Audio Signal Based Enviornment Determination For Mobile Robots By Time-Frequency Analysis

Lovely Yadav [a], Sandeep Raghuwanshi [b], Amit Swami [c]
a, b, c *Department of Information Technology , S.A.T.I. Vidisha (M.P.)*

ABSTRACT

Audio feature extraction and classification are important process for audio signal analysis in many applications, such as content-based audio retrieval, multimedia database, and auditory scene analysis. Efficient feature extraction is required by these applications. To determine the information of environments visually is an important problem for mobile robots. In this paper, we describe statistical features of different sound signals were extracted by using time frequency analysis to classify them by using neural network. In this work, classification of four signals was achieved successfully with classification score up to 0.97.

Keywords-Robotics, Time-Frequency Analysis, Choi-Williams distribution, Artificial Neural Networks.

1. INTRODUCTION

In the case of mobile robots, we expect them to be intelligent enough to recognize environment by using the concept of pattern recognition. Some view based mobile robot navigation are explained in [1, 2], but it is computationally expensive task. So here, we need some other technique to make this task easy. As most of the places or objects have a particular sound signal and its features can be determined to recognize different auditory environments.

Many robotic applications are being utilized for navigation in unstructured environments [3, 4]. There are other tasks that require knowing the environment. In [5] Yanco introduced a robotic wheelchair system that switches automatically between control modes for indoor and outdoor environments. Also, laser range-finder can track people in an outdoor environment [6]. In order to use any of these capabilities, we first have to determine the current context, e.g., the location type (outdoor or indoor environment, etc).

Characterizing the scene or environment is the first step to choosing which modality of interaction a robot should engage in. Furthermore, environments are dynamic, and the setting might change even in the same area. 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. Even with a GPS system and a well-defined map, without clear images, it is difficult to discern different characteristics of the environment. It is relatively easy for most people to make sense of what they hear or to discriminate where they are located in the environment based on sound alone. However, this is typically not the case with a robot. Surprisingly little research has been done on audio scene analysis in robots. With increasing number of robots being built for service and social settings, it is ever more important for the robots not only to identify locations, but to comprehend and characterize their auditory features [7]. In [12] different types of signals were classified using Time-Frequency features in combination with Mel-frequency cepstral coefficients features. In this paper, a new method is proposed where mean and standard deviation were calculated by windowing the instantaneous frequency and instantaneous power sequences determined in different frequency bands from time frequency analysis by using Choi-Williams distribution. Then signals were classified by neural network using these features.

2. ENVIRONMENTAL DATA COLLECTION

The four sound signals in wav format were downloaded from [8]. They are:

a. Sound signal of repeated church bell.
b. Sound signal of seashore waves.

c. Sound signal of traffic noise.
d. Sound signal of a passing by train.

The audio data samples collected were mono-channel, 16 bits per sample with a sampling rate of 44 kHz and of varying lengths.

3. TIME-FREQUENCY REPRESENTATION

For continuous time and frequency variables, the Wigner distribution (WD) for signal $x(t)$ is defined as;

$$W_x(t,f) = \int_{-\infty}^{\infty} x(t+\frac{\tau}{2})\; x\left(t-\frac{\tau}{2}\right) e^{-j2\pi f\tau}\; d\tau$$

(1)

Beside the WD, many other quadratic TFR exist with an energetic interpretation [9]. The class of all time-frequency shift-invariant, quadratic TFRs is known as the quadratic Cohen's class. Prominent members of Cohen's class are the spectrogram and the WD. Every member of Cohen's class may be interpreted as a 2-D filtered WD. In fact, it can be shown that $T_x(t,f)$ is a member of Cohen's class, if and only if it can be derived from the WD of the signal $x(t)$ via a time-frequency convolution.

$$T_x(t,f) = \int\int_{-\infty}^{\infty} h(t-t',f-f')\, W_x(t',f')dt'\, df'$$

(2)

