USE OF SUPRASEGMENTAL FEATURES PRESENT IN LP RESIDUAL FOR AUDIO CLIP CLASSIFICATION

Anvita Bajpai

Applied Research Group
Satyam Computer Services Ltd.
Bangalore, INDIA

anvinota@gmail.com

ABSTRACT

In this paper, presence of the audio-specific suprasegmental information in the Linear Prediction (LP) residual signal is demonstrated. The LP residual signal is obtained after removing the predictable part of the audio signal. This information, if added to existing audio classification systems based on segmental and subsegmental features, can result in better performing combined system. The audio specific suprasegmental information can not only be perceived by listening to the residual, but can also be seen in the form of excitation peaks in the residual waveform. However, the challenge lies in capturing this information from the residual signal. Higher order correlations among samples of the residual are not known to be captured using standard signal processing and statistical techniques. The Hilbert envelope of residual is shown to further enhance the excitation peaks present in the residual signal. A pattern specific to an audio class is also observed in the autocorrelation sequence of the Hilbert envelope. An audio-specific pattern is also observed the statistics of this autocorrelation sequence. This indicates the presence of the audio-specific suprasegmantal information in the residual signal. Support Vector Machine (SVM) is used to classify the patterns in the variance of the autocorrelation sequence for the audio clip classification task.

***Index Terms*—** Audio classification, Suprasegmental features, Linear prediction residual, Hilbert envelope, Support Vector Machines (SVMs)

1. INTRODUCTION

Large volume of the multimedia data is in use today for various applications. Statistics show that the volume of the multimedia data in use is about 10^5 times to that of the text data (`http://www.sims.berkeley.edu/research/projects/how-much-info/summary.html`). Content-based analysis of the multimedia data [1] is important for targeting and personalization applications. However, opaque nature of this data (data stored in form of bits and sampling frequency) does not convey any signification information about its contents to the user. These facts make the task more challenging. Audio plays an important role in handling the multimedia data as it is easier to process when compared to the video data, and also the audio data contains perceptually significant information. Audio indexing is the task of dealing with analysis, storage and retrieval of the multimedia data based on its audio content [2, 3]. Classification of the audio data into different categories is one important step in building an audio indexing system.

The information about the audio category is contained in the excitation source (subsegmental), system/physiological (segmental) and behavioral (suprasegmental) characteristics of the audio data. For humans, the information about the audio category is perceived by listening to a longer segment of audio signal. This other level of information contained in the audio signal is the suprasegmental information, that is the variation of the signal over long duration (typically 50 ms to 200 ms) in case of speech. These behavioral characteristics of the audio data are perceived in the Linear Prediction (LP) residual [4] of audio signal also. Sometimes this difference may not be noticed in the waveform, but can be perceived while listening to the signal. The residual of a signal may be less affected by channel degradations as compared to the spectral information [5]. Hence, it is worthwhile to explore these features for audio clip classification task. This paper emphasizes the importance of the suprasegmental information present in the LP residual of the audio signals. An audio classification system based on suprasegmental features, if combined [6] with existing systems based on the segmental [7, 8, 9] and subsegmental [10, 11] features, can give a better performing audio classification system.

The classes considered for study are advertisement, cartoon, cricket, football and news. The behavioral characteristics of the audio is also perceived in the form of the sequence of the excitation peaks in the LP residual of signal for the five audio classes considered for study. The excitation peaks can further be enhanced using the Hilbert envelope [12] of the residual signal. The gap between the excitation peaks corresponds to the pitch period in the case of speech. The pitch period varies for different audio signals. The categories considered for the study are combination of various audio components, like, speech and music [10]. Hence, the patterns in the excitation peaks in the Hilbert envelope for different audio categories are also a combination of periodicities of the audio components, which varies for different categories. The pattern in these excitation peaks leads to a pattern in the peaks in the autocorrelation sequences of the Hilbert envelope for the five audio categories. It further leads to different statistical distribution of autocorrelation peaks for different audio categories. This emphasizes the presence of the audio-specific suprasegmental information in the LP residual signal. Support Vector Machines (SVMs) are used to classify the pattern in the variance of the autocorrelation sequence for the audio clip classification task. The classification accuracy achieved is 60%, on a test database of 100 audio clips (20 clips for each class) recorded from TV broadcast. Each clip is recorded for about 10 sec. duration with 8 KHz sampling frequency.

