Content Analysis for Acoustic Environment Classification in Mobile Robots

Selina Chu[†], Shrikanth Narayanan[*†], and C.-C. Jay Kuo[*†]

Signal Analysis and Interpretation Lab, Integrated Media Systems Center
Department of Computer Science[†] and Department of Electrical Engineering[*]
University of Southern California, Los Angeles, CA 90089, USA
{selinach, shri, cckuo}@sipi.usc.edu

Abstract

We consider the task of recognizing and learning the environments for mobile robot using audio information. Environments are mainly characterized by different types of specific sounds. Using audio enables the system to capture a semantically richer environment, as compared to using visual information alone. The goal of this paper is to investigate suitable features and the design feasibility of an acoustic environment recognition system. We performed statistical analysis of promising frequency- and time-domain based audio features. We show that even from unstructured environmental sounds, we can predict with fairly accurate results the type of environment that the robot is positioned.

Introduction

Recognizing the environment from sounds is a basic problem in audio and has important applications in robotics and scene recognition. The method in which robotic systems navigates depends on their environment. Current approaches for robotic navigation mostly focus on vision-based systems, for example model-based [DeSouza] and view-based [Matsumoto]. These approaches lose their robustness or their utility if visual indicators are compromised or totally absent. With the loss of certain landmarks, a vision-based robot might not be able to recover from its displacement because it is unable to determine the environment that it is in. Knowing the scene provides a coarse and efficient way to prune out irrelevant scenarios. To alleviate the system's dependency on vision alone, we can incorporate audio information into the scene recognition process. A stream of audio data contains a significant amount of information, enabling the system to capture a semantically richer environment. Audio data could be obtained at any moment the robot is functioning, neglecting any external condition, e.g. lack of lights, and is also computationally cheaper than most visual recognition algorithms. Thus, the fusion of audio and visual information can be advantageous, such as in disambiguation of environment and object types.

Many robotic applications are being utilized for navigation in unstructured environments [Pineau, Thrun]. There are tasks that require the knowledge of the environment, for example determining if indoor or outdoor [Yanco, Fod]. In order to use any of these capabilities, we first have to determine the current ambient context. A context denotes a location with different acoustic characteristics, such as a coffee shop, outside street, or a quiet hallway. Differences in the acoustic characteristics could be caused by the physical environment or activities from humans and nature.

Research on general unstructured audio-based scene recognition has received little attention as compared to applications such as music or speech recognition, with exception of works from [Eronen, Malkin, Peltonen, Bregman]. There are areas related to scene recognition that have been researched to various degrees. Examples of such audio classification problems include music type classification, noise type detection, content-based audio retrieval, and discrimination between speech/music, noisy/silent background, etc. [Cano, Essid, De Santo, Zhang].

It is relatively easy for most people to make sense of what they hear or to discriminate where they are located in the environment largely based on sound alone. However, this is typically not the case with a robot. Even with the capacity to capture audio sounds, how does it make sense of the input? A question that arises at this point is: is it meaningful to base environment recognition on acoustic information?

In this paper, we consider environment recognition using acoustic information only. We investigate a variety of audio features to recognize different auditory environments, specifically focusing on environment types we encounter commonly. We take a supervised learning approach to this problem, where we begin by collecting sound samples of the environment and their corresponding ground-truth maps. We begin by examining audio features, such as energy and spectral moments, gathered from a mobile robot and apply to scene characterization. We describe the collection of evaluation data representing the common everyday sound environment, allowing us to access the feasibility of building context aware applications using audio. We apply supervised learning to predict the class as a function of sounds we collected. We show that even from unstructured environment, it is possible to predict with fairly accurate results the environment that the robot is positioned.

Data Collection

We would like to capture actual scenarios of situations where a robot might find itself, including any environmental sounds, along with additional noise generated by the robot. To simplify the problem, we restricted the number of scenes we examined and enforced each type of environmental sound not to overlap each other. The locations we considered are recorded within and around a multipurpose engineering building on the USC campus. The diverse locations that were focused include: 1) café area, 2) hallways where research labs are housed, 3) around and inside elevator areas, 4) lobby area, and 5) along the street on the south side of the building.

