Audio Data Mining Using Multi-perceptron Artificial Neural Network

[1]Surendra Shetty, [2]K.K. Achary
[1]Dept of Computer Applications, NMAM Institute Of Technology, Nitte,Udupi, Karnataka, India, PIN- 574110,

[2]Dept of Statistics, Mangalore University, Mangalagangothri, India, PIN-574199,

***Abstract*--**Data mining is the activity of analyzing a given set of data. It is the process of finding patterns from large relational databases. Data mining includes: extract, transform, and load transaction data onto the data warehouse system, store and manage the data in a multidimensional database system, provides data, analyze the data by application software and visual presentation. Audio data contains information of each audio file such as signal processing component- power spectrum, cepstral values that is representative of particular audio file. The relationship among patterns provides information. It can be converted into knowledge about historical patterns and future trends. This work involves in implementing an artificial neural network (ANN) approach for audio data mining. Acquired audio is preprocessed to remove noise followed by feature extraction using cepstral method. The ANN is trained with the cepstral values to produce a set of final weights. During testing process (audio mining), these weights are used to mine the audio file. In this work, 50 audio files have been used as an initial attempt to train the ANN. The ANN is able to produce only about 90% accuracy of mining due to less correlation of audio data.

Index Terms—ANN, Backpropagation Algorithm, Cepstrum, Feature Extraction, FFT, LPC, Perceptron, Testing, Training, Weights

I. INTRODUCTION

Data mining is concerned with discovering patterns meaningfully from data. Data mining has deep roots in the fields of statistics, artificial intelligence, and machine learning. With the advent of inexpensive storage space and faster processing over the past decade, the research has started to penetrate new grounds in areas of speech and audio processing as well as spoken language dialog. It has gained interest due to audio data that are available in plenty. Algorithmic advances in automatic speech recognition have also been a major, enabling technology behind the growth in data mining. Currently, large vocabulary, continuous speech recognizers are now trained on a record amount of data such as several hundreds of millions of words and thousands of hours of speech. Pioneering research in robust speech processing, large-scale discriminative training, finite state automata, and statistical hidden Markov modeling have resulted in real-time recognizers that are able to transcribe spontaneous speech. The technology is now highly attractive for a variety of speech mining applications.

Audio mining research includes many ways of applying machine learning, speech processing, and language processing algorithms [1]. It helps in the areas of prediction, search, explanation, learning, and language understanding. These basic challenges are becoming increasingly important in revolutionizing business processes by providing essential sales and marketing information about services, customers, and product offerings. A new class of learning systems can be created that can infer knowledge and trends automatically from data, analyze and report application performance, and adapt and improve over time with minimal or zero human involvement. Effective techniques for mining speech, audio, and dialog data can impact numerous business and government applications. The technology for monitoring conversational audio to discover patterns, capture useful trends, and generate alarms is essential for intelligence and law enforcement organizations as well as for enhancing call center operation. It is useful for a digital object identifier analyzing, monitoring, and tracking customer preferences and interactions to better establish customized sales and technical support strategies. It is also an essential tool in media content management for searching through large volumes of audio warehouses to find information, documents, and news.

II. TECHNICAL WORK PREPARATION

A. *Problem statement*

Audio files are to be mined properly with high accuracy given partial audio information. This can be very much achieved using ANN. This work involves in implementing supervised backpropagation algorithm(BPA). The BPA is trained with the features of audio data for different number of nodes in the hidden layer. The layer with optimal number of nodes has to be chosen for proper audio mining.

Manuscript received October 5, 2008
Manuscript revised October 20, 2008

B. *Overview of Audio mining*

Audio recognition is a classic example of things that the human brain does well, but digital computers do poorly. Digital computers can store and recall vast amounts of data perform mathematical calculation at blazing speeds and do repetitive tasks without becoming bored or inefficient. Computer performs very poorly when faced with raw sensory data. Teaching the same computer to understand audio is a major undertaking. Digital signal processing generally approaches the problem of audio recognition into two steps, 1) Feature extraction, 2) Feature matching

Each word in the incoming audio signal is isolated and then analyzed to identify the type of excitation and resonate frequency [2]. These parameters are then compared with previous example of spoken words to identify the closest match. Often, these systems are limited to few hundred words; can only accept signals with distinct pauses between words; and must be retrained. While this is adequate for many commercial applications, these limitations are humbling when compared to the abilities of human hearing.

There are two main approaches to audio mining.