Section 2 discusses the presence of the suprasegmental information in the LP residual signal. Section 3 discusses the Hilbert envelope, and also discusses the presence of the audio-specific suprasegmental information in the Hilbert envelope. The methods to extract suprasegmental information from the Hilbert envelope are discussed in Section 4. In Section 5, SVMs have been discussed for pattern classification task. Section 6 presents the experimental results. Section 7 concludes the paper.

2. SUPRASEGMENTAL FEATURES IN THE LP RESIDUAL SIGNAL

2.1. Computation of LP Residual from Audio Signal

The first step is to extract the LP residual from the audio signal using linear prediction (LP) analysis [4]. 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, 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.

2.2. Suprasegmental Features in the LP Residual Signal

The behavioral characteristics of one audio category differ from that of the other. In Fig. 1, the LP residual signals for five audio categories are shown. In Fig. 1, one may notice some differences in the patterns in the residual signals of five audio categories. Sometimes this difference may not be noticed in the waveform, but could be perceived while listening to the residual signal. Hence, it is worthwhile to explore these features for audio clip classification task.

Patterns in the LP residual signal are in the form of a sequence of excitation peaks. These excitation peaks can be considered as event markers. The sequence of these events contain important perceptual information about the source of excitation and behavioral characteristics of audio. By listening to the residuals of different types of audio clips, one can distinguish between speakers, music, instruments, etc. Excitation peaks can further be enhanced by taking the Hilbert envelope of residual signal. The Hilbert envelope computation removes the phase information present in the residual, thereby leading to better identification of the excitation peaks.

3. THE HILBERT ENVELOPE OF THE LP RESIDUAL SIGNAL

3.1. Computation of the Hilbert Envelope from Residual Signal

The residual signal is used to compute the Hilbert envelope, where the excitation peaks show up prominently. The Hilbert envelope is defined as,

$$h_e(n) = \sqrt{e^2(n) + h^2(n)} \tag{3}$$

where $h_e(n)$ is the Hilbert envelope, $e(n)$ is the LP residual and $h(n)$ is the Hilbert transform of the residual. The Hilbert transform of a signal is the 90^0 phase shifted version of the original signal. Therefore, the Hilbert envelope represents the magnitude of the analytic signal,

$$x(n) = e(n) + ih(n) \tag{4}$$

where $x(n)$ is the analytic signal, $e(n)$ is the residual and $h(n)$ is the Hilbert Transform of the residual.

The Hilbert envelope computation removes the phase information present in the residual. This leads to emphasis of the excitation peaks. The excitation peaks are further emphasized by using the neighborhood information of each sample in the Hilbert envelope. The modified Hilbert envelope is

