CLAP DETECTION AND DISCRIMINATION FOR RHYTHM THERAPY

Nathan Lesser

LabROSA, Dept. of Elec. Eng.
Columbia University
New York NY 10027 USA
nathan@ee.columbia.edu

Daniel P.W. Ellis

LabROSA, Dept. of Elec. Eng.
Columbia University
New York NY 10027 USA
dpwe@ee.columbia.edu

ABSTRACT

An auditory training system relies on determining how well individual users can clap their hands together 'in time' with a prompt. Because the system is intended for a scenario in which an entire class of students is simultaneously engaged in this training, each system must distinguish between the claps of a single user and background claps from other nearby users. Available cues for this discrimination include the absolute energy of the clap sound, its source azimuth (estimated from stereo microphones), and its range as conveyed by the direct-to-reverberant energy balance. We present a set of features to capture these cues, and report our results on detecting and distinguishing 'near-field' and 'far-field' claps in a corpus of 1650 claps recorded in realistic classroom environments. When room and location are matched between training and test data, the classification error rate falls as low as 0.13%; when training data is recorded from a separate room, the error rate is still below 4.8% in the worst case.

1. INTRODUCTION

Automatic detection of the spatial origin of sound sources is a challenging problem, particularly in real-world, reverberant situations. The most difficult spatial dimension to recover is *range* i.e. the absolute distance between sensors and source. Unlike azimuth and elevation, range has little or no discernable effect on relative timing at closely-spaced sensors such as a stereo microphoine or the ears of a listener. Listeners show a surprising ability to judge range, which appears to depend on the relationship between direct-path sound and reflections from walls – the "direct-to-reverberant sound ratio" [1].

This paper describes a system where automatic classification of source range was required, and measures designed to capture the balance of direct and reverberant energy proved the most successful. The application stems from a novel therapy based on an observed link between rhythmic skill and ability to focus, and, more specifically, that improving a child's rhythmic acuity through rhythmic training can cause improvements in attention and general motor planning skills. An exercise can be as simple as clapping in time with a periodic stimulus, but a real-time 'score' giving feedback on how accurately the beat is being followed will provide extra motivation to improve. Detecting the times of claps from a single child in front of a computer is relatively easy, but it would be most convenient to build a system that could be used independently but simultaneously by several children in a single classroom. In this case, the system has the problem of discriminating between the claps of the nearest child (the one interacting directly with that machine), and other claps from the other children in the room.

Our focus, then, is on detecting clap sounds recorded by a simple, portable microphone setup in a typical classroom, and discriminating between 'near-field' claps (originating directly in front of the microphones), and 'far-field' claps (coming from other locations, as if from other students). The next section describes our overall system and the cues we devised to help this discrimination. Section 3 describes the evaluation data we collected in two real classrooms, and section 4 gives the results of our evaluation experiments.

2. SYSTEM DESCRIPTION

The system breaks down the problem of detecting near-field claps into two stages: first, candidate clap events are extracted by a simple process of looking for a rapid transient in energy. Secondly, these candidate events are classified by a statistical model as either near-field or far-field claps - which, in our experiments, were the only transients present in the recordings. A fast and simple first-stage transient detector minimizes the expense involved in executing a more complex second-stage classifier.

2.1. Transient Detection

The first stage of our detector differentiates between silence and clap events, with the latter being passed forward to the statistical classifier. We calculate the energy in 25 ms rectangular windows every 20 ms, then calculate the ratio between energy in successive windows. A simple threshold on this value detects the transients with high reliability. For the experiments reported below, the threshold was set to give exactly the known number of claps in each recording, which resulted in zero errors. In a practical system, an adaptive threshold would be devised.

2.2. Clap Classification

In the second stage of the detector, we need to discriminate between near-field and far-field claps i.e. those originating from the user directly in front of the microphones, and those coming from students elsewhere in the classroom. We approach this as a standard statistical pattern recognition task [2], and our attention is focused on the definition of cues that will compactly and reliably distinguish between these cases.

