Occupancy Estimation using Ultrasonic Chirps

Oliver Shih Anthony Rowe
Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh PA, USA
{oshih, agr}@andrew.cmu.edu

ABSTRACT

Estimating the number of people within a room is important for a wide variety of applications including: HVAC load management, scheduling room allocations and guiding first responders to areas with trapped people. In this paper, we present an active sensing technique that uses changes in a room's acoustic properties to estimate the number of occupants. Frequency dependent models of reverberation and room capacity are often used when designing auditoriums and concert halls. We leverage this property by using measured changes in the ultrasonic spectrum reflected back from a wide-band transmitter to estimate occupancy. A centrally located beacon transmits an ultrasonic chirp and then records how the signal dissipates over time. By analyzing the frequency response over the chirp's bandwidth at a few known occupancy levels, we are able to extrapolate the response as the number of people in the room changes. We explore the design of an excitation signal that best senses the environment with the fewest number of training samples. Through experimentation, we show that our approach is able to capture the number of people in a wide-variety of room configurations with people counting accuracy below 10% of the maximum room capacity count with as few as two training points. Finally, we provide a simple mechanism that allows our system to recalibrate when we know the room is empty so that it can adapt dynamically over time.

Categories and Subject Descriptors

C.3 [**Special-purpose and application-based system**]: Real-time and embedded systems

General Terms

Algorithm, Design, Experimentation

Keywords

Occupancy detection, ultrasonic sensing, machine learning

ICCPS '15, April 14 - 16, 2015, Seattle, WA, USA

http://dx.doi.org/10.1145/2735960.2735969

1. INTRODUCTION

Being able to accurately count the number of people in a space has high utility for a number of applications. In building automation systems, knowing if a room is occupied or not can be used to control zone heating and cooling or simply disable unused lighting. In more advanced systems with variable drive air handling units, knowing the number of people (load) can be used to more accurately control temperature and ventilation to save energy. In the context of large facilities like conference centers or in the retail space, knowing how many people are in certain locations and how long they dwell can be used to value shelf-space or storefront locations. In the event of an emergency, first responders often need to know if people are trapped and where they might be located in large buildings. These applications require a sensor capable of counting how many occupants are within a space.

There are currently many approaches for measuring occupancy in spaces including: passive infra-red (PIR) sensors, ultrasonic ranging sensors, microwave sensors, smart cameras, break beam sensors and laser range-finders. These devices span across a wide spectrum of cost and performance. Lower-cost alternatives, like PIR and ultrasonic ranging sensors, are typically error-prone and usually only detect binary occupancy values rather then estimating load. More expensive sensors tend to require sophisticated site-specific installation and calibration approaches.

In this paper, we introduce an active ultrasonic sensing approach for estimating the number of people in a space. It is well known from the acoustics community that the number of people within a room impacts the reverberation of sound. Reverberation is typically defined by the RT_{60} time constant which is measured as the amount of time it takes for a signal to decrease by 60dB [1] (in early experiments by Sabine at Harvard, this was the amount sound decrease before organ pipes became inaudible). When designing concert halls, musicians quickly realized that not only did the number of people in the audience significantly impact reverberation, it was also frequency dependent. People in the audience act like sound absorber which reduce the amplitude of reflections. As early as the 1890's, Sabine began to model the impact of people, frequency and the geometry of spaces on reverberation [2]. Many concert halls have been designed to sound their best when full of people and don't sound nearly as good when empty. Sabine often modelled rooms in terms of per-person audience absorption. We propose leveraging the change in this reverberation phenomena in the ultrasonic frequency range as a way to silently (to humans) sense occupancy.

Reverberation is both frequency dependent and changes based on the room geometry, wall materials and furniture material. Making accurate and generalizable models of reverberation is quite challenging. For this reason, we propose an approach where the reverberation is trained on a per-room basis using a machine learning approach. Instead of measuring and classifying the reverberation at discrete frequencies like what is done for concert halls, we use ultrasonic chirps that sweep across a frequency range to rapidly measure the space since we are not concerned with exactly quantifying reverberation. Chirps can also be constructed using fade-in and fade-out periods that prevent audible artifacts in low-cost speakers that could be detected by humans [3]. Since the reflections coming back from these signals are room specific, we apply a semi-supervised machine learning approach that is able to model the characteristics of the room under multiple loads in order to estimate how reverberation changes with respect to number of people. Typically this requires taking samples when the room is empty as well as when the room has enough people to make a significant difference in reverberation times. Alternative signal characteristics like Doppler shift or simply time of day schedules can be used to determine when the room is empty for periodic re-calibration of the zero point.

Figure 1 shows an overview of our proposed system where a tweeter transmits an ultrasonic chirp into a room and a co-located microphone is used to receive the reflected signal. An electronics package is responsible for generating the signal and then processing the reflected signal. Our prototype system uses a computer for this purpose, but we show that the actual run-time computation of the system is simple enough to execute entirely from a platform based on a micro-controller.

There are four main research challenges associated with our proposed system. First, we need to design an appropriate excitation signal that is both inaudible to humans and also excites the room in a manner that can clearly distinguish changes as the number of people increase. Second, we need a technique that can sample quickly and efficiently so that occupancy can be estimated before the dynamics of the room change. This approach also requires a transducer that is able to uniformly distribute the ultrasonic signal. Third, we need algorithms that can classify received signals in order to estimate load. Finally, we need an approach that can periodically retrain in order to adapt to slight changes in the environment over time.

2. RELATED WORK

In this section, we discuss the background related to acoustics followed by similar approaches that have been used to measure both presence and occupancy. Common commercially available occupancy sensors like PIR motion detectors; ultrasonic motion detectors and microwave sensors usually only detect presence (if one or more people are in a room). Cameras and more advanced infrared systems attempt to estimate the actual number of people in a space, but are typically expensive, difficult to train and suffer from occlusion. Our proposed approach is comparatively low-cost, relatively easy to train and has the advantage of filling an entire space with sound making it more immune to obstacles.

