Exploring Features for Audio Clip Classification using LP Residual and AANN Models

Anvita Bajpai and B. Yegnanarayana
Speech and Vision Laboratory
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 linear prediction (LP) residual, obtained after removing the predictable part of the signal. It is known that the residual of a signal is less subject to channel degradations as compared to spectral information. So systems built using the residual may be robust against degradations. This emphasizes the importance of information present in the LP residual of audio signals. But it is difficult to extract information from the residual using known signal processing algorithms. Autoassociative neural networks (AANN) models have been used to capture the distribution of feature vectors for pattern recognition tasks. In this paper, AANN models have been shown to capture audio-specific information from the LP residual of signals to classify audio data.

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, is opaque in nature (in the form of bits), and conveys no significant information in order to organize and use it. The volume of data is large, and is increasing continuously. 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. Classification of data into different categories is one important step for building an audio indexing system. Audio plays an important role in classifying multimedia data, as it is easier to process when compared to video data, and also the audio data contains significant information. For these reasons, commercial products of audio retrieval are emerging, e.g., (http://www.musclefish.com) [1].

In the traditional approach of audio indexing, audio is first converted to text, and given to text-based search engines [2]. But this approach is not applicable for non-speech data like music. Even for speech data the information in the form of prosody cannot be used once speech is converted into text. Also, the accuracy in retrieval depends on the performance of the speech recognizer. Feiten et al. [3] present the audio retrieval work based on content capability. In a content-based audio indexing and retrieval system, the most important task is to identify the category of audio automatically. Depending on application, different categorization could be applied. An elaborate audio content categorization is proposed by Wold et al. [1], which divides the audio content into ten groups at the first level. Furthermore, sounds of one group (music) are classified into six categories of musical instruments at the second level. To characterize the differences among these audio groups, the authors have used mean, variance and autocorrelation of loudness, pitch and bandwidth as audio features. A nearest neighborhood classifier based on weighted Euclidean distance measure was employed. The classification accuracy was about 81% for an audio database with 400 sound files. To classify the same database in above stated method, G. Guo et al. [4] have further refined the work by using perceptual features, composed of total power, subband powers, 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 [5].

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 [6]. But it is shown that perceptually significant information of audio data is present in the form of sequence of events, which is 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 reason for considering residual data for study is that the residual part of the signal is subject to less degradation as compared to the system part of the same [7]. 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 im-