We used a Pioneer DX mobile robot from ActivMedia, running Playerjoy and Playerv [player]. The robot was manually controlled using a laptop computer. To train and test our algorithm, we collected about 3 hours of audio recordings of the five aforementioned types of environmental locations. We used an Edirol USB audio interface, along with a Sennheiser microphone mounted to the chassis of the robot. Several recordings were taken at each location, each about 10-15 minutes long, taken on multiple days and at various times. This was done in order to introduce a variety of sounds and to prevent biases in recording. The robot was deliberately driven around with its sonar sensors turned on (and sometimes off) to resemble a more realistic situation and to include noises obtained from the onboard motors and sonar. We did not use the laser and camera because they produce little, if any, noticeable sounds. Recordings were manually labeled and assigned to one of the five classes listed previously to aid the experiments described below.

Our general observations about the sounds encountered at the different locations are:

- *Hallway*: mostly quiet, with occasional doors opening/closing, distant sound from the elevators, and individuals quietly talking, some footsteps.
- *Café*: many people talking, ringing of the cash registers, moving of chairs.
- *Lobby*: footsteps with echos, different from hallways due to the type of flooring, people talking, sounds of rolling dollies from deliveries being made.
- *Elevators*: bells and alerts from the elevator, footsteps, rolling of dollies on the floor of the elevator.
- *Street*: footsteps on concrete, traffic noise from buses and cars, bicycles, and occasional planes and helicopters.

We chose for this study to focus on a few simple, yet robust signal features, which can be extracted in a straightforward manner.

The audio data samples collected were mono-channel, 16 bits per sample with a sampling rate of 44 kHz and of varying lengths. The input signal was down-sampled to a 22050 Hz sampling rate. Each clip was further divided into 4-second segments. Features were calculated from a 20 msec rectangular window with 10 msec overlap. Each 4 sec segment makes up an instance for training/testing. All spectra were computed with a 512-point Fast Fourier Transformation (FFT). All data were normalized to zero mean and unit variance.

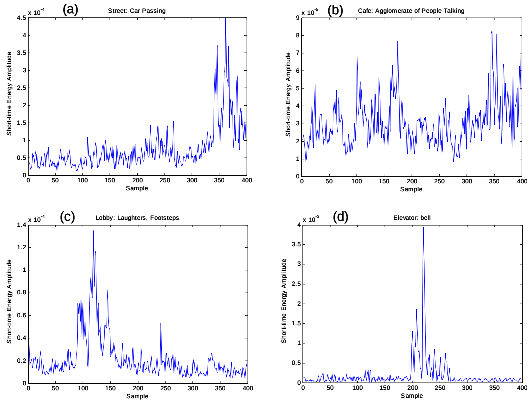

Figure 1. Short-time energy function of (a)car passing, (b)people talking, (c) laughter, footsteps, (d) elevator bell

Audio Feature Analysis

One major issue in building a recognition system for multimedia data is the choice of proper signal features that are likely to result in effective discrimination between different auditory environments. Environmental sounds are considered unstructured data, where the differences in the characteristics to each of these contexts are caused by random physical environment or activities from humans or nature. Unlike music or speech, there exist neither predictable repetitions nor harmonic sounds. Because of the nature of unstructured data, it is very difficult to form a generalization to quantify them. In order to obtain insights into these data, we performed an analysis to evaluate the characteristics from a signal processing point of view. There are many features that can be used to describe audio signals. The choice of these features is crucial in building a pattern recognition system. Therefore, we examined a

