Using Audio and Haptic Feedback to Detect Errors in Humanoid Musical Performances

Alyssa M. Batula, Manu Colacot, David K. Grunberg, and Youngmoo E. Kim
Music and Entertainment Technology Laboratory (MET-lab)
Electrical and Computer Engineering Department
Drexel University
Philadelphia, PA 19104, USA
{batulaa, m.colacot, dgrunberg, ykim}@drexel.edu

ABSTRACT

We present a system that determines whether an adult-sized humanoid has correctly played a pitched percussive instrument in real time. Human musicians utilize sensory feedback to determine if they are playing their instruments correctly and robot performers should be capable of the same feat. We present a classification algorithm that uses auditory and haptic feedback to decide if a note was well- or poorly-struck. This system is demonstrated using Hubo, an adult-sized humanoid, which is able to play pitched pipes using paddles. We show that this system is able to determine whether a note was played correctly with 97% accuracy.

Keywords

Musical robots, humanoids, auditory feedback, haptic feedback

1. INTRODUCTION

As robots have become more sophisticated, it has become possible for them to participate in increasingly complex and intricate musical activities [1]. In particular, we are interested in developing general-purpose humanoids capable of performing autonomously alongside humans in music ensembles. While the literature is replete with examples of machines that can sing, dance, and play music, there is more to a live musical performance than simply moving in a predefined way to produce the desired sounds from an instrument [2, 3]. While being able to perform a pre-rendered performance is one thing, live shows inevitably feature variations and differences from prior events that require the performer to adapt and adjust. Humans can accomplish this, in part, by using their senses to obtain feedback from the performance around them and determine what is different. For robots to accomplish the same goal, they must also use sensor feedback from their surroundings to adjust their playing to suit the environment.

One particularly useful capability for a musical robot would be the ability to determine if it is successfully playing its instrument. In order to produce the correct notes or sounds, instruments must be played in a particular way. For example, a performer may need to pluck a string at a certain location, strike a drumhead in a particular place, or move his or her arm into a specific position relative to a theremin. Should the musician play the instrument in a different manner, the resulting sound may be completely different from the desired note. Because the instrument, musician, and even environment may shift between or during performances, the musician cannot assume that just because the previous song was performed correctly, the next one will be accurate as well [4]. When humans perform, they pay attention to how their instrument sounds and feels as they play it in order to gauge their performance. To accomplish this same task, a musical robot should also be able to use sensor feedback to estimate whether, when it attempts to play a note on an instrument, it successfully does so.

In order to maximize the flexibility of such a performance robot, the final system requires several key features. First, it should be able to utilize different types of sensory information. A robot in one situation may be able to hear its own performance very clearly, and thus should be able to use auditory feedback to determine if it is playing correctly. A drum-playing robot might be able to feel the forces exerted by the impact of its stick on a drumhead, and should have the ability to utilize haptic feedback for similar purposes. Ideally, the system will be able to combine different types of feedback to produce more reliable results. The system should also be able to work with physical instruments and function in real acoustic environments. While certain pitch detection algorithms that use clean digital audio have been developed for robots, these have limited utility in a live concert environment when the instruments are acoustic and the auditory source is contaminated by noise [5]. Robots themselves are often noisy, compounding the problem [6]. As such, the auditory portion of the algorithm should be able to function, even when using microphones to record acoustic instruments.

Our efforts focus on Hubo, an adult-sized humanoid developed by the Korean Advanced Institute for Science and Technology (KAIST) (Figure 1). This robot has over forty degrees of freedom and is capable of smooth and fluid motion [7]. In order to allow the robot to play musical notes, Hubo has been provided with a PVC instrument referred to as a "Hubophone" (Figure 2), which is similar to instruments used by the performance artists *Blue Man Group*[1]. Each Hubophone contains multiple pitched pipes that are struck with a foam paddle. The system is designed to determine if the robot successfully played a particlular note.