Each member T_x of Cohen's class is associated with a unique, signal-independent kernel function $h(t,f)$ (or 2-D filter). Clearly, the convolution will be transformed into a simple multiplication in the Fourier transform domain. For the analysis of each signal, a particular Cohen's class distribution Choi-Williams distribution (CWD) was chosen [10]. It was applied to numerical sequences, after the calculation of the analytical signal, by using the equation (2) with the WD by equation (1) and the Choi-Williams exponential by equation (3).

$$h(t,f) = \sqrt{\frac{4\pi}{\sigma_c}}\, e^{-4\pi^2\frac{(tf)^2}{\sigma_c}}$$

(3)

The function (3) preserves the properties of the WD [10], such as the marginal properties and instantaneous frequency. Moreover, it is able to reduce the WD interferences by estimating an adequate parameter σ_c for the working band. The proposed estimation criterion was,

$$1 - \frac{A_{mb}}{A_{mt}} \geq 0.95$$

(4)

Where A_{mb} represents the mean value of the spectral amplitude between test tones and A_{mt} *is* the mean value of the spectral amplitude at the test tones [10].

4. AUDIO SIGNAL FEATURE EXTRACTION

To calculate time-frequency distribution, the time-frequency toolbox for MATLAB was obtained from [12]. Here in time frequency analysis normalized frequency is used and the frequency band is characterized in different bands i.e. VLF (0 to 0.1 Hz), LF (0.1 to 0.2 Hz), HF (0.2 to 0.4 Hz) and VHF (0.4 to 0.5 Hz).

The value of parameter σ_c of the CW exponential was estimated at 0.005 in order to eliminate the interferences produced by a test sinusoidal signal made by four frequency components at 0.1 Hz, 0.2 Hz, 0.3Hz, 0.4 Hz as shown in figure (1) and (2). Instantaneous power and instantaneous frequency was calculated for each signal by equation (5) and (6) respectively.

$$InstPower(t) = \frac{\int_{f1}^{f2} T_x(t,f)df}{\int_{-\infty}^{\infty} T_x(t,f)df}$$

(5)

Where f_1 and f_2 are the limits of the frequency bands.

$$InstFreq(t) = \frac{\int_{-\infty}^{-\infty} f\, T_x(t,f)df}{\int_{-\infty}^{\infty} T_x(t,f)df}$$

(6)

From figure, (3) to (6) are the time frequency distribution of first 800 points of the signals to be analyzed just to observe the difference in their distributions. As the difference between signals is more, less complicated will be the classification.