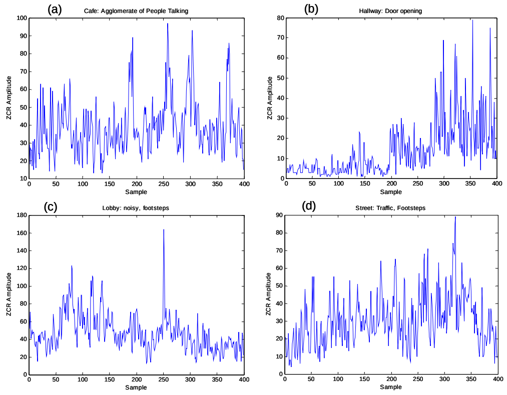

Figure 2. Short-time zero-crossing rates of (a)car passing, (b)people talking, (c) laughter, footsteps, (d) elevator bell

wide range of features in order to evaluate the effect of each feature and to select a suitable feature set to discriminate between the classes. All features are measured with a short-time feature analysis, where each frame size is 20ms with 10ms overlap.

Acoustic features can be grouped into two categories according to the domain in which they are extracted from: frequency-domain (spectral features) and time-domain (temporal features). The temporal information is obtained by reading the amplitudes of the raw samples. Two common measures are the energy and zero-crossing rates.

Short-time energy:

$$E_n = \frac{1}{N}\sum_m \left[x(m)w(n-m)\right]^2$$

where

x(m) is the discrete time audio signal, *n* is the time index of the short-time energy, and *w(m)* is the window of length *N*

Energy provides a convenient representation of the amplitude variation over time. Zero-crossings occur when successive samples have different signs. It is a simple measure of the frequency content of a signal.

Short-time average zero-crossing rate (ZCR):

$$Z_n = \frac{1}{2}\sum_m \left|\operatorname{sgn}\left[x(m)\right] - \operatorname{sgn}\left[x(m-1)\right]\right| w(n-m)$$

where $\operatorname{sgn}[x(n)] = \begin{cases} 1, & x(n) \geq 0, \\ -1, & x(n) < 0 \end{cases}$

Similarly, *w(m)* is the window of length *N*.

Since the energy level varies depending on the distance from the sound source, we use the range of the short-time energy as a measure and feature, instead of the average. An example of how the range might be more prominent, we observe the example between *café* and *hallway*. Because the *café* was noisy throughout, it produced the lowest energy ranges among the five classes. On the other hand, the *hallway* has a quiet background with some loud sounds like doors opening and shutting, which results in high range values. A similar characteristic was observed in the *elevator* class, where people do not usually talk in elevators when other people (or robot) are present, leaving the environment quiet with beeping sounds from the elevator and sounds from the elevator door opening and closing.

Spectral features are obtained by first performing a 512-point FFT on each short-time window of the audio signal. The following spectral features are examined in this work:

Mel-frequency cepstral coefficients (MFCC) [Rabiner]: the spectrum envelope is computed from energy averaged over each mel-scaled filter.

Spectral flux [Tzanetakis]: measures the change in the shape of the power spectrum. It is calculated as a spectral amplitude difference between successive frames:

$$SF_k = \sum_{m-1} \left|\left|X_k(n)\right| - \left|X_{k-1}(n)\right|\right|$$

where $|X(n)|$ is the magnitude spectrum of the k^{th} frame.

Statistical moments: obtained from the audio signal's spectrum (i.e. spectral centroid, spectral bandwidth, spectral asymmetry, and spectral flatness) [Essid]:

Spectral centroid: measures the brightness of a sound; the higher the centroid, the brighter the sound.

Signal bandwidth: measures the width of the range of frequencies that the signal occupies.

Spectral flatness: quantifies the tonal quality; how much tone-like the sound is as opposed to being noise-like.

Spectral rolloff: measures the frequency where a specific amount of the power spectrum resides. A commonly used value for the threshold is 0.95. We measured the rate at which the accumulative magnitude of the frequency response is equal to that of 95% of the total magnitude.