Ellis is supported by the National Science Foundation under Grant No. IIS-0238301.

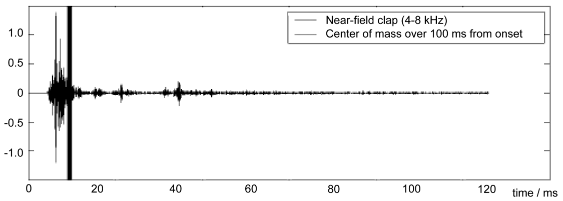

Fig. 1. Example near-field clap showing the center-of-mass (over the 100 ms following the energy peak) as a thick vertical line.

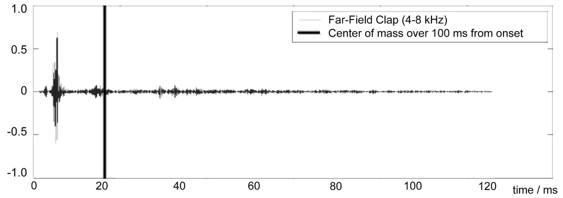

Fig. 2. Example far-field clap with 100 ms center-of-mass.

We employed four types of features, described below. The first two aim to capture direct-to-reverberant information, and the other two look at azimuth and energy respectively.

2.2.1. Center of Mass

The 'center of mass' statistic is simply the first moment of the energy of the signal within some window – the point on the time axis which 'balances' the energy in the window – following the time of the peak energy in the frame where the energy transient is detected, which is taken as the nominal onset time for the clap. Thus, if the local peak energy of waveform $x[n]$ occurs at t_o, the center of mass

$$t_c = \frac{1}{\sum_{n=t_o}^{t_o+t_w} x^2[n]} \sum_{n=t_o}^{t_o+t_w} n \cdot x^2[n] \tag{1}$$

where t_w is the window over which the center of mass is calculated. We use two windows, 20 ms and 100 ms, to capture both the characteristics of the initial clap burst, and the first part of the reverberant tail (and how it balances the energy of the initial burst).

The two t_c values are calculated for the signal in each of three frequency bands, 0-2 kHz, 2-4 kHz, and 4-8 kHz, yielding a total of six values for each clap event. While all these values are passed to the classifier, a pilot investigation showed that most of the discriminability is provided by just two or three of the dimensions. Figures 1 and 2 show example waveforms of near-field and far-field claps, respectively, with the center of mass in the 100 ms post-onset window indicated in each case. As expected, the near-field clap has its energy concentrated much closer to the initial onset point, whereas the far-field clap is centered almost 20 ms after the onset, due to the energy in the reverberant tail.

2.2.2. Slope

The slope calculation is a first order (linear) approximation of shape of the sound after the onset (i.e. the decaying portion). As the distance between clap source and microphones increases, the energy decay becomes more gradual as a result of the relative dominance of reverberation compared to direct sound. As with the center-of-mass, we fit two slopes, over 20 ms and 100 ms following the onset time, to measure the decay rate both of the initial burst, and the longer sound including reflections. Each slope is calculated as:

$$s = \frac{1}{\sum_{n=t_o}^{t_o+t_w} (n-\overline{n})^2} \sum_{n=t_o}^{t_o+t_w} 10log_{10}(x^2[n])(n-\overline{n}) \tag{2}$$

where $\overline{n} = t_o + t_w/2$ i.e. the average value of the time index in the summation. We fit the slopes to the log-domain energies (i.e. in deciBels) so that an exponential decay (as might be expected from a simplified model of reverberation) will be well fit by a straight line. The two slopes are again fit to each of the three frequency bands; combined with the centers of mass, this gives a total of 12 features per clap event.