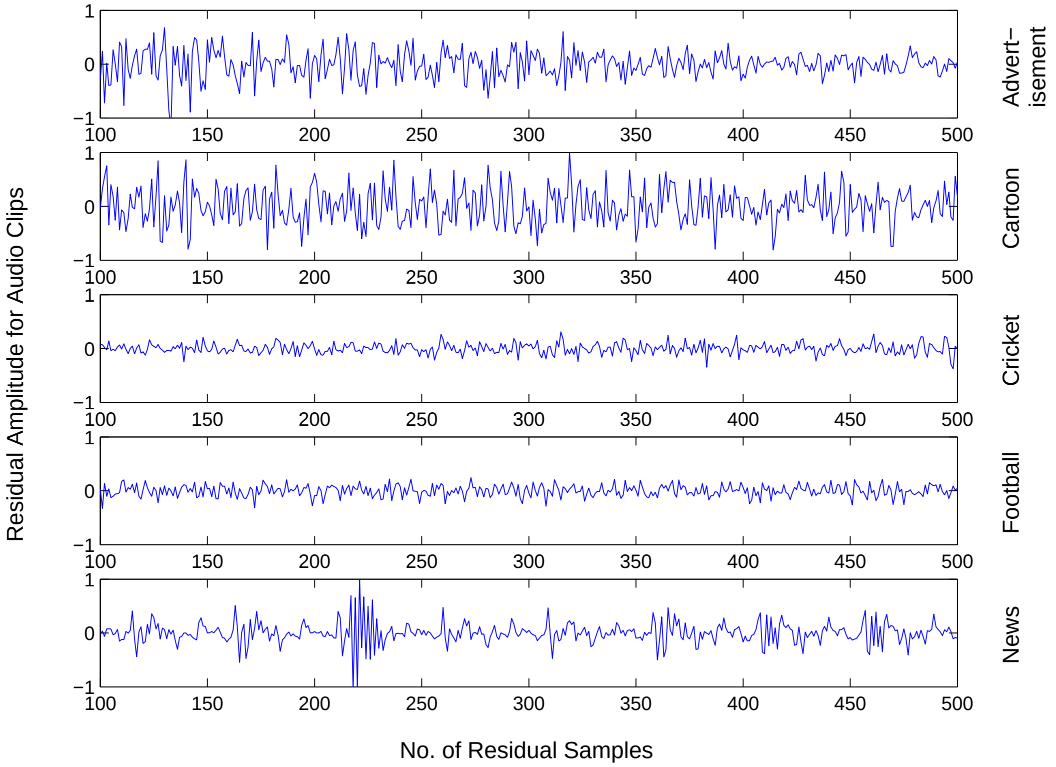

Fig. 1. The LP residual for the segments of audio clips belonging to five audio categories.

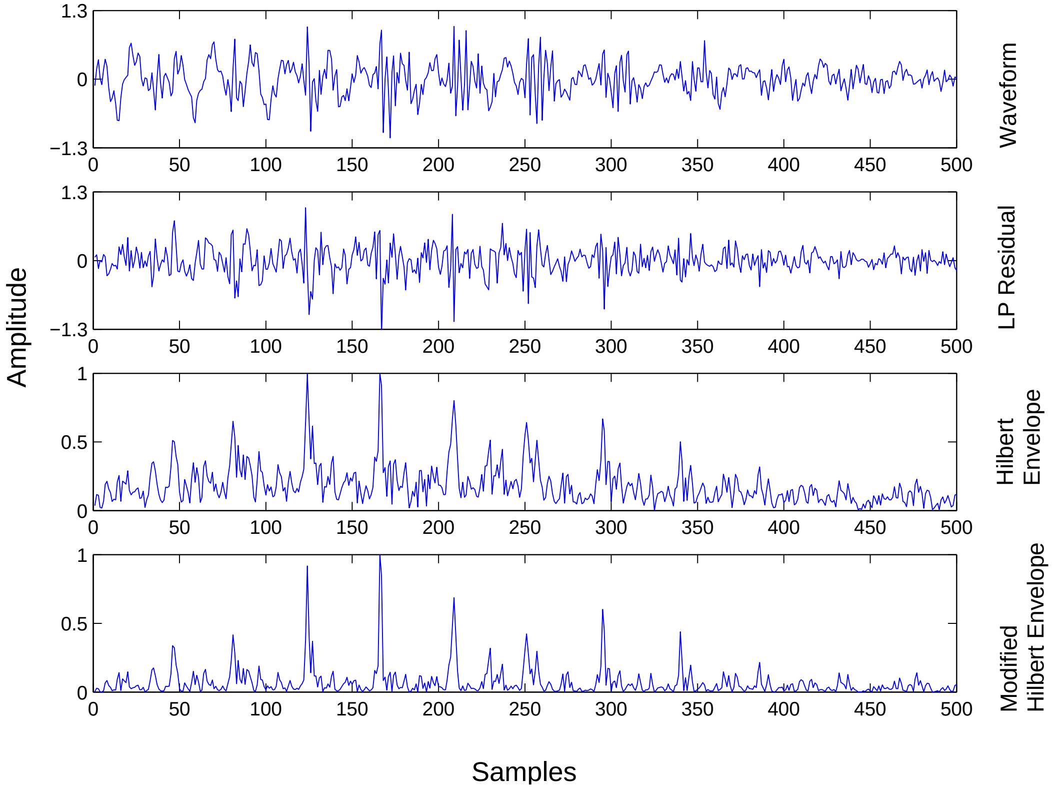

Fig. 2. The waveform, LP residual, Hilbert envelope, modified Hilbert envelope of the residual signal for a noisy audio segment.

computed as,

$$h_{em}(n) = \frac{h_e^2(n)}{\sum_{k=n-l}^{k=n+l} \frac{h_e(k)}{(2l+1)}} \tag{5}$$

where $h_{em}(n)$ is the modified Hilbert envelope, $h_e(n)$ is the Hilbert envelope and l is the number of samples on either side of the neighborhood of current sample n.

