Journal of Naval Science and Engineering
2009, Vol. 5 , No.3, pp. 64-78

UNDERWATER ACOUSTIC SIGNAL RECOGNITION METHODS

Murat KUÇUKBAYRAK Lt., Ozhan GUNES Lt.Jr. Grade, Asst. Prof. Nafiz ARICA, Cdr.
Turkish Naval Academy
Naval Science and Engineering Institute
Tuzla, Istanbul, Turkiye
{mkucukbayrak, ogunes, narica}@dho.edu.tr

Abstract

The term Underwater Acoustic Signal Recognition (UASR) is used for identifying the platforms by some techniques from the acoustic sound signals they produce. In this paper, we propose two different schemes for UASR. In both schemes, the feature extraction is performed using Mel-Frequency Cepstral Coefficients and Linear Predictive Coding derived Cepstral Coefficients which have been extensively utilized in speech recognition. In the first scheme, the features extracted frame by frame are used as a sequence in the representation of the whole signal. The classification of that sequence of vectors is then performed by Hidden Markov Models with various topologies. The second scheme represents the frame features using Bag of Acoustic Words approach. In training stage, all the feature vectors extracted from the input signal are first clustered into a set of acoustic words. Each feature vector is then assigned to an acoustic word. After the frequency of each word is calculated in the input signal, the final representation is performed by the co-occurrence list of the acoustic words.

SU ALTI AKUSTİK SİNYAL TANIMA YÖNTEMLERİ

Özetçe

Su altı Akustik Sinyal Tanıma (SAST) terimi, platformları ürettikleri seslerden bazı teknikler kullanarak tanıma işlemi için kullanılmaktadır. Her gemi makine, pervane, tekne yapısı ve mürettebat alışkanlıklarının birleşiminden meydana gelen kendine özgü özelliğe sahiptir. Bu makalede, SAST için iki *değişik yöntem önermekteyiz. Her iki yöntemde özellik çıkarımı işlemi konuşma tanıma konusunda yarar sağladığı kanıtlanmış Mel-Frekans Kepstral Katsayıları ve Doğrusal Kestirimci Kodlama ile türetilmiş Kepstral Katsayılar ile hesaplanmaktadır. İlk yöntem de öznitelik çıkarımından sonra sinyal vektör dizisi olarak ifade edilir. Vektör dizilerinin sınıflandırılması daha sonra değişik topolojilere sahip Saklı Markov Modelleri ile yapılmaktadır. İkinci yöntem çerçeve özelliklerini Akustik ses kümesi yaklaşımını kullanarak temsil eder. Eğitme safhasında, giriş sinyalinin çerçevelerinden çıkarılan tüm öznitelik vektörleri önce bir akustik kelimeler kümesine gruplandırılır. Öznitelik vektörlerinin her biri bir akustik kelimeye atanmaktadır.*

Keywords: *Bag Of Acoustic Words, Circular HMM, Underwater Acoustics, Time-shift invariant, Sound Recognition, and MFCC.*
Anahtar Kelimeler*: Akustik Kelimeler Kümesi, Çevrimsel Saklı Markov Model, Su altı Akustiği, Zaman-Kayması Bağımsız, Ses Tanıma ve MFCC.*

1. INTRODUCTION

People can easily guess who is singing on the radio from the singer’s voice. It is a simple model that includes ears, cochlea and brain with nerves for humankind. Every person has a unique voice; this characteristic can be used as one of the key factors in recognition and identification processes. Similarly, every ship has a unique combination of engines, propeller, hull shape and crew habits etc. so the noise made by it is unique, and thus can be considered as its "acoustic signature". The term Underwater Acoustic Signal Recognition (UASR) is used for identifying the platforms using their acoustic signature. Unfortunately, that "signature" is heavily distorted by propagation and masked in sea/ocean noise [1]. Due to the interference-filled under sea recognizing an underwater target precisely is always a very difficult task for the navy.

In this paper, we propose two different schemes for UASR. In both schemes, the feature extraction is performed using Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding derived Cepstral Coefficients (LPCCC) which have been extensively utilized in speech recognition. In the first scheme, the features extracted frame by frame are used as a sequence in the representation of the whole signal. The classification of that sequence of vectors is then done by Hidden Markov Models with various topologies. In this paper, we propose to use Circular HMM, which is previously developed for boundary based shape recognition in the computer vision area. The second scheme represents the frame features using Bag of Acoustic Words approach. In training stage, all the feature vectors extracted from the frames of the input signal are first clustered into a set of acoustic words. Given a sequence of feature vectors, each of them is then assigned to an acoustic word. After the frequency of each word is calculated in the input signal, the final representation is performed by the co-occurrence list of the acoustic words. The proposed schemes are summarized in Figure 1-1

In the experiments circular HMM is tested against conventional HMM topologies. It is shown that the circular HMM outperforms the classical HMM topologies both in time and success rates. It is also shown that The Bag of Acoustic Words approach is also performed better than moment based features extracted from the acoustic signal.

