AUDIO CLIP CLASSIFICATION USING LP RESIDUAL AND NEURAL NETWORKS MODELS

Anvita Bajpai and B. Yegnanarayana

Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai- 600 036, India
{anvita, yegna}@cs.iitm.ernet.in

ABSTRACT

In this paper, we demonstrate the presence of audio-specific information in the linear prediction (LP) residual, obtained after removing the predictable part of the signal. We emphasize the importance of information present in the LP residual of audio signals, which if added to the spectral information, can give a better performing system. Since it is difficult to extract information from the residual using known signal processing algorithms, neural networks (NN) models are proposed. In this paper, autoassociative neural networks (AANN) models are used to capture the audio-specific information from the LP residual of signals. Multilayer feedforward neural networks (MLFFNN) models or multilayer perceptron (MLP) are used to classify the audio data using the audio-specific information captured by AANN models.

1. INTRODUCTION

In this era of information technology, the data that we use is mostly in the form of audio, video and multimedia. The data, once recorded and stored digitally, conveys no significant information in order to organize and use it. The volume of data is large, and is increasing daily. Therefore it is difficult to organize the data manually. We need to have an automatic method to index the data, for further search and retrieval. Audio plays an important role in classifying multimedia data as it contains significant information, and is easier to process when compared to video data. For these reasons, commercial products of audio retrieval are emerging, e.g., (http://www.musclefish.com) [1]. Content-based classification of data into different categories is one important step for building an audio indexing system.

In the traditional approach of audio indexing, audio is first converted to text, and then it is given to text-based search engines [2]. Drawbacks of this approach are: (a) not having accurate speech recognizer, (b) not using speech information present in form of prosody, and (c) not applicable for non-speech data like music. An elaborate audio content categorization is proposed by Wold *et al.* [1], which divides the audio content into sixteen groups. The authors have used mean, variance and autocorrelation of loudness, pitch and bandwidth as audio features and a nearest neighborhood classifier for the task. The authors quote 81% classification accuracy for an audio database with 400 sound files. Guo *et al.* [3] have used features consisting of total power, subband energies, bandwidth, pitch and MFCCs, and support vector machines (SVMs) for classification. Wang *et al.* classify audio into five categories of television (TV) programs using spectral features [4]. Features based on amplitude, zero-crossing, bandwidth, band energy in the subbands, spectrum and periodicity properties, along with hidden Markov model (HMM) for classification are explored for audio indexing applications in [5]. But it was shown that perceptually significant information of audio data is present in the form of sequence of events, which can be obtained after removing the predictable part in the audio data. Perceptually, there are some discriminating features present in the residual which could help in various audio indexing tasks. The challenge lies in developing algorithms to capture these perceptually significant features from the residual, as it is difficult to extract information using known signal processing algorithms.

Objective of this study is to explore the features in addition to the features that are currently used to improve the performance of an audio indexing system. In particular, features not used explicitly or implicitly in the current system are being investigated. Many interesting and perceptually important features are present in the residual signal obtained after removing the predictable part. Thus the main objective of this study is to explore the features present in the linear prediction (LP) residual for audio clip classification task. The reason for considering the residual data for study is that the residual part of the signal is generally subject to less degradation as compared to the system part [6]. The residual data contains higher order correlation among samples. As known signal processing and statistical techniques are not suitable to capture this correlation, an autoassociative neural networks (AANN) model is proposed to capture these higher order correlations among samples of the residual of the audio data. AANN have already been studied to capture information from the residual data for tasks such as speaker recognition [7]. Further, multilayer feedforward neural networks (MLFFNN) models or multilayer perceptron (MLP) are proposed for decision making task using the audio-specific information captured by AANN models.

The paper is organized as follows: Section 2 discusses extraction of the LP residual from audio data. Section 3 discusses AANN models for capturing features in LP residual for the audio clip classification. Section 4 discusses MLP models for decision making. Section 5 presents the workflow of the system. The results of the experimental studies are presented in Section 6. Various issues addressed in this paper and possible directions for the future study are summarized in Section 7.

2. SIGNIFICANCE OF LP RESIDUAL FOR AUDIO CLIP CLASSIFICATION

The first step is to extract the LP residual from the audio signal using linear prediction (LP) analysis [8]. In the LP analysis each sample is predicted as a linear weighted sum of the past p samples, where p represents the order for prediction.

If $s(n)$ is the present sample, then it is predicted by the past p samples as,

$$s'(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

The difference between the actual, and predictable sample value is termed as prediction error or residual, which is given by,

$$e(n) = s(n) - s'(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

The linear prediction coefficients $\{a_k\}$ are determined by minimizing the mean squared error over an analysis frame.

