Synthetic Emotions for Humanoids:
Perceptual Effects of Size and Number of Robot Platforms

David K. Grunberg, Drexel University, USA

Alyssa M. Batula, Drexel University, USA

Erik M. Schmidt, Drexel University, USA

Youngmoo E. Kim, Drexel University, USA

ABSTRACT

The recognition and display of synthetic emotions in humanoid robots is a critical attribute for facilitating natural human-robot interaction. The authors utilize an efficient algorithm to estimate the mood in acoustic music, and then use the results of that algorithm to drive movement generation systems to provide motions for the robot that are suitable for the music. This system is evaluated on multiple sets of humanoid robots to determine if the choice of robot platform or number of robots influences the perceived emotional content of the motions. Their tests verify that the authors' system can accurately identify the emotional content of acoustic music and produce motions that convey a similar emotion to that in the audio. They also determine the perceptual effects of using different sized or different numbers of robots in the motion performances.

Keywords: *Beat Tracking, Humanoids, Mood Tracking, Music Information Retrieval, Robot Emotion, Synthetic Dance*

INTRODUCTION

We are seeking to enable humanoid robots to participate alongside humans in live musical performances. Such robots could be useful for a wide variety of tasks, both relating to research and to musical performance. They could be used, for example, to study the precise affect of certain parameters (such as force, momentum, or center of mass) on the human perception of ensemble musical performances. These platforms could also be used to help produce novel and interesting presentations, as they could play in different ways than humans. Certain robots might be able to play a series of notes more quickly than human performers, for example, and so could make certain compositions feasible to perform at faster speeds.

In order for the robots to be able to intelligently participate in musical ensembles,

DOI: 10.4018/jse.2012070104

they must be able to extract certain high-level features from the audio and incorporate those features into their performances. It is not sufficient for the robots to simply perform a series of pre-recorded or choreographed motions. Live performances inevitably vary from one show to the next, and even if there are no major changes in the music, there will always be small shifts in tempo, phrasing, and other factors. If the robot can only play according to a predetermined sequence, its responses will not necessarily mesh with the live performance. Additionally, we do not want the musicians to be forced to play a certain way in order to conform to the robot. The robot, rather, should be capable of following the musicians, even if their performance is completely different from previous ones.

We have performed substantial prior work on enabling robots to step or dance in response to music (Grunberg, Lofaro, Oh, & Kim, 2011; Kim et al., 2010). These performances, however, only considered the beat locations and tempo of the audio, and disregarded other aspects. We now turn our attention to another crucial feature: the emotional content, which we also refer to in this paper as 'mood', of music. Humans modify their dance motions so that their performances match the mood of the music they dance with, and when the dance motion and musical audio do not convey the same emotion, the resulting performance can look jumbled or confused. Therefore, to optimize their dances, the robots should be capable of determining an appropriate mood and then communicating it via gestures.

The choice of robot platform will influence the types of motions and emotions that the robot is able to produce. Smaller humanoids with more cartoonish looks for example, may be more suitable for happier or 'cute' emotions. They are also generally more rugged and easier and cheaper to repair than larger humanoids, so they offer practical advantages as well. Conversely, adult sized humanoids are often more human-like in appearance, so they may be able to display gestures with a higher degree of fidelity than the cartoonish miniature platforms. These robots also tend to boast more computational power and more sophisticated motion algorithms. As both types of robots offer different advantages, we would like to determine which ones are better at communicating which emotions to human audiences.

Another important decision is the number of robots to be used in a performance. Some human dance troupes have several members perform identical or similar motions to communicate the desired emotional effect to their audience. Using multiple robots instead of one might increase some of the emotional intensity, but it could also encourage the viewer to focus less specifically on any one single robot and thus be less moved by whatever emotion is being expressed. We would therefore like to more precisely identify if using multiple humanoids for dance performances influences how humans perceive the mood of the dances.

LITERATURE REVIEW

The proposed work draws on several different areas of engineering. Several signal processing algorithms are required to analyze the acoustic music, dance theory is needed to determine how best to represent emotions with gestures, and knowledge of robots is needed to accurately control the humanoid platforms.

Beat Tracking

In order for the robot's motions to appear to be congruent with a song, it must move in synchrony with the music's beats. It is thus necessary that the final system include an algorithm for identifying the tempo and beat locations of acoustic music. Numerous such algorithms, called beat trackers, have been proposed in the literature. Some of the better performing of these are highly complex systems that use sophisticated machine learning algorithms such as Hidden Markov Models or Linear Discriminant Analysis to precisely locate beat positions (Klapuri, Eronen, & Astola, 2006; Peeters & Papadopoulos, 2011). These systems are not, unfortunately, as useful in situations with restricted computational

power or the requirement of causality. Some simpler beat trackers work by calculating low-level features, then performing relatively low cost operations such as autocorrelation to find periodicities in these features that can be used to find tempi and beat locations. Systems taking this approach include those of Scheirer (1998), which calculates the subband spectral envelopes for its low-level feature, and of Davies and Plumbley (2007), which uses the signal's complex spectral difference.

Mood Tracking

Several different algorithms for estimating the emotional content of audio have also been developed. Systems that our group has studied range from using conditional random fields to model emotion probabilistically (Schmidt & Kim, 2011a) and using deep belief networks to determine complicated optimal features for mood detection (Schmidt & Kim, 2011b), to simply mapping easily calculated low-level features to an emotion space (Schmidt, Turnbull, & Kim, 2010). Both spectral contrast (Jiang, Lu, Zhang, Tao, & Cai, 2002) and Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980) have been found to be useful features for the latter sort of algorithm. MFCCs represent the audio spectrum warped according to a perceptual scale that reflects how the human ear perceives audio, and spectral contrast is a measure of the peaks and valleys in the acoustic spectrum. Both of these features can be mapped to a space representing emotions, allowing mood to be quickly estimated (Schmidt, Turnbull, & Kim, 2010).