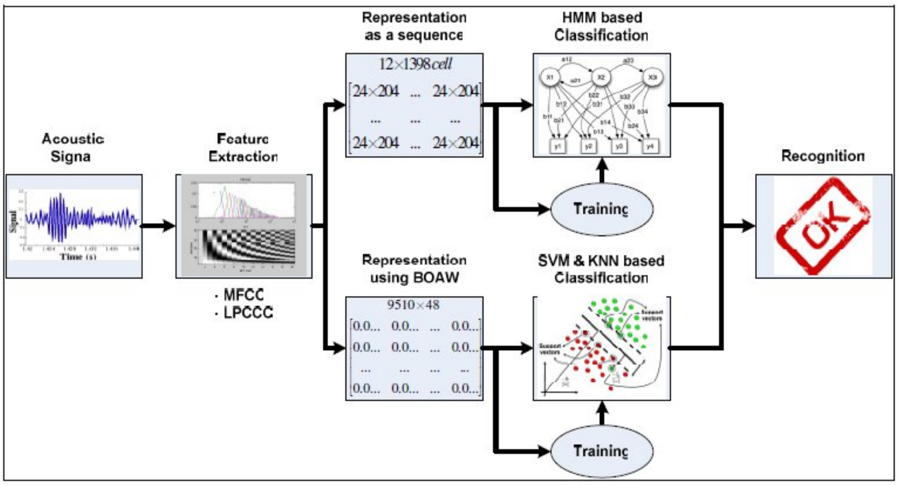


Figure 1 The system which established for classification.

UASR is a critical problem which must be solved in a quite fast and accurate manner for all navies. It is actually an application of Sound Recognition Systems (SRS). The other examples of SRS's are as follows; speaker recognition systems [10] for validating a person's entrance to a restricted building, speech recognition systems for differentiating the given commands in a noisy environment [2], vehicle recognition systems [3] in traffic management, machine error/fault diagnosis/inspection systems [4], finding the species of animals from their sounds [5], [6], [7], inspection of noise, analyzing the earthquake audio waves, finding the mines under water, systems distinguishing music or instruments and lastly environmental sound classification systems.

Although there are many studies for general purpose SRS in the literature, the available UASR studies are very limited. One example work developed in [8] tries to recognize the ship acoustic sounds using MFCC and Support Vector Machines (SVM). In another study [9], the acoustic signals are modeled by Hidden Markov Trees (HMT) in wavelet domain.

The proposed HMM topology is very different from the available topologies used for sound recognition. In our implementation there is no segmentation process to overcome the time shifting problem, because CHMM is shift invariant and gives accurate success rates. The other method which uses the bag of acoustic words is also new to sound signal processing. By the combination with SVM classifier it gives satisfactory results.

2. FEATURE EXTRACTION AND REPRESENTATION OF ACOUSTIC SIGNAL

Obtaining the acoustic characteristics of the speech signal is referred to as Feature Extraction. Linear Predictive Coding (LPC) coefficients, Reflection coefficients (RC), Cepstral coefficients (CC), LPC-derived Cepstral coefficients (LPCCC), Mel-frequency Cepstral coefficients (MFCC) and their variations are commonly used as feature extraction methods especially for speech. Among all popular sound feature extraction techniques, the most functional and efficient ones extract spectral information (in the frequency domain) from sound, because a more concise and easier analysis of acoustic signal can be performed spectrally rather than temporarily (in the time domain) [12]. Spectral analysis is preferred over temporal analysis to discriminate between structures like phonemes and extract source independent features from sound signals.

LPC has been popular for speech compression, synthesis and as well as recognition since its introduction in the late 1960s. It was because; it offered a reasonable engineering approach for especially speech signal analysis. Linear predictive models are widely used in speech processing applications such as low–bit–rate speech coding in cellular telephony, speech enhancement and speech recognition [13].

