Chapter 5

Indexing and Retrieval of Environmental Sounds

In the previous chapter, various models are analyzed for classifying the sound source separated environmental sound data into their corresponding categories. In this chapter, a method is proposed for indexing and retrieval of the classified environmental sound using Mel Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction Coefficients (PLP) features. Environmental sound clip extraction, feature extraction, creation of index and retrieval of the query clip are the major issues in automatic environmental sound indexing and retrieval. Section 5.5 explains the method of extracting the clip for indexing the classified environmental sound. Acoustic feature extraction for indexing and retrieval is described in Section 5.3. Section 5.4 describes the creation of index and retrieval of the corresponding query clip for environmental sound using GMM.

5.1 Introduction

Multimedia information retrieval is a growing research field that gained importance in recent years, due to the increasing number of available digital media. Traditionally, research is focused on visual information retrieval. The rise of audio information retrieval (AIR) was motivated by the development of efficient audio compression tech-
niques that support the distribution of digital audio [133]. This work concerns with the retrieval of natural sounds. Natural sounds are a subset of environmental sounds.

Environmental sounds are those sounds which fill our everyday acoustic environment. The panoply of environmental sounds is vast; it includes the sounds generated in, e.g., domestic, business and outdoor environments [134]. The investigation surveys a broad set of audio features and several similarity measures. Additionally, a new set of features are introduced in this work and their quality is evaluated by a selected set of classes of Natural sounds [135]. Due to the complexity and nearly infinite domain of non-speech sounds, the quality of retrieval is typically lower than the quality of speech retrieval, which is already well understood. Retrieval results presented in this work for the domain of Natural sounds are comparable to that of state-of-the-art research in the area of environmental sound retrieval [136], [137]. Fig. 5.1 shows the proposed method for environmental sound indexing and retrieval system.

For the environmental indexing and retrieval problems, the objective is that users should be able to retrieve all sound events of interest in an intuitive and efficient manner without too many false positives, i.e. the user can quickly form the query and can quickly execute the retrieval. A convenient strategy is Query-By-Example (QBE), where users input recordings they consider similar to the desired retrieval sounds. Users can either upload the query from a file or present it orally. Oral query is quite prevalent in melody retrieval in the form of Query-By-Humming (QBH), where the sound objects to be retrieved concern melodies. During retrieval, the database considered most similar to the query are obtained. As an explicit model of query behavior, it adopts a likelihood-based QBE strategy that computes the likelihood over all possible queries that arise as a result of each sound in the database.

In this chapter, a new approach to index and retrieve environmental sound clips
is presented using static MFCC and PLP coefficients, which prominently discriminate one sound group from another. Kullback-Leibler Divergence is used as a measure for retrieving the environmental sound clips.

5.2 Environmental Sound Clip Extraction for Retrieval

The primary issue in environmental audio indexing is the storage of audio data in a form suitable for retrieval. Environmental audio query can be classified into two different approaches: a-whole-object search and in-object search [138], [139]. Each approach generates a different type of query result [140]. A-whole-object search approach searches for data that are globally similar to the query input; on the other
hand, an in-object search approach searches for a large piece of data that contains a
fragment that is similar to the query [141], [142].

5.3 Features for Environmental Sound Indexing

There are many features that can be used to characterize audio signals. Usually
audio features are extracted in two levels: frame level and clip level. For a feature to
reveal the semantic meaning of an audio signal, analysis over a much longer period is
necessary, usually from one second to several tens seconds. Such an interval is known
as an audio clip. A clip consists of a sequence of frames, and clip-level features usually
characterize how frame-level features change over a clip. The clip boundaries may be
the result of audio segmentation and classification such that the frame features within
each clip are similar. Alternatively, fixed length clips, usually 1 to 2 seconds, may be
used. In this work, fixed duration clips have been considered for environmental sound
clip retrieval.

5.3.1 Mel Frequency Cepstral Coefficients

If the audio data can be efficiently characterized by a set of parameters which capture
essentially all the key attributes of the original data, this can be used as a feature set.
These features must occur frequently, should be easily measurable and should not vary
with time. MFCCs are obtained through a frame-based analysis of a signal where the
waveform is divided into a sequence of frames, the purpose is to smooth the frequency
spectra and reduce the effects of acoustic variation [119], [143], [144]. A sinusoidal
transform (DFT) is performed using a hamming window overlapping each frame to
obtain an amplitude spectrum, which is then converted to a Mel-scale spectrum using
triangular filters, emphasizing frequencies according to their perceptual importance
on this scale [145], [146]. In this work, MFCCs features are extracted for each clip as
described in Section 3.2.2 to index the audio clips.

