Exploring Features for Audio Indexing

Gaurav Aggarwal, Anvita Bajpai, A. Nayeemulla Khan and B.Yegnanarayana

Abstract— **The objective of this paper is to emphasize the need for identifying and extracting suitable features for storage and efficient retrieval of data in the context of audio indexing. There are a wide variety of audio indexing tasks such as, identifying the speaker, language, music of different classes (Pop, Classical, Jazz, etc.), music from different instruments and classification of audio clips (commentary, news, football, advertisement etc.). Each type of task requires features specific to that task. Normally gross and measurable parameters or features based on amplitude, zero-crossing, bandwidth, band energy in the sub-bands, spectrum and periodicity properties are used in audio indexing applications. But it is shown that perceptually significant information of the audio data is present in the form of sequence of events. It is a challenge to extract features from these sequences of events. In this paper we demonstrate through various illustrations, the importance of residual data obtained after removing the predictable part in the audio data. This residual data seems to contain perceptually significant information in the audio data. But it is difficult to extract that information using known signal processing algorithms.**

I. Introduction

The problem of audio information retrieval is familiar to the one faced by a person returning from a vacation to find his answering machine full of messages. Trying to retrieve the important call from his boss, if any, he has to go through a whole lot of messages, may be the entire tape. A wide variety of audio data, ranging from music, speeches by individuals, conferences, news broadcasts, TV programs and telephone conversations are being recorded and stored in large volumes worldwide, and made available online for public access, as there are no longer any constraints on data storage. To classify and index such huge collections of audio data manually is impractical. This highlights the need for automatic methods of indexing.

A. Requirements of Audio Indexing

The goal of an audio indexing system is to provide the user with capabilities to locate, manipulate, browse, retrieve and index the audio data in an efficient manner, and to achieve real-time performance. The system should be able to present relevant results based on the user queries. It should contain capabilities for skimming through the results (playback audio at a faster rate without changing the pitch), so that the user finds the desired audio file quickly.

Most type of manual audio classification is based on meta-data like the name of the composer or band, the song title, the type of music, the year, and sometimes based on the lyrics. The description of a piece of audio (e.g. type of music) by one person may not match the description by another (there are 350 different varieties of music categorized at *http://www.mp3.com*). If the user knows some of these meta-data, he would be able to locate the required audio clip. More often than not the user does not know these. If the user is interested in Rock/Jazz/Classical music, then the organization of the audio files should be in terms of such classes of data for efficient retrieval.

B. Categories of Audio Indexing tasks

For efficient retrieval, some sort of audio similarity measure for organizing similar sounds into one category is desirable. Neural network based clustering of audio files [1] and self organizing maps (SOM) [2] have been used to organize similar sounds. Likewise at Muscle Fish (*http://www.musclefish.com*), Wold *et al.* have used a vector of attributes like loudness, pitch, bandwidth and harmonicity [3] to organize the sounds into similar classes. Such type of classification comes under unsupervised categorization of audio files. Supervised categorization of audio implies that the user is searching for a known/familiar audio segment. In doing so, an exact clip of the original file may be presented or the tune of the music clip may be hummed. In such cases, distances between the histograms of a discriminately trained vector quantizer may be used [4] to categorize different training classes or query by humming technique developed by Ghais *et al.* [5].

C. Examples of Audio Indexing Tasks

Audio indexing encompasses speaker identification, language identification, musical instrument identification, music classification (Jazz/Pop/Classical), music/speech discrimination (music/speech/silence), audio classification (category of audio file - commentary/advertisement/interview etc.) among others.

In speaker identification, we are interested in classifying an unknown speech utterance, as belonging to one of the set of M reference speakers. The primary use being to retrieve speech by a specific speaker (say Gandhiji's speech). Language identification is an im-

Speech and Vision Laboratory, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, Chennai - 600036, India. E-mail: {gaurav@neuron, {anvita@, nayeem@,yegna@}}cs.iitm.ernet.in

portant issue when we are retrieving a discourse in a particular language. (It is known that Sathya Sai Baba normally speaks only in Telugu). This will help in eliminating all other audio files not of this language, and thus narrow down the search.

As the volume of music recorded, stored and made available is large when compared to any other audio data, locating a particular audio track of say, a particular instrument is a difficult task. Algorithms to identify a particular instrument based on its special characteristics would improve music retrieval performance. This type of classification falls under musical instrument identification. More often than not a user is interested in a particular class of music like Pop, Jazz, Rock, Heavy metal, Classical, Instrumental etc. In such cases, classification of music into appropriate classes is an important aspect of audio indexing. For speech recognition or audio content retrieval, one needs to analyze the audio clip only in the regions of interest say, music or speech. For this one needs to distinguish between such regions. This is called music/speech discrimination.

Another important aspect of audio indexing is to classify whether the audio clip contains music, commentary, advertisement, cartoon, interview or some similar content. This is termed as audio clip classification. It is normally a closed-set problem, where we identify the audio clip as one among the known audio classes. We explain this task briefly in Section II.