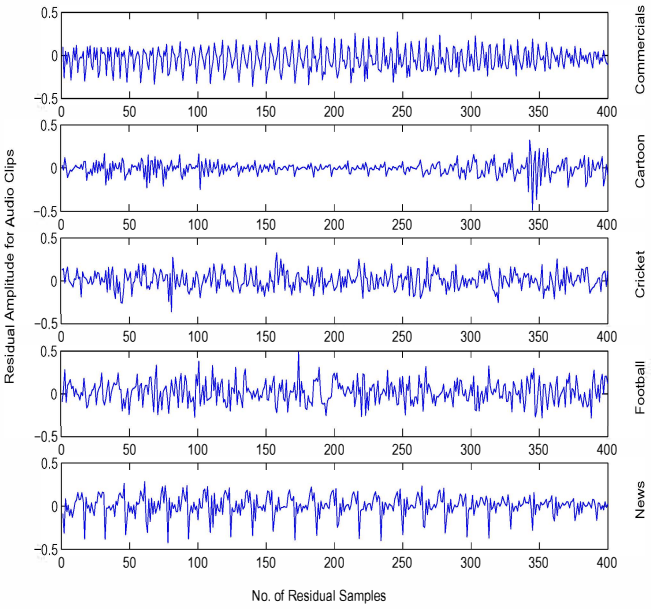

Figure 1: LP residual for the segments of audio clips belonging to five different categories.

The five classes considered for the present study show variations among them. News audio (in closed studio environment) has clean speech, while speech for cartoon category differs from the news speech in terms of prosody. Music is a part of cartoon audio. Cricket and football have casual speech and other background sounds, like noise. Noise is more in the case of football audio. Advertisement audio is the most difficult class to study, as it has many variations within it. The residual signal of the five different classes are shown in Figure 1. For some cases even if the difference cannot be observed in the residual signal, the audio-specific information could be perceived while listening. In the next section, we discuss methods to capture the audio-specific information from the LP residual.

3. AANN MODELS FOR CAPTURING AUDIO INFORMATION IN LP RESIDUAL

Since LP analysis extracts the second order statistical features through the autocorrelation functions, the LP residual does not contain any significant second order correlations of the audio production system. But the excitation source characteristics are present in the LP residual. We conjecture that the audio features may be present in the higher order relations among the samples of the residual signal. Since specific set of parameters to represent the audio information in the LP residual is not clear, and also since the extraction of such an information may involve non-linear processing, we propose neural network models to capture the audio-specific information from the LP residual [9].

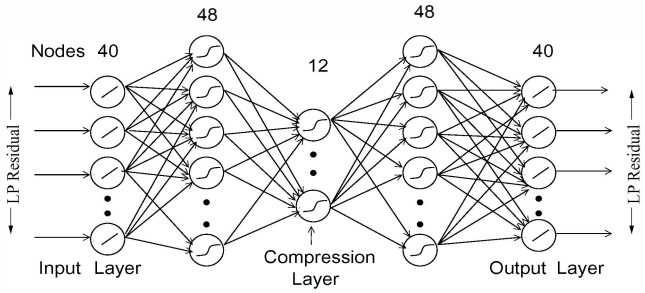

Figure 2: Structure of AANN model used for capturing audio-specific information.

AANN models are feedforward neural networks performing an identity mapping of the input space [10]. AANN models have been shown to capture excitation source features specific to the speaker in the LP residual [7]. For capturing the audio information present in the LP residual signal, a five layer AANN model as shown in Figure 2 is used. The structure of the network used in our study is 40L 48N 12N 40N 40L, where L refers to linear units, and N to non-linear units. A *tanh* is used as the non-linear activation function. Since, we are interested in the source features, the 5 ms duration of the residual signal is chosen. The signal is recorded at 8 kHz sampling frequency and hence the number of nodes in input and output layer is 40. However, this the duration of residual signal to be considered for study has been derived using speech knowledge (which is suitable for the categories considered for study), and may vary for other audio. The number of nodes in hidden layers have been decided experimentally. The performance of the network does not critically depend on the structure of the network [7]. In the following subsection, we discuss the use of audio component knowledge for building AANN models for capturing the audio-specific information present in LP residual signal.

3.1 Use of Audio Component Knowledge for Building AANN Models

In an audio category, there could be one or more audio components. For example, news audio has clean speech, while advertisement audio has music and speech as major components. The knowledge of these components present in audio is used for classification task. The components selected for the study are speech, music and noise (audio which is non-speech and non-music). There are significant

variations in speech and noise in the categories considered for the study. So further three types of speech (clean, conversation and cartoon speech) and two types of noise (cricket and football kind of noise) have been considered as components. We build an AANN model for each of these six components.