5.3.2 Perceptual Linear Prediction Coefficients

Hermansky developed a model known as PLP. It is based on the concept of psychophysics theory and discards unwanted information from the human pitch [147]. It resembles the procedure to extract LPC parameters except that the spectral characteristics of the speech signal are transformed to match the human auditory system. The block diagram for PLP computation is shown in Fig. 5.2.

PLP is the approximation of three aspects related to perceptron namely resolution curves of the critical band, curve for equal loudness and the power law relation of intensity loudness. The audio signal is hamming windowed to reduce discontinuities. The Fast Fourier Transform (FFT) transforms the windowed speech segment into the frequency domain. The power spectrum for the Fourier transform coefficients is calculated using Equation 5.1.

\[ P(\omega) = \text{Re}(S(\omega))^2 + \text{Im}(S(\omega))^2 \]  
(5.1)

where \( P(\omega) \) is the Power spectrum, \( \text{Re}(S(\omega)) \) represents the real part and \( \text{Im}(S(\omega)) \)
represents the imaginary part of Fourier transform

The bark transformation is used to warp the spectrum $P(\omega)$ along its frequency axis $\omega$ into the bark frequency $\Omega$ in Equation 5.2:

$$\Omega(w) = 6 \ln \left( \frac{w}{1200\pi} + \sqrt{(w/1200\pi)^2 + 1} \right) \tag{5.2}$$

The auditory warped spectrum is convolved with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. Critical band is the frequency bandwidth created by the cochlea, which acts as an auditory filter. The cochlea is the hearing sense organ in the inner ear. Bark scale corresponds to 1 to 24 critical bands. The power spectrum of the critical band masking curve and auditory warped spectrum are convoluted to simulate the human hearing resolution. The equal loudness pre-emphasis needs to compensate the unequal perception of loudness at varying frequencies.

A weight function is added to the sampled values using an equal loudness curve to simulate the human hearing sensitivity at varying frequencies. The intensity loudness power law is an approximation of the power law of hearing, which relates sound intensity and perceived loudness of the sound. Each intensity is raised to the power of 0.33 as stated by the power law and thus the equalized values are transformed. An all pole model normally applied in Linear Prediction (LP) analysis is used to approximate the spectral samples. Either the coefficients can be used as such for representing the signal or they can further be transformed to Cepstral coefficients.

In this work, a $9^{th}$ order LP analysis is used to approximate the spectral samples and hence obtained a 9-dimensional feature vector for a environmental sound signal of frame size of 20 milliseconds is obtained.
5.4 Techniques for Environmental Sound Indexing and Retrieval

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. Many approaches to audio information retrieval consider similarity in the audio domain by comparing features extracted from the audio signals.

5.4.1 Gaussian Mixture Models (GMM)

In this work, each piece of environmental audio is represented by a Gaussian inference about the content of a particular audio file. Let $D$ be the size of the feature vectors ($MFCCs$ and $PLPs$). Thus when plotting the feature vectors in the space of $R^D$, there are one or several areas where most of the feature vectors are placed. These areas are called clusters, and if there are more than one of these areas it may be assumed that the feature vectors have been generated by a mixture distribution.

The probability distribution of feature vectors is modeled by parametric or non-parametric methods. Models which assume the shape of probability density function are termed parametric. In non-parametric modeling, minimal or no assumptions are made regarding the probability distribution of feature vectors. The potential of Gaussian mixture models to represent an underlying set of acoustic classes by individual Gaussian components, in which the spectral shape of the acoustic class is parameterized by the mean vector and the covariance matrix, is significant.

Also, these models have the ability to form a smooth approximation to the arbitrarily-shaped observation densities in the absence of other information [148]. With Gaussian mixture models, each sound is modeled as a mixture of several Gaussian clusters in the feature space. The basis for using GMM is that the distribution of feature vectors
extracted from a class can be modeled by a mixture of Gaussian densities as shown in 
Fig. 5.3.

![Fig. 5.3: Gaussian mixture models.](image)

For a $D$ dimensional feature vector $\mathbf{x}$, the mixture density function for category $s$ is defined as

$$p(\mathbf{x}/\lambda^s) = \sum_{i=1}^{M} \alpha_i^s f_i^s(\mathbf{x})$$

The mixture density function is a weighted linear combination of $m$ component unimodal Gaussian densities $f_i^s(\cdot)$.