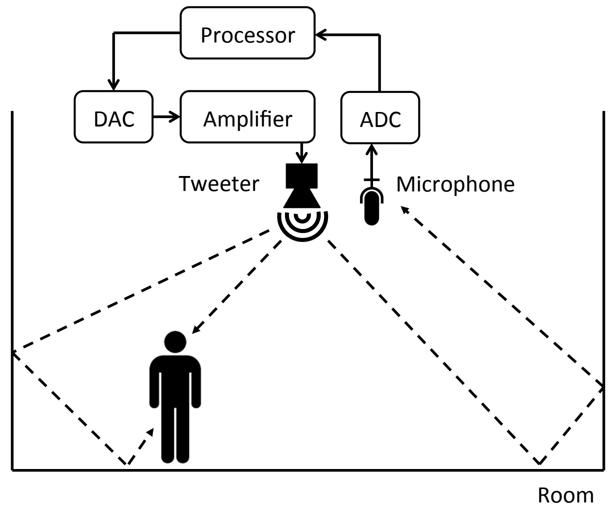

Figure 1: System overview

2.1 Acoustics

Seminal work in acoustics has shown that people in a space significantly impact reverberation and that reverberation is frequency [2] as well as room geometry dependent [4]. Over the last 120 years there have been countless efforts proposed to model these acoustic properties in order to improve concert hall performance. Recent work in this space has used computer simulations [5–8]. It is clear from this large body of research that creating simple generalizable models of reverberation is quite challenging. For this reason, we propose using machine learning techniques to learn and classify the reverberation response on a per-installation basis. In various recent profiles of reverberation [9], it is clear that given a particular room geometry, audience absorption follows relatively distinct curves that make it an ideal feature for occupancy detection.

Active acoustic approaches have shown great potential in multiple forms of sensing. In [10], the authors use a single speaker with multiple microphones to determine the shape of a room based on echoes. In [11], the authors show how reflected Doppler signals can be used to classify anything from speech, to walking motion and even gestures. To the best of our knowledge, this is one of the first system where ultrasound has been used to directly estimate occupancy.

2.2 Occupancy

Aside from the conventional solution of using PIR sensors to detect the presence of people, most other related work has been carried out on using cameras or multiple sensors to measure occupancy level. All of these approaches generally fall into two categories based on slightly different goals. One group focuses on only detecting the presence of people [12] [13] [14] [15], which often comes with analysis of more detailed user behavior and actions. The other categories focuses on people counting [16] [17] [18] [19], usually involving more sophisticated algorithms for learning.

Presence Detection

In the category of presence detection, many approaches fuse data readings from different sensor types. For example in

[20], the authors combines multiple available sensors feeds of data to estimate occupancy. In [21] the authors focus primarily on WiFi signals. In both cases, the approaches do not perform as well in large spaces like auditoriums unless each occupant is carrying a mobile device that cooperates with the system. Two of the recent works use similar approach by utilizing ultrasonic signals [12] [13]. In [12], the author proposed a sonar system using four microphones and a single frequency sinusoid of 20 kHz in order to detect the user's attention state and several pre-defined activities. The classifier is built by characterizing the *echo delta*, namely the variance in intensity, of the reflections from user's body. Their experimental results show supportive evidence that a user's presence impacts the intensity of the echoes, which is a fundamental characteristic we assume in our approach. Nevertheless, this techniques requires copious amount of training data to predict the pre-defined activity, and assumes the environment to be free from interference.

Similar work in [13] proposed an ultrasonic array sensor and tracking algorithm to detect presence and capture the movement of targets. This is achieved by taking the difference in the received echo signal to estimate direction-of-arrival (DoA) with the array of sensors, and utilizing the received signal to noise ratio (SNR) as an indicator of occupancy. A simple tracking algorithm is also proposed to increase performance of presence detection. While this method shows better performance than PIR sensors, the detection zone is limited to a certain area and confined by DoA angle. Other approaches proposed in [14] and [15] take advantages of using multiple co-located sensors. In [14], $TelosB$ motes are deployed with pressure sensors, PIR sensors, and audio sensors. The system is able to predict pre-defined activities by correlating the binary readings from multiple sensors. The overall classification accuracy is more than 90%, but it requires careful deployment of multiple sensors at different locations in the room. Similar in the choice of sensors, the author in [15] adopts additional light and CO_2 sensors. Classification is done using a decision tree in order to determine which sensors are most important. The results indicate that the motion sensor is dominant, and accounts for 97% of accuracy even when used alone.

Although most of the presence detection techniques have the advantage of low-cost and low-complexity, their applications are limited due to the coarse resolution. Based on the proposed methods, they also suffers from scalability and deployment difficulties due to the confined detection area of the sensors.

People Counting

The most common solutions for people counting tend to use cameras [16] [17] [18]. An early work for fine-grained indoor people counting is presented in [16], where the locations of the objects are first measured by their silhouettes from image sensors deployed around the room. The system shows accurate results up to 12 people moving in a room, but requires careful placement of multiple image sensors. Also, the computational complexity grows proportionally to the number of sensors. For counting larger groups of people, a crowd counting algorithm proposed in [17] shows accurate results for tens of pedestrians with an error of less then 2 people. The algorithm also claims to be privacy preserving by segmenting the crowd into groups using low-level features, and then using a regression model to count people within each

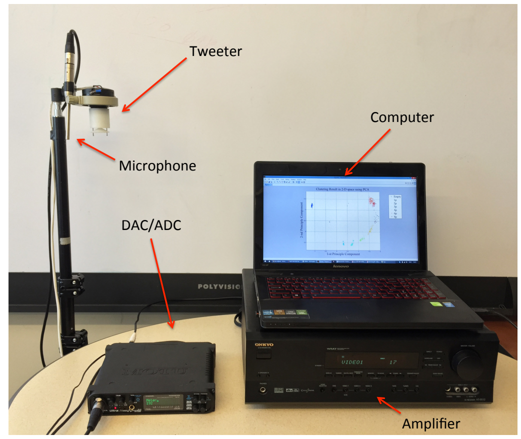

Figure 2: Experimental setup

segment. A pedestrian database is required for providing a large number of training images, which is often costly and thus makes it less feasible in more constrained use cases like on an embedded sensor. Recently, Chen *et al.* [18] proposed a semi-supervised learning method for crowd counting to reduce the effort in acquiring labeled data. The algorithm first performs a spectral clustering on the unlabeled data to pick out the most representative data for labelling and then uses feature mapping to facilitate learning of a new target model. This concept enables the use of knowledge from a previous scene and thus reduces required training data for bootstrapping learning in the new scene, but the assumption is that the two scenes must share similar manifold representations. In [19], the authors evaluated three different learning methods Support Vector Machine (SVM), Neural Network (NN), and Hidden Markov Model (HMM) over dozens of different sensor inputs, and are able to estimate $0-3$ occupants in an open office area with 75 % accuracy.

