An efficient similarity estimation for audio retrieval

ATULYA VELIVELLI

1 Introduction

There has been a large increase in the amount of digital Multimedia data over the past one decade, specifically in the case of digital audio data. Hence, there is a need for being able to efficiently index the audio data, and for the user to be able to retrieve similar audio segments by specifying the category. In [1] radio broadcast is classified into 2 categories, speech and music based on statistical properties like zero crossing rate. [2] performs an unsupervised search by presenting an audio query example, uses a Gaussian mixture model(GMM) pdf based metric as a similarity measure.

The goal of this project is to detect the best N , number of music segments in an audio file of any length in a likelihood based ranked order. The Hidden Markov Model captures the temporal dimension in addition to capturing the audio pattern of a GMM. We use the continuous observation density Hidden Markov Model(HMM), to model the acoustic properties of the music segments we wish to detect. For this we first collect the training data. The video training data that we use is the NIST Special Database 26. Then we train a Music HMM and a General audio HMM over the annotated data. Using these 2 HMM's we synthesize a new General+Music+General HMM , and using this HMM we detect music segments as the states corresponding to music in the viterbi decoding of the audio data. This procedure is performed again on the neighboring segments until we obtain N music detections.

2 General Audio HMM

We use the mediaconvert audio decoder software on SGI workstation to extract the sound track of all the training files to different audio files, and using the HTK toolkit we represent these audio files using Mel-frequency cepstrum coefficients+energy coefficient(Mfcc+E) vectors as features. Using the HTK toolkit we model this Observation sequence of (Mfcc+E) vectors as a 5 state continuous observation density Hidden Markov model. This HMM is Left to right as shown in Figure 1, though the figure does not show all the transition arcs in the left to right directions, all the transitions in the left to right direction are possible. Each state observation density $b_j(\mathbf{O})$ is modeled as a Gaussian pdf.

$b_j(\mathbf{O}) = \mathcal{N}(\mathbf{O}, \mu_\mathbf{j}, \mathbf{U_j}), 1 \leq j \leq N.$

$\mu_\mathbf{j}$, $\mathbf{U_j}$ are the mean vector and the Covariance matrix for state j. The first state and the last state are non-emitting.

3 Music HMM

From the audio files, we manually extract only those segments that correspond to music and represent using (Mfcc+E) vectors. Using the HTK toolkit we model this observation sequence of (Mfcc+E) vectors as a 4 state continuous observation density Hidden Markov model. Similar to the General audio HMM , each state observation density is modeled as a Gaussian pdf. This HMM is Left to right as shown in Figure 1, though the figure does not show all the transition arcs in the left to right directions, all the transitions in the left to right direction are possible.

4 Synthesized HMM

Detecting music in most cases is essentially detecting the General audio +music +General audio pattern in the audio data. For this we first synthesize a General+Music+General HMM as shown in Figure 1. We first combine the General audio HMM, and music HMM as shown in Figure 2, and then combine the General+Music model with the General model. P is the number of states of the General HMM, while Q is the number of states of the music HMM. In our case P=5 and Q=4.

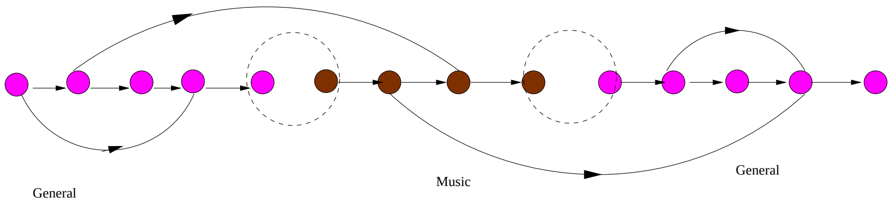

Figure 1: The General+Music+General HMM is synthesized by combining by combining the non-emitting states of the General and Music HMM's as shown by the circles

We first combine the General and Music HMM's transition probabilities as shown below

1. For $1 \leq i \leq P-1, \quad 1 \leq j \leq P-1$

   $gm_{i,j} = g_{i,j}$

   For $1 \leq i \leq P-1, \quad P \leq j \leq P+Q-2$

   $gm_{i,j} = g_{i,P} \;.\; m_{1,j-P+2}$

   For $P \leq i \leq P+Q-2, \quad P \leq j \leq P+Q-2$

   $gm_{i,j} = m_{i-P+2,j-P+2}$

We next combine the General+Music and General HMM's transition probabilities

2. For $1 \leq i \leq P+Q-3, \quad 1 \leq j \leq P+Q-3$

   $gmg_{i,j} = gm_{i,j}$

   For $1 \leq i \leq P+Q-3, \quad P+Q-2 \leq j \leq 2P+Q-4$

   $gmg_{i,j} = gm_{i,P+Q-2} \;.\; m_{1\;j-P-Q+4}$

   For $P+Q-2 \leq i \leq 2P+Q-4, \quad P+Q-2 \leq j \leq 2P+Q-4$

   $gmg_{i,j} = m_{i-P-Q+4,j-P-Q+4}$

Figure 2: Procedure for synthesizing the elements of the transition matrix $gmg_{i,j}$ of General+Music+General HMM.