A significant concern with developing systems to identify mood in musical audio is that it is difficult to obtain ground-truth values. People may disagree on the mood of various clips of music, and without a corpus of songs with emotion labels; it is difficult to train systems to identify the same emotions in music as humans do. Our group has previously developed multiple computer programs to help collect this ground truth data. One such program, called MoodSwings, requires players to mark the emotion of mood in conjunction with a peer (Kim, Schmidt, & Emelle, 2009). We also developed a similar task for Amazon's Mechanical Turk, to validate the MoodSwings results (Speck, Schmidt, Morton, & Kim, 2011). The gathered data was richly labeled in terms of emotional content, and usable for training mood estimation algorithms.

Gesture Production

Once the mood of a song is known, the robot must modify its motions based on that knowledge. Dance motions convey emotions in and of themselves, even to small children (Boone & Cunningham, 1998), and this emotion must match that of the audio to avoid producing a confusing or disorganized performance. As such, it is crucial to incorporate knowledge and research from dance studies to determine how best to make the robot move in order to communicate the required emotion.

Research by Camurri, Lagerlof, and Volpe (2003) indicates several important factors in dance motions that determine what mood they will convey. For example, angry gestures tend to take less time than sad gestures, contain more tempo changes, and carry more dynamic tension. Another group enabled a robot to identify the emotional content of a user based on their gestures (Lourens, van Berkel, & Barakova, 2010) by modeling the emotions with Laban notation, a type of notation used specifically to record dances (von Laban, 1956). Importantly, this group verified that the values of certain elements of Laban notation such as 'weight' and 'effort' correspond strongly to specific emotions. It was later independently verified that robots are capable of inducing emotion if they move according to Laban principles (Nakata, Mori, & Sato, 2002). Clay, Couture, and Nigay (2009) used a similar model to capture emotions from ballet performances.

Musical Robots

Several other dancing or musically expressive robots have been developed by various groups. Keepon is able to move its head and respond to

the beat of audio, although it lacks the ability to incorporate mood information into its motions (Michalowski, Sabanovic, & Kozima, 2007). Haile, a drumming robot, is able to listen to drum sequences and respond with complimentary notes (Weinberg & Driscoll, 2006). Neither of these two robots are humanoid, though, which limits their ability outside of these narrow contexts. Asimo, a humanoid robot developed by Honda, is able to step in response to music (Murata et al., 2008; Yoshii et al., 2007). This system demonstrates both beat tracking and the ability to respond to music, though mood is neglected. Similarly, the humanoid robot HRP-2 can produce human-like dance gestures obtained via motion capture technology (Nakaoka et al., 2007), again without considering emotional content.

Shiratori and Ikeuchi (2008) have performed research on synthesizing dance performances based on human perceptions of musical mood. This work also incorporates Laban analysis of motion, particularly the 'weight' and 'effort' features. Features of a performance such as 'intensity' are calculated separately for a piece of music and a human dancer's response to that music, and then are incorporated into a synthesized dance to be performed by an artificial dancer. This modeling, however, is more useful for offline performances than live ones, because the dance motions need to incorporate a human dancing to the same music as will be used in the performance. This system cannot work as well if the human musicians want to add a new section or change the music slightly.

ROBOT PLATFORMS

In order to evaluate the impact of the appearance of a humanoid on the emotional content of that humanoid's dances, we have selected two different platforms to perform a series of gestures. The first of these is a small humanoid called DARwIn (Figure 1 left). Developed by Virginia Polytechnic Institute and State University, DARwIn is an open platform robot measuring 45.5 cm and possessing 20 degrees of freedom (DoF). It has a sleek body, with catlike ears and very large eyes, which gives the robot a cartoonish look. This robot is extremely capable for a miniature humanoid, and was victorious at the RoboCup 2011 competition. Because the platform is open source, we are able to modify it to perform the gestures and parameterizations that we want. It is thus a suitable miniature robot platform for our project.

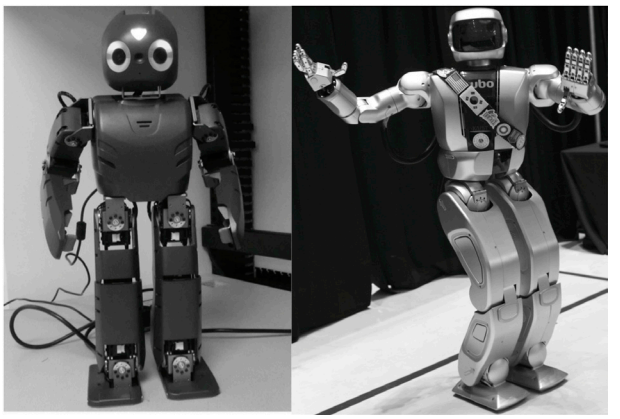
Our adult sized robot platform is Hubo, from the Korean Advanced Institute for Science and Technology (KAIST) (Figure 1 right). Hubo is 125 cm tall and has over 40 DoF (Park, Kim, Lee, & Oh, 2007). Hubo can perform dances such as tai chi with grace and fluidity, indicating that it has the potential to dance or gesture well enough to communicate emotion. Additionally, our lab has recently obtained six Hubos as part of an international collaboration between several American and Korean universities. The large number of these robots makes it practical to perform tests comparing performances of one robot to those produced by multiple humanoids, to determine if the additional robots have any effect on the perception of the performance's emotional content.

MOOD SPACE

One difficulty in creating systems to analyze mood and emotion is selecting a representation of these perceptual features. We have opted to use the Arousal-Valence, or A-V, model of mood (Thayer, 1989). Mood is mapped onto a two-dimensional plane, with the dimensions of the plane being arousal and valence. Arousal represents the intensity of a particular emotion; strong emotions such as "rage" or "ebullience" have large arousal values, while weak ones such as "tranquility" or "moroseness" have smaller values. Valence represents if an emotion is positive, such as "happiness" or "peace," or negative, such as "grief" or "despair." An example A-V plane with some emotions labeled is depicted in Figure 2.