In spoken document retrieval (SDR), the goal is to retrieve the desired document from a large collection of spoken documents like news bulletins. The task is to retrieve documents that have similar content or the keywords specified in the query by the user. Conventional approach is to convert the document to text before indexing and retrieval.

The documents are converted into text and indexed as follows. Initially the audio file is segmented into parts containing speech by a single speaker (speaker segmentation). The identity of the speaker among a set of M speakers is established (speaker identification). Thereafter using a speech/non-speech detection module speech regions are extracted. Speaker adapted models are used by the speech recognizer on these speech segments to produce the corresponding text. The boundaries of a story in a document are identified by story segmentation techniques. Next, the possible topics that could be associated with the story are identified (topic classification). For example "cyclonic storm in Chennai ...", the topics that may be associated with this story are, cyclone, weather, rain and Chennai. Names are important keywords as they occur less frequently and may uniquely identify a story. The user mostly queries using such names. A name spotting module spots the names. This procedure results in an indexed database. Based on the queries from the user, the relevant documents are retrieved from the indexed database. In combination with information retrieval techniques, the task of SDR can be represented schematically as shown in Figure 1.

D. Issues in Audio Indexing

The primary issue in audio indexing is the storage of audio data in a form suitable for indexing and retrieval. Using segments of data as indexing terms is infeasible due to vast amount of data. It is essential to extract features from this data which are representative of the data. For example, when a human reads a car number sequence, he reads the number as a whole, and doesn't scan pixel by pixel of the number plate. This represents the characteristic of extracting salient information from the data. In the audio domain, the feature to extract is a perceptual attribute. Humans perceive certain key events and anchor their recognition of the sound through these events. The aim of this paper is to highlight the presence of these perceptual features, suitable for different audio indexing tasks.

The organization of the paper is as follows. In Section II the importance of features and feature extraction is highlighted. In section III the importance of the excitation information for audio indexing is explained. Derivation of features from the residual is discussed in Section IV. Section V summarizes the ideas discussed in this paper.

II. Importance of Features and Feature Extraction

If the audio data can be efficiently characterized by a set of parameters which capture essentially all the key attributes of the original data, we can use this as a feature set. These features must occur frequently, should be easily measurable, should not vary with time, and must not be distorted by channel distortion and background noise [6]. These features may be derived from the data or may be based on the knowledge of the characteristics of the data. For example, if the data is telephone speech then maximum frequency content is 4kHz; if the language spoken is Tamil then aspirated sounds cannot occur; if the data is of a particular instrument then the frequency band would be in a particular range. This knowledge of the data guides in deriving appropriate features. But in all such cases, we cannot derive a feature that has universal discrimination capability, to discriminate between speakers, languages, music, musical instruments etc. Each feature has its own limitations on the ability to capture the information in the given data. This necessitates the need for a separate set of features depending on

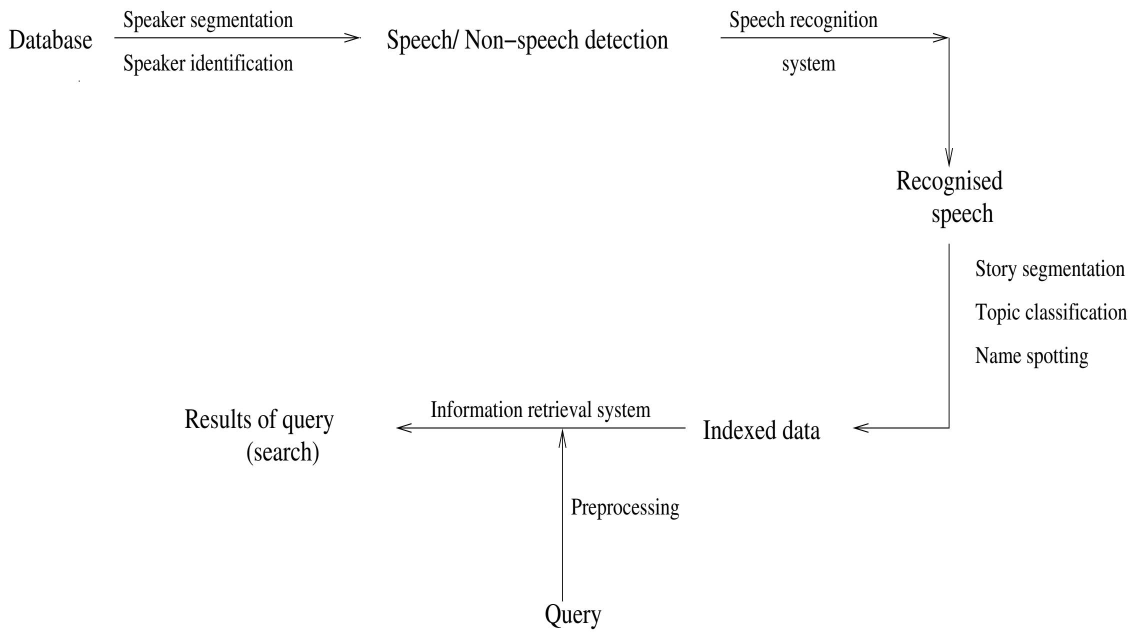

Fig. 1. Schematic diagram showing the task of SDR

the task at hand.