Cepstral analysis is a special case of homomorphic signal processing. A homomorphic system is defined as a nonlinear system whose output is a linear superposition of the input signals under a nonlinear transformation. Cepstral analysis has become popular in sound recognition since its discovery in the late 1960s, due to the powerful yet simple engineering model of human speech-production behind it [14], [15].

There are no special methods for feature extraction of ship underwater acoustic signals. But if we think of the best method for identifying them is the human ear as we mentioned earlier, then using these analysis for feature extraction will not be wrong because especially MFCC models the human ear with various triangular filters.

3. BAG OF ACOUSTIC WORDS (BoAW) REPRESENTATION

This is a representation model introduced as the "Bag of words model" (BoW) first in natural language processing (NLP) and then it is being used in computer vision, especially for object categorization. In NLP the BoW is a popular method for representing documents. BoW ignores the word orders. The BoW model allows a dictionary-based modelling, and each document looks like a "bag" (thus the order is not considered), which contains some words from the dictionary. Computer vision researchers uses a similar idea for image representation. The BoW representation serves as the basic element for further processing, such as object categorization [16].

The proposed algorithm starts with frame base feature extraction. The features extracted from all the frames of the acoustic signals in the training stage are clustered using the k-means clustering algorithm. Given a set of features, k-means partitions them into k clusters, in which each feature belongs to the cluster with the nearest center. The center of each cluster is chosen as the representative of the features in that cluster. Finally, those representatives are considered to be the acoustic words in he vocabulary. The acoustic signals are then represented by their frequencies of "acoustic words". A histogram of descriptor occurrences is built to characterize an underwater signal. The illustration of proposed approach is given in figure 2.

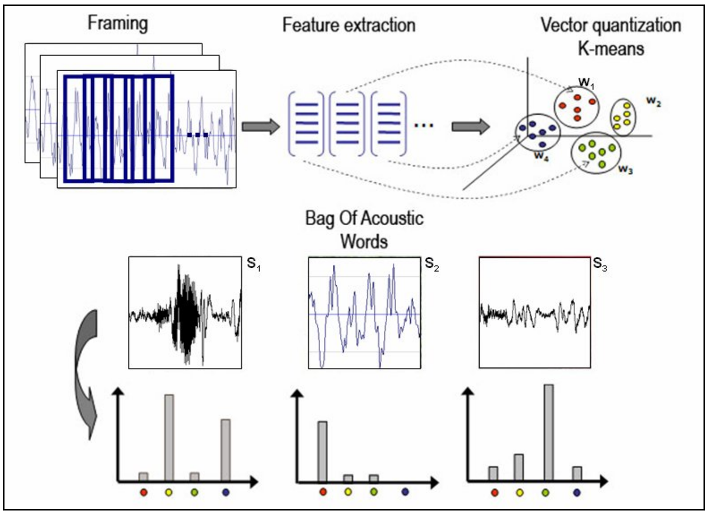


Figure 2: BOAW algorithm illustration. [1]

[1] Graphic adapted from [Bosch, 2008]

The BoAW algorithm is very new to acoustic signal processing field. In the literature review no study was come across. After establishing the BoAW vectors, they are classified using both KNN and SVM recognizers.

4. CLASSIFICATION ALGORITHMS

4.1. Hidden Markov Model (HMM) Recognizer

Signal models can be categorized into two main classes; which we can name deterministic models and statistical models [18]. If we look at the first class, the Dynamic Time Warping (DTW) approaches provide a non-parametric, straight-forward and template-based models which aim to match the incoming acoustic signal to the distinct templates generated by normalizing time variations of speech sounds. Although the DTW approach is relatively simple and straightforward; it has some known limitations in sound recognition. This is because of the variable status of the acoustic signal patterns [19], [20]. The second mentioned class of statistical acoustic signal models much better for non-linear variability in sound signal waveform. For this reason it is outperforming the models of the first class which is deterministic. At this point the use of Artificial Neural Networks (ANN) and the Hidden Markov Models techniques have mainly constituted relatively elaborate and yet extensively-used statistical models in sound recognition field thus far [20].

The basic theory behind the Hidden Markov Models (HMM) goes back to the late 1900s when the Russian statistician Andrei Markov first presented the Markov chains. Baum and his colleagues introduced the Hidden Markov Model as an extension to the first-order stochastic Markov process and developed an efficient method for optimizing the HMM parameter estimation in the late 1960s [21], [22]. Jelinek at IBM and Baker at Carnegie Mellon University outlined the first HMM implementations to use in speech recognition applications in the early 1970s [23].