1. Text-based indexing: Text-based indexing, also known as large-vocabulary continuous speech recognition, converts speech to text and then identifies words in a dictionary that can contain several hundred thousand entries.
2. Phoneme-based indexing: Phoneme based indexing doesn't convert speech to text but instead works only with sounds. The system first analyzes and identifies sounds in a piece of audio content to create a phonetic-based index. It then uses a dictionary of several dozen phonemes to convert a user's search term to the correct phoneme string. (Phonemes are the smallest unit of speech in a language, such as the long "a" sound that distinguishes one utterance from another. All words are sets of phonemes). Finally, the system looks for the search terms in the index. A phonetic system requires a more proprietary search tool because it must phoneticize the query term, and then try to match it with the existing phonetic string output.

Although audio mining developers have overcome numerous challenges, several important hurdles remain. Precision is improving but it is still a key issue impeding the technology's widespread adoption, particularly in such accuracy-critical applications as court reporting and medical dictation. Audio mining error rates vary widely depending on factors such as background noise and cross talk. Processing conversational speech can be particularly difficult because of such factors as overlapping words and background noise [3][4]. Breakthroughs in natural language understanding will eventually lead to big improvements.

The problem of audio mining is an area with many different applications. Audio identification techniques include,

Channel vocoder
Linear prediction
Formant vocoding
Cepstral analysis

There are many current and future applications for audio mining. Examples include telephone speech recognition systems, or voice dialers on car phones.

C. Schematic diagram

The sequence of Audio mining can be schematically shown as below.

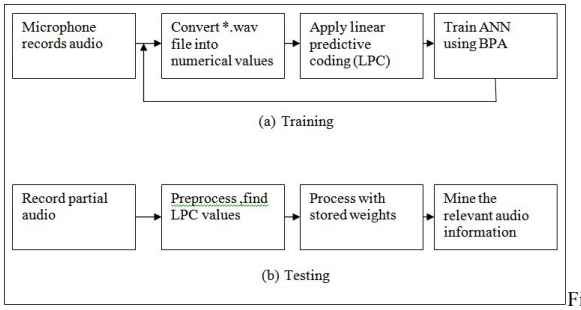

Fig .1. Sequence of audio processing

D. *Artificial neural network*

A neural network is constructed by highly interconnected processing units (nodes or neurons) which perform simple mathematical operations [5]. Neural networks are characterized by their topologies, weight vectors and activation function which are used in the hidden layers and output layer [6]. The topology refers to the number of hidden layers and connection between nodes in the hidden layers. The activation functions that can be used are sigmoid, hyperbolic tangent and sine [7]. A very good account of neural networks can be found in [11]. The network models can be static or dynamic [8]. Static networks include single layer perceptrons and multilayer perceptrons. A perceptron or adaptive linear element (ADALINE) [9] refers to a computing unit. This forms the basic building block for neural networks. The input to a perceptron is the summation of input pattern vectors by weight vectors. In Figure 2, the basic function of a single layer perceptron is shown.

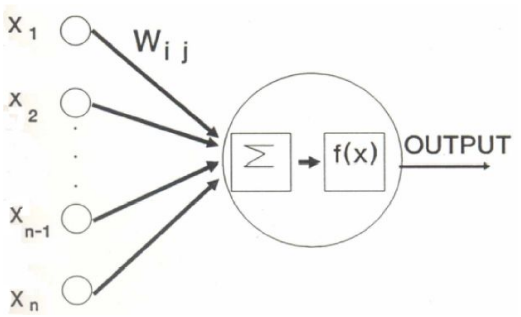

Fig. 2. Operation of a neuron

In Figure 3, a multi layer perceptron is shown schematically. Information flows in a feed-forward manner from input layer to the output layer through hidden layers. The number of nodes in the input layer and output layer is fixed. It depends upon the number of input variables and the number of output variables in a pattern. In this work, there are six input variables and one output variable. The number of nodes in a hidden layer and the number of hidden layers are variable. Depending upon the type of application, the network parameters such as the number of nodes in the hidden layers and the number of hidden layers are found by trial and error method [10]