Each Gaussian density function $f_i^s(\cdot)$ is parameterized by the mean vector $\mu_i^s$ and the covariance matrix $\Sigma_i^s$ using

$$f_i^s(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_i^s|}} \exp(-\frac{1}{2}(\mathbf{x} - \mu_i^s)^T(\Sigma_i^s)^{-1}(\mathbf{x} - \mu_i^s)),$$

where $(\Sigma_i^s)^{-1}$ and $|\Sigma_i^s|$ denote the inverse and determinant of the covariance matrix $\Sigma_i^s$, respectively. The mixture weights $(\alpha_1^s, \alpha_2^s, \ldots, \alpha_M^s)$ satisfy the constraint $\sum_{i=1}^{M} \alpha_i^s = 1$. Collectively, the parameters of the model $\lambda^s$ are denoted as $\lambda^s = \{\alpha_i^s, \mu_i^s, \Sigma_i^s\}$. 

---

96
The number of mixture components is chosen empirically for a given data set. The parameters of GMM are estimated using the iterative expectation-maximization algorithm [149].

The motivation for using Gaussian densities as the representation of audio features is the potential of GMMs to represent an underlying set of acoustic classes by individual Gaussian components in which the spectral shape of the acoustic class is parameterized by the mean vector and the covariance matrix. Also, GMMs have the ability to form a smooth approximation to the arbitrarily-shaped observation densities in the absence of other information [148]. With GMMs, each sound is modeled as a mixture of several Gaussian clusters in the feature space.

5.4.2 Kullback-Leibler Divergence (KL)

Kullback-Leibler Divergence produces a measure of “distance” or “divergence” between two statistically distributed populations in terms of their similarity information. The symmetric Kullback-Leibler divergence between two Gaussian models can be computed as:

\[
d = (\text{cov}_q \ast \text{icov}_d) + (\text{cov}_d \ast \text{icov}_q) + \left( (\text{icov}_q + \text{icov}_d) \ast (\text{mean}_q - \text{mean}_d) \ast (\text{mean}_q - \text{mean}_d)' \right)
\]

(5.3)

The KLD is the only measure of difference between probability distributions that satisfies some desiderata, which are the canonical extension to those appearing in a commonly used characterization of entropy. Consequently, mutual information is the only measure of mutual dependence that obeys certain related conditions, since it can be defined in terms of KLD. In probability theory and information theory, the Kullback-Leibler divergence, also called discrimination information, information divergence, information gain, relative entropy, KLIC, KL divergence, is a measure of the
difference between two probability distributions $P$ and $Q$. It is not symmetric in $P$ and $Q$. In applications, $P$ typically represents the “true” distribution of data, observations, or a precisely calculated theoretical distribution, while $Q$ typically represents a theory, model, description, or approximation of $P$.

Specifically, the KLD from $Q$ to $P$, denoted $D_{KL}(P\|Q)$, is a measure of the information gained when one revises one’s beliefs from the prior probability distribution $Q$ to the posterior probability distribution $P$. In other words, it is the amount of information lost when $Q$ is used to approximate $P$. The KLD also measures the expected number of extra bits required to code samples from $P$ using a code optimized for $Q$ rather than the code optimized for $P$.

Although it is often intuited as a way of measuring the distance between probability distributions, the KLD is not a true metric. It does not obey the triangle inequality, and in general $D_{KL}(P\|Q)$ does not equal $D_{KL}(Q\|P)$. However, its infinitesimal form, specifically its Hessian, gives a metric tensor known as the Fisher information metric. The KLD is a special case of a broader class of divergences called $f$-divergences as well as the class of Bregman divergences. It is the only such divergence over probabilities that is a member of both classes. The distance is rescaled to obtain improved results the distance between the feature vectors with other information.

$$d = \exp(1/\text{fact} \ast d) \quad (5.4)$$

Where $\text{fact} = 450$ is the rescaling factor which gave better results in the combinations.

5.5 Proposed Method for Indexing and Retrieval of Environmental Sound Clip

The algorithm for indexing and retrieval of Environmental sound clips is described below:
5.5.1 Algorithm for Creation of Index using GMM

1. Collect 200 environmental sound clips $e_1, e_2, e_3, \ldots, e_n$ each of 3 seconds duration from different environmental sounds.

2. 13-dimensional MFCC features are extracted from all n environmental sound clips to form the environment sound clips index.