To summarize, although most of the presence detection techniques have the advantage of low-cost and low-complexity, they only provide a coarse resolution of people within a space. In contrast, most of the people counting techniques are either more expensive in terms of cost and complexity, suffer from privacy issue, or require large labelled databases. To the best of our knowledge there is no existing framework that can perform wide area people counting with a single cost-effective and versatile sensor.

3. SYSTEM ARCHITECTURE

The overall system is composed of three main parts: (1) the transducer with microphone and tweeter, (2) the signal that is transmitted into the room being tested and (3) the algorithm that classifies the response. Figure 2 shows our experimental setup that consists of a laptop computer, Motu UltraLight MK3 DAC and ADC, an audio amplifier with a flat frequency response up to 100kHz, an omni-directional tweeter and a measurement microphone sensitive up to 30kHz. The DAC is capable of 24bit 192kHz audio input and output. Figure 3 shows an enlarged view of the measurement microphone (Audix TM1) along with a custom ultrasonic horn tweeter. For testing purposes, we use relatively high-end audio equipment but as described in our evaluation section when the signals were artificially degraded in terms

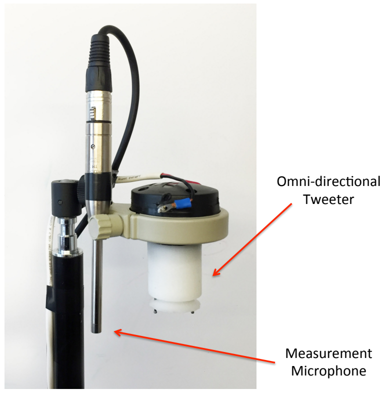

Figure 3: Microphone and tweeter detail

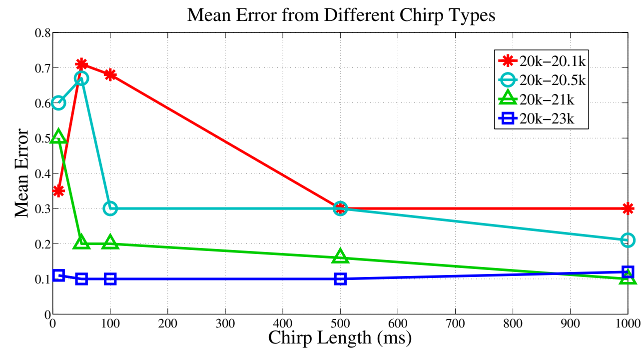

Figure 4: Impact of chirp length on classification error

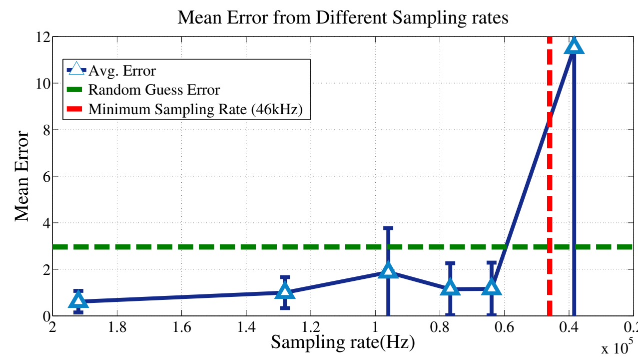

Figure 5: Impact of sampling rate on classification error

of sampling rate and bit resolution we see little impact on performance. We believe that this system could eventually run on a low-cost module consisting of a MEMs ultrasonic microphone and a consumer-grade audio DAC and amplifier. Our motivation for using audio equipment just above the human hearing range is predominantly that of cost and electronic design complexity. Between 20kHz and 24kHz, it is possible to use standard 24kHz audio amplifiers, DACs and ADCs. Higher frequencies tend to become more directional, but can likely still perform well given adequate speaker geometry.

3.1 Impulse Signal

In order to test the response of the environment over a range of frequencies, we utilize *chirps* (sinusoidal signals that linearly increase in frequency).

Ultrasonic Chirps

Chirps exhibit pulse compression which is a common technique often used in RADAR systems to improve the ranging resolution. Chirps have a high correlation with themselves, and therefore can be easily detected with an increased SNR. Since the chirps naturally sweep across a frequency range, this allows us to conveniently collect the reverberation characteristics across a larger bandwidth in a single operation. In fact, the same approach can also be observed in nature. A number of bat species emit short but broad-band signals to differentiate the texture of their prey by the interference pattern reflected in echoes. As described in [3], many tweeter speakers exhibit non-ideal impulse responses that can result in audible artifacts like clicking sounds. To alleviate these problems, we add 10 ms of fade-in and fade-out time to the chirp's ramp up and ramp out time.

Bandwidth and Chirp Length

One would expect that a chirps' frequency and duration should have a direct impact on the performance of the system. Given more bandwidth, we should be able to collect more reverberation characteristics as the signal dissipates. The length of the chirps define the resolution of the frequencies we can acquire given a fixed sampling rate. In order to test bandwidth and chirp length, we collected 100 points of data for 0-5 people at four different bandwidths and five different chirp lengths for a total of 8000 samples. In Figure 4, we show the sensitivity of chirp length and bandwidth on our classifier. We will review the details of the classifier in the next section, but the important trend to see is that the performance is proportional to a bandwidth and time product. Based on these tests, we choose to use a chirp of at least 200 ms and a bandwidth of at least 1 kHz. Picking the minimum length and bandwidth helps scope the hardware requirements and maximizes sensing rate.