Table 1. Time-domain features (averaged)

Class	Energy (x10-4)	Energy Range	Zero-Crossing
Street	0.145	0.919	14.19
Elevator	0.123	0.917	18.60
Café	0.048	0.866	29.99
Hallway	0.062	0.977	11.90
Lobby	0.064	0.932	29.09

Table 2. Frequency range for each class

Class	Frequency Range (Hz)	Mean (Hz)	Mode (Hz)	% of data w/ mode value
Street	0-172	74.5	86	26.7
Elevator	0-172	72.1	43	43.3
Café	0-603	178.6	129	66.7
Hallway	0-172	66.0	58	64.4
Lobby	0-560	164.3	129	56.7

We further analyzed the frequency spectrum by examining the power spectrum distribution for each class. This was accomplished by taking a 512-point FFT of each short-time window and locating the frequency where the power is maximized and its corresponding amplitude. The summary of the findings are give in Table 2 and Figure 3. We discover the *café* and *lobby* class to contain a wider distribution of frequencies, 0-603 Hz and 0-560 Hz respectively. The rest of the classes (*street, elevator, and hallway*) have their ranges between 0 and 172Hz. A high percentage of data from each class tends to concentrate at specific frequencies. Both *café* and *lobby* class centers at 129Hz, with 67% and 57% of the data respectively.

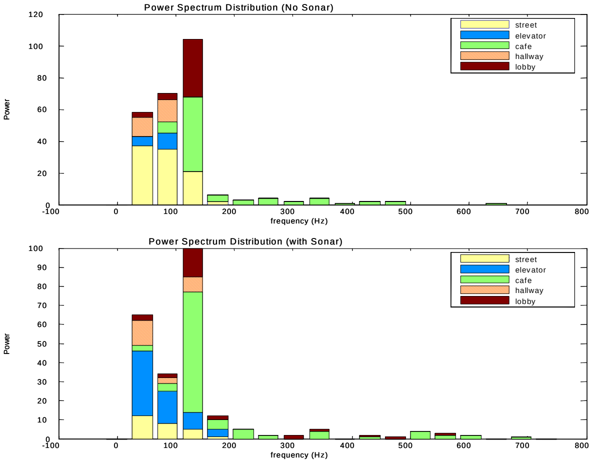

Figure 3. Power spectrum distribution. The power was minimal for higher than 800 Hz and is not shown in the figures. (Top) no sonar, (Bottom) with sonar

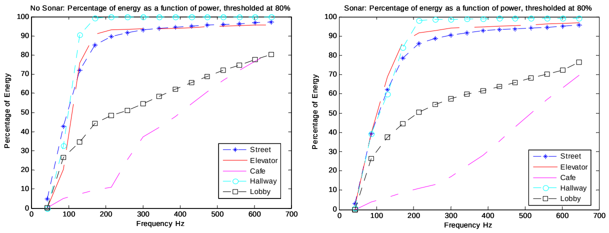

Figure 4. Percentage of energy as a function of power: (Left) no sonar, (right) with sonar

The main distinction between these two classes is in the range, where the distribution for *café* is much wider than that of *lobby*. The frequencies for the *street* class are gathered within the range, 43-129 Hz, which made up for 78% of the data and are evenly apportioned at 43, 86, and 129Hz (with 22%, 24%, and 24% respectively).

To explore the relationship combining temporal and spectral information, we observed the amount of energy when the accumulated power is greater than certain threshold at each frequency band. For each short-time window and at each frequency band, we record the average energy when the accumulated power is above 75% of the total power. Then, we found the percentage of energy relative to the total energy. The result is visualized in Figure 4. For the *street*, *hallway*, and *elevator* class, the percentage of the energy is relatively high at lower frequencies, as compared to *café* and *lobby* where the curve is flatter. When the sonar is turned on (right plot in figure 4), the amount of energy is pushed to higher frequencies. A viable explanation for this occurrence is due to the different amount of human voices contained in each of the classes. When there are more voices, the energy seemed to be more spread out among the spectrum, which corresponded to flatter lines. Examining figure 4 more closely, we can observe the lines from low to high. The flattest, being the *café* class, contains mostly voices. It is followed by *lobby*, which contains both human voices and sounds made by various physical objects. *Street* typically contains infrequent human voices and is mostly composed of traffic from vehicles. There was very little talking in the elevator class and almost none in the hallway; these correspond to the higher lines in the plot.