There are several advantages to utilizing this particular representation of mood. Primar-

Figure 1. The DARwIn (left) and Hubo (right) robots



ily, this mapping allows for a continuous representation of emotion. Instead of having to select from a discrete array of emotions or moods, the system can map any new acoustic musical signal to a point on the A-V plane. This also allows for a logical parameterization, in which gestures that are near each other on the plane can naturally use similar parameter values. This mapping has been found suitable for gathering emotion labels on MoodSwings (Speck et al., 2011), and these labels can in turn be used as ground-truth values for evaluation of our system.

MUSIC INFORMATION RETRIEVAL

In order for the robot platform to be able to respond to audio, it must be able to extract the tempo, beat location, and AV values from acoustic music. These algorithms, developed in the field of music information retrieval (M-IR), are the backbone of any system that proposes to enable an intelligent response to musical audio. A brief overview of our algorithms is presented here; the interested reader is directed to the appropriate references.

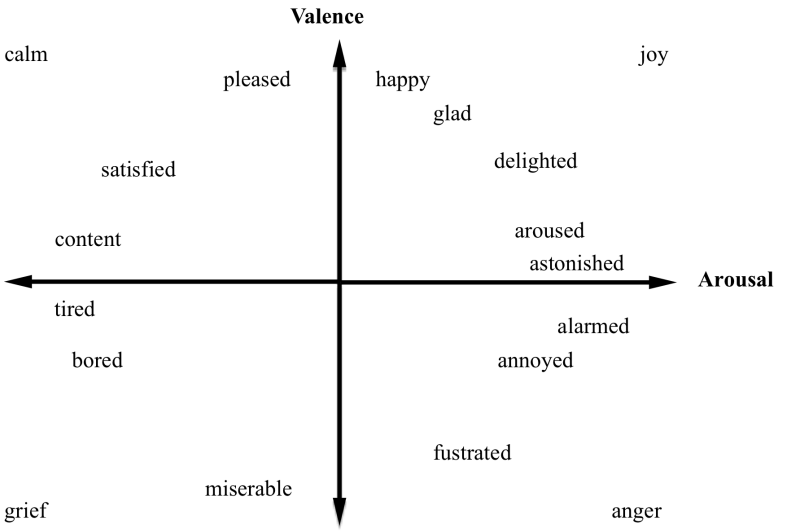
Constraints

In order that the system be useful for live audio performances, it must be able to function without any future knowledge of the music (Grunberg et al., 2011). Live performances have inherent vagaries that cannot be accounted for if the robot simply assumes that the new music will be identical to a previous recording. Such an assumption would also limit the musicians, who could not choose to modify or extend their songs because the robot would not be able to react appropriately. As such, the final system must be able to function at any point in time using only audio played before that point and no knowledge of the music that is yet to be played; in other words, it must be 'causal.'

It is also necessary that our algorithms be robust regarding acoustic noise. While it may be possible in some contexts to pass audio to the robot directly with an audio cable, this will not always be feasible. A better solution is for the robot to pick up audio over an acoustic channel with one or more microphones. Such audio will invariably be contaminated by noise, however (Ince, Nakadai, Rodemann, Tsujino, & Imura, 2010). Not only are robot platforms themselves often noisy, but it is not always possible to remove all sources of noise contamination in the room. Therefore, the system must be able to robustly extract the required audio features even when the audio is clouded by other acoustic sources.

Lastly, the system must require minimal computation. We would eventually like for the entire system to run onboard the robot platforms,

Figure 2. An A-V plane with several labeled emotions



eliminating the need for an external computer to run any of the algorithms. However, robot computers are often limited. Robots generally have limited power, physical space, and capacity for heat dissipation, so their computers are often less powerful than normal. Furthermore, these computers must also handle all of the tasks required to running the basic systems of the robot, and the M-IR algorithms must function with only the leftover capacity of the computers. Thus, the M-IR algorithms must be extremely computationally efficient.

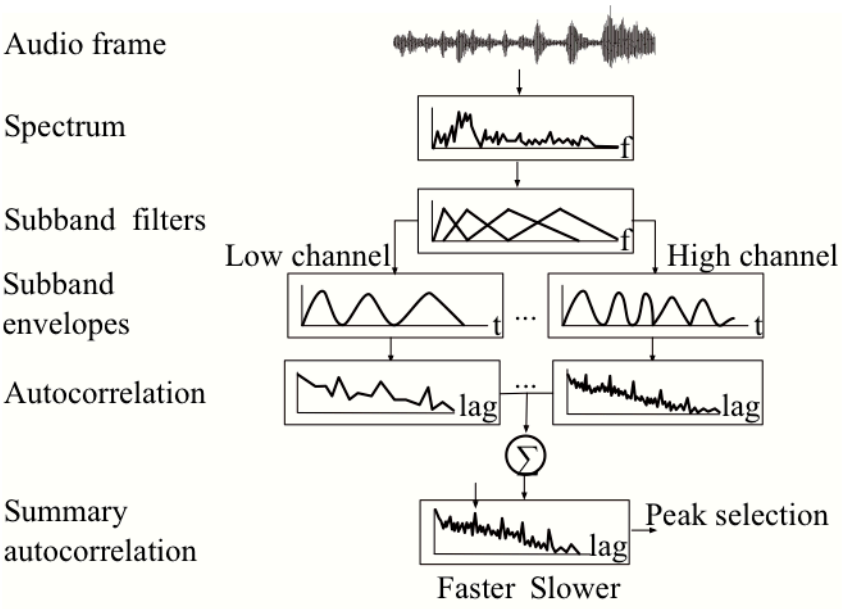
Tempo Identification and Beat Tracking

In order for the robot's motions to be synchronized with the music, the robot must know the tempo and the beat locations in the music. When humans dance, their motions are often spaced according to the tempo and apex on beat locations, and we want our robot dancers to be able to do the same.