The audio data for the classification experiments is collected from TV programs (Indian commercial broadcast channels). For building component AANN models we have collected data of 25 sec duration for each of the six components. The signal is sampled at 8 kHz, and is stored as 16 bit integers. Various parameters for study have been decided experimentally. LP residual is extracted using 12^{th} order LP analysis, and the residual is normalized to unit magnitude before feeding to AANN models. Residual samples have been given in blocks of 40 samples for every sample shift. In different sets of experiments the component AANN models are trained up to 3500 epochs using backpropagation learning algorithm [7].The training error curves for all the six components models are given in Figure 3. Training error reduces as the number of epochs is increased. The rate of convergence of the network is different for different audio components. Depending on the percentage and importance of corresponding component, we could decide the number of epochs.

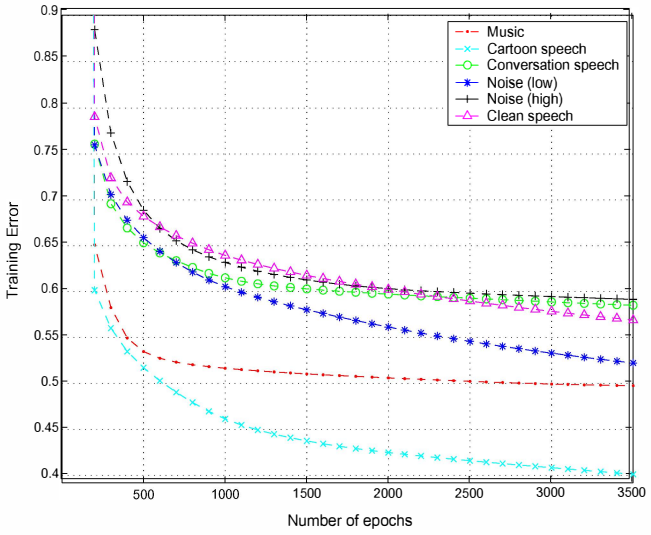

Figure 3: Training error curves for six components used to train the network for audio clip classification.

3.2 Use of AANN Models for Capturing Audio-specific Information in LP Residual

During verification, a test clip of 10 sec duration is used. Test data is processed in the same way as the training data. Blocks of 40 samples of the LP residual are presented with one sample shift to the model. The output of each model is compared with its input to compute the squared error for each block. The error E_i for the i^{th} block is transformed into a confidence value using $C_i = \exp(-\lambda E_i)$, where the constant $\lambda = 1$. The average confidence scores (of 10 samples) of all the six component models for a segments of news test clip are shown in Figure 4. As shown in the Figures 4 for a test clip having a particular audio component, the confidence score values corresponding to the same component AANN model are higher as compared to that of other component models. The category of audio could be decided using this knowledge. This shows the presence of features in the residual part of the audio data for classification task.

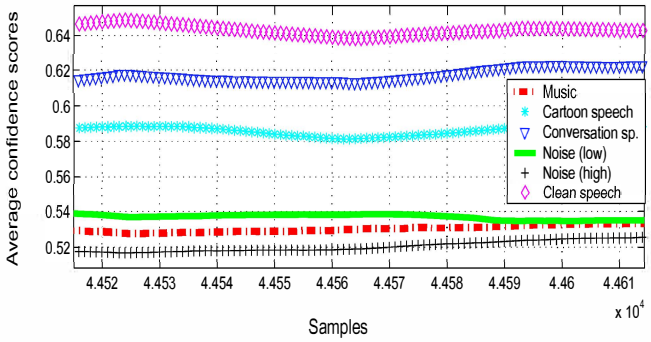

Figure 4: Average confidence score (for 10 samples) values with respect to six components for a segment of news test clip.

The other noticeable observation of the study is that in the case of test clip having speech as a component, models built using speech components give significantly higher confidence scores as compared to non-speech components. But in the case of test clips having non-speech components there is relatively lesser discrimination. Also in the case of speech test clips, the range of confidence score values for speech components is higher as compared to that of non-speech component, as shown in Figure 4. This trend in confidence scores values was observed in 90% of the clips considered for study, and is almost uniform for the 10 sec duration of the clip. This emphasizes the usefulness of the LP residual for audio clip classification, and capability of AANN models to capture the audio-specific information present in the LP residual. The following section discusses the role of MLP for decision making task using the audio-specific information captured by AANN models.

4. MLP MODEL FOR AUDIO CLIP CLASSIFICATION TASK

It is difficult to decide the audio category automatically using the information captured by AANN models. Therefore, we propose MLP to classify the audio clips by using this information. Artificial neural networks have been shown to be suitable for pattern recognition tasks because of their ability to form complex decision surface by using discriminating learning algorithms [9]. The structure of the MLP used in the present study is 6L 24N 12N 5N, where L refers to a linear unit and N to a non-linear unit. The number of nodes in input and output layers are 6 and 5 respectively, as 6 AANN component models are considered to classify audio in 5 categories. The number of nodes in hidden layers have been decided experimentally.