Hubo has previously been programmed to perform a musical sequence on pipe instruments. Specifically, this system allowed a Hubo quartet to use three Hubophones and one drum kit to play *Come Together* by The Beatles[2] (Figure

NIME'13, May 27 – 30, 2013, KAIST, Daejeon, Korea.

[1] `blueman.com`
[2] `youtube.com/watch?v=UMQLX-aw_dc`

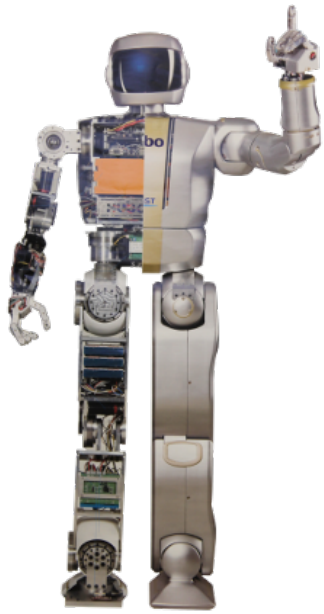

Figure 1: Hubo, an adult-sized humanoid robot

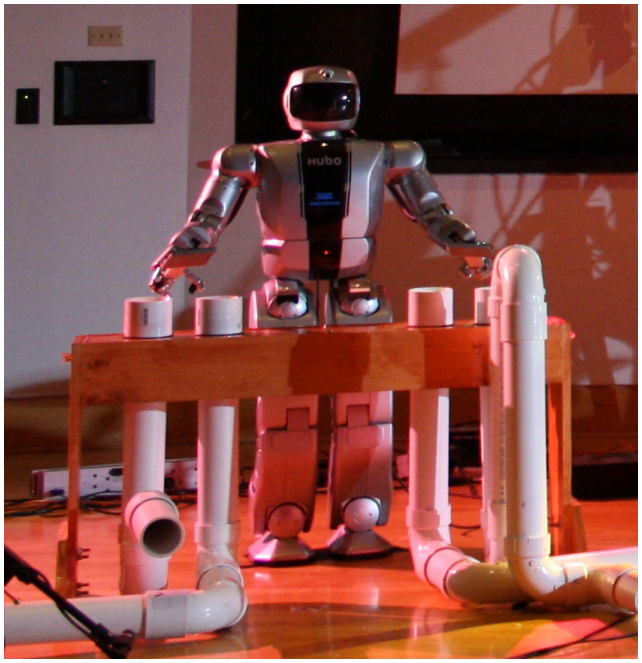

Figure 2: A "Hubophone" as set up for a performance

3). The robots, however, had to be precisely positioned to play each instrument, a process which took a great deal of time. An automatic calibration system would make such performances more viable by drastically reducing the set up time. Automatically determining whether or not the robots are correctly striking the instruments is an important first step to such a calibration system.

2. LITERATURE REVIEW

Several musical humanoids have been developed by various groups. As one example, the Toyota Partner Robots are a series of humanoids that play musical instruments such as trumpets, violins, and drums [2]. Another is the Advanced Step in Innovative MObility (ASIMO), a humanoid designed and produced by Honda, which can step, scat, and sing in response to music [8]. ASIMO also conducted the Detroit Symphony Orchestra through a performance of *The Impossible Dream* [9]. These robots, however, could not determine if they made mistakes in their performance. This made them less suitable for live performances of unknown works.

The RoboNova, a small humanoid developed by HiTec robotics, has been programmed to play simple melodies on the keyboard using customized 'fingers' [5]. Before playing a piece, this robot could use auditory feedback to determine the position of the keys. This allowed the humanoid to calibrate itself and resulted in better performances. Adding other sources of data, such as a camera, was mentioned as potentially fruitful future work.

Other groups have incorporated feedback and calibration into their systems as well. Mizumoto et al explored using audio feedback to assist HRP-2, a humanoid robot, in playing a theremin [4]. This algorithm induced the robot to move its arms around the instrument, analyzed the resultant pitches, and determined the optimal arm positions to produce a sequence of notes. This system, however, could assume that there would always be a note being played, and did not consider the possibility of missing the theremin entirely. We are interested in implementing a system that can determine whether or not a note was played, as this information would also be useful for calibrating the robot.