Sampling Rate

The minimum sampling rate to support the system is also an important factor driving both the cost of the hardware components and the computational requirements of receiving the signal. Generally speaking, normal commodity audio equipment designed for music only supports sampling rates up to 48 kHz. Also, the dispersion pattern of a lower ultrasonic frequency tends to be more omni-directional. As shown in Figure 5, a higher sampling rate has a slightly better overall performance and large error is expected when the sampling rate drops below the Nyquist limit. The interesting point to note is that the performance does not significantly increase when you go to much higher sampling rates than the input audio signal. This support the notion that our feature is likely based on the decay within our frequency band.

3.2 Preprocessing

Before attempting to classify data, the raw signals are preprocessed to minimize noises caused by multi-path or any audio sources to improve prediction accuracy. We apply the following filters to the received signal.

Matched Filter

We assume that the transmitted signal goes through an additive white Gaussian noise (AWGN) channel while disseminate in the room. In this case, the matched filter is known to be the optimal receiver filter to increase the signal-to-noise ratio (SNR) of the received signal. Here the signals can be represented as

$$y(t) = h(t) * x(t) + n(t) \tag{1}$$

where $y(t), x(t)$ is the received signal and the transmitted signal, $h(t)$ is the impulse response of the room, and $n(t)$ are the background noises. Since the transmitted signal is known and $h(t)$ is the target of interest, we match filter the received signal with the original transmitted signal to maximize the SNR. A high SNR of the received signals is vital for the later analysis with machine learning techniques, which identifies the most important characteristics in the frequency changes that differentiate the signals of different occupancy levels.

Bandpass Filter

The matched-filtered signal is then transform into frequency domain using Fast Fourier Transform (FFT), and is then band-pass filtered to remove noise from other acoustics sources. The filter's bandwidth is exactly the same as the chirps' sweeping bandwidth. Transforming into the frequency domain also helps to reduce the dimensions of the collected data and minimize the training time and complexity.

3.3 Occupancy Estimation Algorithm

The people counting algorithm is composed of two parts. In the first part, we categorize all of the collected data into clusters using the density-based spatial clustering of applications with noise (DBSCAN) algorithm. In the next step, a regression model is built based on the two training points (*e.g.* empty room and 1 person measurement) and the clustering result. As explained below, two different regression model are used depending on the room size and the maximum capacity of the room.

Figure 6 provides a simple example of the types of features the algorithm is trying to identify. The top row shows the filtered spectrum after matched filtering of an empty room, a half-full room and a full room. The bottom row of the image shows the difference between each top image and the empty room sample. For example, an empty room shares little difference with another empty room and hence you see almost no changes in the signal. However, in the case of a half full and full room, we see a significant difference. It is worth noting that the difference between a half empty room and a full room is much more subtle.

In order to build a computationally effective model, principal component analysis (PCA) is applied to ensure all the pre-processed n-dimensional signals are projected into a n'-dimensional space, where $n' \leq n$ and all variables in the new space are linearly uncorrelated with each other. Furthermore, when projected into n'-dimensions, we are using the first n' principal components for transformation, where the first principle component is defined as the variable that gives the maximum possible variance in the dataset. Note that while lowering the dimensions of data reduces the overall complexity, more information is lost during the transformation process. Ideally, $5 \leq n'$ gives the best performance in clustering based on our empirical experiment, and the corresponding eigenvalue ratio representing the ratio of variance kept after transformation is around 25%.

Clustering

Once we have found the principle components of the signal, we need to cluster each identifier. The DBSCAN clustering algorithm [22] has been widely used in this manner due to its robustness to outliers and zero prior knowledge of the number of clusters. Moreover, we do not want to assume any prior distribution of people in the room since the real distribution can vary from day-to-day and largely depends on the usage and functionality of the room. These properties of DBSCAN allow us to cope with noise caused by different distribution of bodies in the room and successfully categorize the data with high accuracy. A limitation of using DBSCAN is that the clustering results are sensitive to the minimum neighborhood points and neighbor distance ϵ. In order to reduce the indeterministic outcomes and improve the quality of DBSCAN, each collected data point consists of multiple samples with a known number of chirps. Different neighborhood distances ϵ are also evaluated based on the intra-cluster distance derived from the training data, and the most frequent combination is selected as the clustering result.

The primary reason to cluster data before performing regression is to improve the prediction accuracy especially for smaller room environments. In most of the scenarios, the overall dataset are quite noisy and often overlapped with each other even in a high dimensional space. By clustering the data into groups and removing outliers, the accuracy in regression is drastically increased especially in cases with few people where we expect high granularity. Also, the computation complexity is greatly reduced since only the mean of each cluster is needed in building the regression model instead of computing on the whole dataset. The clustering algorithm also benefit from the chirps' physical characteristic. When using chirps with larger bandwidth, more reverberation information across the frequency band is learned in the training process. As a result, the density of each cluster is higher and inter-cluster distance is greatly increased in the observed data.

On the other hand, in larger rooms such as an auditorium, DBSCAN can failed to give a conclusive clustering result due to excessive scattered data points. However, these cases are often the ones where clustering algorithm will contribute the least to the results because the granularity of the estimation is relatively less important. The estimate will then rely mainly on the regression model, as discussed as follows.