3.2. Presence of Suprasegmental Features in the Hilbert Envelope

Fig. 2 shows the residual, the Hilbert envelope and the corresponding modified Hilbert envelope for a noisy audio (speech) segment. It can be noticed that the excitation peaks are clearly visible in the modified Hilbert envelope[1]. The gap between excitation peaks is the pitch period in case of speech signal, which varies for different audio signals. The categories considered for the study are combination of various audio components. Hence the patterns in the excitation peaks in the Hilbert envelope for different audio categories are also a combination of events of the audio components. As shown in Fig. 3, the pattern over a longer duration of segment, in excitation peaks is different for different audio categories, hence it could be utilized for audio clip classification task.

4. EXTRACTION OF SUPRASEGMENTAL FEATURES PRESENT IN THE HILBERT ENVELOPE

As discussed in previous section, there is audio-specific information at the suprasegmental level in the LP residual signal, which can be perceived by listening to the signal, and it can also be observed in the Hilbert envelope of the LP residual signal, as shown in Fig. 3. One method to capture this pattern in the Hilbert envelope is by taking the autocorrelation of the Hilbert envelope. For a segment of 100 ms of the Hilbert envelope the autocorrelation sequence is calculated. The reason for choosing 100 ms window size for calculation of the autocorrelation is that the long term characteristics of the audio signal are of interest. The window is further shifted by 50 ms, and calculation of the autocorrelation is repeated till whole length of the audio clip is considered. These autocorrelation sequences (starting from 3^{rd} sample from the center peak to 400^{th} sample, normalized with respect to the central peak) are plotted in Fig. 4 for a clip for each of the five audio categories considered.

It can be seen in Fig. 4 that for a news audio there is a sharp peak in autocorrelation sequence around 60^{th} sample, and the pattern is relatively uniform. While for cartoon audio the peaks occur at an interval of (around) 30 samples. No clear peak distribution is found in football, advertisement, and noisy regions of cricket audio clips, but peak strengths are different in autocorrelation sequences for these three categories. In clear commentary regions of cricket audio, the pattern is similar to that of news clip. Hence the distribution of these peaks gives an evidence of suprasegmental characteristics of different audio categories. The variance of autocorrelation sequences along frame sequence axis for each of 3^{rd} to 200^{th} sample are calculated. The variance for five test clips belonging to different audio classes, is plotted in Fig. 5. A pattern specific to audio classes is observed in variance plots. This pattern in variance of autocorrelation sequences is utilized for audio clip classification using SVMs. SVMs are well-known for their good generalization performance [13].

5. SUPPORT VECTOR MACHINES FOR CLASSIFICATION

Support vector machines [13] for pattern classification are built by mapping the input patterns into a higher dimensional feature space using a nonlinear transformation (kernel function), and then optimal hyperplanes are built in the feature space as decision surfaces between classes. Nonlinear transformation of input patterns should be such that the pattern classes are linearly separable in the feature space. According to Cover's theorem, nonlinearly separable patterns in a multidimensional space, when transformed into a new feature space are likely to be linearly separable with high probability, provided the transformation is nonlinear, and the dimension of the feature space is high enough [14]. The separation between the hyperplane and the closest data point is called the margin of separation, and the goal of a support vector machine is to find an optimal hyperplane for which the margin of separation is maximized. Construction of this hyperplane is performed in accordance with the principle of structural risk minimization that is rooted in Vapnik- Chervonenkis (VC) dimension theory [15]. By using an optimal separating hyperplane the VC dimension is minimized and generalization is achieved. The number of examples needed to learn a class of interest reliably is proportional to the VC dimension of that class. Thus, in order to have a less complex classification system, it is preferable to have those features which lead to lesser number of support vectors. The performance of the pattern classification problem depends on the type of kernel function chosen. In this work, we have used radial basis function as the kernel, since it is empirically observed to perform better than the other types of kernel functions.

