SVM-BASED AUDIO CLASSIFICATION FOR INSTRUCTIONAL VIDEO ANALYSIS

Ying Li and Chitra Dorai

IBM T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598

E-mail:{yingli,dorai}@us.ibm.com

ABSTRACT

Automatic content analysis and annotation for efficient search and browsing of topics in instructional videos are current challenges in the management of e-learning content repositories. This paper presents our current work on classifying the soundtrack of instructional videos into seven distinct audio classes using the Support Vector Machine (SVM) technology. The classification results are then used to partition a video into homogeneous audio segments, which forms the fundamental basis for its higher-level content analysis and exploration. Initial experiments carried out on three education and four training videos totalling to 185 minutes have yielded an average 97.9% classification accuracy. The performance comparisons between the SVM-based, the decision tree (DT)-based and the threshold-based audio classification schemes further demonstrates the superiority of the proposed scheme.

1. INTRODUCTION

Web-based learning is rapidly emerging as an alternative to traditional classroom-based education. Many universities and industrial organizations have started offering remote education and training programs. As a result, the amount of instructional videos available on corporate intranets and the Internet is dramatically increasing. However, so far we still lack powerful tools to analyze and index their content so as to facilitate the searching and browsing of multimedia data. This paper describes our current efforts on structuralizing instructional media by classifying their audio tracks into distinct audio classes which are meaningful in this domain. The classification results are then employed to partition them into homogeneous audio segments, which form the fundamental basis for a higher-level content analysis and exploration.

Audio classification has been an active research area in recent years, and various audio features and classification schemes have been proposed. For instance, Kimber and Wilcox applied hidden Markov models (HMMs) to classify audio signals into music, silence or speech using cepstral features [1]. To find an optimal classification scheme, Scheirer and Slaney [2] examined four different classification frameworks including multidimensional Gaussian maximum *a posteriori* estimator, Gaussian mixture model, a spatial partitioning scheme based on k-d trees, and a nearest neighbor classifier for speech/music discrimination.

To take more sound types into consideration, Srinivasan *et al.* proposed to classify audio signals into speech, music, silence or unclassified sound type using fixed thresholds [3]. Zhang and Kuo also used a similar thresholding scheme to distinguish eight different audio types including silence, speech, music, song, environmental sound and their various combinations [4]. A sports audio classification scheme was proposed in [5] where MPEG-7 audio features and MFCC were evaluated using variants of HMM.

In this work, the Support Vector Machine (SVM) technique [6] is employed for the classification task due to its capability in handling complicated feature space and in distinguishing different classes with overlapped or interwoven areas. Lu *et al.* also employed SVMs in their work [7], which hierarchically classified audio signals into five classes. Particularly, it first distinguished silence from non-silence, then non-silence signals were classified into speech or non-speech. Next, non-speech segments were further classified into music and background sound, while speech segments were classified into pure speech and non-pure speech. Promising results have been reported, yet this hierarchical classification scheme has two drawbacks: (i) if the signal is misclassified in an earlier stage, it will never reach the correct type (leaf node); (ii) it does not distinguish speech with noise from speech with music, which are two important audio classes on their own right.

In contrast, seven audio classes are considered in this paper which are pertinent to instructional videos including corporate education videos and professionally produced training videos. They are *speech*, *silence*, *music*, *environmental sound*, *speech with music*, *speech with environmental sound*, and *environmental sound with music*. Four binary SVM classifiers are trained to recognize these audio types in parallel. Preliminary experiments carried out on three education and four training videos have yielded an average 97.9% classification accuracy. Finally, a sophisticated comparison was conducted between the SVM-based, the decision tree (DT)-based and the threshold-based classification schemes. The proposed SVM-based scheme has outperformed the other two.