1. For $1 \leq j \leq P-1$
   $b_j(O) = bg_j(O)$

   For $P \leq j \leq P+Q-3$
   $b_j(O) = bm_{j-P+2}(O)$

   For $P+Q-2 \leq j \leq 2P+Q-4$
   $b_j(O) = bg_{j-P-Q-4}(O)$

Figure 3: Procedure for calculating the observation densities $b_j(O)$ of General+Music+General HMM.

The elements of the transition matrix of the synthesized General+Music +General HMM are calculated as shown in Figure 2, $g_{i,j}$ represent the transition probabilities of the General Audio HMM, $m_{i,j}$ represent the transition probabilities of the Music HMM, $gm_{i,j}$ are the transition probabilities of the intermediate General+Music HMM, and $gmg_{i,j}$ are the transition probabilities of General+Music+General HMM. The observation densities $b_j(O)$ of the states of the General+Music+General model are estimated as shown in Figure 3. $bg_j(O)$ and $bm_j(O)$ represent the observation densities of the General audio and Music HMM's respectively.

5 Detection Scheme

To detect a music segment in the audio file, we do the viterbi decoding of the audio observation sequence **O**, given the General+Music+General HMM parameters. The time where states P to P+Q-3 start and end indicates the music segment. As shown in the Figure 4, the yellow bar indicates the first music segment detected from the output of the viterbi decoding. We again apply viterbi decoding to the entire audio segments adjoining the start and end of the first music segment. This returns two candidate music detections as indicated by the red bars. The next step is to decide which of these red bars should be given a preference as a music detection. For this we calculate

$$ratio = \frac{p(\mathbf{O}/General + Music + General)}{p(\mathbf{O}/General)} \tag{1}$$

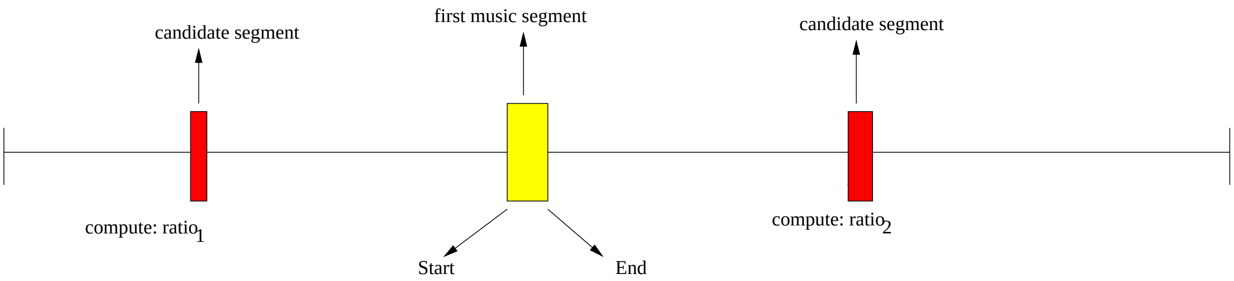

Figure 4: $ratio_1$ and $ratio_2$ are evaluated over the duration of the candidate music detections before and after the first music detection.

for both the candidate segments and select the one with the highest ratio. In general the selection of N best segments is done as shown in Figure 5

```
•$j \leftarrow 1$, $i \leftarrow 1$
• While $j \leq N$
    Compute the likelihood ratio for the candidate segment on the left of $music_{j-1}$: $ratio_i$
    $seg_i \leftarrow [start_i \; end_i]$
    $i \leftarrow i + 1$
    Compute the likelihood ratio for the candidate segment on the right of $music_{j-1}$: $ratio_i$
    $seg_i \leftarrow [start_i \; end_i]$
    $i \leftarrow i + 1$
    $music_j \leftarrow seg_{argmax\{ratio_1, ratio_2, \ldots, ratio_i\}}$
    $ratio_{argmax\{ratio_1, ratio_2, \ldots, ratio_i\}} \leftarrow 0$
    $j \leftarrow j + 1$
end
```