A. Features used for Audio Indexing

The set of features commonly used for speaker recognition are formants of retroflex vowels and nasals, voice-pitch frequency, amplitude spectrum of vowels and nasals, slope of the glottal source spectrum, word duration and pitch frequency [6] [7]. Prosodic features like mean F_0 [8], mean and variance of the pitch periods in voiced segments and energy contours of speech [9] have been used. The first four statistics of the pitch, i.e., mean, variance, skew and kurtosis have also been used [10].

In the traditional approach of indexing spoken documents, where speech is first converted to text prior to indexing, cepstral coefficients have been the predominant choice as features for training Hidden Markov Models (HMMs). For the general case of audio indexing, features like energy contours, zero crossing rate and spectral centroid have been used [1]. Loudness, pitch, bandwidth and harmonicity have been used to classify music [11]. Music/speech discrimination has been done using the average length of time in which peaks exist in a narrow frequency range [12]. To retrieve music, query by humming has been tried by Ghais *et al.* [5].

In the case of language identification prosodic features that have been explored are speech rate, syllable timing etc. [13]. Pitch contour, differential pitch, size (distance between syllables and syllable duration), the differential size (differenced distance between syllables and syllable duration), shape of amplitude contour, differential amplitude, rhythm (low frequency Fourier Transform of the amplitude envelope) and phrase location, were found useful for language identification [13].

The wideband spectrogram for a typical speech waveform is shown in Figure 2. The zero crossing rate, displayed next, is high for unvoiced regions. It is low for voiced speech. The pitch contour plot shows distinct roll off as the syllable ends, and resets to a higher value at the beginning of the next syllable.

For the audio content classification task we use a set of fourteen features. The volume based features are standard deviation of volume, dynamic range of volume, volume undulation [14] and 4Hz modulation energy [11]. The zero crossing rate based features used are standard deviation of zero crossing rate and non-silence ratio (the ratio of number of non-silence frames to the total number of frames in the clip). Pitch based features like pitch contour, pitch standard deviation, similar pitch ratio (the ratio of the number of neighboring frames having similar pitch to the total number of frames), non-pitch ratio (the ratio of frames with zero pitch, i.e., frames for which pitch is not defined), and spectrum based features like frequency centroid, bandwidth and ratio of energy in various frequency sub-bands to the total energy were also used.

Using a 10 ms frame size, the gross features are computed over a one second duration clip. Some of the

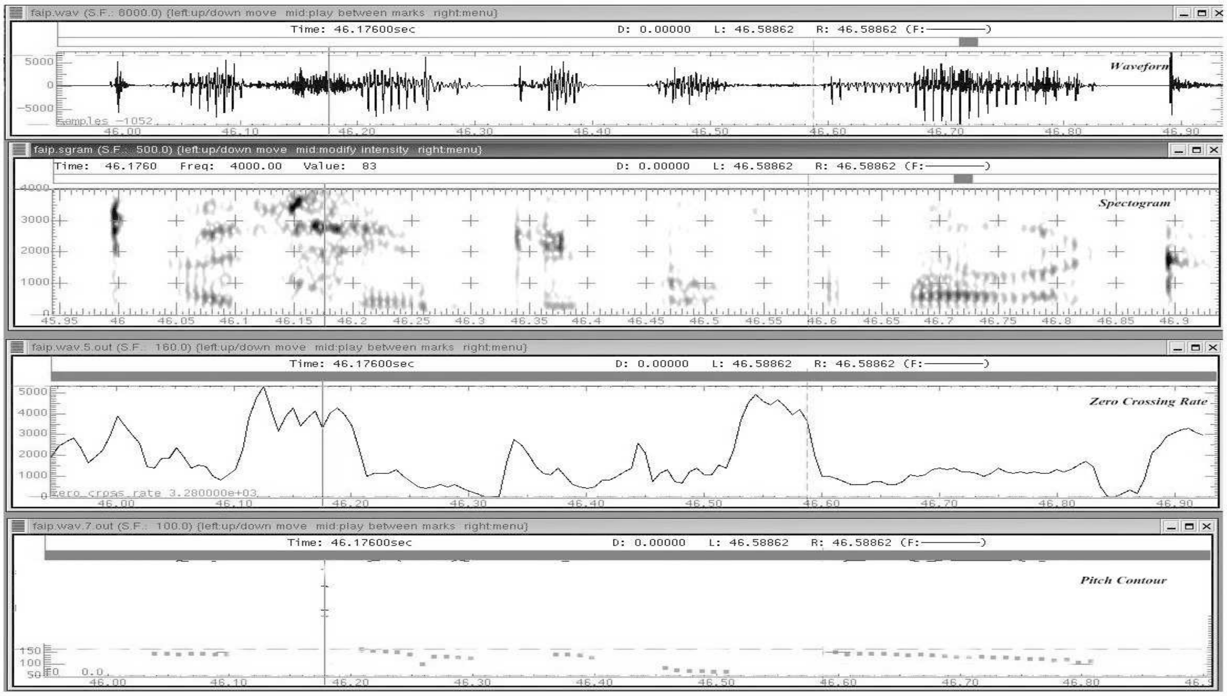

Fig. 2. Waveform, spectrogram,ZCR and pitch contour

features used are explained in brief. The volume of a signal is approximated by its root mean square (rms) value and is given by