We can observe the amount of temporal similarity between events within each class by examining their correlation. For each short-time window, we found the correlation coefficient against every other 20-msec window within the 4-second sample. Then, we find the number of incidents in which the coefficient was higher than 0.8 and average that over the number of windows per 4-sec sample.

Table 4: Percentage having high correlation within each sample.

Classes	No Sonar	Sonar
Street	96.67	97.50
Elevator	100	100
Café	67.09	58.16
Hallway	100	100
Lobby	91.11	86.67

The lowest percentage of data that have correlation within each clip is from the *café* class. One possible explanation is because events happen sporadically in this environment. There is little consistency within each sample. We can see that the *elevator* and the *hallway* class are fairly constant within each sample, especially *hallway*, where it is mostly quiet. Similar trends are observed between the no sonar and sonar class, but the overall correlation is lower for classes that contain sonar.

Classifying Environmental Sounds

We investigated three different classification methods: K-Nearest Neighbors (KNN) [Mitchell], Gaussian Mixture Models (GMM) [Moore], and Support Vector Machine (SVM) [Scholkopf]. For KNN, we used the Euclidean distance as the distance measure and the 1-nearest neighbor queries to obtain the results. As for GMM, we set the number of mixtures for both training and testing to 5. For the SVM classifiers, we used a 2-degree polynomial as its kernel with regularization parameter C=10 and the epsilon ε=1e^{-7}, which controls the width of the e-insensitive zone, which used to fit the training data, affecting the number of support vectors used. Since SVM is a two-class classifier, we used the one-against-the-rest algorithm [Burges] for our multi-class classification in all of the experiments. We performed leave-one-out cross-validation on the data. The recognition accuracy using leave-one-out cross-validation was found from calculating:

$$accuracy = \frac{\#\_of\_correctly\_classified}{Total\_\#\_of\_dataset}$$

There are many features that can be used to describe audio signals. We have examined some of them in the previous section. The feature vector for the experiments consisted of features summarized in Table 5.

Previously, we performed classification on five different contexts using partial of the features in Table 5 and data obtained from a mobile robot to determine how the sonar separated the data into two sets: A) with sonar and B) without sonar. Classifications were performed on set A

Table 5. List of features used in classification

Feature No.	**Types of Features**
1-12	1st – 12th MFCCs
13-24	Standard Deviation of 1st – 12th MFCCs
25	Spectral Centroid
26	Spectral Bandwidth
27	Spectral Asymmetry
28	Spectral Flatness
29	Zero-Crossing
30	Standard Deviation of Zero-Crossing
31	Energy Range, E_r
32	Standard Deviation of Energy Range
33	Frequency Roll-off
34	Standard Deviation of Roll-off
35	Spectral Flux
36-45	Relative energy when the concentration of the power is greater than 75% (for the first 10 frequency bands)

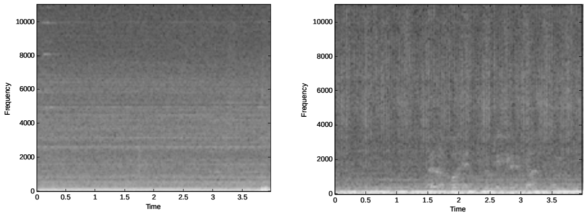

Figure 5. Samples containing sonar sounds (right), vertical streaks indicating the presence of periodic sonar noise are present. When there are no sonar sounds (left), no vertical streaks are present.

only, set B only, and A&B using the first 34 features for KNN, with an accuracy of 90.8%, 91.2%, and 89.5% respectively [Chu]. Since it was determined that the differences between the various sets were minimal, we chose to use set A&B for the rest of the experiments.

We performed different experiments using the three mentioned classification methods: using all 45 features, first 35 features, and a reduced set of features (from feature selection). One problem in using a large number of features is that there are many potentially irrelevant features that could negatively impact the quality of classification. Using feature selection techniques, we can choose a smaller feature set to reduce the computational cost and running time, as well as achieving an acceptable, if not higher, recognition rate. The optimal solution to feature selection is using an exhaustive search of all the features, requiring 2^{34}-1, or roughly 10^{10} combinations. Therefore, we use a greedy forward feature selection method instead. Details are explained in [Chu].

