips of audio were taken from our MoodSwings corpus. These clips were chosen to span the four A-V quadrants evenly, as indicated by the annotations. Using a cable, the audio was passed from a music player into the processing computer that calculated the A-V values and beat locations, and a Hubo robot then produced the parameterized gestures. Human experts listened to the audio and analyzed the resulting performance. The experts, members of the Music & Entertainment Technology Lab at Drexel University, are familiar with the A-V representation of music, having worked with MoodSwings and similar

TABLE II

ANALYSIS OF MUSIC AND ROBOT EMOTION, SYSTEM PREDICTING BEAT ONLY

Quadrant			Arousal: [-.5,.5]			Valence: [-.5,.5]			Congruence: [1,5]	
A	V	Emotion	Music	Robot	Error	Music	Robot	Error	Mood	Beat
+	+	Joy	.30±.12	.28±.13	.08±.10	.27±.18	.30±.14	.12±.15	3.70±1.17	3.85±0.82
+	-	Anger	.35±.17	.20±.14	.16±.13	-.34±.13	-.18±.18	.19±.17	3.22±0.85	2.74±0.81
-	+	Calm	-.21±.19	-.17±.17	.12±.10	.20±.19	.21±.15	.13±.13	3.30±0.91	2.81±0.83
-	-	Grief	-.25±.14	-.04±.20	.23±.21	-.04±.26	-.12±.20	.20±.20	2.93±0.96	3.11±0.93

TABLE III

ANALYSIS OF MUSIC AND ROBOT EMOTION, SYSTEM PREDICTING BEAT AND MOOD

Quadrant			Arousal: [-.5,.5]			Valence: [-.5,.5]			Congruence: [1,5]	
A	V	Emotion	Music	Robot	Error	Music	Robot	Error	Mood	Beat
+	+	Joy	.23±.15	.21±.14	.13±.08	.26±.11	.23±.12	.13±.11	3.22±0.55	2.78±0.94
+	-	Anger	.31±.19	.09±.29	.30±.24	-.22±.27	-.17±.19	.24±.19	2.67±0.91	3.00±0.84
-	+	Calm	-.14±.22	-.09±.23	.20±.21	.08±.22	.08±.23	.17±.18	3.11±0.96	3.06±1.06
-	-	Grief	-.22±.13	-.04±.24	.24±.19	-.07±.23	-.04±.20	.12±.10	3.00±0.97	3.00±1.08

games. Many of these experts are also experienced musicians.

The experts indicated the perceived A-V values of both the music and the robot's motions on a scale of -.5 to .5. Ratings were done with A-V values and not emotions to obtain more precise results, as the A-V values were quantized more finely than the set of four emotions. Additionally, the experts determined if the robot's motions seemed to match both the mood and the beat locations of the audio, on a 5 point scale. 3 was designated as 'average' congruence, so values of more than 3 indicated a better than average (or 'strong') matching between music and motion.

We first analyzed the motions alone (without mood prediction) by explicitly setting the appropriate mood for each clip based on the MoodSwings data. The purpose of this test was to verify that the parameterizations designed by our experts did indeed map to the specified moods, and that the robot was capable of demonstrating those moods with its motions. In a second test, we used the mood-prediction algorithm to parameterize the gestures. Each performance was played twice. If the experts thought that the two performances were dissimilar (indicating an odd fluke, such as packets dropping), a third performance was produced.

VI. RESULTS

The results from the first test, in which the ground-truth mood labels were used to parameterize the robot's motions, are displayed in Table II. Nine human users participated in this test. In the first few columns, the arousal values provided by nine experts for both the music and the robot motions are listed. Values are displayed as the mean ± the standard deviation. The next column is the distance between the arousal ratings for the music and motions (across all experts, and across all songs in the quadrant). The valence values are then displayed. The final two columns display the congruence results.

Our experiment shows that the robot can accurately indicate the desired emotion. The arousal and valence averages of the robot's motions always have the same signs as those of the music, indicating that they belong to the same quadrant. A random classifier, by contrast, would only achieve the correct values 25% of the time. Therefore, in a four-emotion system, the robot can move in such a way as to reliably indicate that emotion to human viewers, to a degree far greater than chance could provide.

The average error of the system is almost always less than .2. This means that, if the true arousal and valence of a piece of audio have a magnitude larger than .2, the system is likely to classify it correctly. We can thus pass a wide variety of music into the system and be confident that the correct emotion will be chosen, enabling it to function very flexibly. Additionally, the congruence values, particularly regarding mood, are generally larger than the 'average congruence' value of 3. The human experts therefore perceived better than average congruence between the motions and the audio. The only emotion with a mood congruence less than 3 is 'Grief', indicating that this parameterization may represent the true emotion less well than the other three. Nonetheless, the congruence values still demonstrate agreement between the robot's movement and its intended mood.