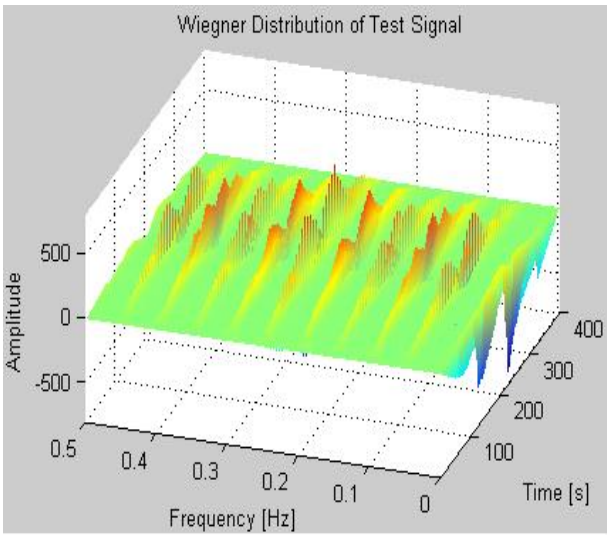

Figure 1.Wigner Distribution of test signal

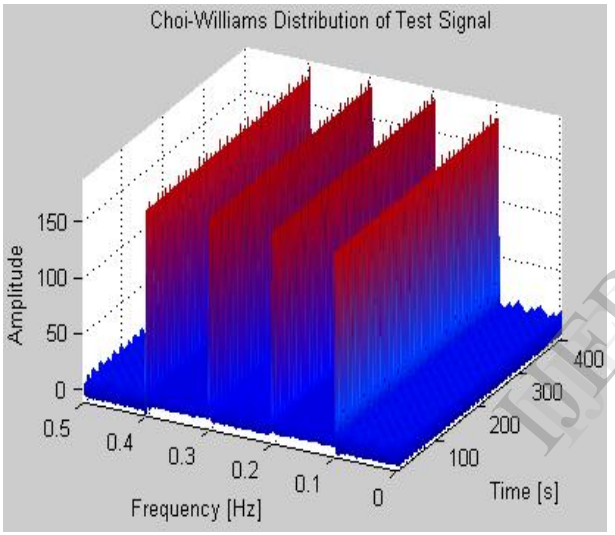

Figure 2.Choi-Williams Distribution of Test signal

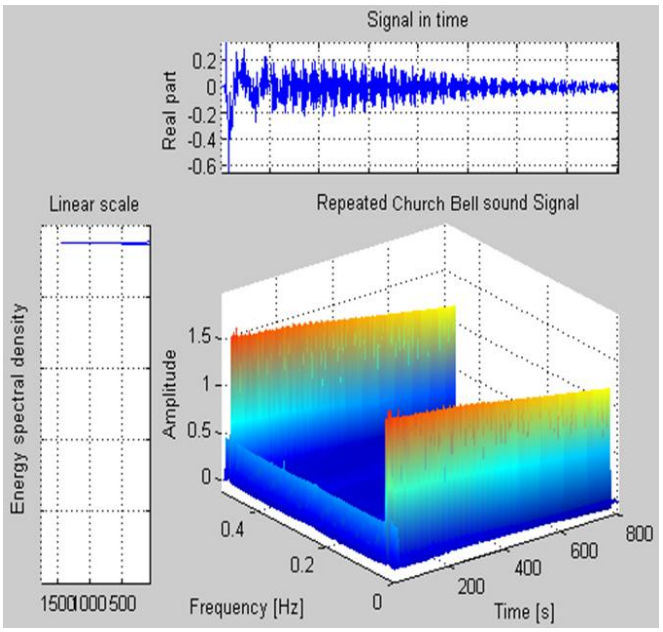

Figure 3.Choi-Williams distribution of repeated church bell sound signal.

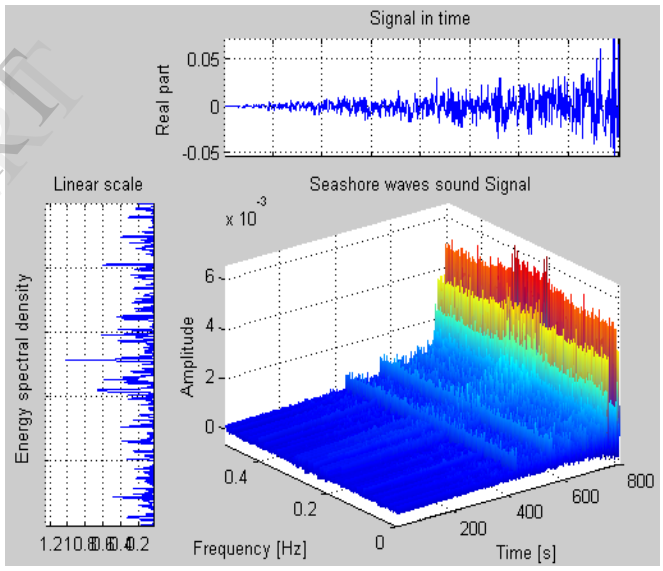

Figure 4.Choi-Williams distribution of Seashore waves.

