# GESTURE ANALYSIS OF RADIODRUM DATA

*Steven R. Ness, Sonmez Methabi,*
*Gabrielle Odowichuk, George Tzanetakis*

University of Victoria
Department of Computer Science

*Andrew W. Schloss*

University of Victoria
School of Music

## ABSTRACT

The radiodrum is a virtual controller/interface that has existed in various forms since its initial design at Bell Laboratories in the 1980's, and it is still being developed. It is a percussion instrument, while at the same time an abstract 3D gesture/position sensor. There are two main modalities of the instrument that are used by composers and performers: the first is similar to a percussive interface, where the performer hits the surface, and the instrument reports position (x,y,z) and velocity (u) of the hit; thereby it has 6 degrees of freedom. The other mode, which is unique to this instrument (at least in the domain of percussive interfaces), is moving the sticks in the space above the pad, whereby the instrument also reports (x,y,z) position in space above the surface.

In this paper we describe techniques for identifying different gestures using the Radio Drum, which could include signals like a circle or square, or other physically intuitive gestures, like the pinch-to-zoom metaphor used on mobile devices such as the iPhone. Two approaches to gesture analysis are explored. The first one is based on feature classification using support vector machines and the second is using Dynamic Time Warping. By allowing users to interact with the system using a complex set of gestures, we have produced a system that will allow for a richer vocabulary for composers and performers of electro-acoustic music. These techniques and vocabulary are useful not only for this particular instrument, but can be modified for other 3D sensors.

## 1. INTRODUCTION

The radiodrum [6] [10] is a virtual controller/interface that has existed in various forms since its initial design at Bell Laboratories in the 1980's, and it is still being developed. Using capacitive sensing, it can detect the position of two augmented drum sticks in a space above a receiver antenna. It is a percussion instrument, while at the same time an abstract 3D gesture/position sensor. Currently, there are two modes of interaction: the first is similar to a percussive interface, where the performer hits the surface, and the instrument reports position (x,y,z) and velocity (u) of the hit; thereby it has 6 degrees of freedom. We call this "whack" mode. Note that this mode does not depend on hitting the surface; it is detected by the change of direction of the stick and not by any impact as in conventional drum pads. The other mode, which we call "continuous" mode, is unique to this instrument (at least in the domain of percussive interfaces). This mode involves moving the sticks through space above the pad, whereby the instrument also reports (x,y,z) position in space above the surface at any desired sampling rate.

In the current work we propose to extend the radiodrum with a new modality of interaction with a system that recognizes gestures made by the performer. These gestures can be as simple as a sweep across the surface of the radiodrum, or can be as complex as the composer or performer desires. These gestures are then sent to a system the produces sound or image. These gestures can either be mapped to actions that send metadata to the patch, for example, a different section of a musical piece could be triggered, or can be mapped directly to sound producing modules. In the case of a direct mapping to sound, the speed of the gesture itself could be mapped to musical properties of the sound.

## 2. BACKGROUND

The word gesture is a highly overloaded term, and has a variety of meanings in a number of different fields. Because of this, the concept of gesture can be viewed as a boundary object [9], that is, a concept that is at once adaptable to different viewpoints, but also robust enough to maintain its identity across different fields. A detailed examination of gestures can be found in Cadoz and Wanderley [2], where they categorize gestures in music as effective gestures, accompanist gesture and figurative gestures. The gestures detected by our system fall into all three of these categories. In a review by Overholt [8], three basic groups of gestural controllers for music are summarized, which include those inspired by instruments, augmented instruments and alternative instruments. Our system is inspired by drums, but can also be viewed as an augmented drum.

## 3. RELATED WORK

There have been many approaches to detecting gestures in the current literature. These include computer vision based methods, and approaches using accelerometer data. Early work in the field of gesture recognition employed

the concept of Space-time gestures [4] in which sets of view models are captured by a Computer Vision system and are matched to stored gesture patterns using Dynamic Time Warping (DTW). Although this system was not used to produce music, it is relevant to the current work because it uses the same technique of Dynamic Time Warping to recognize patterns in a multi-dimensional space.

Another more recent paper [3] also uses DTW for the recognition of a small gesture library, in a non-realtime setting. It takes hand-arm gestures from a small, predefined vocabulary and uses a DTW engine to align these gestures in time and also to perform normalization on them.