Murphy et al developed a system for online calibration of a drumming robot [10]. Many drumming robots use relatively inexpensive and imprecise solenoids to trigger the actuators that generate the drumming motion, but the relationship between the desired volume of the sound and the solenoid's actual output is generally nonlinear in a manner that varies from robot to robot. In order to avoid having to recalibrate for each robot, the authors developed an algorithm for automatic pre-calibration that linearized the solenoids based on feedback from the robot's performances. This system, however, used only a single data source: a piezoelectric sensor mounted on the drum. It also required that the instruments themselves be modified with the piezoelectric sensors, limiting the applicability of the algorithm. It would be preferable for performing robots to be able to use instruments without needing them to have been prepared in advance.

Ness et al consider a similar problem [11]. This group had the robot strike the drum with many different velocities, then calculated the Mel-Frequency Cepstral Coefficients (MFCCs) of the sound produced by the robot. MFCCs are a representation of the frequency content of audio, warped according to how the human ear perceives sound. The first MFCC coefficient was an approximator of the loudness of the signal, and the remaining coefficients were used to evaluate the timbre content of the strike. This data was then used in a mapping to derive a correlation between the robot's velocity and the resultant sound. This system, while impressively accurate, also only analyzed acoustic data, neglecting haptic feedback.

O'Modhrain experimented with adding haptic feedback to virtual instruments to determine whether or not this made them easier for humans to play [12]. This study demonstrated the importance of haptic feedback for playing instruments, and indicated that using this type of sensory data could be useful for robotic applications.

3. ROBOT AND INSTRUMENT DETAILS

We chose the Hubo as our robot platform (Figure 1). Hubo is an adult-sized humanoid that stands over four feet tall

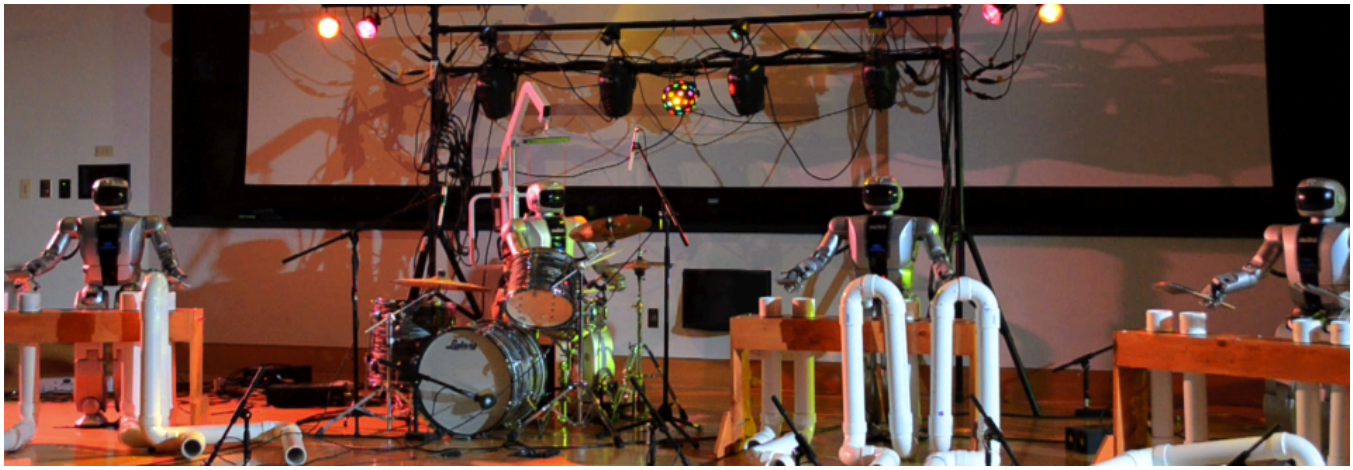

Figure 3: Four Hubo robots performing a rendition of *Come Together* by the Beatles

and possesses over forty degrees of freedom. It is capable of grasping a paddle or drumstick in its hands with the aid of velcro, and uses six motors (three shoulder, one elbow, and two wrist) to move its arm and strike a percussive instrument.