Regression Model

In order to interpolate occupancy beyond the training data, we build a regression model based on only two labelled training points. One data point is when the room is empty, while the other data point should be at a reasonable occupancy level ($\geq 10\%$). Here we derive the relationship between the number of people, which can be seen as the absorption material in a room, and the amplitude difference in frequency with the help of the Sabine equation and reverberation properties found in [2]. As shown by the Sabine acoustic model (2), the duration of the audibility of the residual sound,

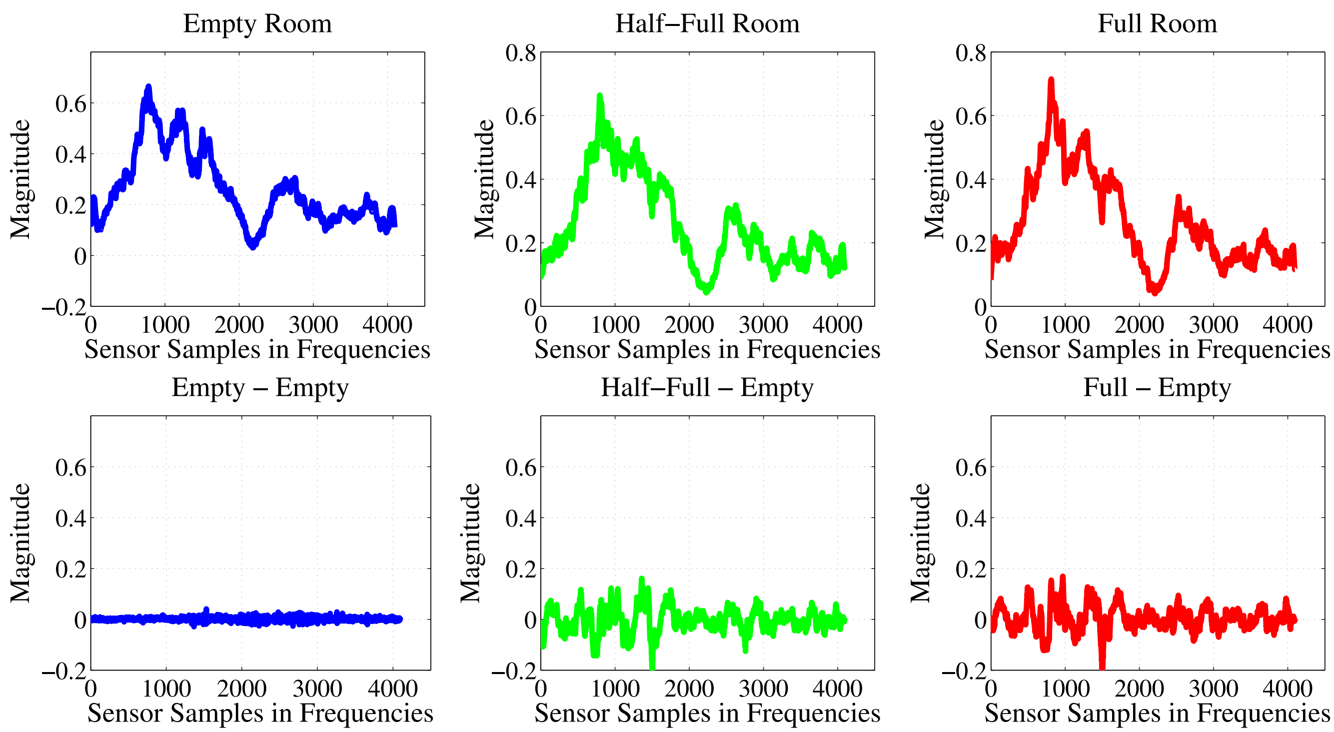

Figure 6: Raw features for empty, half-full, and full room scenarios

namely the reverberation time (RT), follows a rectangular hyperbola curve against the total absorbing material. Here c_{20} is the speed of sound at 20 degree Celsius, V is the volume (m^3) of the room, S is the total surface area (m^2) of a room, and a is the average absorption coefficient of room surface.

$$RT_{60} = \frac{24 \ln 10}{c_{20}} \frac{V}{Sa} \simeq 0.1611 \frac{V}{Sa} \tag{2}$$

Since the RT is defined by the time for a signal to decay by a certain decibel(dB), we get (3)

$$RT \propto \log(\frac{A_0}{A_m}) \tag{3}$$

where A_0 is the constant initial amplitude of the sound source and A_m is the measured amplitude after absorption. Combining equation (2) and (3), we obtain the relationship between the observed frequency amplitude and number of people as (4)

$$A_m \propto e^{\frac{-C_0 V}{Sa}} \tag{4}$$

As plotted in 7, we can see that when the volume of the room is small, the curve tends to be similar to an exponential regression. However, as the volume of the room increases, the curve becomes smoother and more linear in regression. The size of the room can be estimated to help choose the best starting model.

To calculate the amplitude difference, we first re-calibrate the mean of the empty room data as the new origin of the projected space, and for every clusters we calculate how far they are from the origin. We tested with multiple distance metrics and decided that *Chebyshev distance* provided the best fit to regression model shown across our overall data. We use the *Chebyshev distance* defined as,

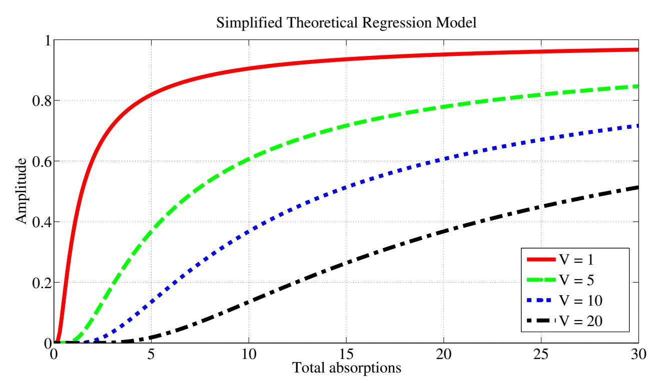

Figure 7: Theoretical regression trends with different room volumes based on equation (4)

$$D_{chebyshev}(a, b) = \max_{1 \le i \le n} (|a_i - b_i|) \tag{5}$$

where a, b are two arbitrary n-dimensional data points. The *unit distance* is further calculate based on the average of the pairwise-distance between the two training datasets, where the *unit distance* is namely the reference distance between N and $(N + 1)$ people instance. Next, we estimate each cluster by fitting its distance to the origin to the regression model. By finding the variable that changes the most among all the data, which noted here is derived from a linear combination of all the variables in the original space, we capture the feature that differentiates the data the most and used it as a measurement to estimate the occupancy level.