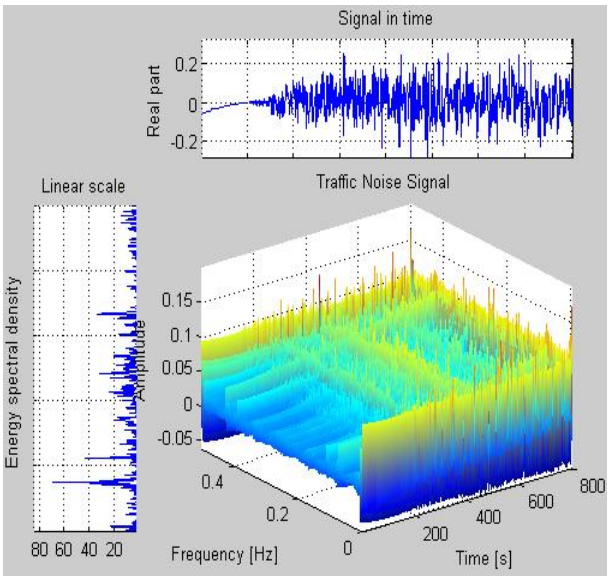

Figure 5.Choi-Williams distribution of Traffic Noise signal.

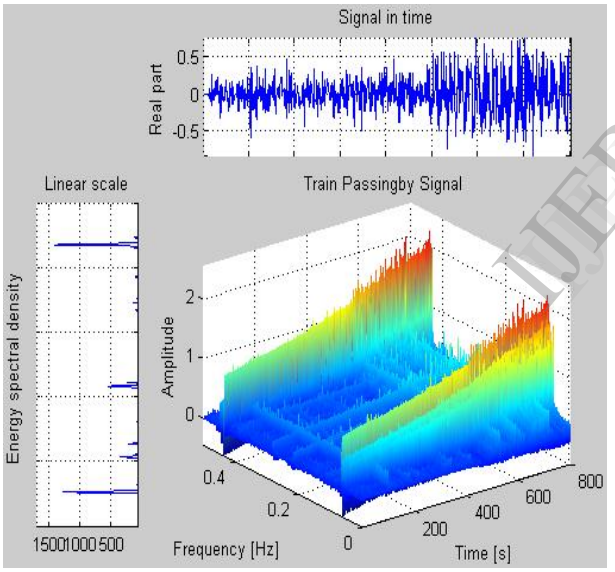

Figure 6.Choi-Williams distribution of Train passing by signal

A window of 500 points was used for time-frequency analysis to calculate mean and standard deviation for instantaneous frequency and instantaneous power in different bands i.e. VLF, LF, HF and VHF. Finally, 16×640 size matrix was obtained for 80000 points of each signal. First 16×400 matrix was used for training a four-layered neural network and remaining 16×240 size matrix was used to test the designed neural network for each signal respectively.

5. CLASSIFICATION

To classify the signal points a four layer Feed Forward Backpropagation network was designed by Neural Network Toolbox from MATLAB with 115 neurons in first layer,55 neurons in second layer,25 neurons in third layer and 1 neuron in last layer along with biases at each neurons shown in figure 7.

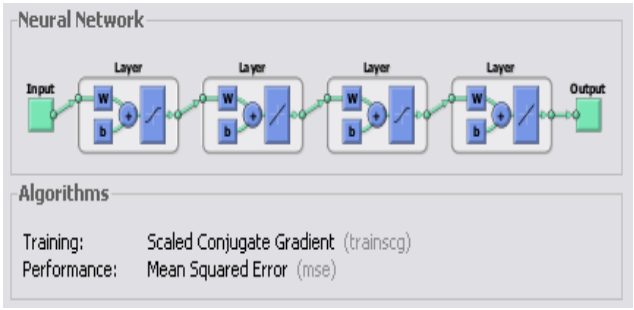

Figure 7.Designed Neural Network

The neural network was trained using scaled conjugate gradient algorithm [13]. The neural network was trained on first 100 point of each signal where each point was obtained by time frequency analysis of 500 signal points. In this way 50000 point out of 80000 points of each sound signal were used in training process. Now remaining 30000 points of each of the four signals were analyzed by time frequency analysis to obtain 60 points of each signal for testing of neural network.