There are many different approaches to beat tracking, but our system requires a causal and computationally efficient approach. We have opted to use a system based off of the work of Scheirer and Klapuri (1998, 2006). We have obtained accurate and robust results with this system in previous work, and have found that it can run quickly enough for our purposes (Grunberg et al., 2010). A flowchart of the system is depicted in Figure 3.

Audio is divided into .025 second frames and split into several subbands (Figure 3). The subbands increase in width as the central frequency increases, to mimic how the human ear divides up the acoustic spectrum. The subband envelopes are calculated and then autocorrelated, and the lag with the maximal autocorrelation value is used to find the tempo of the audio. Autocorrelations have large values at lags that are proportional to the period of the original signal, and musical audio can be treated as having a weak periodicity of its tempo. The system thus identifies the lag corresponding to the maximum value of the autocorrelation, then estimates the tempo based on that lag value. The algorithm then searches for sets of high energy frames that are spaced according to the estimated tempo value. This implementation also uses several smoothing and weighting steps to result in more consistent results.

Figure 3. Flowchart of the beat tracking system



This system was found to be highly accurate in a variety of situations, using both clean and noisy audio (Grunberg et al., 2011). The algorithm obtained an accuracy of a .98 F-Score for clean audio and (with adjustments to eliminate noisy segments) .92 for audio from an acoustic channel. The acoustic audio was captured while the robot itself was moving and generating noise, verifying that the algorithm is accurate even in very noisy environments.

Mood Identification

As with beat tracking, there are many existing methods for estimating the emotional content of audio. Our system was selected to fulfill the constraints of causality and computational efficiency. The algorithm we chose calculates the spectral contrast features of a signal and then maps those to A-V values (Schmidt & Kim, 2010).

Spectral contrast is a measure of the peaks and valleys in a frequency subband, and the spectral contrast of a frame of music influences its timbre. To calculate this feature, our system takes in a frame of audio, divides it into the same subbands as used for beat tracking, and then calculates maximum and minimum values of each subband. For the larger subbands, the largest and smallest few values are averaged to produce a smoothed maximum and minimum estimate. This was found to improve accuracy. The result of this step is fourteen extrema (one maxima and one minima in each of seven subbands) per frame.

The extrema are aggregated and averaged over forty frames to produce fourteen values for every second of audio. These values are multiplied by a 14x2 matrix to map them to A-V values. The mapping matrix is calculated by using least squares regression to optimize it on a large set of 240 clips with A-V annotations. The annotations were obtained via MoodSwings. This large amount of ground-truth data helps ensure a robust mapping matrix that can accurately map the spectral contrast features into the A-V space.

This system was also found to be highly accurate in previous studies. On a music corpus of 240 clips totaling one hour in length, the distance between the moods marked by humans

and those indicated by the system had a mean of only 15.1% of the mood space, with a standard deviation of .8% of the space (Schmidt & Kim, 2010). The system was thus found to be suitably accurate for our purposes.

ROBOT GESTURES

The M-IR values are important in that they can parameterize the gestures that the robot makes, thereby enabling it to respond appropriately to the music. These parameterizations are based on research done by the dance community in determining how human dancers convey emotions via body language (Camurri et al., 2003; Lourens et al., 2010).

For this study, we split the mood space into four quadrants, as dictated by the A-V map (Figure 2) (Thayer, 1989). Starting clockwise from the upper-left, these four areas represent the emotions of 'Joy,' 'Anger,' 'Grief,' and 'Calm.' We then enable the robot to produce a base gesture which could be parameterized to represent any of these emotions.

The base gesture is made as follows (Grunberg, Batula, Schmidt, & Kim, 2012):

- Both arms begin extended down and away from the robot. The head starts turned towards the robot's left side.
- The arms move inward, towards the robot's chest.
- At the same time, they move upwards, towards the robot's head.
- The robot's head moves from left to right as the arms move.
- Once the arms and head have reached their final position, they reverse course and return to their original position.

This gesture is then modified by the M-IR information. The tempo and beat locations found by the beat tracker set the temporal length of the gesture and gesture start times, respectively. The A-V values also influence the gesture, as follows (Grunberg et al., 2012):

- Joy: the arms begin raised relatively high and extended far from the body (Figure 4). The motion takes up the entire beat. As a result, the gesture is wide, expansive, and relatively fast compared to the other emotions (because there is more space to cover in the same amount of time). The head is tilted up as well.
- Calm: the arms are raised as in the Joy emotion, but begin 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. The gesture fills the entire beat.
- Grief: the arms are still close to the body as in the Calm gesture, but are lowered to be closer to the ground at both the beginning and apex of the gesture. The head is lowered as well. The gesture is performed over the full beat.

Figure 4. One Hubo performing the Joy gesture

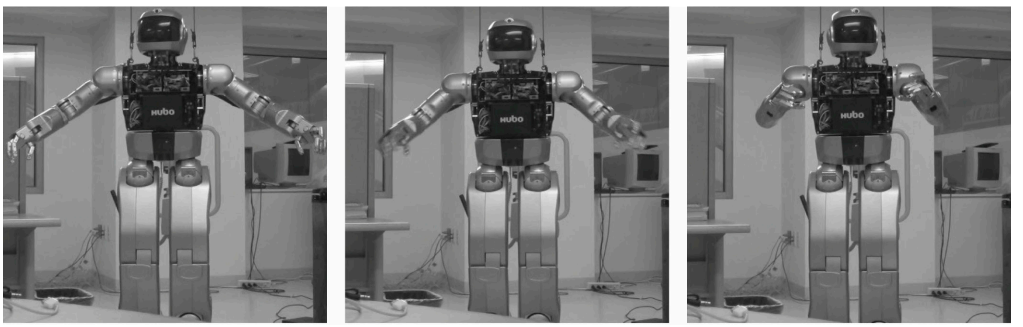


Table 1. Specifications of Hubo and DARwIn gesture parameterizations

Hubo	Joy	Calm	Grief	Anger
Arm distance from body	18″to 5″	13″ to 8″	11″ to 5″	15″ to 3″
Arm height from waist	-3″ to 5″	-2″ to 3″	-2″ to 1″	-5″ to 2″
Tilt of head	15°	15°	-15°	-15°
% of beat used for motion	100	100	100	67
DARwIn	**Joy**	**Calm**	**Grief**	**Anger**
Arm distance from body	6.5″ to 2″	5″ to 3.3″	4.23″ to 2″	5.5″ to 1.8″
Arm height from floor	-.6″ to 6.4″	-.9″ to 4.4″	-1.4″ to 1.6″	-1.9″ to 1.9″
Tilt of head	15°	15°	-15°	-15°
% of beat used for motion	100	100	100	60

- Anger: the arms begin lower to the ground than in any other gesture, and relatively far from the robot's body as well. The gesture additionally only takes up about two-thirds of the beat; the robot's motions are therefore sharper than in the other emotions. The head is at the same tilt as in the Grief gesture.