2. SVM-BASED INSTRUCTIONAL AUDIO CLASSIFICATION

A Support Vector Machine (SVM) is a supervised binary classifier which constructs a linear decision boundary or a hyperplane to optimally separate two classes [6]. Since its inception, the SVM has gained wide attention due to its excellent performance on many real-world problems. It is also reported that SVMs can achieve a generalization performance that is greater than or equal to other classifiers, while requiring significantly less training data to achieve such an outcome [8]. So far, SVMs have been applied to various tasks such as image and video content analysis and annotation, handwritten digit recognition, text classification, speech recognition and speaker identification [8]. However, they have not been well explored in the domain of audio classification, and this is another reason that motivate our investigation of this technology.

2.1. Audio Feature Extraction

To classify the audio track of an instructional video, we first uniformly segment it into non-overlapping 1-second long clips, then various features are extracted from each clip to represent it. Currently, 26 audio features are considered in this work, which are chosen due to their effectiveness in capturing the temporal and spectral structures of different audio classes. A brief description of these features are given below. Readers should refer to [3, 4] for more detailed discussions.

1. Mean and variance of short-time energy (STE). The energy is computed for every 20-ms audio frame which advances for every 10-ms.
2. Low ST-energy ratio (LSTER). LSTER is defined as the ratio of the number of frames whose STE values are less than 0.5 times of the average STE to the total number of frames in a 1-second clip.
3. Mean, variance and range of short-time zero-crossing rate (ZCR). ZCR is also computed for every 20-ms frame which coarsely measures a signal's frequency content.
4. High ZCR ratio (HZCRR). HZCRR is defined as the ratio of the number of frames whose ZCR is above 1.5 fold average ZCR rate to the total number of frames in a 1-second clip.
5. Mean of the spectrum flux (SF). SF is defined as the average variation of the spectrum between adjacent two frames in a 1-second clip.
6. Mean and variance of energies in four frequency subbands. With 11KHZ sampling rate, we define the four frequency subbands to be [0, 700HZ], [700-1400HZ], [1400-2800HZ], and [2800-5512HZ].
7. Mean and variance of energy ratios of the above four frequency subbands. The energy ratio of subband i is the ratio of its energy to the sum of the four subband energies.
8. Harmonic degree (HD). HD is the ratio of the number of frames that have harmonic peaks to the total number of frames. Fundamental frequency is computed for measuring the signal's harmonic feature.
9. Music component ratio (MCR). MCR is determined from the signal's spectral peak tracks which remain at the same frequency level and last for a certain period of time for most of musical sounds [4].

2.2. Audio Classification Using Combinations of SVMs

At this step, every 1-second clip is classified into one of seven audio classes which include four pure audio classes: *speech*, *silence*, *music*, *environmental sound*, and three sound combinations: *speech with music*, *speech with environmental sound*, and *environmental sound with music*. Four binary SVM classifiers are trained for this purpose, which discriminate between speech and non-speech (*spSVM*), silence and non-silence (*silSVM*), music and non-music (*musSVM*), and, environmental sound and non-environmental sound (*envSVM*). A decision value DV is output from each classifier for every test clip, whose sign determines the predicted class. For instance, if the output of the *spSVM* for clip s_i is positive, then it contains speech; otherwise it is non-speech. However, since it is a multi-class classification task, additional processing steps, which are described below, are needed to achieve this goal.

Step 1: Given the four DV values obtained from the four SVM classifiers for clip s_i, we first count the number of positive values and denote it by C. If C equals 0, it normally implies that there is a signal transition (*e.g.* from silence to speech) within this 1 second period. In this case, we will disregard this clip. Otherwise, if C equals 1, we can then confidently assign s_i to the class with the positive DV. When C is greater than 2, which indicates that there possibly exist more than two dominant signals, we proceed as follows.

1. Find the first two largest DV values, and compute their mean m and variance v.
2. If both DVs are larger than threshold T_1, or if the variance v is smaller than threshold T_2 and the mean m is larger than threshold T_3, we say that there are two audio signals co-existing. For instance, if both *spSVM* and *musSVM*'s DVs are larger than T_1, then s_i contains both speech and music. Currently, we set T_1, T_2 and T_3 to be 1.0, 0.75 and 0.35, respectively, which are empirically determined from our experiments and are fixed for all test videos.
3. Otherwise, if the variance v is larger than threshold T_2, *i.e.* one signal dominates the other, we choose the audio class with the larger DV value.