Targets for the neural network were +1.5 for repeated church bell signal, +0.5 for seashore signal, -0.5 for traffic noise signal, and -1.5 for passing by train signal. After training classification thresholds were taken $\text{output} > +0.75$ for repeated church bell signal, $output > 0\ and \leq +0.75$ for seashore signal, $output < 0\ \&\ \geq -0.75$ for traffic noise signal and $output < -0.75$ for passing by train signal. To measure classification a classification score was calculated by following equation.

$$\text{Classification Score} = \frac{\text{Correctly Classified Points of a particular class}}{\text{Total Points of that particular class}}$$

If some frequency components are common between different classes then it results in confusion for neural network during classification. For wrongly classified points of other class with respect to the analyzed signal misclassification score is calculated. Misclassification score is calculated by following equation.

$$\text{Misclassification Score} = \frac{\text{Wrongly Classified Points Of another Particular Class}}{\text{Total Points of that particular class}}$$

6. RESULTS

After 1,80,000 iterations mean square error reduced to 0.00239 as shown in figure (8). Neural network output and classified output for training dataset is shown in figure (9).

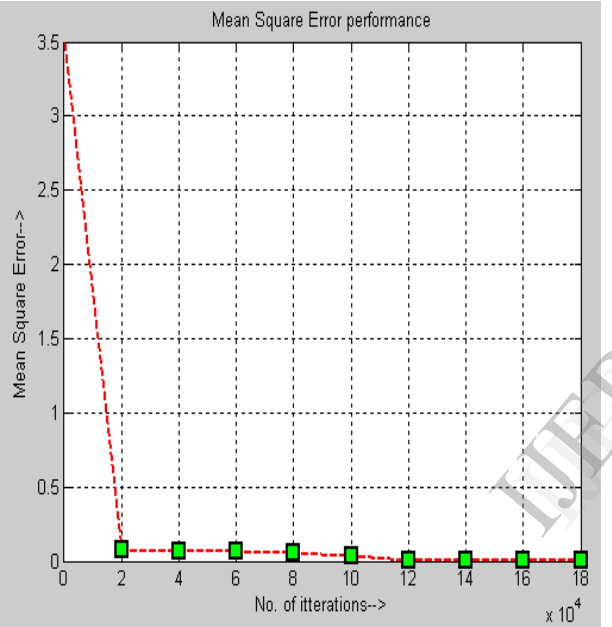

Figure 8.MSE (Mean Square Error) performance graph

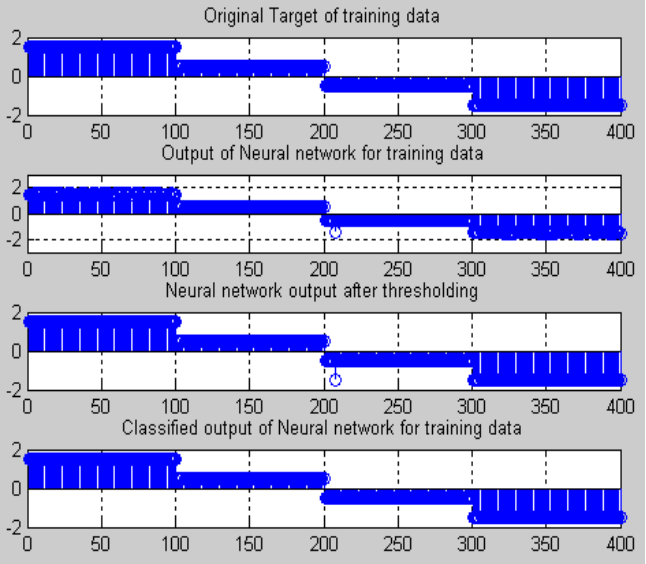

Figure 9.Neural network output for training dataset

a. Confusion matrix for training dataset

Repeated Church Bell signal		Sea Shore signal	Traffic Noise signal	Train Signal
Repeated church Bell signal	1.0000	0	0	0
Seashore signal	0	1.0000	0	0
Traffic Noise signal	0	0	0.9900	0.0100
Train Signal	0	0	0	1.0000

Neural network output and classified output for testing dataset is shown in figure (10).