2.2.3. Cross-Correlation

Whereas the center of mass and slope features attempt to allow the classifier to distinguish based on the range of the clap source (by characterizing the direct-to-reverberant balance), the direction from which the clap originates may also help to reject spurious claps, since it is assumed the user is seated directly in front of the equipment. By cross-correlating the signals from two microphone elements placed side by side on the equipment, then recording the lag corresponding to the peak cross-correlation, an estimate of the best-fitting time delay between the two versions of the clap sound is obtained. For sound sources directly ahead (i.e. in a direction normal to the inter-mic axis) this delay should be approximately zero; larger positive or negative time differences indicate sound from an off-axis direction that should be rejected [3, 4, 5]. This is the only feature that requires a second microphone so we refer to it as the stereo feature.

2.2.4. Energy

We have also tried using features relating to the energy of the clap burst; the energy ratio used in the initial transient detection (section 2.1) may contain further information to help distinguish near and far claps; the main influence here is likely to be that distant claps simply have a lower initial energy (and hence lower ratio compared to the preceding background noise) than the very intense near-field claps.

2.2.5. Classifier

The features for each clap event are concatenated and passed to a Regularized Least-Squares Classifier (RLSC) [6]. This classifier is closely related to the better-known Support Vector Machine (SVM) classifiers [7] in that it finds a discriminating hyperplane for a projection of the data points into some kernel space. RLSCs have been shown capable of performance very similar to SVMs, but they are simpler to create, involving only a matrix inversion rather than a quadratic-programming optimization. We used a publicly-available Matlab implementation [8].

We performed initial experiments to determine how much data was needed for training the classifier. Eventually, we ended up collecting quite large data sets (many hundreds of clap events), and using half of the data in any particular condition for training. The large number of training points compared to the data dimension

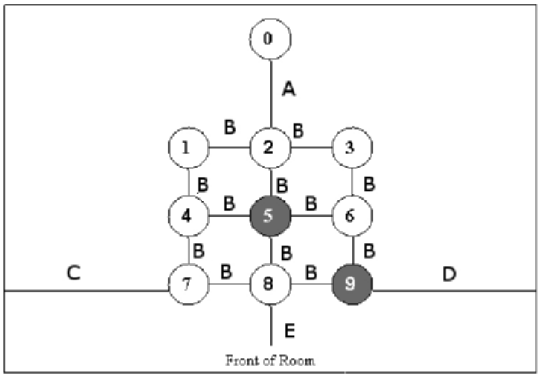

Fig. 3. Recording setup: Source location of claps and location of microphones. The same topology was used in both rooms, but the dimensions were different. For room 1, A = 2.4m, B = 1.4m, C = 2.7m, D = 2.8m, E = 3.7m. For room 2, A = 2.3m, B = 1.3m, C = 2.0m, D = 2.0m, E = 2.3m. Recordings were made at the two shaded locations, 5 and 9; claps originated from all ten locations

allowed the use of a linear kernel and gave robust and stable classifier results.

3. DATA COLLECTION

To train our classifier and evaluate its performance, we collected data from two classrooms with acoustics typical of our intended application. To allow the comparison of near-field and far-field claps, we recorded claps made at a number of points throughout the room, and repeated the recordings with the microphones at two different locations. This layout is illustrated in figure 3.

The microphones were placed at positions 5 (central) and 9 (closest to a corner) to provide good examples of reverberation differences based on location. Claps were generated by a person moving among all 10 numbered locations throughout the room. Near-field claps are those recorded at the same position as the microphones. Thus when the microphones are at position 5 and the person clapping is also at position 5 all resultant claps will be considered near-field. Far-field claps are all others. The clap pattern does not vary and consists of 50 claps, spaced approximately one second apart, at each location.

Recordings were made with a pair of Samson microphones (similar to Shure SM58) connected to an M-Audio MobilePre combined preamp and A/D converter. The digital audio streams were stored directly on a laptop computer. Due to the wide dynamic range of clap events, some care was needed in adjusting recording levels to avoid clipping distortion. We made a number of pilot recordings before achieving successful collection of the datasets reported on here.

