1 RESEARCH METHODOLOGY

The proposed algorithm for audio segmentation segment the audio into different parameters as described before also feature extraction algorithm separates out the different audio features such as MFCC, LPC, SF, SNR, and HZCRR.

Feature extraction

Considering the lower frequency spectrum is too sensitive to even a bit of changes of the scenes and speakers, it could cause segmented clips too small. It will have effects on succeeding audio classification. We, thus, use Multiple sub-Bands spectrum Centroid relative Ratio (MBCR) [5] over 800Hz as basic feature. This feature may depict centroid movement trend in a time frequency-intensity space. Its mathematical description can be described as follows.

\[
\text{SCR}(i,j) = \frac{\text{SC}(i,j)}{\text{Max} \left( \frac{\text{SC}(i,j)}{J} \right)} \quad (1)
\]

\[
\text{SC}(i,j) = \frac{\text{f}(j) \times \text{FrmEn}(i,j)}{N} \sum_{k=1}^{N} \text{FrmEn}(i,k) \quad (2)
\]

Where \( \text{SCR}(i,j) \) is MBCR of the \( i \)thframe and the \( j \)thsub-band, \( \text{SC}(i,j) \) is the frequency centroid of the \( i \)thframe and the \( j \)thsub-band, and \( N \) denotes the number of frequency sub-bands. The element of \( f(j) \) is the normalized central frequency.

\[
\text{FrmEn}(i,j) = \log \left( \int_{\omega_L(j)}^{\omega_H(j)} |F(i,\omega)|d\omega \right) \quad (3)
\]
where $\omega_L(j)$ and $\omega_H(j)$ are lower and upper bound of sub-band $j$ respectively, $F(i,\omega)$ represent denotes the Fast Fourier Transform (FFT) at the frequency $\omega$ and frame $i$, and $|F(i,\omega)|$ is square root of the power at the frequency $\omega$ and frame $i$.

**Results:**

We conducted a series of experiments based on proposed audio segmentation and classification approach. The experimental platform we used is a workstation Pentium4, 2.4G CPU, 1GB RAM memory. The performance was evaluated on the recordings of real TV program. The segmentation and classification results were evaluated by the recall rate $\delta$, accuracy rate $\xi$, and average precision $\eta$. These are defined as

\[
\delta = \frac{\text{the number of correctly objects}}{\text{the number of objects that should be correct}}
\]

\[
\xi = \frac{\text{the number of correctly objects}}{\text{the number of all get objects}}
\]

\[
\eta = \frac{\delta \times \xi}{0.5 \times (\delta + \xi)}
\]

We pre-defined six categories as audio classes, which is pure speech (PS), pure music (PM), song (S), speech with music (SWM), speech with noise (SWN) and silence (SIL).

Table 1: The results of first level classification

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Audio type</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Equal time</strong></td>
<td>Pure Speech (PS)</td>
<td>85.15%</td>
<td>85.63%</td>
<td>87.62%</td>
</tr>
<tr>
<td></td>
<td>Silence (SIL)</td>
<td>97.10%</td>
<td>86.14%</td>
<td>91.29%</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>77.95%</td>
<td>95.08%</td>
<td>85.67%</td>
</tr>
<tr>
<td><strong>MBCR</strong></td>
<td>Pure Speech (PS)</td>
<td>91.33%</td>
<td>93.65%</td>
<td>92.47%</td>
</tr>
<tr>
<td></td>
<td>Silence (SIL)</td>
<td>98.22%</td>
<td>92.97%</td>
<td>95.52%</td>
</tr>
</tbody>
</table>
Table1 gives the result of first level classifying. It firstly puts the audio clips into three classes, pure speech, silence and others. The others are further classified into four classes: speech with music, song, pure music, and speech with noise, in the second level classifying.