6. EXPERIMENTAL RESULTS

The experimental results of audio clip classification based on suprasegmental features present in Hilbert envelope of LP residual using are shown in Table 1. The variance samples sequence, derived using autocorrelation sequences of Hilbert

[1] For this study modified Hilbert envelope is considered. So in following text whenever Hilbert envelope is mentioned, it actually refers to the modified Hilbert envelope.

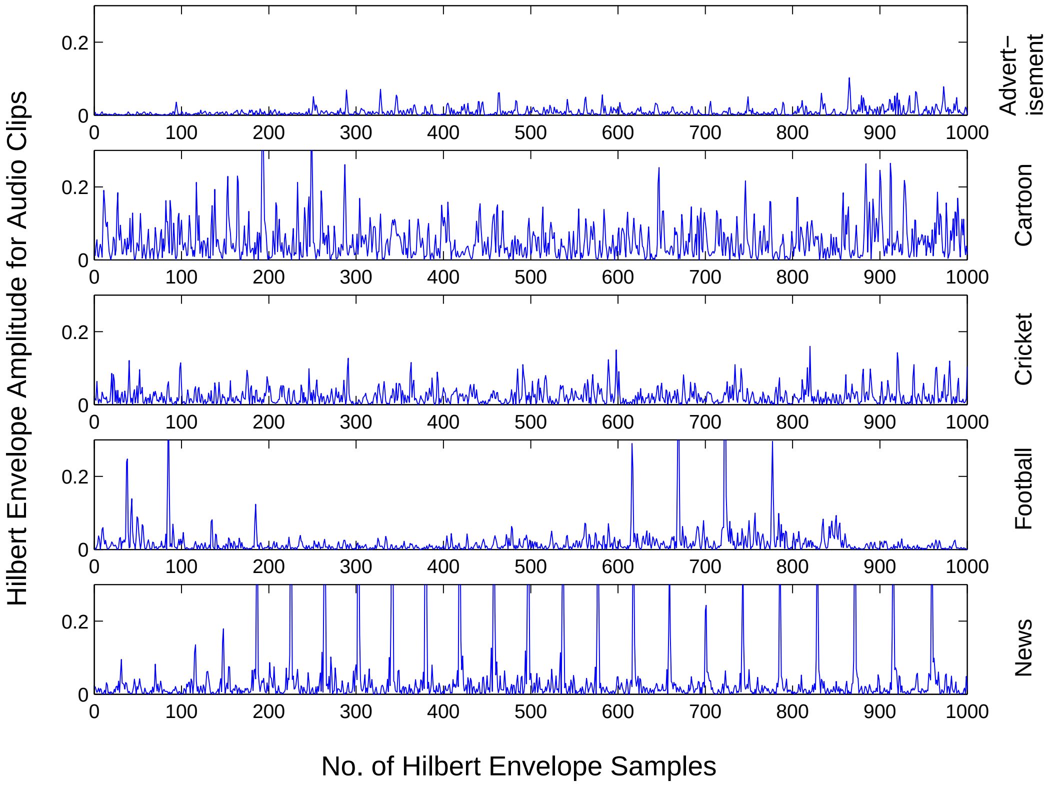

Fig. 3. The Hilbert envelope of the residual signal for the segments of audio clips belonging to five audio categories.