The Hubo gestures were designed first. Subsequently, the DARwIn gestures were crafted to attempt to mimic the Hubo motions, although the robots' different constructions made an exact replication impossible. Table 1 and Table 2 specify the exact parameterization of the gesture for the DARwIn and Hubo robots, respectively.

EXPERIMENTAL SETUP

Both the beat tracking and the A-V value prediction are performed simultaneously in a single M-IR system. This allows the system to share

Table 2. Results of the one-Hubo emotion perception tests

Ground-truth emotion test										
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	3.7±1.2	3.9±0.8
+	_	Anger	.35±.17	.20±.14	.16±.13	-.34±.13	-.18±.18	.19±.17	3.2±0.9	2.7±0.8
-	+	Calm	-.21±.19	-.17±.17	.12±.10	.20±.19	.21±.15	.13±.13	3.3±0.9	2.8±0.8
-	-	Grief	-.25±.14	-.04±.20	.23±.21	-.04±.26	-.12±.20	.20±.20	2.9±1.0	3.1±0.9
M-IR predicted emotion test										
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	.26±.12	.13±.11	3.2±0.5	2.8±0.9
+	_	Anger	.31±.19	.09±.29	.30±.24	-.22±.27	-.22±.19	.24±.19	2.7±0.9	3.0±0.8
-	+	Calm	-.14±.22	-.09±.23	.20±.21	.08±.22	.08±.23	.17±.18	3.1±1.0	3.1±1.1
-	-	Grief	-.22±.13	-.04±.24	.24±.19	-.07±.23	-.07±.20	.12±.10	3.0±1.0	3.0±1.1

common processing elements, such as calculating Fourier Transforms, thereby decreasing the computational cost accrued from using both algorithms together. The M-IR system thus determines the beat locations, tempo, and A-V values for the music being played. These values are then transmitted via Universal Datagram Packet (UDP) to the main computer of the robot(s).

UDP is a communications protocol that is useful for rapid communication between systems. While UDP lacks the error checking of some other algorithms, such as the Transmission Control Protocol (TCP), it is faster and therefore more suited to a system such as this one. It is less problematic if the robot drops a packet than if the latency between the calculation of the M-IR features and the receiving of those values by the robot grows too large, because the latter case could result in an apparent desynchronization between the robot and the music.

Upon receiving a new packet, the robot produces a gesture as instructed by the M-IR values. As long as the packets are not received during a gesture, the robot will continue to perform one motion per packet. If the robot receives the next set of values before it is done with its latest gesture, however, it disregards them. This prevents the robot from trying to do two things at once (finishing a previous gesture and commencing the next), which could result in damage to the platform.

The test music corpus consisted of twelve segments of audio taken from our prior selection of clips that was annotated on MoodSwings. These clips spanned the four A-V quadrants evenly, as indicated by the annotations. The robot then moved in response to these gestures. The length and timing of its gestures was determined by the beat tracker values, and which parameterization it used was based on the mood identification system. Human experts listened to the audio and watched the performance, and then they rated both in terms of arousal and valence on a scale of -.5 to .5. Additionally, they determined if the robot's motions seemed to be congruent (or 'fit') both the mood and the beat locations of the audio, on a 5 point scale.

Three different robot configurations were used for the experiments:

- One Hubo robot (the baseline test).
- One DARwIn robot.
- Four Hubo robots.

These tests would allow us to determine whether or not the number or size of the robots had an effect on the perceived emotion.

For each configuration, two sets of evaluations were performed. The first used ground-truth emotional values obtained from MoodSwings (instead of the values predicted by the M-IR system). This evaluated if the robots could demonstrate the correct emotion in the best possible case, with perfect knowledge of musical mood. The second test (called the 'predicted emotion test') used the mood tracking algorithm. Because our system does not have 100% accuracy for determining A-V values, this introduced error into the robot motions. We therefore wished to evaluate if the robot could still produce the correct emotions despite this error.

RESULTS AND DISCUSSION

The perceptual results for the tests using one Hubo are shown in Table 2. Results have been averaged across all experts and all clips within the quadrant. Nine experts evaluated the system when it used ground-truth emotion, and six when it used the M-IR predicted emotions. The final totals are shown as the mean and standard deviation of the experts' responses. 'Error' is defined as, for each expert and each song, the magnitude of the difference between the arousal and valence values marked for the music and the robot.