Hubo is equipped with multiple sensors, including three to measure force and torque readings in each wrist. These sensors detect changes in the force and torque placed on the hand while playing an instrument. The robot is also capable of reading auditory data through external microphones which, when mounted near the robot's performance space, allow it to hear the instrument that it is playing. Hubo is therefore capable of detecting both auditory and haptic data, which can then be used to help the system determine if it is playing an instrument correctly.

The instrument that we have selected for the robot, the "Hubophone," consists of four open, pitched PVC tubes (Figure 2). When an open cylindrical pipe is struck, it will vibrate at its resonant frequency as well as integer multiples (or *harmonics*) of that frequency. These harmonics can be calculated with the equation:

$$f_n = \frac{nc}{2L} \tag{1}$$

Where n is the harmonic number, f_n is the frequency of that harmonic, c is the speed of sound, and L is the length of the tube.

The pipes in each Hubophone are cut to a particular length L such that the resulting f_1 is the desired pitch. The four-pipe configuration was chosen to allow each hand to play two notes without requiring it to travel too far. This allows the robot to play more rapidly without damaging its motors, making the system more useful for performance environments.

In order for the robot to perform in live concerts and shows, it must be able to determine whether or not its motion actually hits a pipe and plays a proper note. The pipes are large and must be hit with sufficient force to make them resonate. A blow that is too weak, or only glances off the pipe, will not produce noticeable pitched content, and the note will not sound correct. Accuracy is also crucial. If the robot misses the center of the pipe by more than a few inches, the paddle will glance off the side and produce a dull, thumping sound instead of a clear note. Larger inaccuracies could result in the robot missing the pipe completely, hitting the wooden frame or empty air.

4. DATA COLLECTION

In order to gather data to train and test our system, a Hubo robot was recorded while striking the Hubophone pipes. We collected force-torque and audio recordings as the robot repeatedly played two different pipes with its right hand. The robot's position was shifted approximately every 20 recordings in order to obtain more variety in the dataset. We collected a total of 650 recordings for each pipe, evenly distributed between correct and incorrect hits. Immediately after playing each note it was labeled as a good or bad hit by at least one researcher with a musical background. These judgements were based on both the audible quality of the note and the position of the paddle when it struck the pipe. Video recordings of the robot playing were used to assist in the ground-truth labeling if there was any uncertainty.

Each hit was designed to take 0.25 seconds to complete. Audio and haptic data were recorded for two seconds from the start of the playing motion in order to analyze the sustain and decay of the note. Monophonic audio was recorded at 44.1 kHz through a microphone mounted near the struck end of the pipes, and haptic data was recorded from both wrists at 100 Hz. This data included the front-back and left-right torques (denoted as M_x and M_y respectively), and the up-down force (F_z). Both the audio and haptic data were recorded by the same computer that signaled the robot to play in order to ensure that they were synchronous with the robot's motions and each other.

5. SYSTEM ALGORITHM

After the robot plays a note, it must assess the note's quality. This algorithm can be broken into two steps. First, the robot must extract informative features from the auditory and haptic data. Second, those features are used in a classification algorithm to estimate whether or not the robot successfully played the desired note.

5.1 Feature selection

For audio data, we determined that the *harmonic energy* is an informative and useful indicator of the robot's accuracy. A strong hit on a particular pipe will produce a sound that contains more energy at that pipe's harmonics than when the robot misses, deflects off the side of the pipe, or hits the pipe too softly. The first three harmonics in particular should contain noticeably more energy on successful hits than unsuccessful ones. By taking the ratio of the energy in the first three harmonics to the total energy in the audio frame, a feature called the *harmonic ratio* can be calculated.