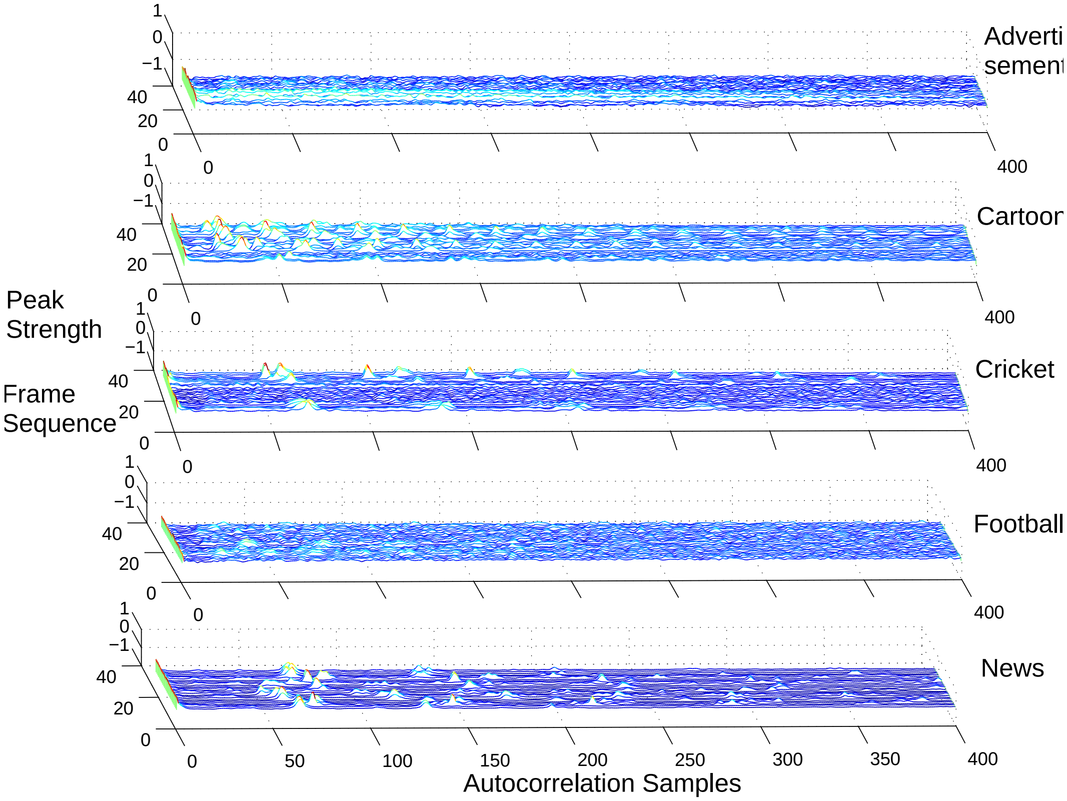

Fig. 4. autocorrelation sequences of the Hilbert envelope of the residual signal of five audio categories.