Our baseline results are promising. The robot performances were judged, on average, to belong to the same quadrants as the music in both the ground-truth and predictive cases. This indicates that a single Hubo can indeed display the appropriate emotions using the specified gesture parameterization, and that the

Table 3. Results of the one-DARwIn emotion perception tests

Ground-truth emotion test										
Quadrant			Arousal: [-.5,.5]			Valence: [-.5,.5]			Congruence: [1,5]	
A	V	Emotion	Music	Robot	Error	Music	Robot	Error	Mood	Beat
+	+	Joy	.35±.13	.25±.11	.10±.10	.35±.15	.30±.10	.10±.10	3.5±1.0	3.1±0.9
+	_	Anger	.37±.19	.42±.09	.11±.18	-.37±.17	-.37±.10	.10±.15	4.1±1.0	4.0±0.9
-	+	Calm	-.15±.20	-.05±.20	.13±.13	.14±.20	.22±.15	.12±.15	3.7±0.9	3.4±1.1
-	-	Grief	-.20±.21	-.07±.24	.14±.11	-.05±.26	-.16±.19	.18±.18	3.3±0.9	3.2±0.9
M-IR predicted emotion test										
Quadrant			Arousal: [-.5,.5]			Valence: [-.5,.5]			Congruence: [1,5]	
A	V	Emotion	Music	Robot	Error	Music	Robot	Error	Mood	Beat
+	+	Joy	.27±.18	.13±.16	.15±.13	.28±.18	.27±.17	.09±.08	2.9±1.0	3.2±1.0
+	_	Anger	.31±.20	.28±.21	.23±.16	-.34±.12	-.21±.14	.19±.15	2.8±1.2	3.0±0.9
-	+	Calm	-.17±.20	.08±.23	.27±.22	.13±.24	.21±.17	.13±.16	3.1±1.2	3.0±1.3
-	-	Grief	-.24±.18	-.07±.27	.18±.19	-.00±.27	-.06±.23	.15±.18	3.3±1.3	3.1±1.1

mood prediction algorithm is accurate enough to maintain this accuracy. The mean error never exceeded .2 (or 20% of the mood space) for either arousal or valence in the ground-truth case, or .25 in the predictive case. If the ground-truth of a piece of music has arousal and valence values of magnitude .25 or greater, then, the system will likely correctly identify its quadrant, and the Hubo will then produce an appropriate gestural response.

The congruency values are also notable. In the ground-truth case, congruence for mood is almost always greater than 3. This indicates that the experts felt there was a good fit between the robot's gestures and the mood of the music. Beat congruent values are also relatively high, especially in the Joy and Grief emotions. While the average congruence values are slightly lower for the predictive case, they are still generally greater than 3, further validating the good fit of the music and the robot's emotions.

The results from the test using one DARwIn as the robot platform are displayed in Table 3. Eight experts participated for both of these tests.

The results from the ground-truth DARwIn tests compare favorably to the Hubo values. The experts again identified the robot gestures as belonging to the same mood quadrant as the audio for all four quadrants. Additionally, the average errors are generally less than in the equivalent Hubo case, particularly regarding the valence dimension, where all four emotions show improvement. Finally, the congruency values are greatly improved for all of the emotions except for Joy. Therefore, when using ground-truth mood information, DARwIn seems better able to communicate most of the selected gestures than Hubo, although Hubo may be better suited for the Joy gesture. Hubo's greater ability at this one gesture may be evidence that the larger robot is better equipped to simulate motions that take up a lot of 'space' (in Laban notation) like Joy.

The results are less in favor of DARwIn when considering the predictive case, however. In particular, the arousal axis of the Calm emotion was poorly identified by the experts; though the emotion has a negative arousal value, the experts identified the robot's motions as being positive. Additionally, the valence of the Grief emotion was only slightly negative. Average error for arousal, and especially valence, is higher as well for the DARwIn using the mood

Table 4. Results of the four-Hubo emotion perception tests

Ground-truth emotion test										
Quadrant			Arousal: [-.5,.5]			Valence: [-.5,.5]			Congruence: [1,5]	
A	V	Emotion	Music	Robot	Error	Music	Robot	Error	Mood	Beat
+	+	Joy	.33±.12	.30±.11	.05±.08	.37±.14	.36±.10	.07±.12	4.1±0.8	3.8±0.8
+	_	Anger	.33±.26	.32±.16	.18±.19	-.39±.11	-.22±.21	.18±.17	3.5±1.0	2.8±0.9
-	+	Calm	-.19±.19	-.15±.15	.09±.11	.19±.23	.25±.13	.14±.16	3.6±1.1	2.9±1.1
-	-	Grief	-.28±.17	-.17±.23	.17±.20	.10±.22	-.08±.19	.24±.20	2.9±1.1	2.6±1.0
M-IR predicted emotion test										
Quadrant			Arousal: [-.5,.5]			Valence: [-.5,.5]			Congruence: [1,5]	
A	V	Emotion	Music	Robot	Error	Music	Robot	Error	Mood	Beat
+	+	Joy	.29±.11	.29±.11	.09±.10	.29±.20	.28±.16	.12±.13	3.6±1.1	3.5±1.2
+	_	Anger	.35±.17	.20±.18	.19±.14	-.39±.11	-.11±.22	.29±.25	2.8±1.0	2.8±0.8
-	+	Calm	-.15±.16	.04±.21	.23±.21	.13±.24	.13±.16	.19±.18	2.8±1.2	2.8±1.0
-	-	Grief	-.23±.19	-.07±.20	.17±.16	.06±.24	-.01±.25	.16±.16	3.0±1.0	3.3±1.0

tracking algorithm than for the Hubo, and congruence is also less.

These results indicate that, while DARwIn may be superior given perfect mood information, the emotional content of its performances can be more easily thrown off by errors. When the M-IR algorithm errs, the DARwIn's erroneous motions are more easily classified in other emotion quadrants by the experts than in the Hubo case. This may be because the DARwIn is smaller and there is less distinction (in absolute distance) between gestures. As the gestures are closer, incorrect ones may disproportionately influence a human's view of the entire performance towards the wrong quadrant.

The results of the multi-Hubo test are displayed in Table 4.

In the ground-truth test, error values were slightly decreased in the four-Hubo case as compared to the one-Hubo case. Positive arousal gestures held the most improvement, although the Grief emotion contained a substantial decrease in the arousal error as well (changing by a full .06, or 6% of the total space). The robot performance average A-V values were also in the correct quadrant for every emotion. This demonstrated that four Hubos better conveyed the correct mood than one. Furthermore, mood congruence is increased for every emotion except for Grief, which remained constant. The multiple Hubos thus seem to have a reinforcing effect on the emotional content of the performance, when perfect mood data is passed to the robot. Lastly, beat congruence values are approximately the same for the first three gestures, though worse for Grief.