In order to calculate the total harmonic ratio for a hit, the robot breaks the two seconds of input audio into 100 msec frames and processes each one separately. The frequency content of each frame is calculated with a Short-Time Fourier Transform, and the bins corresponding to the first three harmonics of the desired note are identified. The energy in these bins, total energy of the signal, and ratio of these two values are calculated. Once all frames have been processed, the average ratio for the entire signal is determined and used as the audio feature.

Figure 4 shows the harmonic energy in each example plotted against the total energy for that example. These plots show that the audio features labeled as 'good' or 'bad' tend to cluster together, allowing classification with relatively high accuracy in spite of the overlap between classes. The plots also show different clustering patterns for the two pipes, most likely resulting from the different motions required to play the notes.

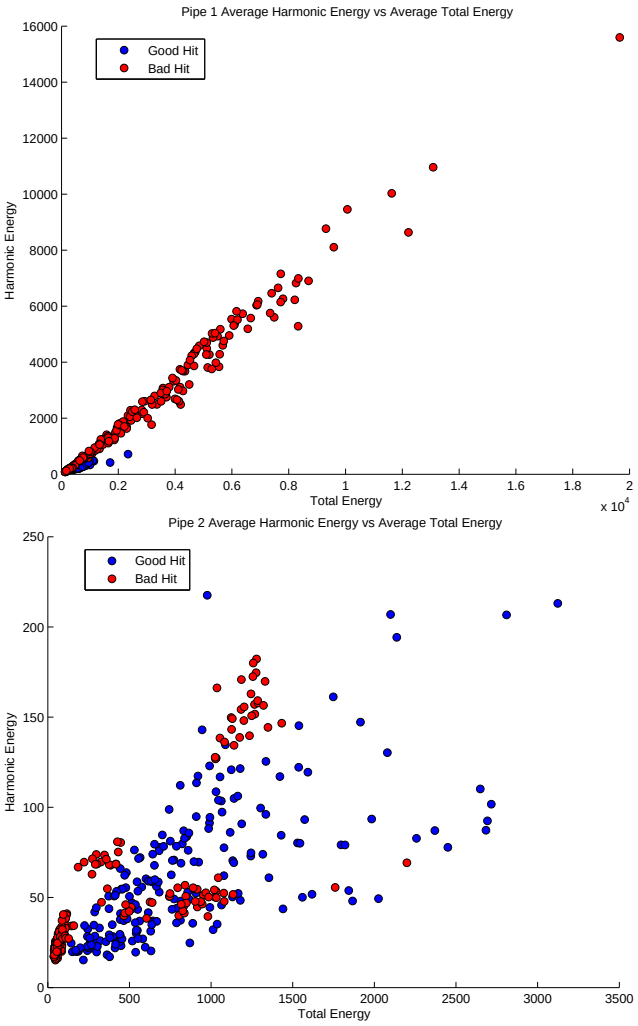

Figure 4: Audio features with ground truth labels

Features are also extracted from the force and torque sensor readings. Each set of sensor readings is averaged, creating a set of three features $[\bar{F_z}, \bar{Mx}, \bar{My}]$. This set of features represents the average magnitudes of the force and torque on the wrist. As the robot plays a pipe with varying force and at different locations, the force and torque readings should reflect the differences in hit technique and quality. Figure 5 shows three-dimensional plots of the training dataset with ground truth labels for both pipes. Like with the audio data, examples labeled as 'good' or 'bad' tend to cluster together.