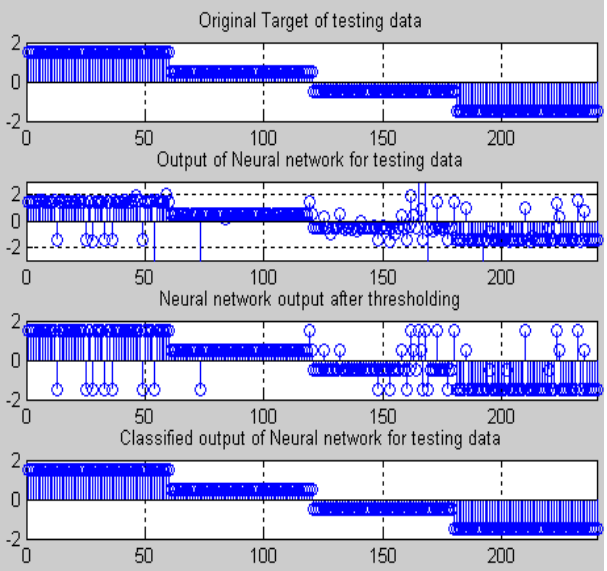

Figure 10.Neural network output for testing dataset

b. Confusion matrix for testing dataset

	Repeated Church Bell signal	Sea Shore signal	Traffic Noise signal	Train Signal
Repeated church Bell signal	0.8833	0	0	0.1167
Seashore signal	0.0167	0.9667	0	0.0167
Traffic Noise signal	0.0833	0.0833	0.7333	0.1000
Train Signal	0.0500	0.0500	0.0500	0.8500

The results show that different environmental signals have different frequency components at different time instants, which make them unique. Hence, by Time-Frequency analysis complete information of signal can be determined. The trained network was able to classify training dataset up to classification score from 0.99 to 1 and testing dataset with classification score from 0.73 to 0.97. So here after analyzing 80,000 points signal were classified easily. Seashore signal have very les similarity with other three signals and traffic noise signal is more similar to other signals due to its very random nature, which contains nearly all frequency components present in other signals.

7. CONCLUSIONS AND FUTURE WORK

As time frequency distribution for each signal is different, the features derived from them can be used to train a neural network and classify the particular signal. This approach is applied to four signals only here but it can also be applied to more signals of different classes and increasing length of training dataset can also increase the classification accuracy. In future, it will be tried to classify more signals with larger datasets to validate the method completely.

REFERENCES

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

[2.] Matsumoto, Y., Inaba, M. and Inoue, H. "View-based approach to robot navigation," Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems, pp. 1702-1708, 2000.

[3.] Pineau, J., Montemerlo, M., Pollack, M., Roy, N., and Thrun, S. "Towards robotic assistants in nursing homes: challenges and results," Robotics and Autonomous Systems, Volume 42, Issues 3-4, pages 271-281, 2003.

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

[5.] Yanco, H.A. "Wheelesley, A Robotic Wheelchair System: Indoor Navigation and User Interface," Lecture Notes in Artificial Intelligence: Assistive Technology and Artificial Intelligence, Springer-Verlag, 1998.

[6.] Fod, A., Howard, A., and Mataric´, M. J., "Laser-Based People Tracking", Proc. of IEEE Int. Conf. Robotics and Automation, pages 3024-3029, 2002.

[7.] Selina Chu, Shrikanth Narayanan, C.-C. Jay Kuo and Maja J. Matarić," Where am i? Scene recognition for mobile robots using audio features".

[8.] www.freesound.org.

[9.] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," IEEE Signal Process. Mag., 1: 21-67, 1992.

[10.] L. Cohen, Time-frequency analysis, Prentice Hall Signal Processing Series, 1995.

[11.] Behnaz Ghoraani and Sridhar Krishnan, "Time-Frequency matrix feature extraction and classification of environmental audio signals", Transactions on audio, speech and language processing, Vol. 19, September 2011.

[12.] http://crttsn.univ-nantes.fr/%7Eauger/tftb.html

[13.] Martin F. Moller,"A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning", Computer Science Department, University of Aarhus, Denmark, November 13, 1990.