0-7803-8243-9/04/$17.00 © 2004 IEEE

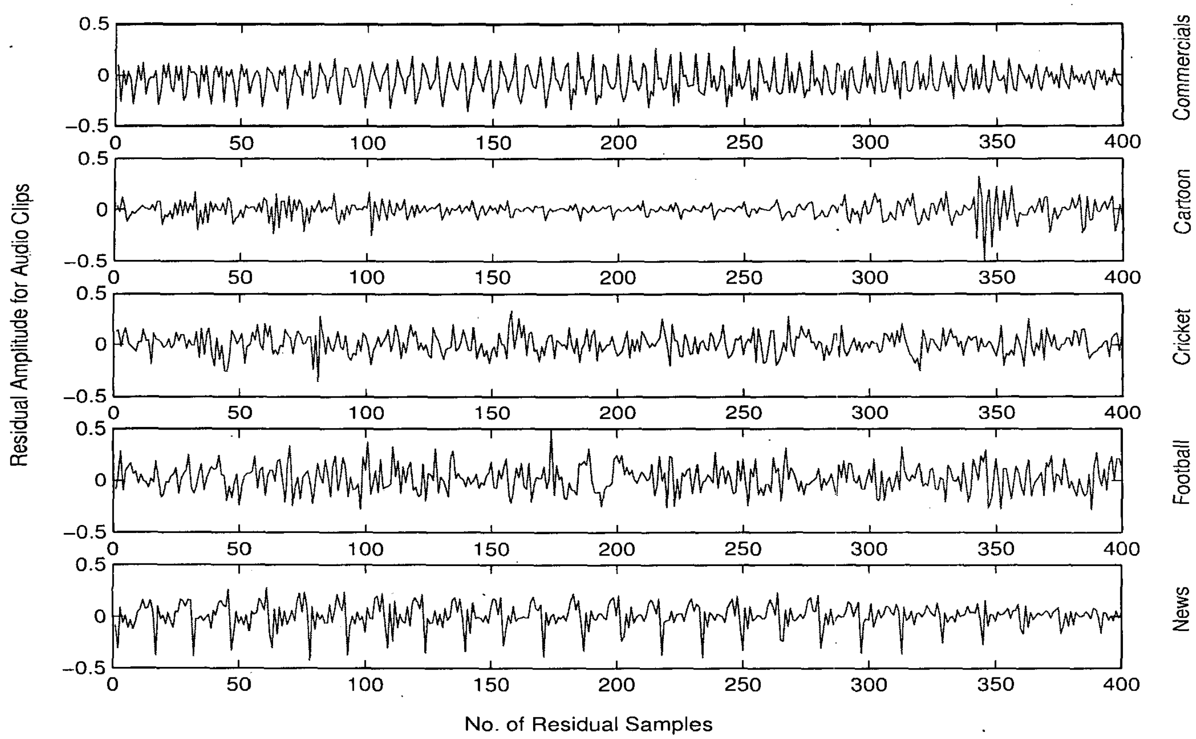

Figure 1. LP residual for the segments of audio clips belonging to five different categories.

portant 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 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 [8].

Organization of the paper is as follows: Section 2 discusses the extraction of LP residual from audio data. Section 3 discusses AANN models for the audio clip classification task. Section 4 presents results of the experimental studies. Various issues addressed in this paper and possible directions for the future study are summarized in section 5.

2. SIGNIFICANCE OF LP RESIDUAL FOR AUDIO CLIP CLASSIFICATION

The first step in using the LP residual for audio clip classification is to extract it from the audio signal. One method of doing it by using linear prediction (LP) analysis [9]. In 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,

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

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

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

The linear prediction coefficients $\{a_k\}$ are determined by minimizing the mean squared error over an analysis frame. It has been known that the LP order used for extracting the residual plays a crucial role on the performance of audio, speech and speaker recognition system [10].

The five classes considered for the present study show variation among them. News audio 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. And advertisement is the most difficult class to study, as it has too many variations within it. Variation in the residual of five different classes is shown in Figure 1. But for some cases even if the difference cannot be observed, 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 matrix, the LP residual does not contain any significant second order statistics corresponding to 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 nonlinear processing, we propose neural network models to capture the audio information from the LP residual [11].

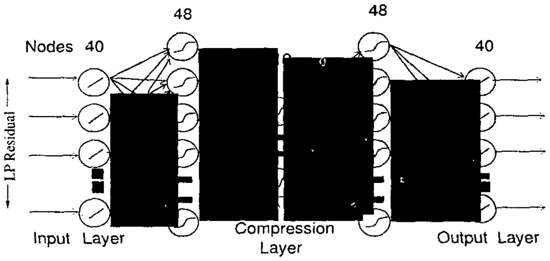

Figure 2. Structure of AANN model used for capturing audio-specific source information.

AANN models are feedforward neural networks performing an identity mapping of the input space [12]. AANN models have been shown to capture excitation source features specific to the speaker in the LP residual [8]. 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 nonlinear units. A *tanh* is used as the nonlinear activation function. The performance of the network does not critically depend on the structure of the network [13].

4. AUDIO CLIP CLASSIFICATION STUDIES

In an audio category, there could be one or more audio components. The knowledge of these components present in each audio category is used for classification task. The components selected for the study are speech, noise and music. But there are significant variations in speech and noise of different types in the categories considered for the study. So further three types of speech (clean, conversation and cartoon speech) and two types of noise (one with speech, and other pure noise) have been considered as components for this study.

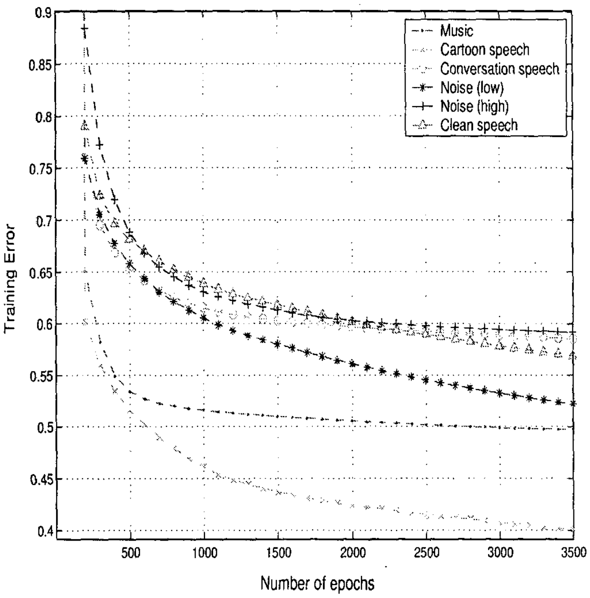

Figure 3. Training error curves for six components used to train the network for audio clip classification.