$$volume = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} s_n^2(i)} \qquad (1)$$

where,

N is the total number of samples in the frame

s_n is the magnitude of the i^{th} sample of the frame

The dynamic range of the volume (VDR) is defined by

$$VDR = \frac{\nu_{max} - \nu_{min}}{\nu_{max}} \qquad (2)$$

where,

ν_{max} is the maximum value of the volume over the entire clip

ν_{min} is the minimum value of the volume over the entire clip

The volume undulation (the accumulation of the differences between volume of successive frames) and the standard deviation of the volume are computed. The zero crossing rate (ZCR) is the rate at which the sign of the signal changes. Typically ZCR is high for unvoiced speech regions. Using the ZCR contour, standard deviation of the ZCR is also computed.

The volume information gives an indication of the silence and unvoiced regions, since unvoiced regions are typically low energy regions. Combining the volume and ZCR evidences, we can distinguish between silence and non-silence regions in a reliable manner.

Figure 3 shows the 4 Hz modulation energy, calculated as the Fourier transform of the volume contour. It is based on the premise that normal speech consists of around four syllables per second. This can be used to distinguish between speech segments and other high energy segments like music and noise. We notice in Figure 3 that for the initial noisy segment of the cricket clip and the later half of the noisy football clip, the 4 Hz modulation energy is low.

The similar pitch ratio (SPR) plots for audio clips belonging to three different categories viz., cricket, football and news are shown in Figure 4. The value of SPR depends on the density of the voiced regions. As the density of the voiced regions is nearly uniform in the news clip, the SPR is also more or less uniform.

The fourteen features mentioned above were computed for every one second duration clip and combined to form a feature vector. Feature vectors were derived for a 10 second duration clip. Hidden Markov Mod-