3. A GMM is fit for all 200 environment sound clips using the MFCC features.

4. Repeat step 2 and step 3 for PLP and MFCC-PLP combined features.

5.5.2 Algorithm for Retrieval of Environmental Audio for a Given Query

1. An environmental query sound clip of 3 seconds duration is extracted.

2. MFCC features are extracted from the environmental sound clip query.

3. A GMM is fit for environment sound query clip using the MFCC features.

4. Finally, Kullback-Leibler (KL) distance between given query Gaussian and every indexed Gaussians are found and minimum distanced results are retrieved.

5. Repeat steps 2 to 4 for PLP and MFCC-PLP combined features.

5.6 Performance Measures

The performance of environmental sound indexing and retrieval system is measured in terms of either one of the following measures:

- **Precision**: It can be seen as a measure of exactness or quality. High precision means that an algorithm returned substantially more relevant results than irrelevant.

  \[
  Precision = \frac{|\{\text{relevant audios}\} \cap \{\text{retrieved audios}\}|}{|\{\text{retrieved audios}\}|} \quad (5.5)
  \]
• **Recall**: It is a measure of completeness or quantity. High recall means that an algorithm returned most of the relevant results.

\[
Recall = \frac{|\{\text{relevant audios}\} \cap \{\text{retrieved audios}\}|}{|\{\text{relevant audios}\}|}
\]  
(5.6)

• **F-measure**: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score.

\[
F\text{-}measure = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]  
(5.7)

F-measure is a measure that combines precision and recall. The closer the value of F-measure to 1, the better the accuracy of retrieval.

• **Accuracy of retrieval**: Accuracy of retrieval is a performance measure for environmental sound indexing system. It is measured using Equation 5.8.

\[
R_e = \frac{K_m}{Q_m} \times 100
\]  
(5.8)

where \(R_m\) is accuracy of retrieval, \(K_m\) is the number of environmental sound clips retrieved in the top 'n' ranked list and \(Q_m\) is the total number queries.

• **Average number of clips retrieved for each query based on a threshold**:

The environmental retrieval performance of the system is measured with \(T\) as the number of clips retrieved on an average for each query based on a predefined threshold as shown in the Equation 5.9.

\[
T = \frac{\sum^N_{i=1} \text{Retrieved}_i}{N}
\]  
(5.9)

where \(T\) is the average number of clips retrieved based on threshold and \(\text{Retrieved}_i\) is the number of clips retrieved for a given query environmental audio clip.
5.7 Experimental Results

- **Database for environmental sound:** The experiments are conducted for indexing environmental sound using the natural sound data collected from BBC database. This database consists of samples from 10 different natural scenes, each having 4 source recordings of 5 minutes each. The recordings are categorized into general classes according to common characteristics of the scenes (kitchen noises, living room noises, laundry sounds, meeting sounds, office sounds) and events (pan boiling, steel plate, music player, paper scrap, washing machine, flush, overlapped speech, footsteps, typewriter, dust bin, etc.,). The recordings are manually cropped and are separated into 2, 5 and 10 seconds query fragments, which is sampled at 16kHz and encoded by 16-bit.

- **Acoustic feature extraction:** Each environmental sound clip with 3 seconds of duration for MFCC, PLP and MFCC-PLP combined are extracted as described in Section 5.3. A frame size of 20 ms and a frame shift of 10 ms are used. Thereby 13 MFCC features are extracted for each environmental sound clip of 3 seconds. Hence $300 \times 13$ feature vectors are arrived at for each of the 3 second clip and this procedure is repeated for all 200 clips. Similarly, experiments are conducted to extract PLP features of $300 \times 9$ and MFCC-PLP combined features of $300 \times 22$ respectively. Same procedure is repeated for all the 200 clips.

- **Forming the training set:** For the 200 environmental sound clips MFCC and PLP features are extracted and this results in a training set of $200 \times 28$ dimensional vector. Since there are 100 such clips, extracting one random feature vector from a clip results in a training set of $100 \times 28$. 
• Creation of Index: GMMs are constructed for 200 environmental sound clips using MFCC features which forms the index. Experiments are also conducted with GMMs using for PLP and MFCC-PLP combined features to create index.

• Retrieval of a Clip using Index: For retrieval, the 3 seconds environmental sound clip is used as query. MFCC and PLP features are extracted. GMMs are constructed for environmental sound query clip using MFCC features. Experiments are also conducted with GMMs using PLP and MFCC-PLP combined features to create index. Kullback-Leibler (KL) distance between given query Gaussian and every indexed Gaussians is found and results are obtained. Retrieval is based on minimum distance.