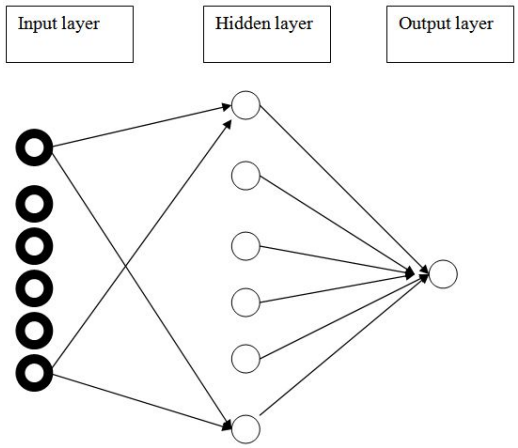

Fig. 3 Multilayer perceptron

In most of the applications one hidden layer is sufficient. The activation function which is used to train the ANN, is the sigmoid function and it is given by:

$$f(x) = \frac{1}{1+\exp(-x)} \qquad (1)$$

where , $f(x)$ is a non - linear differentiable function,

$$x = \sum_{i=1}^{N_n} W_{ij}(p)\, x_i^n(p) + \theta(p),$$

where ,

N_n is the total number of nodes in the n^{th} layer

W_{ij} is the weight vector connecting i^{th} neuron of a layer with the j^{th} neuron in the next layer.

θ is the threshold applied to the nodes in the hidden layers and output layer and

p is the pattern number.

In the first hidden layer, x_i is treated as an input pattern vector and for the successive layers, x_i is the output of the i^{th} neuron of the proceeding layer. The output x_i of a neuron in the hidden layers and in the output layer is calculated by:

$$X_i^{n+1}(p) = \frac{1}{1+\exp(-x+\theta(p))} \qquad (2)$$

For each pattern, error E (p) in the output layers is calculated by:

$$E(p) = \frac{1}{2}\sum_{i=1}^{N_M} (d_i(p) - x_i^M(p))^2 \qquad (3)$$

where

M is the total number of layer which include the input layer and the output layer,

N_M is the number of nodes in the output layer.

$d_i(p)$ is the desired output of a pattern and

$X_i^M(p)$ is the calculated output of the network for the same pattern at the output layer.

The total error E for all patterns is calculated by:

$$E = \sum_{p=1}^{L} E(p) \qquad (4)$$

where, L is the total number of patterns.

E. *Implementation*

The flowchart, fig. 4 explains the sequence of implementation of audio mining. Fifty audio files were chosen. The feature extraction procedure is applied Pre-emphasizing and windowing. Audio is intrinsically a highly non-stationary signal. Signal analysis, FFT-based or Linear Predictor Coefficients (LPC) based, must be carried out on short segments across which the audio signal is assumed to be stationary.

The feature extraction is performed on 20 to 30 ms windows with 10 to 15 ms shift between two consecutive windows. To avoid problems due to the truncation of the signal, a weighting window with the appropriate spectral properties must be applied to the analyzed chunk of signal. Some windows are Hamming, Hanning and Blackman.

Normalization

Feature normalization can be used to reduce the mismatch between signals recorded in different conditions. Normalization consists in mean removal and eventually variance normalization. Cepstral mean subtraction (CMS) is a good compensation technique for convolutive distortions. Variance normalization consists in normalizing the feature variance to one and in signal recognition to deal with noises and channel mismatch. Normalization can be global or local. In the first case, the mean and standard deviation are computed globally while in the second case, they are computed on a window centered on the current time.

Feature extraction method

LPC starts with the assumption that an audio signal is produced by a buzzer at the end of a tube, with occasional added hissing and popping sounds. Although apparently crude, this model is actually a close approximation to the reality of signal production. LPC analyzes the signal by estimating the formants, removing their effects from the signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue.

The numbers, which describe the intensity and frequency of the buzz, the formants, and the residue signal, can be stored or transmitted somewhere else. LPC synthesizes the signal by reversing the process: use the buzz parameters and the residue to create a source signal, use the formants to create a filter, and run the source through the filter, resulting in audio.

Steps:

1. Audio files –in mono or stereo recorded in natural or inside lab, or taken from a standard data base
2. Extracting features by removing noise provided it is a fresh audio, otherwise for existing audio, noise removal is not required
3. Two phases have to be adopted: Training phase and testing phase
4. Training Phase: In this phase, a set of representative numbers are to be obtained from an initial set of numbers. BPA is used for learning the audio files
5. Testing phase: In this phase, the representative numbers obtained in step 4 has to be used along with the features obtained from a test audio file to obtain, an activation value. This value is compared with a threshold and final decision is taken to retrieve an audio file or offer solution to take further action which can be activating a system in a mobile phone, etc.