In the literature usually HMMs are grouped into two main groups based on transition characteristics of model states. The fully connected which is also called ergodic model and left to right model which is known as Baki's model as well. The ergodic model allows each state to be fully connected to the rest of the other states. In Baki's model the next observation vector is allowed to stay in the same state or advance to the next state [24]. It is not allowed to skip states or go back in time generally. In addition to these we will discuss another type of HMM named Circular HMM, implemented in [25].

It is obvious and a well known issue that when you increase the number of states and the non-zero state transitions as well, the complexity of the HMM training process and the recognition process increases exponentially [26].

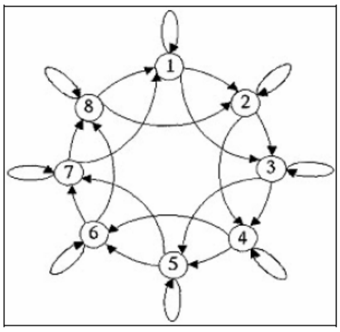


Figure 3: Circular HMM for State=8, Transition=2. [2]

In this work we used Circular HMM in order to defeat time shifting problems and for efficient complexity matters. Circular HMM is a simple modification of left-to-right HMM model. In Baki's model the initial and terminal states are connected through the state transition probabilities. This connection eliminates the need to define a starting point of a cut acoustic signal, in the recognition problem. Since it has no starting and terminating

[2] Graphic courtesy of [ARICA, 2000].

state, it is insensitive to the starting point of the acoustic signal. The Circular HMM topology is both temporal and ergodic [26].

The computational complexity is the same with the left-right model, but the circular HMM has much superiority compared to it. First of all, the circular HMM does not require increasing the number of states as the size of the acoustic signal increases. Therefore, it is size invariant. Secondly, circular HMM does not require as many non-zero state transition probabilities as the classical topologies. As a result of this; the computational complexity of the circular HMM is relatively less than the other models for the same recognition rates [25].

4.2. K Nearest Neighbor (KNN) Recognizer

K-Nearest Neighbour or KNN is an instance based learning algorithm. KNN is part of supervised learning that has been used in many applications in the field of data mining, statistical pattern recognition, image processing and many others. It works based on minimum distance from the query instance to the training samples to determine the K-nearest neighbours. After we gather K nearest neighbours, we take simple majority of these K-nearest neighbours to be the prediction of the query instance [27]. It is a universal approximator. It can model any many to one mapping arbitrarily well. But it has disadvantages; it can be easily confused in high dimensional spaces. That’s why dimensionality reduction techniques are needed and often used. The Model can be slow to evaluate for large training sets [28].

4.3. Support Vector Machines (SVM) Recognizer

Support Vector Machine (SVM) term was first introduced by Boser, Guyon, and Vapnik in 1992. SVM is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over fitting to the data. At the very beginning SVMs were developed for solving the classification problems, but nowadays they have been redesigned to solve regression problems as well [29]. SVM is a classifier that discriminates the data by creating boundaries between classes rather than making estimations about the class conditional densities. That's why it needs considerably less data to perform reliable classification in most of the cases. On the contrary, SVMs are binary classifiers, so some type of strategy must be employed to be able to use them in the multi class problems. There exists some other ways of applying SVMs to the multi-class problem as well [30].

The kernel function plays a critical role in SVM and its performance. The idea of the kernel function is to enable operations to be performed in the input space rather than the potentially high dimensional feature space. Hence the inner product does not need to be evaluated in the feature space [31].

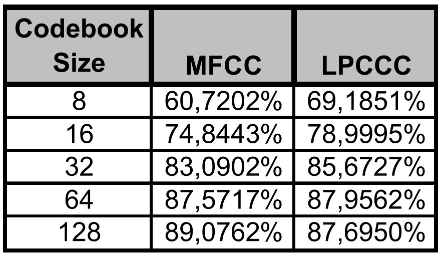
5. PERFORMANCE EVALUATION

The data set contains 116 pieces of audio files from 12 different types of ships. The durations of the files vary between 30 seconds and 4 minutes long. In order to increase the size of the dataset, we perform Additive White Gaussian Noise Addition with random variances to all the audio signals. By this way, we double the number of recordings. In total we have 368 minutes recording. Each recording is in wave PCM signed, 16 bit, 256 kbps and mono format. The playback rate is 16 kHz.