For each microphone position, we collected a total of 900 claps in room 1, divided into two sets of 225 near-field and 225 far-field claps (i.e. 25 at each of 9 far-field locations), so that one set can be used for training and the other for test. In room 2 we collected two sets of 150 near-field plus 225 far-field claps, for a total of 750 clap events. The results reported below are the average, for each particular train/test configuration, of training on the first set / testing on the second set, and then training on the second set and testing on the first set. Thus, although all claps serve as both train and test examples, there is a strict disjunction between train and test data in any given experiment.

4. EXPERIMENTS

Tables 1 through 4 and the top-left quadrant of table 5 compare the effect of using different features. As a baseline, we used just the energy ratio features to see how well this simple feature could discriminate near from far-field. Table 1 shows four error rate percentages for each combination of training and testing with the data recorded at positions 5 and 9 in room 1. Thus, the values on the diagonal represent matched conditions where a classifier trained on near and farfield claps from a particular point in the room is tested on (disjoint) data recorded from the same position, and the off-diagonal values describe the two cases in which a classifier is tested on data recorded at a different point (in the same room). For these simple energy cues, we see little influence of the position at which data was recorded, and a large error rate of more than 20%.

Table 2 gives the results of classification based on the 12 slope and center of mass features that reflect the direct-to-reverberant ratio (2 centers and 2 slopes for each of 3 frequency bands). Accuracies are greatly improved, down to below 0.5% error in the best case. Now we notice a strong influence of the test set, with the claps recorded at position 5 (room center) resulting in significantly greater error rates than those from position 9 (corner). (Statistical significance requires a difference of around 2% absolute for these results).

Tables 3 and 4 extend the feature vectors with the cross-correlation (stereo) and energy ratio features, respectively. We see little improvement from this extra information, and the stereo features actually worsen the performance in some cases, albeit insignificantly. The top-left corner of table 5 gives the results using all features (slope, center, stereo, energy) which are best overall, although insignificantly different from the slope and center features alone.

The remainder of table 5 uses the same set of features but introduces data recorded from the second room. The bottom right quadrant corresponds to the top left where both train and test data are taken from room 2: we see significantly lower error rates for the more difficult position 5 test data (less than half the comparable values in room 1), but otherwise a very similar trend, with position 9 test data much easier, and little difference effect of the position used for training data. The two remaining quadrants, shown shaded in the table, are the most interesting cases since they represent cross-room conditions, where test data is taken from a different room than the training data. We see little impact of this mismatch, although training on room 2 data actually improves the recognition of data recorded at the more difficult position 5 in room 1, with these improvements on the edge of statistical significance.

As the results in table 5 show that the test set is the most significant factor in overall performance, table 6 pools the results of all four classifiers (rows of table 5) to break down the sources of error in a little more detail. We see that the majority of misclassifications come from false accepts of the room 1 position 5 data; a further examination of these errors show they arise mostly from the two flanking positions. Position 9 test data had error rates up to an order of magnitude lower, and room 2's position 5 test data was significantly easier to discriminate than the corresponding position in room 1.

		TEST	
		Pos 5	Pos 9
TRAIN	Pos 5	22.67	21.78
	Pos 9	22.67	21.00

Table 1. Baseline classification error rate percentages for room 1 data using energy ratio cues.

		TEST	
		Pos 5	Pos 9
TRAIN	Pos 5	6.00	0.44
	Pos 9	7.45	0.67

Table 2. As table 1, but using the center of mass and slope features (only).

		TEST	
		Pos 5	Pos 9
TRAIN	Pos 5	5.78	0.56
	Pos 9	7.45	0.78

Table 3. As table 2, but including the cross-correlation (stereo) feature.

		TEST	
		Pos 5	Pos 9
TRAIN	Pos 5	5.45	0.33
	Pos 9	6.67	0.56

Table 4. As table 2, but including the energy ratio feature (no stereo feature).