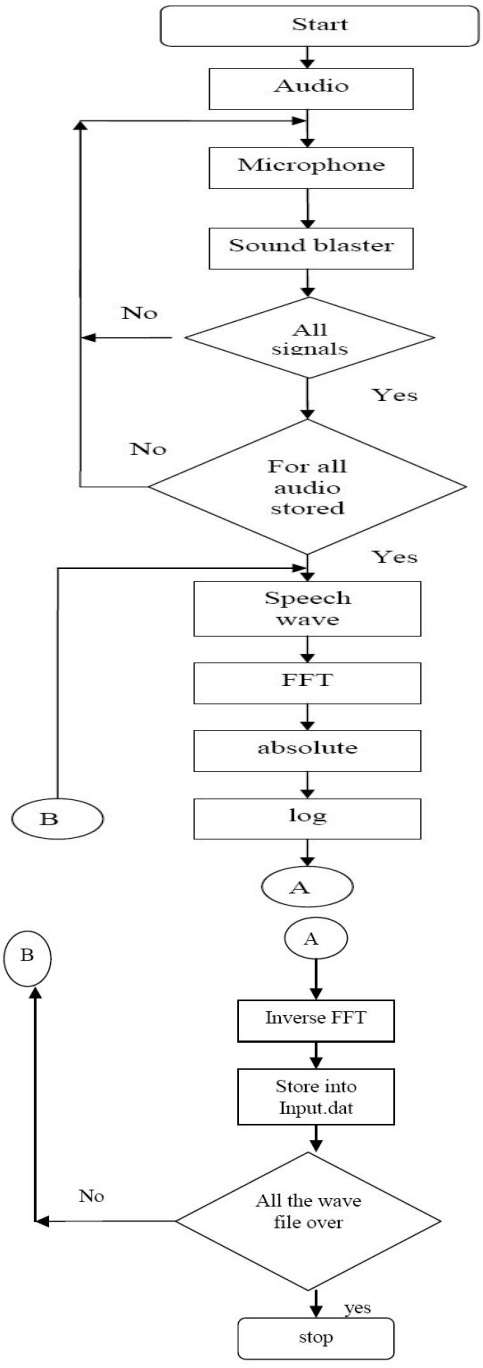

Fig.4 Flowchart for LPC analysis

F. *Results and Discussion*

Cepstrum analysis is a nonlinear signal processing technique with a variety of applications in areas such as speech and image processing. The complex cepstrum for a sequence x is calculated by finding the complex

Natural logarithm of the Fourier transform of x, then the inverse Fourier transform of the resulting sequence. The complex cepstrum transformation is central to the theory and application of homomorphic systems, that is, systems that obey certain general rules of superposition. The real cepstrum of a signal x, sometimes called simply the cepstrum, is calculated by determining the natural logarithm of magnitude of the Fourier transform of x, then obtaining the inverse Fourier transform of the resulting sequence. It is difficult to reconstruct the original sequence from its real cepstrum transformation, as the real cepstrum is based only on the magnitude of the Fourier transform for the sequence. Table 1, gives the cepstral coefficients for 25 sample audio files.

Each row is a pattern used for training the ANN with BPA. The topology of the ANN used is 6 × 6 × 1. In this, 6 nodes in the input layer, 6 nodes in the hidden layer and 1 node in the output layer is used for proper training of ANN followed by audio mining.

TABLE 1 CEPSTRAL FEATURES OBTAINED FROM SAMPLE AUDIO FILES

	Inputs						**Target Labeling**
S. No	F1	F2	F3	F4	F5	F6	
1	1.0509	0.2882	1.3973	1.0009	0.4728	1.4385	**.01**
2	0.5265	0.5915	1.1231	0.4965	0.6709	1.2166	**.02**
3	0.7240	0.4771	1.1208	0.6214	0.4891	1.0801	**.03**
4	0.5834	0.5237	1.0096	0.3820	0.5322	0.8734	**.04**
5	1.1890	0.2859	1.3671	0.9001	0.4903	1.2222	**.05**
6	1.6472	0.2045	1.7316	0.9678	0.7314	1.5301	**.06**
7	0.5673	0.5900	1.1411	0.5513	0.6330	1.2625	**.07**
8	0.5608	0.6301	1.0655	0.5731	0.5404	1.0951	**.08**
9	0.8946	0.3685	1.2516	0.8690	0.4926	1.3202	**.09**
10	3.2865	1.4042	2.8429	0.3122	0.4480	0.0497	**.10**
11	0.5844	0.5030	0.9480	0.5972	0.3445	1.0805	**.11**
12	0.5526	0.5753	1.1038	0.4919	0.6018	1.1156	**.12**
13	0.1067	0.5291	0.7697	0.2701	0.5747	0.8090	**.13**
14	0.4723	0.5623	0.9182	0.3272	0.4989	0.7941	**.14**
15	1.3434	1.3344	0.7108	1.7522	0.5459	1.0386	**.15**
16	1.2278	0.2356	1.5131	1.2207	0.4238	1.6181	**.16**
17	0.7858	0.4885	1.2054	0.6380	0.5747	1.1776	**.17**
18	0.2584	1.3039	0.1095	0.4740	0.2051	0.0947	**.18**
19	0.6000	0.6270	1.0688	0.3905	0.6478	1.0870	**.19**
20	0.8567	0.3006	1.1244	0.7313	0.5416	1.2119	**.20**
21	1.1814	0.3790	1.6468	0.9274	0.6741	1.5006	**.21**
22	3.0731	1.0311	1.4705	2.7696	0.4590	2.3783	**.22**
23	1.2440	0.2540	1.3620	0.6791	0.6099	0.9259	**.23**
24	0.0384	0.5341	0.4769	0.3072	0.1552	0.7388	**.24**
25	0.0384	0.5341	0.4769	0.3072	0.1552	0.7388	**.25**