More relevant to the current work are papers that perform gesture recognition on the data from accelerometers, such as those now commonly found on devices such as the iPhone and Wii-mote. An early work in this field was described in "SICIB: An Interactive Music Composition System Using Body Movements" [7] where a rule based coupling mechanism linking the position, velocity, acceleration, curvature, torsion of movements and jumps of dancers is mapped to intensity and tone in music sequences.

Another recent paper relevant to current work describes uWave[5], an accelerometer-based personalized gesture based system. This system is unusual in that the system uses a single prototypical example of each gesture, and uses DTW to recognize a simple gesture vocabulary containing eight signs, derived from the Nokia gesture alphabet. Akl et al. [1] describe an approach that extends this work using DTW, affinity propagation and compressive sensing. In this paper, 18 gestures are recognized, as opposed to 8 for the uWave[5] paper. The addition of compressive sensing allows gestures to be reduced to a more sparse representation that is then matched using DTW.

Another interesting new approach [11] does not use either DTW or heuristics, but instead uses the machine learning technique of Support Vector Machines (SVM) and does gesture recognition with 3D accelerometer data. In this paper, a system that uses a frame-based descriptor and a multi-class SVM is described. With frame-based descriptors, instead of using the entire time series, a reduced representation is used, and in this paper, spectral descriptors calculated by a Fast Fourier Transform are used. This paper describes how this approach outperforms DTW, Naive Bayes, C4.5 and Hidden Markov Model (HMM) machine learning systems.

The approaches described in the previous paragraphs have positive attributes as well as challenges. Computer vision based approaches have the advantage that electronic cameras are now cheap commodity hardware and are easily available. However, computer vision based approaches are fundamentally limited by their hardware requirements of cameras and transmitters and high computational load [5]. They also have a low acquisition framerate as compared to other approaches.

Accelerometer based approaches have the advantage that with MEMS technology small cheap solutions are becoming common [11] and they provide a high speed and continuous source of data, ideal for machine learning approaches. However, accelerometers suffers from abrupt changes due to hand shaking[1].

In the past three years, we collaborated with Bob Boie, the original designer of the radiodrum at Bell Labs in the 1980's, in the design of a new instrument that does not use MIDI at all. Boie worked with Max Mathews in the 1980's, and is known for being one of the original designers of capacitive sensing tablets, including very early versions of trackpads and similar devices. There is considerable confusion in the computer music community about the radiodrum vs. Radio Drum vs. Radio Baton and the history thereof. We will not go into detail here, but suffice it to say that this new instrument is more accurate and satisfying to use in performance that the older versions. At the moment, we only have one prototype, but we are in the process of building more instruments.

MIDI works reasonably well for asynchronous data like drum-hits, but it is badly suited for continuous position data. The new instrument that Boie designed is entirely analog, and we collect data from it by amplitude-modulating the analog sensor data in order to take it out of the DC frequency range. Then we digitize it using an ordinary audio interface, and this allows us to use almost any sampling rate so we get very precise temporal data with no jitter or throughput issues, orders of magnitude faster than data obtained by computer vision techniques.

Our radiodrum-based approach has the advantage that absolute position data is transmitted, and these data are in the form of (x,y,z) three-tuples that have a direct relation to the position of the sticks in the real world. They also have the advantage that extremely high rates of data with excellent time resolution can be obtained. Since the radiodrum is a 3D sensor, we know exactly where the sticks are at all times, not just when they strike the surface.

## 4. SYSTEM DESCRIPTION

Our system works by matching gestures to a library of template gestures, usually provided by the performer or composer. It then performs real-time matching of gesture data from the radiodrum in the form of a list of tuples of x,y,z values and matches this stream of data to the template library, returning the matching template gesture.

The mapping of gestures to sounds is in the domain of the composer, who in our system would be responsible for creating this mapping of gestures to sounds. To use our system, the composer and performer would define an alphabet of gestures that will be used.

During the performance of the piece of music, when the musician wants a gesture to be recognized, they then push a foot switch which activates the gesture recognition system, and then execute a gesture. A similarity matrix between this gesture and all examples of all gestures in the system is then calculated. A similarity matrix of two examples of the gesture for "A" is shown in Figure 1.

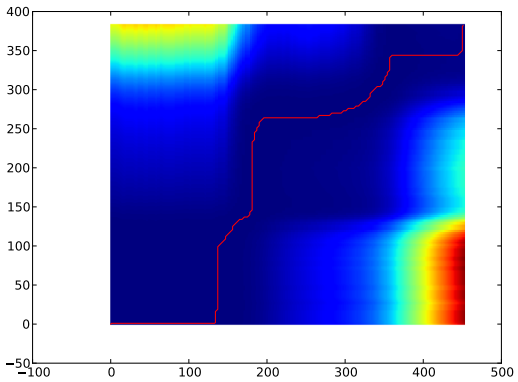


**Figure 1**. Similarity matrix for two examples of the sign "A" along with the best path as determined by Dynamic Time Warping shown in red.