The audio data for the classification experiments is collected from TV programs. For building component models we have collected data of 25 sec duration for each of the 6 components considered for study. The signal is sampled at 8kHz, and is stored as 16 bit integers. 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 components models are trained for 1 to 3500 epochs using back-propagation

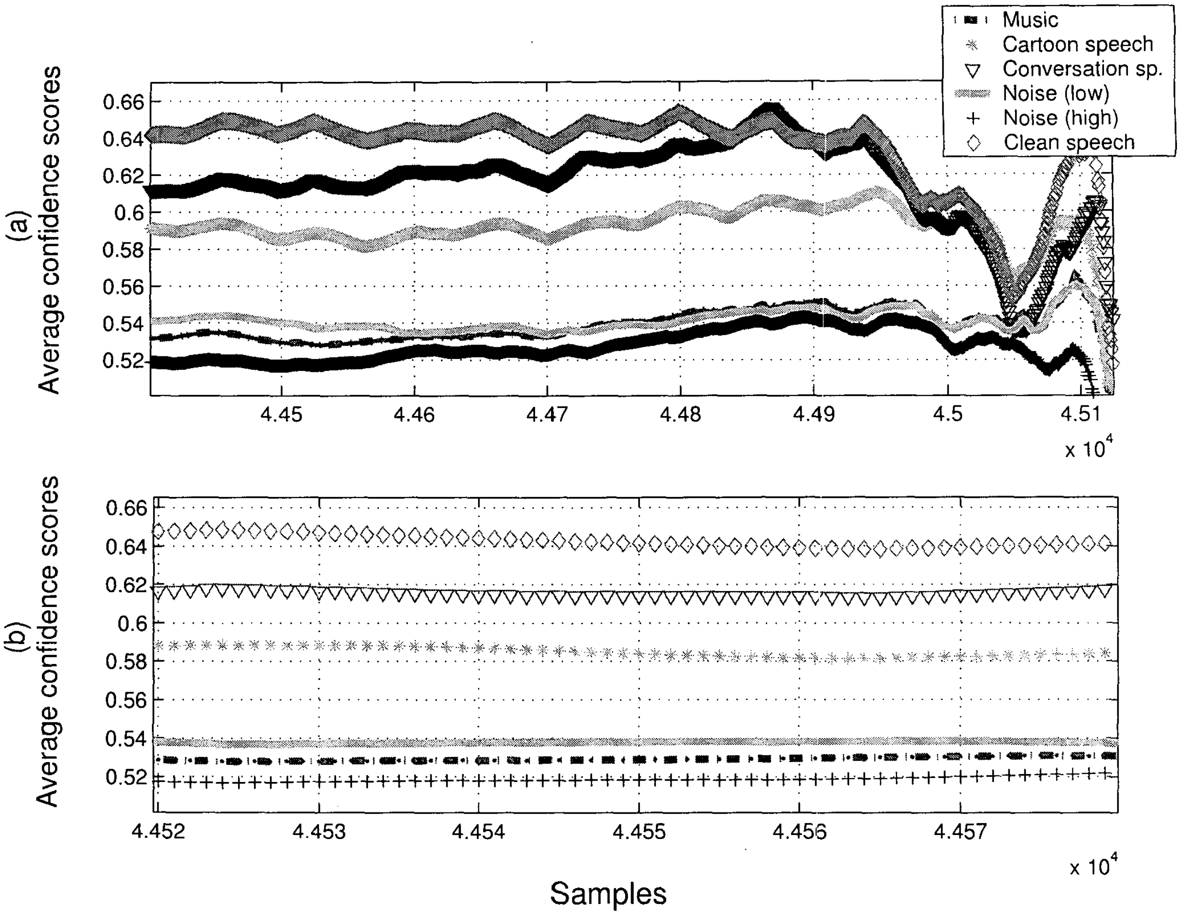

Figure 4. (a)Average confidence score (for 10 samples) values with respect to six components for a segment of news test clip (b) Expended version for a segment.

learning algorithm [10]. The training error curves for all the six components of audio are given in Figure 3. Training error comes down as we increase the number of epochs. The rate of convergence of the network is different for different audio categories. Depending on the percentage and importance of corresponding component, we could decide the number of epochs.