We found that using all 45 features for KNN resulted in 92.3% accuracy, but ignoring features 36-45 resulted in an accuracy of 89.5%. However when we used forward feature selection, none of the features from 36-45 was selected as effective features for classification. We also performed classification on all 45 features using GMM and SVM, but due to time limitation we only used the first 35 features for the forward feature selection. The results and features selected are summarized in Table 6. SVM have the highest accuracy rate over the other two methods, but is very expensive to train, even with five classes. The average running times required for training and classification of a single query with the full feature set is 1.1, 148.9, and 1681 sec for KNN, GMM, and SVM respectively. The KNN classifier works well overall, outperforming GMM and is roughly 1000 times faster than SVM. Figure 6 above describes the recognition accuracies using increasing number of features.

Table 6: Summary of classification accuracy

Classifiers	**Features Used**	**Recognition Accuracy**
KNN	All 45 features	92.3%
	1-35	89.5%
Forward FS	1-3, 5, 7-10, 12, 13, 16, 17, 28, 31, 33, 35	94.9%
GMM	All 45 features	87.9 %
	1-35	89.5%
Forward FS	1-10, 12-16, 20-22, 25, 26, 28,31-35	93.2%
SVM	All 45 features	94.3 %
	1-35	95.1%
Forward FS	1-3, 5-10, 13, 15, 18, 28, 31-33	96.6%

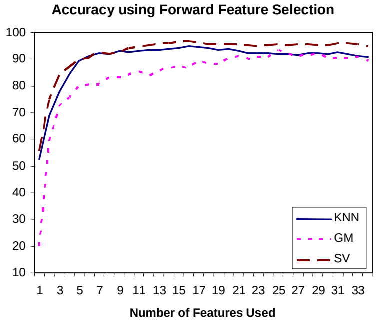

Figure 6: The classification results with KNN, GMM, and SVM respectively.

A final note on the various learning algorithms studied here: unlike KNN, both GMM and SVM require a careful choice in choosing the correct parameters. There are at least two degrees of freedom for GMM and four for SVM. Therefore, even minor changes, such as number of training samples, requires fine tuning of parameters.

The confusion matrix in Table 7 shows the misclassified classes for the KNN classifier using 16 features. It can be seen that the worst performance was from the *elevator* class and had most misclassification from *hallway*. One reason is because the area where the robot was being driven around for the *elevator* class was actually part of the hallway as well; there was less separation between the two areas. However, *hallway* gave the best performance due to its distinct characteristic that it was relatively quiet most of the time. We can also observe from the same table that there are confusion between *lobby* and *street*. Both of these classes contain many sharp footstep noises, but on different flooring. The lobby uses some kind of granite tiling, while the street is of concrete. There are footstep noises in the *hallway* class as well, but the flooring for the hallways uses plastic material. Therefore, the footsteps created were less prominent than from *lobby* or *street* and created less confusion. Footsteps in café were drown out by other noises, such as crowds of people talking and shuffling of furniture.

Table 7: Confusion matrix of the KNN classification using forward feature selection with 16 features

	Street	Elevator	Café	Hallway	Lobby
Street	94.4	0	0	0	5.6
Elevator	0	90.0	1.1	7.8	1.1
Café	0	0	95.6	0	4.4
Hallway	0	0	0	100	0
Lobby	2.2	0	3.3	0	94.4

Conclusion and Future Work

This paper investigates techniques for developing an environment scene classification system using audio features. We have described our experience in analyzing the different environment types for mobile robot using audio information. We explored and investigated suitable features and the feasibility of designing an automatic acoustic environment recognition system. The classification system was successful in classifying five classes of contexts using real data obtained from a mobile robot. We show that even from unstructured environmental sounds, we can predict with fairly accurate results the environment that the robot is positioned. Note however that the type of environments were limited to five, but were based on their general sound characteristics.