Figure 5: Music Detection Scheme

6 Results

For testing purpose, we use 4 audio files in the database which were not used for training the music HMM. Considering the duration of the test documents

Table 1: Detection results for the top 4 music segments

Audio	Music1	Music2	Music3	Music4	Ground Truth
1	hit	miss	hit	hit	4 significant music segments in this doc
2	hit	hit	hit	miss	4 significant music segments in this doc
3	hit	hit	miss	miss	3 significant music segments in this doc
4	hit	miss	miss	miss	1 significant music segment in this doc

we decide that N=4 is an appropriate value. Hence, we test the effectiveness of our scheme for the top 4 music detections. A Hit in Table 1 means that the detected segment was music and a miss means that the detected segment was not music. Table 1 indicates that in the case of the first audio document the second music segment that was detected was not music, though as indicated by the ground truth there were actually 4 significant music segments present. In the case of the fourth audio document, the first music segment is hit while the other music segments are missed. But, the Ground Truth indicates that there is only one significant music segment present . Hence it is inferred that the only music segment that was present was detected.

Figure 6 shows a plot of the precision for different number of music segments detected. We define:

$$precision = \frac{Number\ of\ correctly\ detected\ Musicsegments}{Total\ number\ of\ detected\ music\ segments} \tag{2}$$

It can be noticed from Figure 6 that initially for one detected music segment the precision is 1, and towards the end for 4 detected music segments it is around 0.56.

7 Conclusion

The first music segment detected corresponds to music in all the cases. The subsequent detections do not have the same quality as the first detection in some cases. In the future we could develop a detection scheme which takes a decision based on the length of the music segment and the likelihood ratio.

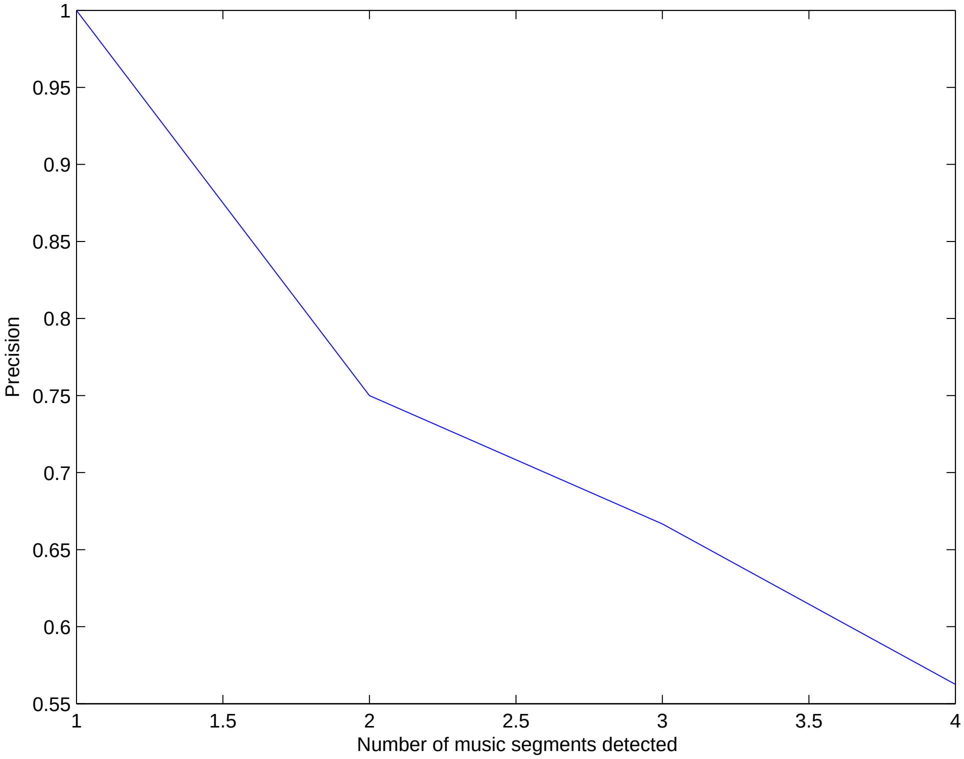

Figure 6: A plot of precision for different number of music segments detected.

References

[1] J. Saunders, “Real-Time dsicrmination of Broadcast Speech/Music,” in *ICASSP 1996,* Atlanta, May 1996, Vol 2, pp. 993-996.

[2] Z. Liu, Q. Huang, “Content-based Indexing and Retrieval-by-Example in Audio,” *ICME 2000,* New York, NY, July 30 - Aug. 2, 2000.