In addition, smaller parts of each recording captures the features needed for classification. For this reason, we partition each recording into the intervals with 2 seconds duration. As a result the final database contains 12390 samples of acoustic signals. We used 2880 pieces that is 240x12 of each class for training our system. The remaining 9510 pieces were used for testing. Approximately 24.44% of the database was arranged for training and 76.76% part was left for testing the implementation. In the experiments, various codebook sizes are tried as shown in Table-1. Another important parameter selection is the state number of the HMM implementations. The number of states and the number of jumps are found empirically as 20 and

10 respectively. The two well known feature extraction methods are compared and the performances are displayed in Table -1.

From these experiments it is clear that both feature extraction methods give similar results. However MFCC is emphasized in this study since it is really robust against tonal changes and it is modeling the human ear. In addition MFCC outperforms LPCCC in the Continuous Time HMM. So we have chosen MFCC as the feature extraction method to feed our recognizer implementation.



Codebook Size	MFCC	LPCCC
8	60,7202%	69,1851%
16	74,8443%	78,9995%
32	83,0902%	85,6727%
64	87,5717%	87,9562%
128	89,0762%	87,6950%

Table 1: MFCC and LPCCC success rates in DHMM.

The performance scores of CHMM and EHMM are very close to each other. Generally CHMM performs 1% greater than EHMM. It should not be forgotten that if the computation times are inspected, CHMM is far away the winner. The elapsed time of CHMM is 29.31 minutes, on the other hand EHMM take 68.86 minutes to complete the computation. In CHMM experiments, the number of Gaussian Mixture Model (GMM) is evaluated. After several experiments approximately the same results are achieved both using 4 GMMs and 6 GMMs. As a result 4 GMMs are preferred in order to save up from computation time. The Continuous time HMM implementation results are highest of all the other implementations. However, the processing time of CHMM is longer than DHMM. Therefore, a selection can be made in time critical cases.

HMM Implementation	HMM TYPE		
	CHMM	EHMM	LRHMM
Discrete time HMM	91,44%	91,05%	88,31%
Continuous time HMM	95,65%	95,94%	93,46%

Table 2: Comparison results of Discrete and Continuous HMM.

We compared our method of BoAW with the general statistical moment (mean and standard deviation) feature extraction method. And approximately 8% higher results are achieved with the BoAW algorithm. The KNN recognizer is used to compare these two methods (Table 3). In the tests done it is clearly observed that the Bag of Acoustic Words (BoAW) method is performing better than the commonly used Moment method. An SVM implementation is also used for a more robust comparison. We realized that the kernel function used in this SVM implementation is very important depending on the dataset in use. Since BoAW is a collection of histograms, histogram intersection kernel gave the most promising results.

Test No	KNN(10)		KNN(5)		KNN(3)		KNN(1)	
	BoAW	Moment	BoAW	Moment	BoAW	Moment	BoAW	Moment
1	82,74%	76,92%	84,15%	79,53%	85,20%	77,23%	87,94%	80,37%
2	87,74%	80,22%	89,48%	82,12%	90,56%	81,49%	91,84%	84,51%
3	89,65%	81,79%	90,83%	82,67%	91,77%	82,82%	93,19%	84,96%
Average	**86,71%**	**79,64%**	**88,16%**	**81,44%**	**89,17%**	**80,51%**	**90,99%**	**83,28%**
Average Success rates (%)								
Average BOAW Success					**88,76%**			
Average Moment Success					**81,22%**			

Table 3: Recognition rates BoAW- Moment feature extraction Tests.

The kernel parameter selection plays a very important role on the performance of the system. However in this study the one against one option of the SVM is experimented, using one against all may perform better.

6. CONCLUSION

The main procedures are focused on developing a speed, environment and platform independent implementation that can be used as a robust solution for the UASR problem. The concept of UASR using HMM with MFCC and BoAW has been investigated throughout this paper. We propose two approaches which are novel in acoustic signal recognition. The first method is Circular HMM topology which is constituted by making simple changes in left to right model. This topology eliminates the need to define a starting point in the acoustic signal leading us to an UASR system without segmentation need of the audio signal.

To our knowledge, BoAW concept is introduced for the first time in acoustic classification. BoAW modeling is based on histogram representations of signals sequences of features. BoAW is simple to implement and use and have the capacity of working coupled with other modern recognition and classification techniques.

The tests indicate that MFCC is quite stable under noisy conditions. It is observed that the CHMM based recognition is insensitive to time shifting. The proposed methods performed considerably well. The circular HMM gives higher recognition rates than the conventional HMM methodologies in UASR. The simulation results show that the proposed CHMM and BoAW algorithms used for UASR system has successfully achieved the desired goal.

