EASIER Sampling for Audio Event Identification

Surong Wang   Min Xu   Liang-Tien Chia   Manonranjan Dash

Center for Multimedia and Network Technology
School of Computer Engineering
Nanyang Technological University, Singapore 639798
{pg02759741, mxu, asltchia, asmdash}@ntu.edu.sg

Abstract

An audio event refers to some specific audio sound which plays important role for video content analysis. In our previous work [3], we have established audio event identification as an audio classification task. Due to the large size of audio database, representative samples are necessary for training the classifier. However, the commonly used random selection of training samples is often not adequate in selecting representative samples. In this paper we present EASIER *sampling algorithm to select those data which more efficiently represent audio data characters for audio event identifier training.* EASIER *deterministically produces a subsample whose "distance" from the complete database is minimal. Experiments in the context of audio event identification show that* EASIER *outperforms simple random sampling significantly.*

1. Introduction

Audio, that includes voice, music, and various kinds of environmental sounds, is an important type of media, and also a significant part of video. Recently people have begun to realize the importance of effective audio content analysis which provides important cues for semantics. Effective audio analysis techniques can provide convincing results. In consideration of computational efficiency, some research efforts have been done for audio content analysis [1][2].

In [3][4], we present our research on audio event identification by using Hidden Markov Models (HMM). We randomly selected samples to train audio event identifier without considering sample efficiency and computation time for training. When the audio database is large, we have to consider the following two issues: 1) With audio data coming from various audio sequences, simple random sampling (SRS) may not generate good representative of all sequences. 2) There is a trade-off between the size of training data set and accuracy.

As the representativeness of a training sample influences the classification performance significantly and SRS is often found to be inadequate in this aspect, in this paper we present a new improved training set selection algorithm, called EASIER, for producing such a small yet representative sample. EASIER deterministically produces a subsample whose "distance" - appropriately defined - from the complete database is minimal. EASIER can produce samples of any given ratio with almost fixed amount of time. It requires only one scan of the data, hence suitable for online applications such as streaming data processing. Experiments in the context of audio event identification using HMM show that EASIER outperforms SRS significantly.

In Section 2, we review the procedure of audio events identification based on HMM. In Section 3, we describe the EASIER algorithm and how to apply EASIER for audio data. In Section 4 experimental results are presented. Conclusion and future work are given in Section 5.

2. Audio Event Identification

Audio event is defined as some specific audio sound having strong hints to interesting video events or video highlights. Especially in sports video, some game-specific audio sounds (e.g., excited audience sounds, excited commentator speech, etc.) have strong relationships to the actions of players, referees, commentators and audience. In this paper, we use basketball audio to demonstrate how efficiently the proposed method works.

2.1. Basketball Audio Event Identification

Basketball games have compact structure. Generally, the offence and the defense, that are the highlights of basketball game, take place alternatively. These highlights, which attract most audience's interests, are significant and should be detected for future basketball video editing. Fortunately, excited commentator speech and excited audience sounds play important roles in highlight detection of sports video. Therefore, the basketball audio event identification focus on identifying excited commentator speech (EC) and excited audience sounds (EA). Besides EC and EA, there are two other basketball audio events, plain commentator speech (PC) and plain audience sounds (PA). These four kinds of

0-7803-9332-5/05/$20.00 ©2005 IEEE

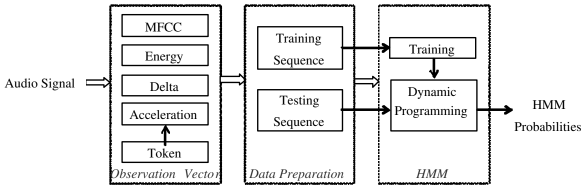


Figure 1: Proposed audio events generation system

events almost cover the full basketball game. A classification task is to classify audio samples into these four predefined classes (EA, PA, EC and PC). There are some other audio events in basketball game, such as whistling, etc., that have small number of samples and are easy to identify. In order to test the efficiency of proposed sampling algorithm, we use only EA, PA, EC and PC.