Figs. 5.4 and 5.5 show the snapshot of Environmental Sound indexing and retrieval system respectively.

![Image of the Environment Sound indexing system](image_url)

**Fig. 5.4:** Snapshot of the Environmental Sound indexing system.
The performance of Environmental sound indexing system, both on level of representation of audio in indexing and on level of retrieval of relevant audio retrieval from database. The retrieval results using the proposed feature set is shown in Fig. 5.6 for various query durations. From the results, it is clear that the overall retrieval accuracy is high for 10 seconds sample when compared to other duration. Experimental are conducted for 13 dimensional MFCC and 9 dimensional PLP. Similarly the performance is studied for combined feature set. The results show that combined feature set shows an improved in accuracy of retrieval when compared to MFCC and PLP features. The overall retrieval rates are shown in Fig. 5.6. Retrieval performance peaks around combined features and the performance slowly degrades as the feature varies. The optimum retrieval rate of 86.0% is achieved with retrieval time of 9 minutes using 1 minutes index and combined feature set.
Table 5.1: Performance Matrix for 5-Seconds Query using Combined Features Set (Out of First 15 \( R_q \) Ranked.

<table>
<thead>
<tr>
<th>Scene</th>
<th>( R_d )</th>
<th>( R_d \cap R_q )</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitchen</td>
<td>20</td>
<td>14</td>
<td>0.93</td>
<td>0.70</td>
<td>0.79</td>
</tr>
<tr>
<td>Living room</td>
<td>20</td>
<td>13</td>
<td>0.87</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td>Laundry</td>
<td>20</td>
<td>12</td>
<td>0.80</td>
<td>0.60</td>
<td>0.80</td>
</tr>
<tr>
<td>Meeting</td>
<td>20</td>
<td>13</td>
<td>0.87</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td>Office</td>
<td>20</td>
<td>13</td>
<td>0.87</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td>Holi</td>
<td>20</td>
<td>12</td>
<td>0.80</td>
<td>0.60</td>
<td>0.80</td>
</tr>
<tr>
<td>Ocean</td>
<td>20</td>
<td>13</td>
<td>0.87</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td>Thunder storm</td>
<td>20</td>
<td>14</td>
<td>0.93</td>
<td>0.70</td>
<td>0.79</td>
</tr>
<tr>
<td>Waterfall</td>
<td>20</td>
<td>14</td>
<td>0.93</td>
<td>0.70</td>
<td>0.79</td>
</tr>
<tr>
<td>Prayer</td>
<td>20</td>
<td>15</td>
<td>1.00</td>
<td>0.75</td>
<td>0.86</td>
</tr>
</tbody>
</table>

\( R_d \)-Relevant data in database

\( R_d \cap R_q \)- Correctly Retrieved data from database

5.8 Summary

In the thesis, new methods are proposed for environmental sound indexing and retrieval. MFCC and PLP features are extracted. For the environmental sound clips, GMMs are used to create an index based on the features extracted. For retrieval, a GMM is fit for environment sound query clip using the MFCC and PLP features. Finally, Kullback-Leibler (KL) distance between given query Gaussian and every indexed Gaussians are found and minimum distanced results are retrieved. The retrieval performance is studied for different features. Efficiency of environmental sound retrieval
Table 5.2: Accuracy of Retrieval of Environmental Sound clips in $n$ Ranked List.

<table>
<thead>
<tr>
<th></th>
<th>Features</th>
<th>Out of 5 ranked</th>
<th>Out of 10 ranked</th>
<th>Out of 15 ranked</th>
<th>Out of 20 ranked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition Accuracy</td>
<td>MFCCs</td>
<td>32%</td>
<td>62%</td>
<td>70%</td>
<td>69%</td>
</tr>
<tr>
<td></td>
<td>PLPs</td>
<td>28%</td>
<td>60%</td>
<td>78%</td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td>Combined (MFCCs &amp; PLPs)</td>
<td>47%</td>
<td>79%</td>
<td>86%</td>
<td>78%</td>
</tr>
</tbody>
</table>

Fig. 5.6: Performance of indexing and retrieval for different durations of Environmental sound clips.

System is evaluated for 100 clips and the method achieves about 86.0% accuracy using MFCC and PLP features. This method is sensitive to the duration of the query clip.