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 are presented with one sample shift to each of the models. 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(-λE_i), where the constant λ = 1. The average confidence (of 10 samples) for a segment of all the components are shown in Figure 4. As shown in the Figures 4 and 5, for a test clip having a particular audio component, the confidence score values corresponding to the same audio component are higher as compared to other audio components. The category of audio could be decided using this knowledge. This shows that there are some features in the residual part of the audio data, which could be explored for audio clip classification task.

The other noticeable observation of the study is that in the case of test clip having speech as a component, 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 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 5.

35 test clips belonging to 5 different categories, 7 test clips

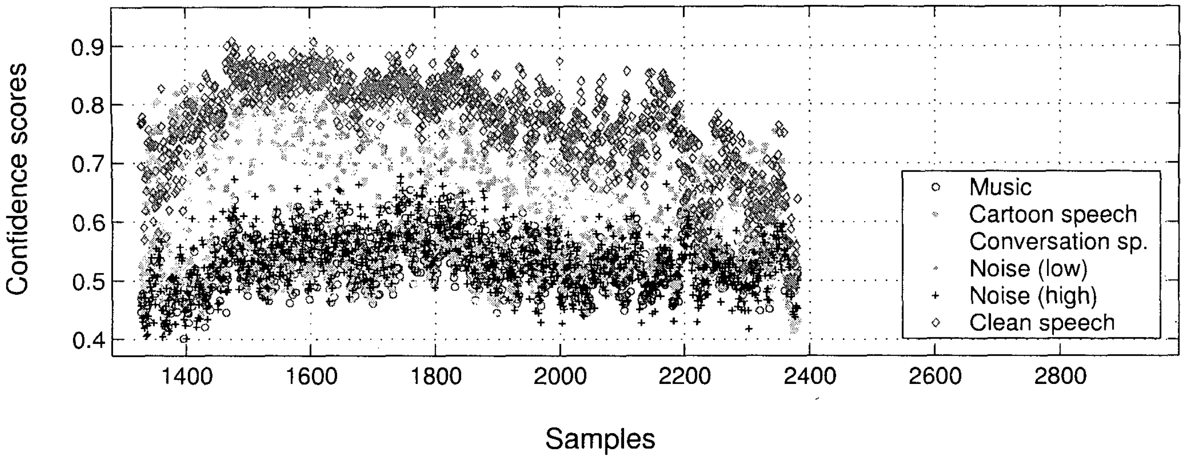

Figure 5. Confidence scores values with respect to six components for a segment of news test clip.

for each category, each of 10 sec duration were given as test utterances. The above stated trend in the values of confidence scores was observed in 90% of the cases. The other noticeable point is that the trend is almost uniform for the 10 sec duration of the clip. It means that even by considering a small duration of the audio clip it may be possible to classify it. 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.

5. 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 into different categories for building a system for indexing, and retrieval relevant to audio data. To effectively represent the data in compact form, significant events of the signal need to be captured, and represented as a set of features. Normally significant part of the audio might be contained in the LP residual. Since the residual does not contain any significant spectral information, systems built using this method might provide robustness against channel and device effects, which are known to degrade the performance of the systems built using spectral information.

In this study we have shown the importance of audio information in the LP residual, and capability of an AANN model to capture the audio-specific information in the residual. Further study is needed to explore the combination of features from the residual and spectrum to obtain a 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] B. Feiten, and S. Gunzel, "Automatic Indexing of a Sound Database using Self-organizing Neural Nets.," Computer Music Journal, vol. 18(3), pp. 53–65, 1994.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Autoassociative Neural Network Models for Speaker Verification using Source Features," in Proc. Int. Conf. Cognitive and Neural Systems, Boston, USA, 2002.

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

[10] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Source Information from Linear Prediction Residual for Speaker Recognition," communicated to J. Acoust. Soc. Amer., 2002.

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

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

[13] K. S. Reddy, "Source and System Features for Speaker Recognition," Sep. 2001.