## 5. RESULTS

We have implemented two independent systems for doing gesture recognition of radiodrum data. The first of these is a system based on Dynamic Time Warping (DTW) [4] and the second is a system based on a frame based implementation using a Support Vector Machine (SVM) machine learning classifier.

### 5.1. Dynamic Time Warping

In our DTW implementation, we take the X and Y pairs of a performed gesture and compare each of the X and Y time series to all the examples of the gesture in our gesture alphabet. Specifically, we first take the X time series of the performed gesture and calculate a similarity matrix between all points of this gesture and each gesture in our database. We then run the DTW algorithm on this similarity matrix and find the cost of the path as determined by the DTW algorithm, where a lower cost path indicates a more likely gesture match. We then repeat this for the Y time series and sum the scores for the X and Y values. The lowest cost path over all the gestures in our alphabet is then returned as the most likely match.

We performed a series of experiments to evaluate the performance of the DTW algorithm. For the first of these, we compared three very distinct signs, those of "A", "K" and "C" as used in the Palm Pilot graffiti alphabet. Exemplars of these three letters are shown in Figure 2. We took 8 to 10 examples of each of these letters and performed DTW against all 30 examples in the gesture database. We then calculated the precision and recall of the top 8 returned gestures. The results of this experiment are shown in Table 1. As one can see, the precision for each of these is high, for A and K the precision is 1, which means all of the 8 returned gestures matched the query gesture correctly. For all three examples, the recall was also high, and varied between 0.727 and 0.888. In addition, for all three gestures the top result (Top-1) was correct in all cases, this is the most important measure because it matches more closely the behaviour of the system in a performance implementation.

In doing these experiments we noticed that we can get a higher recall and consequently a higher F-Measure if we specify a cut-off threshold for DTW score, since this way we eliminate more irrelevant gestures in the retrieved gestures set. In order to investigate this further, first we found the optimum threshold for a training set and then to evaluate the system with obtained thresholds, we tested the system on the testing set and the result is shown in Table 2 2.

### 5.2. Support Vector Machine

In order to extract features for each gesture, first, we divided each gesture into N equal sized segments. Then frames are formed by putting two adjacent segments (each with the length $Ls = L/(N))$ together and in a way that every two adjacent frame have a segment in common. In other words, the frame $i$ consists of the segments i and segment $i+1$. So, we have $N-1$ frames each with the length $2*Ls$ and each frame consists of two time series axis $x_t$ and $y_t$. The feature vector of each gesture consists of the feature vectors of all of its frames. So if the feature vector of the frame i is called $f_i$ then the feature vector of each gesture is: $F = f_0 + ... + f_{N-1}$. So we need to calculate the feature vector of each frame. In order to do that we input the x and y axis into FFT function separately, and in the frequency domain, we calculated the mean and the energy feature:

$$Mean = \mu_{T,K} = t^0_{T,K0}$$

$$Energy = \varepsilon_{T,k} = \frac{\sum_{n=1}^{Ls,2-1} |t^n_{T,k}|^2}{|Ls,2-1|}$$

Where the vector t is the output of the FFT function, $T = x,\ y\ and\ k = 0,...,N-1$. After calculating the features for all the frames, we end up with the feature vector of the gesture:

$$\tau = (\mu_{T,K}, \varepsilon T, k)$$

Now the gesture i can be represented by $(g_i, \tau_i)$ and fed to the binary SVM to train the classifier to recognize the new features.

In this experiment, we first trained the classifier with the training set that includes 7 samples of the gesture "A" and 7 samples of the gesture "B". Then in order to evaluate the model, we used a test set including 3 samples of the gesture "A" and 3 samples of the gesture "B".