2.2. HMM-based Audio Event Identification

Audio signal exhibits consecutive changes in values over a period of time, where variables may be predicted from earlier values. That means, strong context exists in audio data. In consideration of the success of HMM in speech recognition, we propose our HMM-based audio event generation system. The proposed system includes three stages, which are feature extraction, data preparation and HMM learning, as illustrated in Figure1. Selected low-level features are extracted from audio streams and tokens are added to create observation vectors. These data are then separated into two sets for training and testing. After that, HMM is trained and then re-estimated by using dynamic programming. Finally, according to the maximum posterior probability, the audio keyword with the largest probability is selected to label the corresponding testing data. Details can be found in [3][6].

3. EASIER for Audio Application

SRS samples may not represent the characteristics of all audio sequence particularly for small sample ratios. As shown in Figure 4, when only 10% random samples are used for training, the classification performance is very low for SRS, especially for small classes, "Excited Audience" and "Excited Commentator". Therefore, to obtain stable results and achieve better performance using small size of training data, we propose EASIER sampling algorithm to select representative samples. It is applied for efficient audio event identification.

3.1 EASIER Sampling

EASIER is based on its predecessor EASE (Epsilon Approximation Sampling Enabled) algorithm which is proposed by one of our co-authors in [5]. Given an $\varepsilon > 0$, EASE determines a sample set S_0 which is an ε-approximation of the original dataset S, i.e., its discrepancy satisfies $Dist(S_0, S) \leq \varepsilon$ [5]. Starting with S, EASE performs repeated halving to obtain the final sample S_0 according to a penalty function. One way of computing the discrepancy is to calculate the distance of 1-itemset frequencies between subset S_0 and superset S:

$$Dist_\infty(S_0, S) = \max_{A \in I_1(S)} |f(A; S_0) - f(A; S)| \quad (1)$$

where $f(A; S_0) = n(A; S_0)/|S_0|$, $f(A; S) = n(A; S)/|S|$. $n(A; S_0)(n(A; S))$ is the number of transactions in $S_0(S)$ that contain item A in 1-itemset of S, i.e., $I_1(S)$.

However, EASE has some critical limitations. In order to obtain a small sample ratio, the halving procedure is repeated several times. This costs extra time and memory. Besides, due to its halving nature EASE has certain granularity and cannot achieve all sample ratios for a given S. In order to overcome these limitations, we proposed EASIER. Its key innovation is that the halving loop is eliminated. In EASIER, the penalty function is modified to accommodate any sampling ratio, not just 0.5 (halving). Although the new innovation does not guarantee the upper bound distance, experiments in various domains show that its performance is better or very close to that of EASE. Briefly, EASIER works as follows:

1. At the beginning, all transactions uncolored.
2. Each transaction in S is colored as red or blue. Red (blue) means the transaction is selected (rejected). r_i (b_i) is the number of red (blue) transactions in S^i where S^i is the set of all transactions in S that contain item A_i and $r_i + b_i = |S^i|$.
3. The coloring decision is based on a penalty function Q_i for item A_i. Let f_r be the ratio of red transactions, i.e., the sample ratio, so the ratio of blue transactions is $f_b = 1 - f_r$. In EASIER, Q_i is low when $r_i/(2f_r) = b_i/(2f_b)$ approximately, otherwise Q_i increases exponentially in $|r_i/(2f_r) - b_i/(2f_b)|$. Figure 2 shows the shape of the penalty function. The objective of EASIER is to minimize $|r_i/(2f_r) - b_i/(2f_b)|$. After coloring the first j transactions, the value of Q_j for item A_i is:

$$\begin{aligned} Q_i &= Q_i^{(j)} = Q_{i,1}^{(j)} + Q_{i,2}^{(j)} \\ Q_{i,1}^{(j)} &= (1+\delta_i)^{\frac{r_i}{2f_r}} (1-\delta_i)^{\frac{b_i}{2f_b}} \\ Q_{i,2}^{(j)} &= (1-\delta_i)^{\frac{r_i}{2f_r}} (1+\delta_i)^{\frac{b_i}{2f_b}} \end{aligned} \quad (2)$$

where $Q_i^{(j)}$ means the penalty of ith item in jth transaction and δ_i controls how steeply the penalty increases as shown in Figure 2. The initial values of $Q_{i,1}$ and $Q_{i,2}$ are both 1. The initial value of δ_i is $\sqrt{1 - \exp(-\ln(2m)/n)}$ where m is the number of items in the original dataset [5].

If $(j+1)$-th transaction is colored red (or blue), the corresponding penalty function $Q_i^{(j||r)}$ (or $Q_i^{(j||b)}$) will be:

$$\begin{aligned} Q_{i,1}^{(j||r)} &= (1+\delta_i)^{\frac{1}{2f_r}} Q_{i,1}^{(j)} \quad Q_{i,2}^{(j||r)} = (1-\delta_i)^{\frac{1}{2f_r}} Q_{i,2}^{(j)} \\ Q_{i,1}^{(j||b)} &= (1-\delta_i)^{\frac{1}{2f_b}} Q_{i,1}^{(j)} \quad Q_{i,2}^{(j||b)} = (1+\delta_i)^{\frac{1}{2f_b}} Q_{i,2}^{(j)} \end{aligned} \quad (3)$$

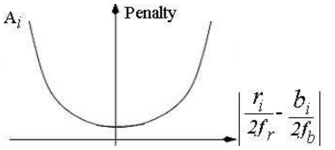


Figure 2: The penalty function for EASIER.

Algorithm 1 EASIER Sampling

```
Input: $S, m, f_r$
Output: $S_0$, the transactions in red color
for each item $i$ in $S$ do
    $\delta_i = \sqrt{1 - \exp(-\frac{\ln(2m)}{n})}$
    $Q_{i,1} = 1 \qquad Q_{i,2} = 1$
end for
for each transaction $j$ in $S$ do
    color transaction $j$ red;
    $Q^{(r)} = 0 \qquad Q^{(b)} = 0$
    for each item $i$ contained in $j$ do
        $Q_{i,1}^{(r)} = (1+\delta_i)^{\frac{1}{2f_r}} Q_{i,1} \qquad Q_{i,2}^{(r)} = (1-\delta_i)^{\frac{1}{2f_r}} Q_{i,2}$
        $Q_{i,1}^{(b)} = (1-\delta_i)^{\frac{1}{2f_b}} Q_{i,1} \qquad Q_{i,2}^{(b)} = (1+\delta_i)^{\frac{1}{2f_b}} Q_{i,2}$
        $Q^{(r)} += Q_{i,1}^{(r)} + Q_{i,2}^{(r)} \qquad Q^{(b)} += Q_{i,1}^{(b)} + Q_{i,2}^{(b)}$
    end for
    if $Q^{(r)} < Q^{(b)}$ then
        $Q_{i,1} = Q_{i,1}^{(r)} \qquad Q_{i,2} = Q_{i,2}^{(r)}$
    else
        color transaction $j$ blue;
        $Q_{i,1} = Q_{i,1}^{(b)} \qquad Q_{i,2} = Q_{i,2}^{(b)}$
    end if
    if transaction $j$ is red then
        set $S_0 = S_0 + \{j\}$
    end if
end for
```

If the overall penalty for current transaction $Q^{(j||r)} = \sum_i Q_i^{(j||r)}$ is more than $Q^{(j||b)} = \sum_i Q_i^{(j||b)}$, the $(j+1)$-th transaction is colored blue and rejected. Otherwise, it is colored red and added to the final sample set. In Algorithm 1 the completed EASIER algorithm is given. The memory complexity of EASIER is only $O(m)$. As far as the computation time is concerned, the time for processing one transaction is bounded by $O(T_{max})$ where T_{max} is the maximum transaction length. Thus, EASIER takes almost fixed amount of time even when sample ratio varies.