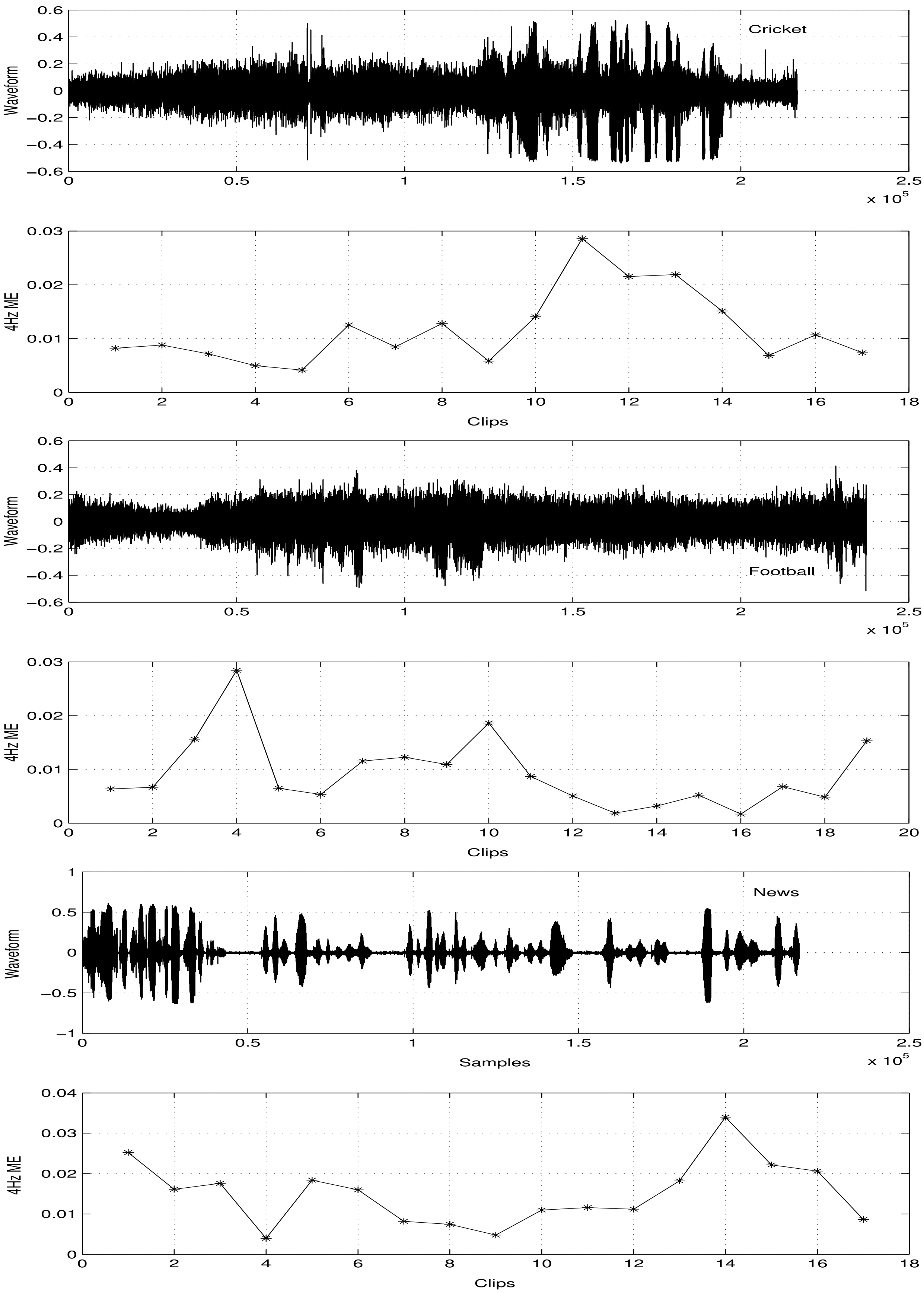

Fig. 3. The 4 Hz modulation energy for cricket, football and news clips

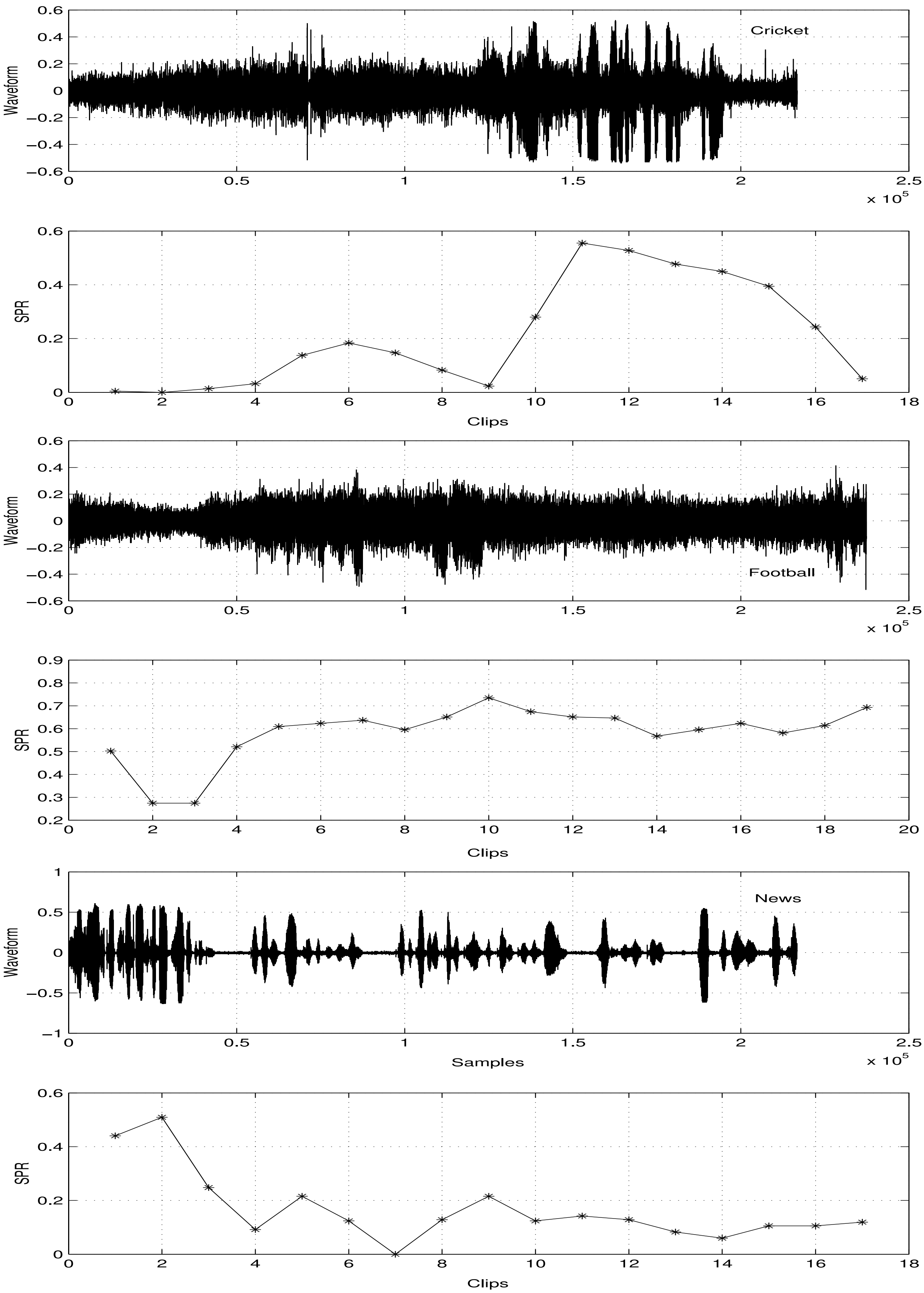

Fig. 4. The similar pitch ratio for cricket, football and news clips

els were trained using these feature vectors to represent the classes, football, cartoon, advertisement, news and cricket. The models were five state ergodic models with first and last states being entry and exit states, respectively. The recognition accuracy of the models trained using features derived from the residual was around 95%.