The results from the second test are shown in Table III. Six human experts were used for this part of the study. They reported that the robot motions and music were from the same mood quadrant for all quadrants, thus validating the system for a four-emotion classifier. Furthermore, the average error is below .25 in almost every case. Only the arousal error for the 'Anger' mood is greater than this value, implying that the spectral contrast feature may

be less adept at indicating this particular emotion. These results still indicate that we could pass a wide variety of songs into the system, and as long as the true arousal and valence values had magnitudes larger than .25, they would likely be classified correctly.

Comparing the two tests, error is smaller, and mood congruence larger, in the case where the mood is known in advance than when the MIR mood system is used. This is due to the mood identification system being imperfect, and its errors propagating through to the rest of the system. However, even when the system does not have ground-truth emotion values, the mood congruence is generally still above 3 (indicating above-average matching between the motions and the acoustic mood) and error remains relatively small.

Across both tests, error is consistently the smallest in the quadrant with positive arousal and valence. This may indicate that this quadrant is the easiest to represent with motions. When both arousal and valence are high, the corresponding motions should be large, excited, and obvious [11]. Other quadrants may require motions which are smaller and more subtle, and thus harder to analyze.

The congruence values for beat tracking are also near 3. Incongruent ratings are due to the robot not moving on every beat [17]. To prevent the robot from trying to be in two places at once, it disregards movement requests that arrive during existing motions, which leads to short periods where the robot stands still.

VII. Conclusion

We have demonstrated a system that can robustly allow a robot to move in a manner congruent to the mood of musical audio. Our results indicate that the system can reliably extract the mood from a song and produce the corresponding motions on a robot platform. This validates our algorithms for mood detection and communication.

We are interested in continuing this research in multiple directions. On one front, we will modify our motions to better demonstrate emotions. Incorporating more of the body into the movements could better convey certain moods, for example, as could using other movement parameters (such as velocity and acceleration). By making the motions more complex, we hope to make the system better able to convey emotional content.

We also seek to enhance our representation of the mood space, perhaps even making it continuous, to allow the robot to display more subtle emotions. This could involve enhancing our mood-prediction system, such as by using features found directly from magnitude spectra via machine learning [9]. We would also like to be able to 'tune' the mood system to represent different groups of people. People from two different countries might react differently to a piece of music, for instance, and it would be interesting if the robot could anticipate the responses of both groups.

References

[1] C. C. Pratt, *Music as the language of emotion*. The Library of Congress, December 1950.

[2] Y. E. Kim, A. M. Batula, D. Grunberg, D. M. Lofaro, J. Oh, and P. Y. Oh, "Developing humanoids for musical interaction," in *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2010.

[3] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[4] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009 –1020, march 2007.

[5] G. Peeters and H. Papadopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: Theory and large-scale evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1754 –1769, aug. 2011.

[6] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford, UK: Oxford University Press, 1989.

[7] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *ACM MIR*, Philadelphia, PA, 2010.

[8] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proceedings of the 2010 International Conference on Music Information Retrieval*, 2010.

[9] ——, "Learning emotion-based acoustic features with deep belief networks," in *Proceedings of the 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.

[10] A. Camurri, I. Lagerloef, and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International Journal of Human-Computer Studies*, vol. 59, no. 1Äì2, pp. 213 – 225, 2003.

[11] T. Lourens, R. van Berkel, and E. Barakova, "Communicating emotions and mental states to robots in a real time parallel framework using laban movement analysis," *Robotics and Autonomous Systems*, vol. 58, no. 12, pp. 1256 – 1265, 2010.

[12] R. Laban, *Principles of Dance and Movement Notation*. New York, USA: Macdonald & Evans, 1956.

[13] T. Nakata, T. Mori, and T. Sato, "Analysis of Impression of Robot Bodily Expression," *Journal of Robotics and Mechatronics*, vol. 14, no. 1, 2002.

[14] K. Yoshii *et al.*, "A biped robot that keeps steps in time with musical beats while listening to music with its own ears," in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA, October-November 2007.

[15] S. Nakaoka, A. Nakazawa, F. Kanehiro, K. Kaneko, M. Morisawa, H. Hirukawa, and K. Ikeuchi, "Learning from observation paradigm: Leg task models for enabling a biped humanoid robot to imitate human dances," *International Journal of Robotics Research*, vol. 26, no. 8, pp. 829–844, Aug. 2007.

[16] T. Shiratori and K. Ikeuchi, "Synthesis of Dance Performance Based on Analyses of Human Motion and Music," *IPSJ Online Transactions*, vol. 1, pp. 80–93, 2008.

[17] D. K. Grunberg, D. M. Lofaro, P. Y. Oh, and Y. E. Kim, "Robot audition and beat identification in noisy environments," in *Proceedings of the International Conference on Intelligent Robots and Systems*, 2011.

[18] E. M. Schmidt and Y. E. Kim, "Modeling musical emotion dynamics with conditional random fields," in *ISMIR*, Miami, Florida, 2011.

[19] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, vol. 1, 2002, pp. 113 – 116 vol.1.