For rooms with a small volume, an exponential regression model (6) is adopted instead of a linear one for estimating the occupancy level based on previous observation. We define an exponential loss function to estimate the most likely capacity combination for each cluster. The loss function is given as (7)(8),

$$f(x) = \alpha e^{\beta x} \tag{6}$$

$$\hat{f} = \underset{\alpha,\beta}{argmin} \sum_{i=1}^{n} e^{W_i \phi(x_i)} \tag{7}$$

$$\phi(x) = f(x) - round(f(x)) \tag{8}$$

where n represents the total number of clusters, W_i is the weight of cluster i, and x_i is the distance between cluster i and origin (the empty room). The weight of each cluster W_i is proportional to the number of members in the cluster, and additional weights are also assigned to the clusters of the two training data. This allows the curve fitted to the most important clusters and prevents over weighting of outliers. Additionally, the function $\phi(x)$ tends to fit the curve in a way that the predicted number of people is close to an integer. By minimizing the loss function, we obtain the best prediction function $\hat{f}$ with corresponding parameter $\hat{\alpha}, \hat{\beta}$, and the estimated occupancy level for cluster i is then assigned accordingly by $\hat{f}(x_i)$. To speed up the process and improve the performance of fitting, we assume the maximum capacity of the room is given and the data collected should contain instances of at least half of the maximum capacity. This can be achieved by setting up a data collecting period, such as a day, in the system for bootstrapping before running the estimator. The idea is to have a self-learning system that requires minimal training effort and capable of training itself as more data is collected and learned over time.

It is worth mentioning that with more given training points, a more sophisticated regression model or semi-supervised learning such as in [18] can be adopted to improve the accuracy of the prediction. However, one of our goals in this paper is to minimize the training effort from the user to improve the feasibility and scalability of the system.

Though not thoroughly evaluated in this work, the problem of selecting the correct regression model for each room size can be determined by parameters derived from(4) or based on echo intensity. For now, we allow the installer to select small or large based roughly on square footage.

Auto Recalibration

To prevent retraining from scratch every time the background environment slightly changes, the system requires a mechanism to slowly recalibrate itself over time. Whenever empty room data is captured, we first project it to the space defined by the current model, and then we calculated the *unit distance* in the current model on the principle component that most significantly differentiates the new data. Since the projection does not alter the magnitude of the raw data, the *unit distance* needs to be further resized by the magnitude difference in the two spaces spanned by the different principle components. The ratio can be easily calculated by the inter-cluster distance of the empty room data, since the background noises should remain constant no matter how the room changes. Once the new principle component and *unit distance* is defined, estimation can be made by applying this delta to the regression model. In this

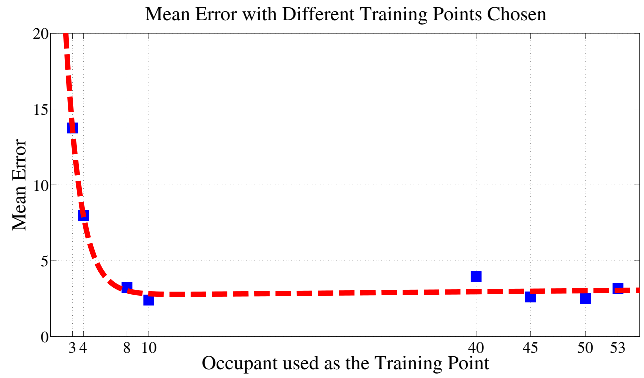

Figure 8: Accuracy based on number of occupants used as training sample in a 150-person room

manner, the system is able to retrain when the environment changes using only empty room training points.

4. EVALUATION

In this section, we evaluate the mechanism and training model based on experimental results with data captured in three different environments. We discuss several of the key design choices and how they affect overall system performance. We test the system's immunity to noises from the environment and its ability to adapt overtime when periodically retrained on an empty room (not the occupied cases).

4.1 Design Parameters

We discussed the sensitivity of different chirp parameters in Section 3.1. Based on our experiments, the chirps with a bandwidth of $20k - 23k$ and a length of greater then 200 ms gives the best performance. To be conservative, we selected 500 ms so as to increase the bandwidth-time product of the signal. Note that the upper bound of $23k$ is also considered as the highest frequency most common (non-ultrasonic) speakers can support. The interval between each chirp is set to 500 ms, allowing the chirp to fully dissipate in the room. This is significantly longer than what is needed as derived from the Sabine and Eyring equation [9]. This results in a 1 second sampling rate. The selection of the second training point can also affect the result dramatically in certain cases. The training point consists of a single person or a group of a few people is typically ideal for small and medium room scenarios. However, as show in Figure 8, using a small group of people as training point in large rooms is likely to cause significant estimation error. The error comes from the fact that such changes in frequency magnitude are not strong enough to be fully captured. A training point of a group of eight people or more in a 150-person room gives similar result with 5% of error in average. Based on our experiments, training points of at least 10% of the maximum capacity works well.

4.2 Experiment Environment

Figure 9 shows photographs of the three rooms where we ran our experiments. The first room is a conference room that seats less then 10 people. The second was a classroom that seats about 24 students and the final room was an 150 person auditorium. In each room, we chose a volume for

(a) Small conference room in Collaborative Innovation Center

(b) Medium-size classroom in Doherty Hall

(c) Auditorium in Hamerschlag Hall

Figure 9: Experiment environment

the transmitter that returned a similar in amplitude first reflection of the signal. We placed the transmitters and the receiver at four different locations around the room including the sides, middle and front of the room. We evaluate several different locations of the transceiver and see similar result as long as the transceiver is at least 1 meter away from the walls. In each room, we collect between 5 and 10 different occupancy levels each with 100 samples. The training points consist of 50 samples for each of the two training levels. While collecting data, the occupants where free to use computers, give presentations, or walk around the room as usual.

4.3 Performance Results

Figure 10 shows the two-dimensional PCA projection results for 0-8 people in the small room. Each color and marker type reflects the clustering of different occupancy levels. Most of the clusters are correctly categorized except for a few points that are associated with the eight person case due to noise. In the figure, we can also see that as the number of people in the room increases, the dynamic distribution of people leads to a higher variance in the clusters.