When using the M-IR values, one of the four emotions was misidentified on average (Calm robot sequences had a positive arousal score). While the error values did generally decrease for arousal (except for the Calm emotion), the values were generally worse for valence. Mood congruence improved for the gestures with positive arousal values, but remained the same or became worse for the others. Clearly, while using multiple robots may present a clear benefit in communicating emotion with perfect ground-truth data, it is not as clearly advantageous when the values are selected using an imperfect mood tracker. Beat congruence values are much improved for Joy and Grief, but worse for the other two emotions.

One other conclusion of note can be seen in this data. First, the Grief valence value of the

music was, on average, marked in the wrong quadrant in both four-Hubo tests. The multiple Hubos seem to be influencing how people are listening to the music and inducing them to think of the songs with the Grief emotion as having larger valence values than they really do.

While using four Hubos does not seem to benefit the system overmuch when using a mood tracking algorithm, there is a clear benefit when the ground-truth values are known. As such, while the multiple robot setup may not be as useful now, if the mood prediction system is improved, several robots could help convey a desired emotion to human viewers more effectively than just one.

CONCLUSION

We have enabled a variety of different sets of robot platforms to respond to the beat locations and mood of musical audio. Our baseline case of one Hubo performed very well, as humans identified its motions to belong on average to the same gesture quadrant as the audio for all quadrants, and found further that its motions were congruent with the audio. Both the DARwIn and Hubos did better than the baseline case when perfect mood information was utilized, but less well (or even worse than the baseline) when the M-IR system was used to determine A-V values. This indicates, while the one-Hubo system is currently the best choice when using music with unknown ground-truth emotions, the other platforms may be able to do better than this system if they use a more accurate mood predicting algorithm.

We aim to continue studying the effects of different robot platforms on the gestures. One aspect of future work is to adjust the gestures on DARwIn and the multiple Hubos to try to more closely simulate the effects of just using one Hubo. This could help provide us with rules that would let us discover more about the relations between the robot platforms and the emotions that they can convey. For example, if we found that emotions must be more expansive on DARwIn to convey the same emotion as more constrained gestures onboard Hubo do, that could be used to help us understand the effects of the robot sizes and shapes better.

We also seek to continue improving our representation of emotion. Currently, we have quantized the emotion space into four quadrants, but there are more than four emotions. This lack of gradation may be why some of the emotions were unclear even when the robots were using the emotional ground-truth data – the music required a combination of arousal and valance values that was not exactly satisfied by any of the four parameterizations. By making the mood space continuous, we hope to enable a greater degree of control over the gestures, and thus performances that are more fluid and human-like.

In terms of the actual robot control system, we are examining systems besides UDP to transmit messages to the robots. While UDP has many advantages, it will sometimes drop packets or even induce noticeable lag. A more reliable or faster control scheme could allow us to more accurately position the robot, and thus more accurately portray a desired emotion.

Finally, we are interested in testing our system on larger and more varied groups of people. All tests for this paper were conducted on experts in a laboratory environment, as these people were already generally well acquainted with the A-V space and the music tasks. However, their reactions may be different from those of the general population. We are investigating making a task that is open to the general public, perhaps on Mechanical Turk. By allowing people with no known prior knowledge of music or emotion to watch videos of the robots dancing and to rate their performances as in the prior tests, we could validate our results on a much wider audience.

REFERENCES

Boone, R. T., & Cunningham, J. G. (1998). Children's decoding of emotion in expressive body movement: The development of cue attunement. *Developmental Psychology*, *34*(5), 1007–1016. doi:10.1037/0012-1649.34.5.1007

Camurri, A., Lagerloef, I., & Volpe, G. (2003). Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, *59*(1), 213–225. doi:10.1016/S1071-5819(03)00050-8

Clay, A., Couture, N., & Nigay, L. (2009). Towards an architecture model for emotion recognition in interactive systems: Application to a ballet dance show. In *Proceedings of the World Conference on Innovative VR* (pp. 19-24).

Davies, M. E. P., & Plumbley, M. D. (2007). Context-dependent beat tracking of musical audio. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(3), 1009–1020. doi:10.1109/TASL.2006.885257

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. doi:10.1109/TASSP.1980.1163420

Grunberg, D., Ellenberg, R., Kim, I. H., Oh, J. H., Oh, P. Y., & Kim, Y. E. (2010). Development of an autonomous dancing robot. *International Journal of Hybrid Information Technology*, *3*(2), 33–44.

Grunberg, D. K., Batula, A., Schmidt, E. M., & Kim, Y. (2012). Emotion recognition and affective gesturing in response to music. In *Proceedings of the International Conference on Intelligent Robots and Systems*.

Grunberg, D. K., Lofaro, D. M., Oh, P. Y., & Kim, Y. E. (2011). Robot audition and beat identification in noisy environments. In *Proceedings of the International Conference on Intelligent Robots and Systems* (pp. 2916-2921).

Ince, G., Nakadai, K., Rodemann, T., Tsujino, H., & Imura, J.-I. (2010). Robust ego noise suppression of a robot. In N. García-Pedrajas, F. Herrera, C. Fyfe, J. Benítez, & M. Ali (Eds.), *Proceedings of the 23rd International Conference on Trends in Applied Intelligent Systems* (LNCS 6096, pp. 62-71).

Jiang, D.-N., Lu, L., Zhang, H.-J., Tao, J.-H., & Cai, L.-H. (2002). Music type classification by spectral contrast feature. In *Proceedings of the IEEE International Conference on Multimedia and Expo* (Vol. 1, pp. 113-116).

Kim, Y. E., Batula, A. M., Grunberg, D., Lofaro, D. M., Oh, J., & Oh, P. Y. (2010). Developing humanoids for musical interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 36-43).

Kim, Y. E., Schmidt, E. M., & Emelle, L. (2009). MoodSwings: A collaborative game for music mood label collection. In *Proceedings of the 10th International Society for Music Information Retrieval* (pp. 231-236).