We repeated this operation for 10 different number of Ns, and three pairs of gestures (a,b), (c,o) and (k,x). The results are shown in Table 3. From this table it can be seen that by choosing the right frame size of the SVM classifier it is possible for this SVM method to outperform the DTW method.

| Gesture | Precision | Recall | Top-1 |
|---|---|---|---|
| A | 1.0 | 0.888 | 9/9 |
| B | 1.0 | 1.0 | 10/10 |
| K | 1.0 | 0.727 | 10/10 |
| C | 0.792 | 0.704 | 8/8 |
| E | 0.818 | 0.595 | 10/10 |
| O | 1.0 | 0.889 | 10/10 |

**Table 1**. Precision and Recall for three different gestures

| Threshold | F-measure "A" | F-measure "K" | F-measure "O " |
|---|---|---|---|
| 0.020 | 0.966 | 0.927 | 0.982 |
| 0.030 | 0.974 | 0.949 | 0.988 |
| 0.031 | 0.974 | 0.951 | 0.989 |
| 0.036 | 0.967 | 0.952 | 0.990 |
| 0.042 | 0.952 | 0.947 | 0.992 |
| 0.050 | 0.921 | 0.929 | 0.986 |

**Table 2**. F-measure for 3 different signs at a variety of training cutoff levels

| Frames | Precision A,B | Precision C,O | Precision K,E |
|---|---|---|---|
| 2 | 1.0 | 0.75 | 1.0 |
| 4 | 1.0 | 0.66 | 1.0 |
| 6 | 0.75 | 1.0 | 1.0 |
| 8 | 1.0 | 1.0 | 1.0 |
| 10 | 1.0 | 1.0 | 1.0 |
| DTW | 1.0 | 0.972 | 0.983 |

**Table 3**. Average Precision and Recall when using the Support Vector Machine testing/training approach to gesture recognition. Shown in the last line of the table are the average results for the DTW algorithm.

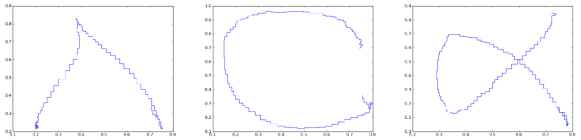

**Figure 2**. Exemplars of the gestures "A", "C" and "K" as X,Y pairs of data, with the Z axis flattened to the plane of the page.

## 6. DISCUSSION

Our system is designed for a professional percussionist for live performances, and thus time and care can be spent in optimizing the dictionary of signs. This would also be dependent on the piece of music and the importance of distinguishing similar signs.

Data from an accelerometer suffers from abrupt changes due to hand shaking [1]. In our case, the user moves a stick through space, and the added mass of the stick helps to mitigate this. In addition, this system is intended to be used by experienced percussionists, who are trained to have good hand coordination when using percussion mallets or sticks, this fact is indirectly utilized by our system because we use actual drum sticks that are augmented by the addition of a small antenna.

Musicians are specifically trained to produce repeatable actions in order to create sounds. By using a natural interface for percussionists, that of a drum stick, we leverage the large amounts of training that these musicians undergo.

## 7. REFERENCES


[1] A. Akl and S. Valaee, "Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, and compressive sensing," in *ICASSP*, 2010, pp. 2270 –2273.

[2] C. Cadoz and M. M. Wanderley, "Gesture-music." 2000.

[3] A. Corradini, "Dynamic time warping for off-line recognition of a small gesture vocabulary," in *RATFG-RTS'01)*, 2001, pp. 82–85.

[4] T. Darrell and A. Pentland, "Space-time gestures," in *CVPR '93*, Jun. 1993, pp. 335 –340.

[5] J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uwave: Accelerometer-based personalized gesture recognition and its applications," *IEEE International Conference on Pervasive Computing and Communications*, pp. 1–9, 2009.

[6] M. Mathews and A. Schloss, "The radio drum as a synthesizer controller," in *ICMC*, 1989.

[7] R. Morales-Manzanares, E. F. Morales, R. F. Dannenberg, and J. F. Berger, "Sicib: An interactive music composition system using body movements," *Comput. Music J.*, vol. 25, pp. 25–36, 2001.

[8] D. J. Overholt, "Musical interface technology: multimodal control of multidimensional parameter spaces for electroacoustic music performance," Ph.D. dissertation, Santa Barbara, CA, USA, 2007.

[9] S. L. Star and J. R. Griesemer, "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39," *Social Studies of Science*, vol. 19, no. 3, pp. 387–420, August 1989.

[10] A. R. Tindale, A. Kapur, G. Tzanetakis, P. Driessen, and A. Schloss, "A comparison of sensor strategies for capturing percussive gestures," in *NIME 2005*, 2005, pp. 200–203.

[11] J. Wu, G. Pan, D. Zhang, G. Qi, and S. Li, "Gesture recognition with a 3-d accelerometer," in *Proceedings of the 6th International Conference on Ubiquitous Intelligence and Computing*, ser. UIC '09, Berlin, Heidelberg, 2009, pp. 25–38.