The features mentioned above, are mostly based on spectrum, amplitude, zero crossing and periodicity. These features describe the system (medium) characteristics of the signal, whereas it is known that the message part of the signal is contained mostly in the excitation source. The characteristics of the excitation source can be observed better by removing the spectral information from the signal. Doing so, we get the residual. Extracting features from this residual is a challenging task.

III. Significance of Residual/Excitation in Audio Indexing tasks

The residual is computed based on the linear prediction model [15], where the signal is predicted based on a linear weighted sum of the past sample values.

Given a speech sample at time n, the signal s_n can be modeled as a linear combination of the past p samples and an excitation term u_n

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{3}$$

where G is a gain factor. Given a particular signal s_n, the problem is to determine the set of predictor coefficients a_k and the gain G in some manner. These can be derived using the method of least squares. We assume that the input u_n is totally unknown, which is the case usually. Therefore, the signal s_n can be predicted only approximately from a linearly weighted summation of past p samples. The approximation $\hat{s}_n$ of s_n is given by

$$\hat{s}_n = -\sum_{k=1}^{p} a_k s_{n-k} \tag{4}$$

The error (e_n) between the actual value s_n and the predicted value $\hat{s}_n$ is given by

$$e_n = s_n - \hat{s}_n = s_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{5}$$

e_n is the desired residual. In the method of least squares, the parameters a_k are obtained as a result of the minimization of the mean or total squared error with respect to each of the parameters.

The residual contains the non-predictable part of the original waveform. For vowels, the spectrum of the residual is approximately flat, since the effects of the formants have been eliminated. This suggests that this error signal (residual) can be utilized to extract information such as pitch (the error signal exhibits sharp pulses corresponding to the pitch periods of the vowels). These pulses (excitation peaks) are the event markers, sequence of these events contain important perceptual information about the source of excitation. By listening to the residuals of different types of audio clips, we can distinguish different speakers, music instruments etc.

A noisy speech waveform and the corresponding residual for a vowel segment /a/ are shown in Figure 5. We observe that the formant regions get deemphasized in the residual. This results in better manifestation of excitation peaks in the residual. The modified residual, shown, is obtained by computing the product $|e_n|.s_n$. This is based on the premise that the residual and speech waveforms are correlated around the regions of excitation peaks. The events are emphasized even for noisy regions of speech as shown in Figure 5.

IV. Features from Residual

In this section we discuss the features inherent in the LP residual and their significance to audio indexing for the task of audio clip classification.

The waveform of the signal shows certain significant excitation but they are hidden by other signal characteristics. The excitation part is emphasized in the decorrelated residual signal. The excitation part can be further highlighted using the Hilbert envelope, bringing out the salient feature of the residual. The Hilbert envelope is defined as

$$h_e(i) = \sqrt{s(i)^2 + h(i)^2} \tag{6}$$

where,
h_e is the Hilbert envelope
s is the given input signal (here LP residual)
h is the Hilbert transform of the given signal

The Hilbert transform of a signal is the 90^0 phase shifted version of the original signal. Therefore, Hilbert envelope represents the magnitude of the analytic signal

$$x_n = e_n + i h_n \tag{7}$$

where,
x_n is the analytic signal

The Hilbert envelope computation removes the phase information present in the input residual. This leads to better identification of the excitation peaks.