3.2 EASIER for Audio Event Identification

We apply EASIER to find the representative training samples from an audio database. In our experiments, 39-dimension (39D) feature vectors are used. Because the amplitude of an audio signal is continuous and EASIER is based on the calculation of the frequency of each item, the format of the features is changed as Figure 3. Firstly the continuous values are non-uniformly quantized to a range [0, 17]. Then, this discretized data is binarized. The binary vector has a length of 18 and it contains all 0's except for an 1 in the position corresponding to the discretized value. After that, each non-zero value in this binary vector is converted back into a new discrete value considering its position. This new vector represents the data and is used for sampling with EASIER.

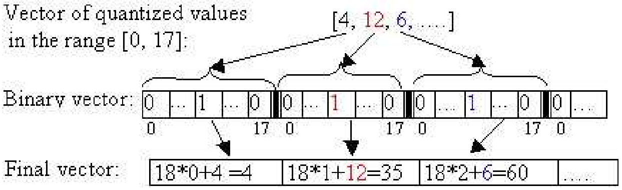


Figure 3: An example of the format modification of features.

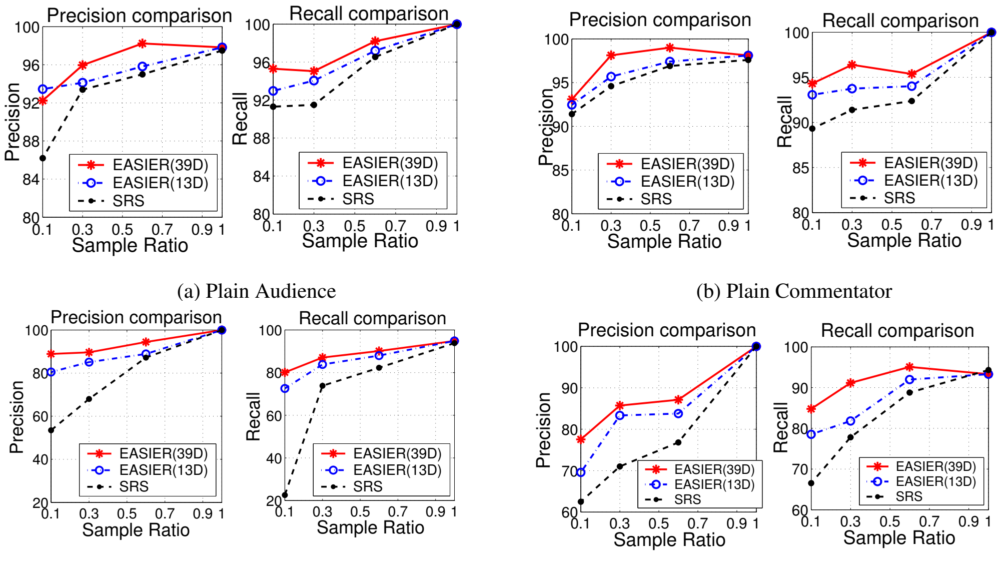
Although experimental results show that using all features give high accuracy, we experimented with reduced dimensionality that requires less space. In order to reduce the number of items, The most dominant 13 dimensions (13D) of feature vectors are used to reduce dimensionality.

These 13 dimensions represent Mel-scale Frequency Cepstral Coefficients (MFCC) and energy of the audio signal. Their efficacy in audio analysis is already shown in [4]. Our experimental results have shown that the classification performance of the 13D data is not significantly far from the original data of 39D.

4. Experimental Results

EASIER and SRS are compared using a database that contains one hour basketball audio in 3600 samples (one sample for each second). Both algorithms run five times and the results are computed as an average over five samples. As earlier discussed, we use HMM as the audio event identifier. Accuracy of HMM is measured using precision and recall. Precision is the ratio of the number of correct results to the total number of results. Recall is the ratio of the number of correct results to the total number of correct data in the database. The results of four audio classes (EA, EC, PA and PC) based on EASIER and SRS are shown in Figure 4 separately. The sampling ratios include 0.6, 0.3 and 0.1.