F1 –F6 are cepstral values. We can choose more than 6 values for an audio file. Target labeling should be less than 1 and greater than zero. When the number of audio file increases, then more decimal values have to be incorporated.

III.CONCLUSION

Audio of common birds and pet animals have been recorded casually. The audio file is suitably preprocessed followed by cepstral analysis and training ANN using BPA. A set of final weights with 6 × 6 × 1 configuration is obtained with 7350 iterations to reach .0125 mean squared error rate. Fifty patterns have been used for training the ANN. Thirty patterns were used for testing (audio mining). The results are close to 90% of mining as the audio was recorded in open. The percentage of recognition and audio mining accuracy has to be tested with large number of new audio files from the same set of birds and pet animals

IV. REFERENCES

[1] Lie Lu and Hong-Jiang Zhang, "Content analysis for audio classification and segmentation.", IEEE Transactions on Speech and Audio Processing, 10:504–516, October 2002.

[2] T. Tolonen and M. Karjalainen, "A computationally efficient multi-pitch analysis model," IEEE Transactions on Speech and Audio Processing, Vol. 8(No. 6):708–716, November 2000.

[3] Haleh Vafaie and Kenneth De Jong, "Feature space transformation using genetic algorithms," IEEE Intelligent Systems, 13(2):57–65, March/April 1998.

[4] Usama M. Fayyad, "Data Mining and Knowledge Discovery: Making Sense Out of Data," IEEE Expert, October 1996, pp. 20-25.

[5] Fortuna L, Graziani S, LoPresti M and Muscato G (1992), "Improving back-propagation learning using auxiliary neural networks," Int. J of Cont. , 55(4), pp 793-807.

[6] Lippmann R P (1987) "An introduction to computing with neural nets," IEEE Trans. On Acoustics, Speech and Signal Processing Magazine, V35, N4, pp.4.-22
[7] Yao Y L and Fang X D (1993), "Assessment of chip forming patterns with tool wear progression in machining via neural networks", Int.J. Mach. Tools & Mfg, 33 (1), pp 89 -102.
[8] Hush D R and Horne B G (1993), "Progress in supervised neural networks", IEEE Signal Proc. Mag., pp 8-38.
[9] Bernard Widrow (1990), 30 Years of adaptive neural networks: Perceptron, madaline and back-propagation, Proc. of the IEEE, 18(9), pp 1415 - 1442.
[10] Hirose Y, Yamashita K Y and Hijiya S (1991), Back-propagation algorithm which varies the number of hidden units, Neural Networks, 4, pp 61-66.
[11] Simon Haykin, Neural Networks-A Comprehensive foundation, 2nd edition.

V. ABOUT THE AUTHORS

Surendra Shetty, is working as Lecturer in MCA Department at NMAMIT, Nitte, Karkala Taluk, Udupi District, Karnataka, INDIA. He has received MCA Degree and pursuing Ph.D degree in the area of Data Mining at Mangalore University. His main research interest includes Data Mining, Artificial Intelligence & Signal Processing.

Dr.K.K.Achary, is a Professor of Statistics in the Dept. of Post-graduate Studies and Research in Statistics, at Mangalore University,India. He holds M.Sc. degree in Statististics and Ph.D. in Applied Mathematics from Indian Institute of Science, Bangalore, India. His current research interests include Stochastic Models, Inventory Theory, Face Recognition Techniques and Audio Data Mining. His research papers have appeared in European Journal of Operational Research, Journal of Operational Research Society, Opsearch, CCERO, International Journal of Information and Management Sciences, Statistical Methods and American Journal of Mathematics and Management.