Klapuri, A. P., Eronen, A. J., & Astola, J. T. (2006). Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 342–355. doi:10.1109/TSA.2005.854090

Lourens, T., van Berkel, R., & Barakova, E. (2010). Communicating emotions and mental states to robots in a real time parallel framework using Laban movement analysis. *Robotics and Autonomous Systems*, *58*(12), 1256–1265. doi:10.1016/j.robot.2010.08.006

Michalowski, M., Sabanovic, S., & Kozima, H. (2007). A dancing robot for rhythmic social interaction. In *Proceedings of the 2nd Annual Conference on Human-Robot Interaction* (pp. 89-96).

Murata, K., Nakadai, K., Yoshii, K., Takeda, R., Torii, T., & Okuno, H. G. …Tsujino, H. (2008). A robot singer with music recognition based on real-time beat tracking. In *Proceedings of the 9th International Conference on Music Information Retrieval* (pp. 199-204).

Nakaoka, S., Nakazawa, A., Kanehiro, F., Kaneko, K., Morisawa, M., Hirukawa, H., & Ikeuchi, K. (2007). Learning from observation paradigm: Leg task models for enabling a biped humanoid robot to imitate human dances. *The International Journal of Robotics Research*, *26*(8), 829–844. doi:10.1177/0278364907079430

Nakata, T., Mori, T., & Sato, T. (2002). Analysis of impression of robot bodily expression. *Journal of Robotics and Mechatronics*, *14*(1), 27–36.

Park, I.-W., Kim, J.-Y., Lee, J., & Oh, J.-H. (2007). Mechanical design of the humanoid robot platform, HUBO. *Advanced Robotics*, *21*(11), 1305–1322. doi:10.1163/156855307781503781

Peeters, G., & Papadopoulos, H. (2011). Simultaneous beat and downbeat-tracking using a probabilistic framework: Theory and large-scale evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(6), 1754–1769. doi:10.1109/TASL.2010.2098869

Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, *103*(1), 588–601. doi:10.1121/1.421129

Schmidt, E. M., & Kim, Y. E. (2010). Prediction of time-varying musical mood distributions from audio. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 465-470).

Schmidt, E. M., & Kim, Y. E. (2011a). Modeling musical emotion dynamics with conditional random fields. In *Proceedings of the International Society for Music Information Retrieval* (pp. 777-782).

Schmidt, E. M., & Kim, Y. E. (2011b). Learning emotion-based acoustic features with deep belief networks. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 65-68).

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010). Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the ACM International Conference on Music Information Retrieval* (pp. 267-273).

Shiratori, T., & Ikeuchi, K. (2008). Synthesis of dance performance based on analyses of human motion and music. *IPSJ Online Transactions*, *1*(1), 80–93. doi:10.2197/ipsjtrans.1.80

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A comparative study of collaborative vs. traditional musical mood annotation. In *Proceedings of the International Society for Music Information Retrieval* (pp. 549-554).

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford, UK: Oxford University Press.

von Laban, R. (1956). *Principles of dance and movement notation*. New York, NY: Macdonald & Evans.

Weinberg, G., & Driscoll, S. (2006). Robot-human interaction with an anthropomorphic percussionist. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1229-1232).

Yoshii, K., Nakadai, K., Torii, T., Hasegawa, Y., Tsujino, H., & Komatani, K. …Okuno, H. G. (2007). A biped robot that keeps steps in time with musical beats while listening to music with its own ears. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1743-1750).

David K. Grunberg received the BS and MS degrees in electrical engineering from Drexel University in Philadelphia, PA in 2010 and 2011, respectively. He is currently a PhD candidate in the Department of Electrical and Computer Engineering at Drexel University in Philadelphia, PA. He is a research assistant in the Music & Entertainment Technology Laboratory (MET-lab) at Drexel University. His current research focus is enabling humanoid robots to extract certain high-level features, such as beat locations, from music, so that they can intelligently respond to the audio. David is on the second year of a National Science Foundation Graduate Research Fellowship. His research interests include signal processing, machine listening, and enabling robots to participate in musical tasks alongside humans.

Alyssa M. Batula received her BS in Electrical and Computer Engineering from Lafayette College in 2009 and her MS in Electrical Engineering from Drexel University in 2011. She is currently pursuing her PhD at Drexel University in the Music & Entertainment Technology Laboratory. Her current research focus is on human-robot interaction and musical robotics. She is on her second year of a NSF-funded STEM GK-12 fellowship, which pairs graduate students with high school teachers in order to teach students engineering concepts. In 2011, Alyssa was awarded the NSF Graduate Research Fellowship. Her research interests are signal processing, machine learning, and robotics.

Erik M. Schmidt received the BS degree in electrical engineering from Temple University in Philadelphia, PA in 2007 and the MS degree in electrical engineering from Drexel University in 2009. He is currently a PhD candidate in the Department of Electrical and Computer Engineering at Drexel University in Philadelphia, PA. He is a research assistant in the Music & Entertainment Technology Laboratory at Drexel University. His current work focuses on the automatic prediction of emotional content in acoustic musical signals. Mr. Schmidt has research interests in the areas of signal processing and machine learning for machine understanding of music audio.

Youngmoo E. Kim is an Associate Professor of Electrical and Computer Engineering and Assistant Dean of Engineering for Media Technologies at Drexel University. He received his PhD from the MIT Media Lab and also holds Master's degrees in Electrical Engineering and Music (Vocal Performance Practice) from Stanford University as well undergraduate degrees in Engineering and Music from Swarthmore College. His research group, the Music & Entertainment Technology Laboratory (MET-lab) pursues machine understanding of sound, interfaces and robotics for expressive interaction, and K-12 outreach for engineering education. He co-chaired the 2008 International Conference on Music Information Retrieval and was invited by the National Academy of Engineering to co-organize the "Engineering and Music" session for the 2010 Frontiers of Engineering conference. His research is supported by the National Science Foundation, including an NSF CAREER award in 2007, and the Knight Foundation.