REFERENCES

[1] Lobo, V., F. M. Pires, ”Ship Noise Classification Using Kohonen Networks”, *EANN 95*, 1995.

[2] Kurcan, R. Serdar, “Isolated Word Recognition from in-Ear Microphone Data Using Hidden Markov Models (Hmm)”, *NPS Master Thesis*, 2006.

[3] Nooralahiyan, A., H. Kirby, “Vehicle Classification by Acoustic Signature”, *Elsevier Science Ltd.,* 1998.

[4] Lin, Jing, “Feature Extraction of Machine Sound Using Wavelet and Its Application in Fault Diagnosis”, *Elsevier Science Ltd.,* 2001.

[5] Bennett, Richard Campbell. ”Classification of Underwater Signals Using a BACK-Propagation Neural Network”, *NPS Master Thesis,* 1997.

[6] Halkias C., Daniel P., "Estimating the Number of Marine Mammals Using Recordings of Clicks from One Microphone", *Columbia University,* 2006

[7] Urazghildiiev, I., C.W. Clark, T. Krein, "Acoustic Detection and Recognition of Fin Whale and North Atlantic Right Whale Sounds", *Bioacoustics Research Prog., Cornell Laboratory of Ornithology,* 2008.

[8] Alkan, Mahmut. "Warship Sound Signature Recognition Using Mel Frequency Cepstral Coefficients", *Naval Science and Engineering Institute MS Thesis*, 2005.

[9] Yue, Z., K. Wei, X. Qing, "A Novel Modelling and Recognition Method for Underwater Sound Based on HMT in Wavelet Domain", *Springer, Verlag Berlin Heidelberg,* 2004.

[10] Rashidul, H., M. Jamil, G. Rabbani, S. Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients", *3rd International Conference on Electrical & Computer Engineering, Dhaka,* 2004.

[11] Bardici, N., B. Skarin, "Speech Recognition using Hidden Markov Model", *Blekinge Institute of Technology MS Thesis,* 2006.

[12] Vergin, R., "An Algorithm for Robust Signal Modeling in Speech Recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98),* 1998.

[13] Vaseghi, Saeed V., "Advanced Digital Signal Processing and Noise Reduction", *Second Edition, John Wiley & Sons Ltd.,* 2000

[14] Rabiner, L. R., R. W. Schafer, "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey,* 1978.

[15] Picone, J. "Signal Modeling Techniques in Speech Recognition," *Proceedings of the IEEE*, 1993.

[16] Wikipedia, the free encyclopedia web site, 2007, (last accessed on 16 July 2009)

[17] Ayats A.R., "Object Recognition for Autonomous Robots: Comparison of two approaches", *Report of the stay at the Autonomous Systems Lab,* 2007.

[18] Rabiner, L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *IEEE*, 1989.

[19] Gold, B., N. Morgan, "Speech and Audio Signal Processing," *John Wiley & Sons,* 2000.

[20] Deller, J. R., J. Hansen, J. Proakis, "Discrete-Time Processing of Speech Signals", *IEEE Press, New York,* 2000.

[21] Baum L.E., T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics,* 1966.

[22] Baum, L.E., T. Petrie, G. Soules, "A maximization technique in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics,* 1970.

[23] RABINER L. R., B-H. Juang, "Fundamentals of Speech Recognition", *Prentice Hall,* 1993.

[24] Kil, D.H., F. Shin, "Pattern Recognition and Prediction with Applications to Signal Processing (Modern Acoustics and Signal Processing)", 1998.

[25] Arica, N., Fatos. T.Y., "A Shape Descriptor Based on Circular Hidden Markov Model", *Department of Computer Engineering, METU,* 2000

[26] Arica, N., Fatos. T.Y., “A New HMM Topology for Shape Recognition”, *Department of Computer Engineering, METU,* 1995.
[27] Teknomo, K., “K Nearest Neighbors Tutorial”, 2006, people.revoledu.com/, (last accessed on 16 July 2009)
[28] Grudic,G. Nearest Neighbor Learning lecture notes, 2005, www.cs.colorado.edu (last accessed on 16 July 2009)
[29] Vapnik, V., Golowich S., Smola A., ”Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing”, *Cambridge,* 1997.
[30] Temko A., C. Nadeu, “Classification of Acoustic Events Using SVM-Based Clustering Schemes”, *Universitat Politècnica de Catalunya*, 2005.
[31] Jakkula, Vikramaditya, “Tutorial on Support Vector Machine (SVM)”, *School of EECS, Washington State University*, 2006.