Figure 12 shows the occupancy estimation made by the exponential and linear regression algorithms respectively in small room and medium-size room scenario, as described in section 3.3. Each data point represents the estimation for an entire cluster, each of of which consists of at least 100 sample points. As we can see in the figure, the error slightly increases as the room size gets larger, but we are still able to achieve an error of less than 2 people from the average ground truth.

In order to test the system on larger rooms, we carried out an experiment in an auditorium before the start of a class. We periodically sample every 10 seconds while students enter the auditorium. Ground truth was captured with a camera that was hand annotated. Figure 13 shows that the estimate tracks the ground truth quite well. Moreover, the system is responsive to rapid dynamics of the environment; the sudden boost in the estimated occupancy level happens right after a large group of students swarmed into the classroom.

Next, we evaluated how the system works in the presence of various error sources. Most importantly, we show how the system performs when a room changes over time. We only evaluate the interference in a small room scenario, since we believe this is where the interference would most

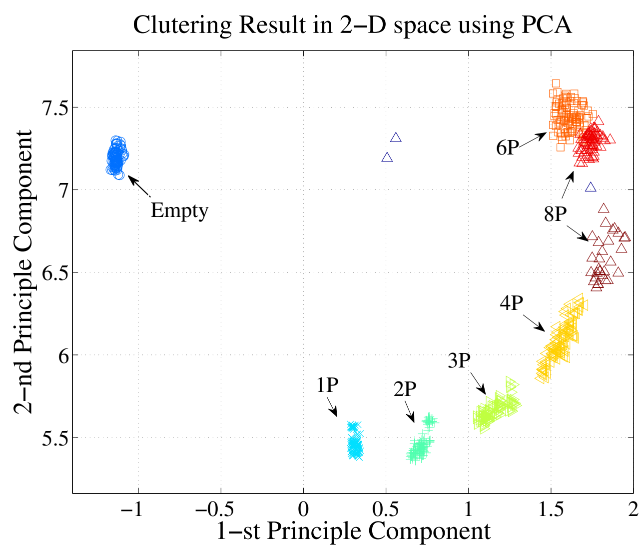

Figure 10: Clusters of different numbers of people in a small conference room shown in 2-D principle component space

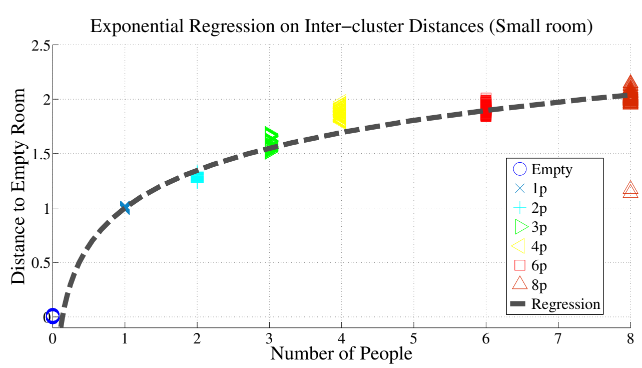

Figure 11: Adaptive exponential regression for occupancy estimation in small room scenario

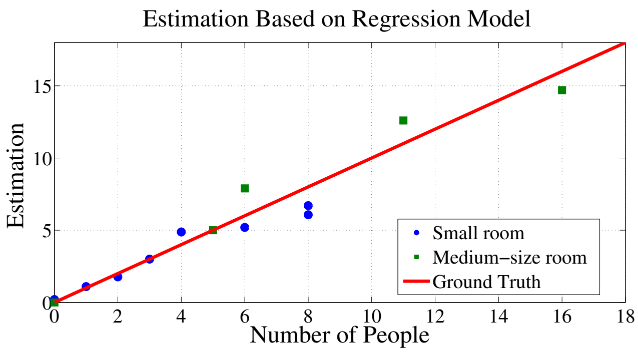

Figure 12: Estimation made by our algorithm compared to ground truth in small and medium size rooms

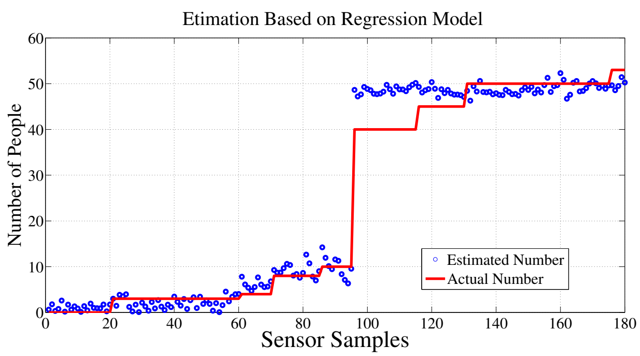

Figure 13: Estimation compared with ground truth as people enter an auditorium

significantly impact the result. We perform tests including opening the door to the room, opening windows in the room, changing the volume of the transmitter, and then testing in the same room one week later. As shown in Table 1, the error was most effected by changes in volume and slightly by opening the windows. Error due to changes in volume are not surprising since the regression model is built around magnitude changes in different frequencies.

To test the system's ability to automatically retrain itself, an experiment is carried out in the same room a week later with slightly different position and volume. Without self-retrain on the new environment, the error increases by 1.2%. This could accumulate and potentially grow worse over time. However, if the baseline and the *unit distance* is correctly calibrated, which can be done if an empty room can be detected, the change in error is negligible. The result again shows in Table 1, less than 1% difference between the calibrated *unit distance* and the ideal one.

Finally, the overall performance of the system is summarized in Table 2, and the comparison with related approaches in people counting is shown in Table 3. The comparison values were extracted from each paper. The number of people estimated by the system is no more than 3 people different from the actual number on average, and the average error in percentage to the maximum capacity of the room is around 5%.