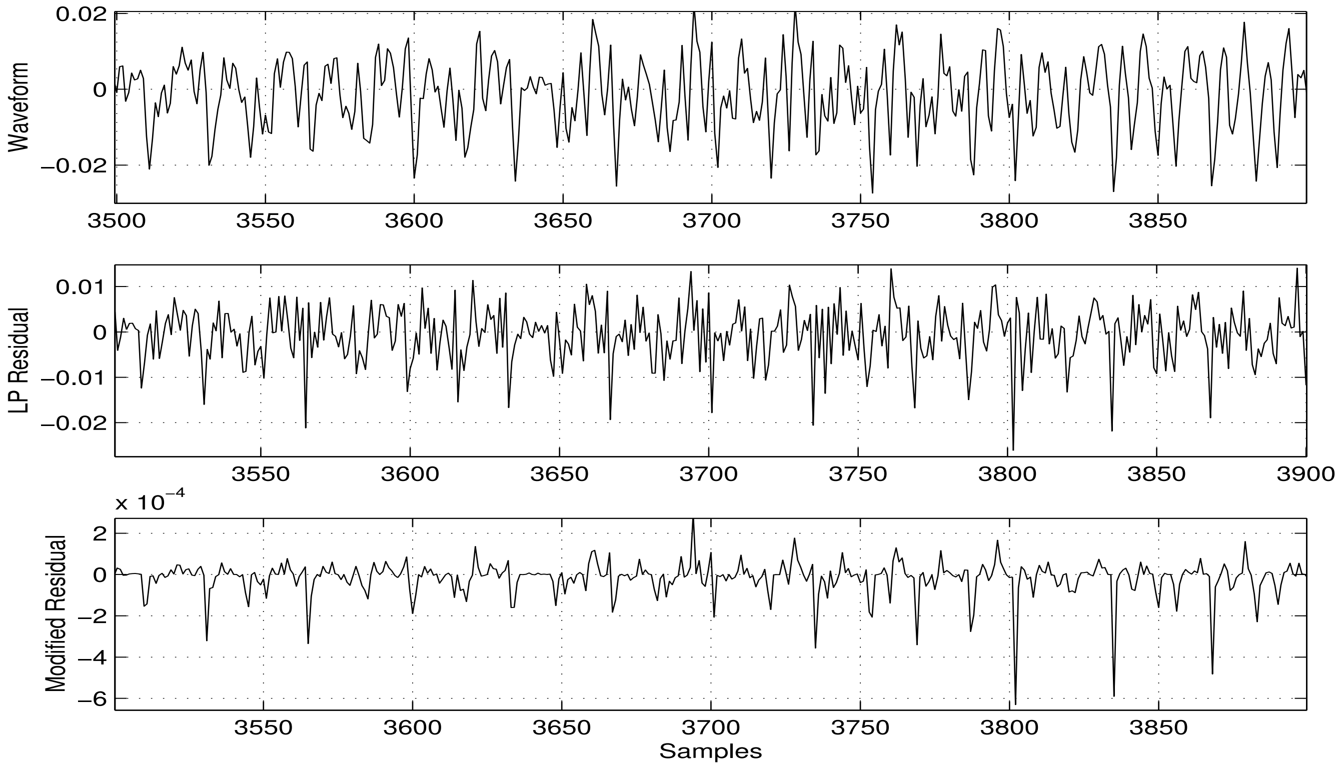

Fig. 5. The waveform, residual and modified residual

The Hilbert envelope is further modified by using the neighborhood information of each sample, and is computed as

$$h_{em}(i) = \frac{h_e^2(i)}{\sum_{k=i-l}^{k=i+l} \frac{h_e(k)}{(2l+1)}} \tag{8}$$

where,

h_e is the Hilbert envelope

l is the number of samples considered in the neighborhood of current sample i

Figure 6 shows the residual, Hilbert envelope of a noisy speech segment and the corresponding modified Hilbert envelope. We see that the excitation peaks are clearly visible in the modified Hilbert envelope. One can use this modified Hilbert envelope to estimate the pitch.

Similar pitch ratio derived from the Hilbert envelope is the ratio of number of neighboring frames having similar pitch to the total number of frames. This is a measure of the number of voiced speech frames in the given clip. It is obtained from the Hilbert envelope. In Figure 4 we observe that the football clip has higher SPR, distinguishing it from the rest of the clips.

Deriving other statistical parameters from the residual signal or its Hilbert envelope may result in a set of discriminating features suitable for audio indexing.

Perceptually, we notice that for different audio indexing tasks, there are some discriminating features present in the residual which can help in the classification of music, audio clips, identification of speakers, language etc. It is obvious that the same set of features from the residual may not help in all audio indexing tasks. The challenge lies in developing algorithms to capture these perceptually significant features from the residual.

V. Conclusion

With ever increasing volumes of audio data being collected, stored and being made online, it is imperative to have an efficient means of indexing and retrieving relevant audio data. To efficiently represent the data in compact form, significant events and characteristics of the signal needs to be captured and represented as a set of features. The set of features needed are different for different audio indexing tasks. Normally the significant part of audio may be contained in the excitation component. This manifests in the residual as a sequence of events. Capturing the sequence of events and deriving the features from these events for different tasks in audio indexing is a challenging task. We have shown how features drawn from the residual can be used in classification of audio clips. Since hu-