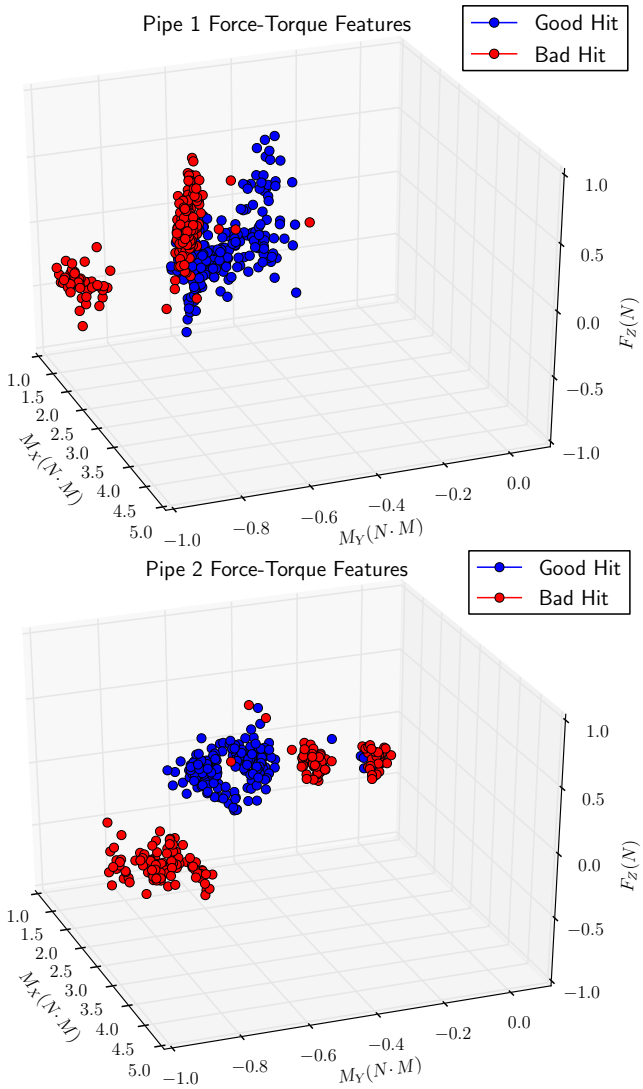

Figure 5: Force-torque values with ground truth labels

5.2 Classification

We used support vector machines (SVMs) to classify examples via the LIBSVM library [13]. SVMs create a multi-dimensional decision boundary in order to classify new data as either a good or bad hit. Prior to classification, each set of features was normalized to have zero mean and unit variance in order to aid in classifier training. A separate SVM was trained for each pipe to account for the distinct motions and forces required to play each note, as well as the different pitch of the notes. In order to determine the optimal features for classification, we also trained multiple classifiers using different SVM kernels and different feature sets (audio only, haptic only, and combined). Once trained, these models could be used to classify new, unlabeled data and determine whether the robot should change its playing motion.

6. EXPERIMENTS

We evaluated our system on a dataset of 1300 examples, 650 per pipe. The samples for each individual pipe were split into two sets, a training set (400 points per pipe) to train our SVM, and a testing set (250 points per pipe) for

Table 1: Training set system accuracy. Units are in %

SVM kernel	Feature	Pipe 1 10.5 in. from mike $f_1 = 49.9$ Hz.	Pipe 2 14.5 in. from mike $f_1 = 80.8$ Hz.	Mean
Linear polynomial	Harmonic ratio	70.0	97.5	83.8
	Force-torque mean	93.3	85.5	89.4
	Combined	97.8	99.3	98.5
2nd order polynomial	Harmonic ratio	70.3	78.5	74.4
	Force-torque mean	76.8	97.5	87.1
	Combined	84.8	98.0	91.4
3nd order polynomial	Harmonic ratio	64.3	92.0	78.1
	Force-torque mean	88.0	80.3	84.1
	Combined	96.3	98.8	97.5
Radial basis function	Harmonic ratio	78.75	97.5	**88.1**
	Force-torque mean	93.8	97.5	**95.6**
	Combined	99.3	99.0	**99.1**

Table 2: Testing set system accuracy. Units are in %

SVM kernel	Feature	Pipe 1 10.5 in. from mike $f_1 = 49.9$ Hz.	Pipe 2 14.5 in. from mike $f_1 = 80.8$ Hz.	Mean
Radial basis function	Harmonic ratio	94.2	95.0	94.6
	Force-torque mean	98.2	76.0	87.1
	Combined	97.1	98.7	**97.9**

verification. Each set contained equal numbers of good and bad hits.