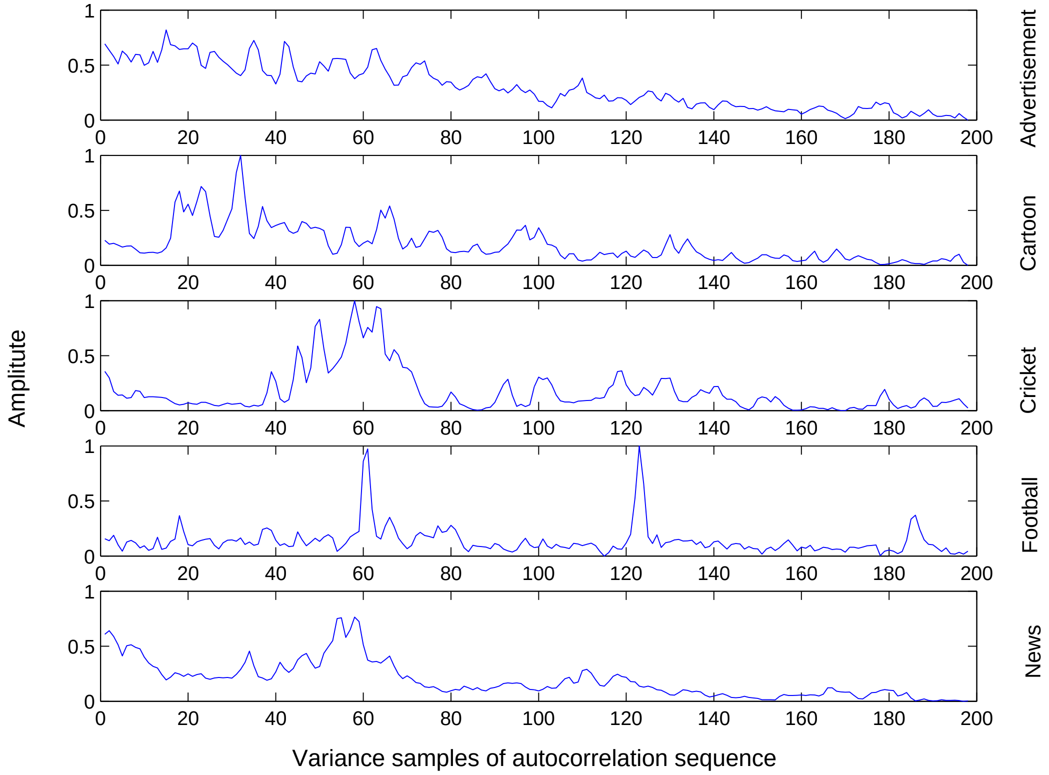

Fig. 5. Variance of autocorrelation sequences of the Hilbert envelope of five audio categories (for 5 test clips).

Table 1. Audio clip classification performance based on suprasegmental features using SVM.

Audio Class	No. of clips correctly classified (out of 20 test clips for each class)
Advertisement	11
Cartoon	19
Cricket	16
Football	04
News	10

envelope, is used as feature vector for SVMs. 75 audio clips (15 clips for each audio class) are used for training SVM. Each clip is recorded from TV broadcast, for about 10 sec. duration with 8 KHz sampling frequency. It can be observed from Table 1 that this technique gives an average accuracy of 60% on a test database of 100 audio clips (20 clips for each class).

The five classes considered for the present study show variations 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. Advertisement audio has has many variations within it, and music is a part of advertisement audio also. In Table 1, it can be observed that for Cartoon class, the performance is maximum. This is because there is a well defined pattern that can be distinguishable perceived in cartoon class. Analysis of data also supports this point. For humans also at times cricket clip may sound like news (when game is slow), or advertisement clip may sound like sports clip (for kids related advertisements), while cartoon clips have distinct suprasegmental characteristics. Results of advertisement, cricket and news data also be supported by the analysis of data and pattern in variance and autocorrelation sequences. The clips belonging to football class are very noisy, which result in poor pattern in the autocorrelation and variance sequences.

The calculation of LP residual, Hilbert envelope and autocorrelation sequence is done based on speech knowledge, for which the results are given in Table 1. However, the parameters used for study may not be appropriate for various audio classes. Hence a detailed study for obtaining the optimal class-specific parameters needs to be conducted to improve the performance of the system.

7. CONCLUSIONS AND FUTURE WORK

The information present in audio signal can be categorized at three levels - subsegmental, segmental and suprasegmental. In this paper the presence of audio-specific suprasegmental features in the LP residual signal is discussed as an additional evidence for audio clip classification task. The pattern in the excitation peaks in the LP residual for different audio categories is enhanced by taking the Hilbert envelope of the residual signal. The statistics of the peak distribution in the autocorrelation sequences of the Hilbert envelopes are noticed to be different for the five audio categories. The variance of autocorrelation sequences is utilized for audio clip classification using SVMs. For future work, the 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.

8. REFERENCES

[1] 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.

[2] N. V. Patel and I. K. Sethi, "Audio Characterization for Video Indexing," in *Storage and Retrieval for Image and Video Databases (SPIE)*, San Jose, USA, Feb. 1996.

[3] J. Boreczky and L. Wilcox, "A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Seattle, WA, May 1998, pp. 3741–3744.

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

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

[6] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah and C. S. Gupta, "Combining Evidence from Source, Suprasegmental and Spectral Features for a Fixed-Text Speaker Verification System," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 4, July 2005.

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

[8] 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.

[9] 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.

[10] Anvita Bajpai and B. Yegnanarayana, "Audio Clip Classification using LP Residual and Neural Networks Models," in *Proc. European Signal and Image Processing Conference*, Vienna, Austria, Sep. 2004.

[11] Anvita Bajpai and B. Yegnanarayana, "Exploring Features for Audio Clip Classification using LP Residual and Neural Networks Models," in *Proc. Int. Conf. Intelligent Signal and Image Processing*, Chennai, India, Jan. 2004.

[12] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.

[13] R. Collobert and S. Bengio, "Svmtorch: Support vector machines for large-scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.

[14] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan College Publishing Company, New York, 1994.

[15] V. Vapnik, *Statistical Learning Theory*, New York: John Wiley and Sons, 1998.