Interference Type	**Error Inc.**(%)
Door opened	1.63
Windows opened	2.38
Change volume	5.38
Change position of the device	2.12
Data collected a week later(no retrain)	1.18
Data collected a week later(auto-retrained)	0.08

Table 1: System performance with error sources in small room

Sizes \ **Parm**	**Max Cap.**	**Avg. Error**	**Error/Max Cap.(%)**
Small room	8	0.61	7.6
Medium room	30	1.6	5.3
Large room	150	2.6	1.7

Table 2: System performance with different room sizes

5. LIMITATIONS

Our proposed technique has a few drawbacks associated with the fact that it is an active sensing system. If multiple of our transducers are placed in the same room, there needs to be a mechanism to coordinate transmissions so that they do not experience cross-talk. For large spaces, there needs to be a proportionally powerful transmitter that will eventually require a larger amplifier and transducer. As the space increases in size, the ability to finely distinguish the exact number of people diminishes. For larger spaces, the system also requires a calibration point with enough people to register as approximately 5-10% of the room load for the best results. This can also be hard to coordinate in certain environments. We imagine in the future that this approach could be coupled with other forms of people counting to help aid in automatic calibration. Finally, ultrasound in our particular frequency is still detectable by animals. Beyond transducer cost (which benefits from being compatible with commodity audio equipment) there is no reason why this approach cannot operate at higher frequencies. At higher frequencies sound becomes more directional, so further investigation would be required to determine if reverberation is still as sensitive to person count.

6. CONCLUSION

In conclusion, this paper introduced an ultrasonic approach for estimating the occupancy level of a room using reverberation across multiple frequencies. The system consists of an omni-directional ultrasonic tweeter with a co-located microphone that first transmits an ultrasonic chirp

Method	**Proposed**	**[16]**	**[17]**	**[19]**
Max. Counts	50	12	35	5
Avg. Error	1.6	0.4	1.3	0.7
Environ.	indoor	indoor	outdoor	indoor
Complexity	low	medium	high	medium
Cost	low	high	medium	low

Table 3: Overall system performance comparison of multiple people counting approaches

into a room and then measures the response over time as the signal decays. When there are more people in a space, the signal decays more rapidly and hence the reverberation time can be used as a feature for estimating occupancy. We apply a clustering followed by regression model to estimate people in the space. With as little as two training points, the system is able to estimate total occupancy with less then 10% error on a wide variety of room sizes. The regression approach lends itself to being able to be retrained with a single updated point when the room is empty.

7. ACKNOWLEDGEMENTS

This research was funded in part by the Bosch Research and Technology Center in Pittsburgh and TerraSwarm, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

8. REFERENCES

[1] M. M. Carroll and C. F. Chien, "Decay of reverberant sound in a spherical enclosure," *The Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1442–1446, 1977.

[2] W. C. Sabine, "Collected papers on acoustics," *Harvard University Press*, 1923.

[3] P. Lazik and A. Rowe, "Indoor pseudo-ranging of mobile devices using ultrasonic chirps," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '12. New York, NY, USA: ACM, 2012, pp. 99–112.

[4] C. A. Andree, "The effect of position on the absorption of materials for the case of a cubical room," *Journal on the Acoustics Society of America*, 1932.

[5] N. N. Hidaka, T. and L. L. Beranek, "Relation of acoustical parameters with and without audiences in concert halls and a simple method for simulating the occupied state," *The Journal of the Acoustical Society of America*, vol. 109, 2001.

[6] T. Hidaka and N. Nishihara, "Reverberation time, mean-free-path, and sound absorption in concert halls-numerical examination by computer simulation," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3430–3430, 2006.

[7] M. R. Schroeder, "Computer models for concert hall acoustics," *American Journal of Physics*, vol. 41, no. 4, pp. 461–471, 1973.

[8] W. J. Davies, Y. W. Lam, and R. J. Orlowski, "Predicting theater chair absorption from reverberation chamber measurements," *Journal of the Acoustical Society of America*, vol. 93, no. 4, pp. 2238–2240, April 1993.

[9] L. L. Beranek, "Analysis of sabine and eyring equations and their application to concert hall audience and chair absorption," *The Journal of the Acoustical Society of America*, 2006.

[10] I. Dokmanic, R. Parhizkara, A. Walthera, Y. M. Lub, and M. Vetterli., "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

[11] B. Raj, K. Kalgaonkar, C. Harrison, and P. Dietz, "Ultrasonic doppler sensing in hci," *Pervasive Computing, IEEE*, vol. 11, no. 2, pp. 24–29, Feb 2012.

[12] S. P. Tarzia, R. P. Dick, P. A. Dinda, and G. Memik, "Sonar-based measurement of user presence and attention," *Ubicomp*, 2009.

[13] D. Caicedo and A. Pandharipande, "Ultrasonic array sensor for indoor presence detection," in *Signal Processing Conference (EUSIPCO)*, 2012.

[14] T. A. Nguyen and M. Aiello, "Beyond indoor presence monitering with simple sensors," in *2nd International Conference on Pervasive and Embedded Computing and Communication Systems*, 2012.

[15] E. Hailemariam, R. Goldstein, R. Attar, and A. Khan, "Real-time occupancy detection using decision trees with multiple sensor types," in *Symposium on Simulation for Architecture and Urban Design*, 2011.

[16] D. B. Yang, H. H. Gonzalez-Banos, and L. J. Guibas, "Counting people in crowds with a real-time network of simple image sensors," in *International Conference on Computer Vision*, 2003.

[17] A. B. Chan, C. La Jolla, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Computer Vision and Pattern Recognition*, 2008.

[18] C. C. Loy, S. Gong, and T. Xiang, "From semi-supervised to transfer counting of crowds," in *International Conference on Computer Vision*, 2013.

[19] K. P. Lam, M. Hoynck, B. Dong, B. Andrews, Y. shang Chiou, D. Benitez, and J. Choi, "Occupancy detection through an extensive environmental sensor network in an open-plan office building," in *Proc. of Building Simulation 09, an IBPSA Conference*, 2009.

[20] L. Yang, K. Ting, and M. Srivastava, "Inferring occupancy from opportunistically available sensor data," in *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, March 2014, pp. 60–68.

[21] D. Li, B. Balaji, Y. Jiang, and K. Singh, "A wi-fi based occupancy sensing approach to smart energy in commercial office buildings," in *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, ser. BuildSys '12. New York, NY, USA: ACM, 2012, pp. 197–198.

[22] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *International Conference on Knowledge Discovery and Data Mining*, 1996.