For the training set, we trained 12 SVMs for each pipe. These classifiers tested four different kernels (linear, polynomial degrees 2 and 3, and radial basis function) on each of three feature sets (audio only, haptic only, and combined). Each classifier was evaluated using leave-one-out cross-validation (LOOCV), in which a classifier is trained on all but one data point, and is then tested on the final example. This process is repeated for all examples and the final classifier accuracy is the average accuracy of all trials. After evaluating the results on the training set, the best kernel for each feature set was used to train new SVMs on the testing set. The testing set accuracy was also obtained using LOOCV.

We also compared our auditory results with other common features to validate our decision to use the harmonic ratio. In particular, we considered the total energy in the signal, as well as several statistical spectrum descriptors (SSDs) commonly used to describe the shape of an acoustic spectrum [14]. These include the spectral centroid (the 'center of gravity' of a spectrum), flatness (a measure of a spectrum's uniformity), flux (a measure of a spectrum's derivative), and rolloff (frequency below which 85% of the spectrum's energy resides). To evaluate our system against the best possible case of the SSDs, the SSD features were classified with an oracle threshold system that was provided with the 400 training examples and determined one or two thresholds to optimally separate the classes. Finally, it was tested on the examples from the testing dataset.

7. RESULTS

The SVM results on the training set are shown in Table 1. As the data shows, the radial basis function (RBF) kernel SVM outperforms the linear and polynomial kernels for all feature sets on the training data. The results for the testing data set, trained using the RBF kernel, are displayed in Table 2. The algorithm achieves 94% accuracy using only audio data and 87% accuracy using only haptic data. This implies that both types of sensory data provide useful, but different, information about the quality of a played note.

The comparison of our auditory classification to an oracle system using various statistical spectrum descriptors is displayed in Table 3. It is clear that, of the spectrum descriptors, the spectral flux and centroid compare best to the harmonic ratio. As the spectral flux measures the rate of change of the spectrogram, and the centroid measures its weighted spectrum, this is likely due to a Hubophone hit abruptly shifting the spectrogram to a new center. The total energy of the signal also proved to classify the signal at better-than-chance accuracy, likely because of the loudness of the successful notes. Particularly for notes further from the microphone, it could be useful to incorporate these three features into a later version of the system. The harmonic ratio, though, still proved to be superior.

The other statistical spectrum descriptors appear to be less useful. Spectral flatness and rolloff classify very poorly. This would imply that the relative uniformity of the spectrum, and the concentration of energy in the lower frequencies, is not overly impacted by the successful performance of a note.

8. CONCLUSION

This work demonstrates that audio and haptic feedback can be utilized, jointly or alone, to determine whether a robot has correctly played a pitched pipe instrument. With results exceeding 97% on average, the classification system performs significantly better than chance (50%). This is the first major requirement for designing an automatic calibration system, which would save time and energy for human operators working with musical robots. This classification algorithm, therefore, is a useful first step in making more flexible and capable performing humanoid robots.

One major aspect of future work will include the uti-

Table 3: Auditory feature comparison. Accuracy is in units of %

Feature	Classification	Pipe 1 10.5 in. $f_1 = 49.9$ Hz.	Pipe 2 14.5 in. $f_1 = 80.8$ Hz.	Mean
Total energy		83.6	58.8	71.2
Spectral centroid		18.0	70.0	44.0
Spectral flatness	Oracle threshold	47.6	47.6	47.6
Spectral flux		84.4	50.0	67.2
Spectral rolloff		48.0	46.4	47.2
Harmonic ratio	SVM	**94.2**	**95.0**	**94.6**

lization of visual data. In addition to haptic and auditory feedback, a robot should be able to use its camera to look down and determine if it is hitting the pipes or not. The introduction of such a system would allow the robot to be more certain that it could truly hit the pipes correctly. Such data could be especially useful in very noisy acoustic environments (such as crowded dance halls), where it may be difficult at first to distinguish between a correct hit and an almost-correct hit based on audio data.