			TEST DATA			
			Room 1		**Room 2**	
			Pos 5	Pos 9	Pos 5	Pos 9
TRAIN DATA	**Room 1**	Pos 5	5.11	0.33	2.40	0.27
		Pos 9	6.78	0.56	3.87	1.20
	Room 2	Pos 5	3.64	0.44	2.40	0.13
		Pos 9	4.78	0.33	2.53	0.40

Table 5. Classification error rate percentages based on all features combined (energy ratio, center of mass, slope, and stereo), and testing both within and between rooms. Shaded cells represent cross-room conditions, where training and test data come from completely separate recording sessions.

	TEST DATA			
	Room 1		**Room 2**	
	Pos 5	Pos 9	Pos 5	Pos 9
’Near’ claps	450	450	300	300
’Far’ claps	450	450	450	450
Total # false accepts	177	2	83	15
Total # false rejects	4	13	1	0
Av. error rate %	5.08	0.42	2.80	0.50

Table 6. Error summary and breakdown for the different test sets, pooled across classifiers trained on all four training sets.

5. DISCUSSION AND CONCLUSIONS

Our results show that accurate discrimination between local and distant clap events can be achieved with high accuracy for our application. Features designed to capture the balance between direct and reverberant sound by looking at the decay of energy after the initial peak – the center of mass and slope estimates over 20 ms and 100 ms windows in three octave subbands – were by far the most useful, as predicted by insights from human range perception. The energy ratio feature gave some improvement, although not statistically significant in this test; the cross-correlation cue appeared to give no benefit at all, possibly because the complex reverb characteristics of the real classrooms made the time delay estimates very noisy.

Our experiments differ from a real application in several respects. Firstly, we have trained on data from only one room at a time; we would expect the best generalization from a model trained on a pool of data from different positions in different rooms. However, even our limited training sets show good generalization: not only do the cross-room tests perform as well as training and testing on data from the same room, but we also observed that accuracy on classifying the training data themselves was little better, indicating the classifier is very far from being overtrained. The biggest departure from a real condition is that our near-field and far-field claps always occurred distinctly – we did not record simultaneous clapping at multiple locations, although we did do some experiments on mixing clap recordings to simulate this. When the claps were distinct in time performance was not affected, but overlapping claps present a more complex challenge. In particular, it might be hard to detect a distant clap that overlaps with near-field claps, although this may not matter for our application. A real implementation would need a more sophisticated way to sort candidate clap events from non-clap transient noises that might be encountered.

We conclude that accurate detection of near-field claps in a realistic, multi-student classroom environment appears feasible, and most of the accuracy can be obtained with just a single microphone per setup, which is often already present on standard laptops. As well as being a useful solution for the specific rhythmic training application that motivated the study, this work has wider interest as a successful example of discrimination of source range from single-microphone cues.

6. REFERENCES

[1] Donald H Mershon and John N Bowers, “Absolute and relative cues for the auditory perception of egocentric distance,” *Perception*, vol. 8, pp. 311–322, 1979.

[2] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification, 2nd Edition*, Wiley, New York, 2001.

[3] A. Grennberg and M. Sandell, “Estimation of subsample time delay differences in narrowband ultrasonic echoes using the hilbert transform correlation,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 1994.

[4] Doh-Hyoung Kim and Youngjin Park, “Development of sound source localization system usning explicit adaptive time delay estimation,” *International Conference on Control, Automation and Systems*, 2002.

[5] H.M. Gross C. Schauer, “A model of horizontal 360 degree object localization based on binaural hearing and monocular vision,” 2001.

[6] Ryan Rifkin, Gene Yeo, and Tomaso Poggio, *Regularized Least-Squares Classification*, vol. 190 of *Advances in Learning Theory: Methods, Model and Applications, NATO Science Series III: Computer and Systems Sciences*, chapter 7, pp. 131–153, IOS Press, Amsterdam, 2003.

[7] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[8] Jason Rennie, “Matlab code for regularized least-squares classification,” 2004, `http://people.csail.mit.edu/u/j/jrennie/public_html/code/`.