Now in order to decide the audio category using the pat-

tern in confidence scores, confidence scores values of six component AANN models have been considered as a feature (component confidence vector (CCV)). A five class classifier (MLP) has been trained for 1000 epochs using these feature vectors of the five audio clips belonging to five different audio categories. The output of MLP is a 5 dimensional vector, $X = \{x_1x_2x_3x_4x_5\}$. During training of MLP, $x_i \in X$, where $i = 1$ to 5, is set to 1 for CCV belonging to the i^{th} audio class, and -1 for all other $x_j \in X$, where $j = 1$ to 5 and $j \neq i$. During testing, the average output vector of the MLP for all the test CCVs are computed. The class of the node having maximum average value is associated with the input test clip.

5. WORK-FLOW OF THE SYSTEM

The work-flow of the system is shown in Figure 5. The LP analysis is performed on the given test audio signal to obtain the LP residual signal. The residual signal is passed through the six AANN models and confidence scores values are obtained for every frame of the input signal, as described in Section 3.2. These CCVs are passed through the MLP, which decides the category of audio using the pattern present in the CCVs, as described in Section 4. Classification results are given in following section.

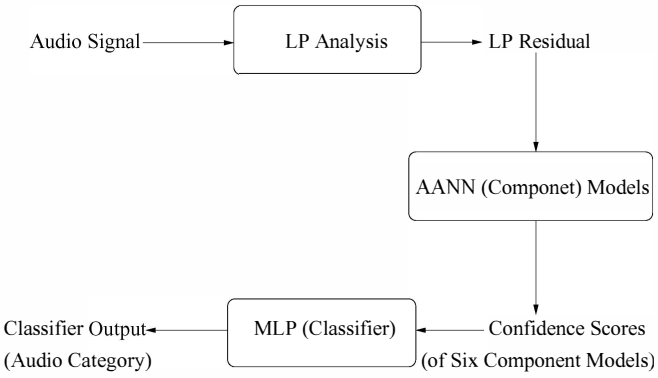

Figure 5: Work-flow of the system for the audio clip classification task.

6. EXPERIMENTAL RESULTS

The results of audio clip classification for 200 test clips are given in Table 1.

Table 1: Audio clip classification results

Audio Class	Number of clips correctly classified out of 200 clips, 40 test clips of each category
Advertisement	34/40
Cartoon	36/40
Cricket	36/40
Football	37/40
News	35/40

7. SUMMARY AND CONCLUSIONS

With ever increasing volume of audio data being collected and used in real life applications, it is imperative to have an efficient means of classifying this audio data for building an audio indexing system. To effectively represent the data in compact form, significant events of the signal need to be captured, and represented as a set of features. Significant part of the audio is present in the LP residual.

In this paper we have shown the importance of audio information in the LP residual. We have shown the capability of AANN models to capture the audio-specific information present in the residual signal, and capability of MLP for deciding audio category using this information. Study needs to be extended to capture more variations in audio categories, and for greater number of audio categories. Further study is needed to explore the combination of features from the residual and spectrum to obtain significantly better performance.

REFERENCES

[1] E. Wold, T. Blum, D. Keslar, and J. Wheaton, "Content-based Classification, Search, and Retrieval of Audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.

[2] J. Makhoul, F. Kubala, R. Leek, D. Lui, L. Nguqen, R. Schwartz, and A. Srivastava, "Speech and Language Technologies for Audio Indexing and Retrieval," *Proc. of the IEEE*, vol. 88, no. 8, pp. 1338–1353, Aug. 2000.

[3] G. Guo, and S. Z. Li, "Content-based Audio Classification and Retrieval by Support Vector Machines," *IEEE Trans. on Neural Networks*, vol. 14, no. 1, pp. 209–215, Jan. 2003.

[4] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia Content Analysis using both Audio and Visual Clues," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, Nov. 2000.

[5] G. Aggarwal, A. Bajpai, A. N. Khan, and B. Yegnanarayana, "Exploring Features for Audio Indexing," Inter-Research Institute Student Seminar, IISc Bangalore, India, Mar. 2002.

[6] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech Enhancement using Excitation Source Information," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Orlando, FL, USA, May 2002.

[7] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source And System Features For Speaker Recognition Using AANN Models," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 2001.

[8] J. Makhoul, "Linear Prediction: A Tutorial Review," *in Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[9] B. Yegnanarayana, *Artificial Neural Networks*, Prentice-Hall of India, New Delhi, India, 1999.

[10] M. A. Kramer, "Nonlinear Principal Component Analysis using Autoassociative Neural Networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, Feb. 1991.