Accuracy and robustness to difficult situations, such as noisy real-world environments, could potentially be improved by incorporating other features into the SVM, particularly the more useful of the statistical spectrum descriptors. It would be interesting to see if adding information about the spectral centroid and flux would improved accuracy for pipes at varying distances from the microphones. We may also consider still other auditory features, such as Mel-Frequency Cepstral Coefficients [11].

The next step after creating an accurate classifier is to develop a motion correction algorithm for the robot. Once the robot has determined that ia note has been played incorrectly, it should be able to change its playing motion to try to correct its playing.

Finally, we want the robot to be able to play with musical expression. When human musicians play instruments, they do not simply strike each note in exactly the same way whenever they play it. They change how they play the instruments in order to create different sounds that evoke certain emotions or moods. By playing the notes in different ways, the robot could demonstrate an expressive performance. This would make it more useful for ensembles that wish to have the robot display and perform the appropriate mood for the music they are playing.

9. ACKNOWLEDGEMENTS

This work was supported by two National Science Foundation Graduate Research Fellowships

10. REFERENCES

[1] Shigeki Sugano and Ichiro Kato, "Wabot-2: Autonomous robot with dexterous finger-arm coordination control in keyboard performance," in *Proceedings of the International Conference on Robotics and Automation*, 1987, pp. 90–97.

[2] S. Takagi, "Toyota partner robots," *Journal of the Robotics Society of Japan*, vol. 24, no. 2, pp. 208–210, 2006.

[3] M. Michalowski, S. Sabanovic, and H. Kozima, "A dancing robot for rhythmic social interaction," in *Proceedings of the 2nd Annual Conference on Human-Robot Interaction (HRI)*, 2007, pp. 89–96.

[4] Takeshi Mizumoto et al., "Thereminist robot: Development of a robot theremin player with feedforward and feedback arm control based on a theremin's pitch model," in *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, October 2009.

[5] Alyssa M. Batula and Youngmoo E. Kim, "Development of a min-humanoid pianist," in *Proceedings of the 10th IEEE-RAS International Conference on Humanoid Robots*, 2010.

[6] Gökhan Ince et al., "Robust ego noise suppression of a robot," in *Trends in Applied Intelligent Systems*, Nicolás García-Pedrajas, Francisco Herrera, Colin Fyfe, José Benítez, and Moonis Ali, Eds., vol. 6096 of *Lecture Notes in Computer Science*, pp. 62–71. Springer Berlin / Heidelberg, 2010.

[7] Ill-Woo Park, Jung-Yup Kim, Jungho Lee, and Jun-Ho Oh, "Mechanical design of humanoid robot platform khr-3 (kaist humanoid robot-3: Hubo)," in *Proceedings of the 5th IEEE/RAS International Conference on Humanoid Robots*, 2005, pp. 321 – 326.

[8] Kazumasa Murata et al., "A robot uses its own microphone to synchronize its steps to musical beats while scatting and singing," in *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nice, France, September 2008.

[9] Brain Geuther, Autumn Breese, and Yunfeng Wang, "A study on musical conducting robots and their users," in *Proceedings of the 2010 IEEE-RAS International Conference on Humanoid Robots*, Nashville, TN, USA, December 2010, pp. 124–129.

[10] Jim Murphy, Ajay Kapur, and Dale Carnegie, "Better drumming through calibration: Techniques for pre-performance robotic percussion optimization," in *Proceedings of the International Conference on New Interfaces and Musical Expression*, 2012.

[11] Steven Ness, Shawn Trail, Peter Driessen, Andrew Schloss, and George Tzanetakis, "Music information robotics: Coping strategies for musically challenged robots," in *Proceedings of the International Symposium on Music Information Retrieval*, 2011.

[12] Maura S. O'Modhrain, *Playing by Feel: Incorporating Haptic Feedback into Computer-Based Musical Instruments*, Ph.D. thesis, Stanford University, 2000.

[13] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[14] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.