We believe this work to be the first step in building a recognition system for auditory environment recognition. This current work opens up a doorway to other open challenges. Here we focus on global characterization of the environment; we should also examine localization and effects of sound sources. Other issues include robustness to new environment types and how this scales with current systems using vision and other sensing mechanisms. Our next step is to increase the number of classes, as well as investigating on combining the use of audio and visual features.

Acknowledgements

The authors would like to thank Dylan Shell for his help on the interface to the Pioneer robot and Wan-Chi Lee for his helpful comments and suggestions. We would also like to thank Professor Maja Mataric´ for supporting the initial work. We acknowledge support from NSF and DARPA.

References

DeSouza, G. N. and Kak, A. C. 2002. Vision for mobile robot navigation: A survey. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 2, pp. 237-267.

Matsumoto, Y., Inaba, M. and Inoue, H. 2000. View-based approach to robot navigation. In *Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems*.

Pineau, J., Montemerlo, M., Pollack, M., Roy, N., and Thrun, S. 2003. Towards robotic assistants in nursing homes: challenges and results. *Robotics and Autonomous Systems*, Vol. 42.

Thrun, S., Bennewitz, M., Burgard W., Cremers, A.B., Dellaert, F., Fox, D., Haehnel, D. Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. 1999. MINERVA: A second generation mobile tour-guide robot. In *Proc. of IEEE International Conference on Robotics and Automation*.

Yanco, H.A. 1998. Wheelesley, A Robotic Wheelchair System: Indoor Navigation and User Interface. Lecture notes in *Artificial Intelligence: Assistive Technology and Artificial Intelligence*, Springer-Verlag.

Fod, A., Howard, A., and Mataric´, M. J. 2002. Laser-Based People Tracking. In *Proc. of IEEE Int. Conf. Robotics and Automation*.

Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T. Lorho, G., Huopaniemi, J. 2006. Audio-based context recognition. In *IEEE Transactions on Speech and Audio Processing.*

Malkin, R., Waibel, A. 2005. Classifying User Environment for Mobile Applications using Linear Autoencoding of Ambient Audio. In *Proc. of IEEE Int. Conf on Acoustics, Speech, and Signal Processing*, 2005.

Peltonen, V. 2001. Computational Auditory Scene Recognition, M.S. thesis Dept of Information Technology, Tampere University of Technology, Tampere Finland.

Zhang, T. and Kuo, C.-C. 2001. Audio content analysis for online audiovisual data segmentation and classification, In *IEEE Transactions on Speech and Audio Processing*, 9(4).

Bregman, A. 1990. *Auditory scene analysis*. Cambridge, USA: MIT Press.

Cano, P., Koppenberger, M., Le Groux, S., Ricard, J., Wack, N. and Herrera, P. 2004. Nearest-neighbor generic sound classification with a wordnet-based taxonomy. In *Proc. of AES 116th Convention*.

De Santo, M., Percannella, G., Sansone, C., Vento, M. 2001. Classifying Audio Streams of Movies by a Multi-Expert System. In *Proc. of International Conference on Image Analysis and Processing.*

Player/Stage, *http://playerstage.sourceforge.net/.*

Rabiner, L. and Juang, B.-H. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall.

Tzanetakis, G. and Cook, P. 2000. Sound analysis using MPEG compressed audio. In *Proc. of Int. Conf. Audio, Speech and Signal Processing*.

Essid, S., Richard, G., and David, B. 2004. Musical instrument recognition by pairwise classification strategies, In *IEEE Transactions on Speech and Audio Processing*.

Chu, S., Narayanan, S., Kuo, C.-C., and Mataric´, M. J. 2006. Where Am I? Scene Recognition for Mobile Robots using Audio Features. In *Proc. of IEEE International Conference on Multimedia & Expo*.

Mitchell, T. M. *Machine Learning*. Mc Graw-Hill, 1997

Moore, A. Statistical Data Mining Tutorial on Gaussian Mixture Models, www.cs.cmu.edu/~awm/tutorials, CMU.

Scholkopf, B. and Smola, A. J. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press.

Burges, C.J. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. In Data Mining and Knowledge Discovery.