As demonstrated in Figure 4, sampling with EASIER can achieve high performance with less training samples, especially for the smaller classes EA and EC. By using EASIER, identification improves in two aspects: 1) To achieve similar recall and precision, EASIER sampling needs relatively less training data than SRS. For example, for a precision of 85% for EA, SRS needs 60% training data whereas EASIER needs only 30%. 2) For the same training set, EASIER gives higher performance. For sample ratio 0.1 and "Excited Audience" class, the precision of EASIER is 90% which is significantly higher than that of SRS (54%). Although expected precision and recall values for 13D data are a little smaller than those



(a) Plain Audience

(b) Plain Commentator

(c) Excited Audience

(d) Excited Commentator

Figure 4: The performance of audio event identification by SRS and EASIER.

for 39D, they are significantly larger than those for SRS.

Note that for identification of an event in basketball, the smaller classes of EA and EC are more important than PA and PC. EASIER gives higher accuracy for these two important classes signifying its choice over SRS. In addition, EASIER's performance *vis-a-vis* SRS improves further as we reduce the sample ratio. This has been shown in other experiments not reported here. Due to small size of EA and EC (approx. 10% each) classes, we could not show results of EASIER *vis-a-vis* SRS for sample ratios less than 10%.

Figure 5 shows the computation time taken by EASIER and SRS to produce the samples. For 39D data the sampling time is about 1.8s. It is acceptable considering the long training and classification time. When 13D data is used, the sampling time is reduced greatly to only 0.4s. Note that for different sample ratios, EASIER requires almost fixed amount of time whereas SRS requires various time.

5. Conclusion and Future Work

The proposed EASIER sampling algorithm works efficiently for audio event identification. EASIER is an online algorithm where the incoming transactions are processed once and a decision is taken regarding its participation in the final sample. It can select representative samples of any ratio. Comparison with SRS shows that, EASIER algorithm improves audio event identification as follows: 1) It effectively finds the relative representative data for training and consequently improves the recall and precision significantly. 2) It provides a feasible way to train identifiers for large audio database by using only a small training dataset. Especially for small sample ratios, EASIER achieves better result than SRS. Although in this paper we showed the efficacy of EASIER over only audio data, it is also shown to work efficiently for image data and transactional data. Currently we are extendind it further to other audio and video data.

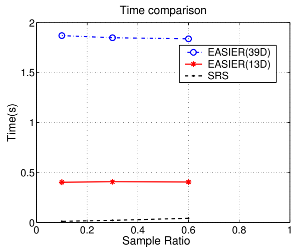


Figure 5: The sampling time for different sample ratios.

References

[1] Y. Rui, A. Gupta and A. Acero, ”Automatically Extracting Highlights for TV Baseball Programs”, In *Proc. of ACM Multimedia*, Los Angeles, CA, pp. 105-115, 2000.

[2] S. Nepal, U. Srinivasan, and G. Reynolds, ”Automatic Detection of Goal Segments in Basketball Videos”, In *Proc. of ACM Multimedia*, 2001.

[3] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C.-S. Xu and Q. Tian: HMM-Based Audio Keyword Generation, In *Proc. of PCM (3)*, 2004: 566-574.

[4] M. Xu, L.-Y. Duan, L.-T. Chia and C.-S.Xu, ”Audio Keywords Generation for Sports Video Analysis”, In *Proc. of ACM Multimedia*, 2004.

[5] H. Brönnimann, B. Chen, M. Dash, P Haas and P Scheuermann, Efficient data reduction with EASE, In *Proc. 9th Int. Conf. on KDD*, 8(2003), pp. 59–68.

[6] S. Young, et al, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, 2002