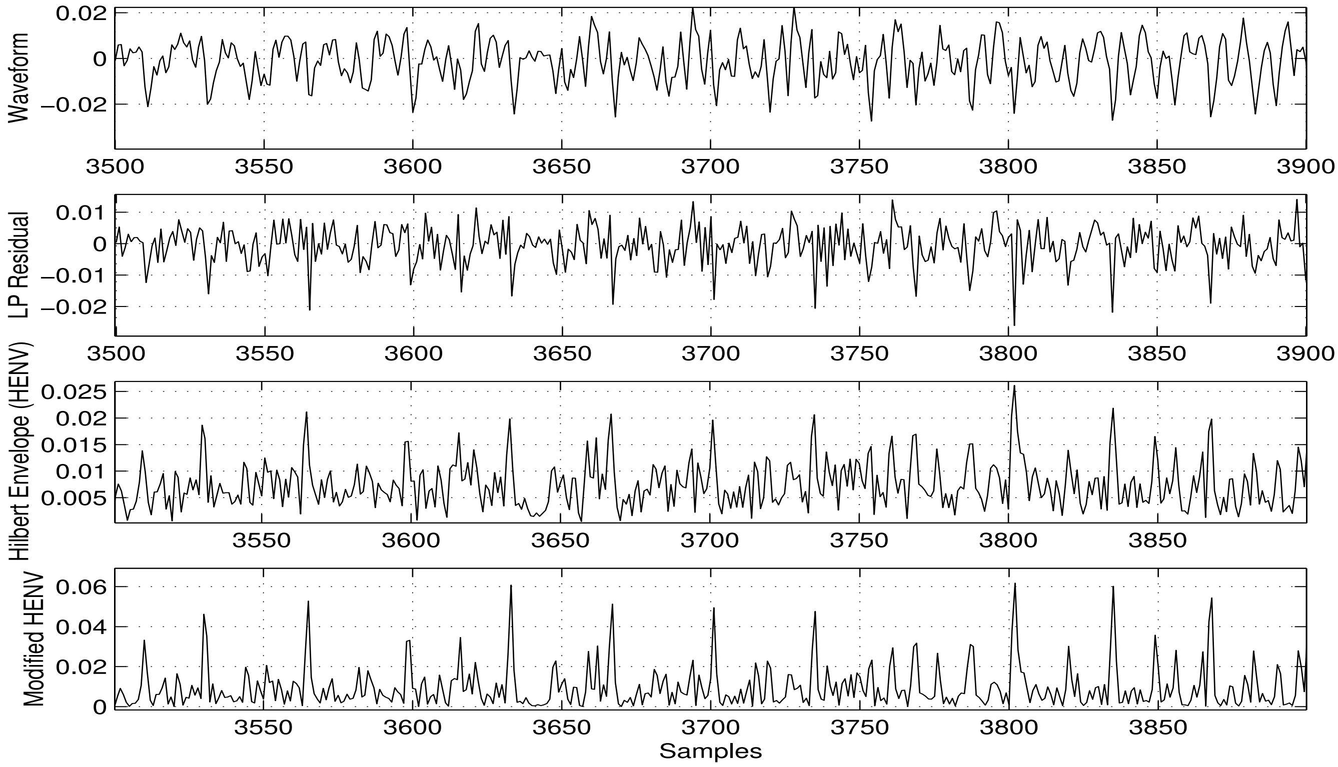

Fig. 6. The waveform, residual, Hilbert envelope and modified Hilbert envelope for a noisy audio segment

man perception is based on the sequence of events, it is necessary to capture this information for realization of the audio indexing tasks on a machine.

References

[1] B. Feiten and T. Ungvary, "Organizing sounds with neural nets," *in Proc. Int. Computer Music Conf.*, 1991.

[2] 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.

[3] E. Wold, T. Blum, D. Keslar and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, pp. 27–36, Fall 1996.

[4] J. T. Foote, "Content-based retrieval of music and audio. In C.-C. J. Kuo et al., editor," *Multimedia, Storage and Archiving Systems II, Proc. of SPIE*, vol. 3229, pp. 138–147, 1997.

[5] A. Ghias, J. Logan and D. Chamberlin, "Query by humming-musical information retrieval in an audio database," *Proc. ACM Multimedia Conf.*, pp. 231–235, 1995.

[6] Jared J. Wolf, "Efficient acoustic parameters for speaker recognition," *The Journal of the Acoustic Society of America*, vol. 51, no. 6(part 2), pp. 2044–2055, March 1972.

[7] George R. Doddington, "Speaker recognition - identifying people by their voices," *Proceedings of IEEE*, vol. 73, no. 11, pp. 1651–1664, November 1985.

[8] Lummis, R., "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Audio and Electroac, AU-21*, pp. 80–89, 1973.

[9] B. S. Atal, "Automatic speaker recognition based on pitch contours," *JASA*, vol. 52, no. 6, pp. 1687–1697, 1972.

[10] Michael J. Carey, Eluned S. Parris, Harvey Lloyd-Thomas and Stephen Bennett, "Robust prosodic features for speaker identification," *Intl. Conf. on Spoken Language Processing*, vol. 3, pp. 1800–1803, 1996.

[11] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 1331–1334.

[12] L. Wyse and S. Smoliar, "Toward content-based audio indexing and retrieval and new speaker discrimination technique," *Inst. Syst. Sci.*, Dec. 1995.

[13] Ann E. Thyme-Gobbel and Sandra E. Hutchins, "On using prosodic cues in automatic language identification," *Intl. Conf. on Spoken Language Processing*, vol. 3, pp. 1768–1771, 1996.

[14] Yao Wang, Zhu Liu and Jin-Cheng Huang, "Multimedia Content Analysis," in *IEEE Signal Processing Magazine*, November 2000, pp. 12–36.

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