Step 2: In this step, we smooth the above classification results by removing isolated audio types since a continuous audio stream does not have abruptly and frequently changed audio content. For instance, given the short duration of each clip, it is impossible that s_i contains *music*, while its preceding and succeeding clips do not.

Step 3: Finally, we group temporally adjoining 1-second clips together if they share the same sound type. As a result, the entire audio stream will be partitioned into homogeneous segments with each having a distinct audio class label.

3. EXPERIMENTAL RESULTS

To speed up the implementation of the classification process, we readily use the SVM^{light} software package [9] for SVM training and testing.

3.1. Experimental Setup

The training data test includes 40-min of speech, 15-min of environmental sound, 7-min of music and 7-min of silence, which are collected from various corporate education and training videos. All data are sampled at 11KHZ rate, with mono channel and 16 bits per sample.

Moreover, to find the optimal SVM parameters such as *kernels*, *variance*, *margin* and *cost factor*, we have also hand-labelled approximately 9-min of speech, 2-min of silence, 3-min of music and 1-min of environmental sound as validation data. Based on the validation results, we choose the radial basis function (RBF) as kernel and set parameters γ to 5 and C to 10. Approximately 99.1% and 98.6% accuracies have been achieved on the training and validation data, respectively. The entire training process takes approximately 2 minutes. Figure 1 shows the process of training the speech SVM classifier, where Pf_i, Nf_i, and Vf_i stand for the positive, negative training and validation clips respectively.

The test set is collected from three education and four training videos which amount to 185 minutes in total. Various types of sounds such as background noise, speech over background noise

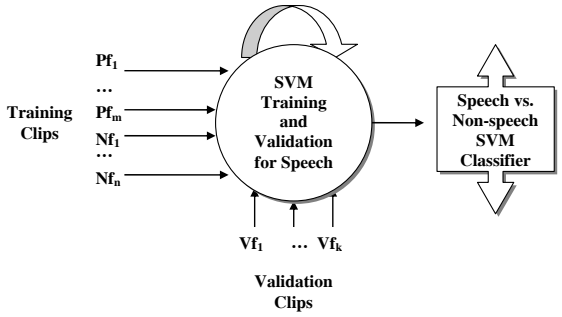

Fig. 1. Training process for the speech SVM classifier.

and speech over music, are contained in these videos. There is no overlap between the test and training sets.

3.2. Audio Classification Results Using SVMs

Table 1 shows the classification result of the seven sound types in the form of a confusion matrix, where names in the leftmost column represents the actual classes while the ones in the top line are the classes predicted by the classification scheme. The classification accuracy is measured for each sound class which is defined as the ratio of correctly classified samples over all predicted samples of the class.

From this table, we can see that all seven audio classes have been well discriminated from each other. Particularly, classes such as speech, silence, speech with music, and speech with environmental sound have achieved classification accuracies as high as 95%. The relatively low classification accuracy of music is caused by the insufficiency of training data relative to feature dimensionality. Moreover, since there is an increased degree of subjectivity in transcribing environmental sounds than labelling other sounds, we have comparatively not as high an accuracy on this class.

The overall classification accuracies for the seven test videos are reported in Table 2. Good classification performances have been observed, especially for "Train2", "Train4" and educational videos, which have clean speech and less noisy audio background.

3.3. Comparisons Between Different Classification Schemes

For comparison purposes, we have also carried out the audio classification task using other popular approaches, specifically, the decision tree (DT)-based and the threshold-based classification schemes. Moreover, as an effort to avoid any uncertainty caused by mixed sound types, we will only consider pure sound types in this comparison, which are speech, silence, music and environmental sound.

To train the decision tree with the above four target classes, we have used the popular C4.5 release 8 software package [10] with the same training set described in Section 3.1. The same process of extracting features from every 1-second clip and forming the training samples was also repeated. A tree with 277 nodes was obtained after pruning.

With the threshold-based classification scheme, we attempted to classify each 1-second clip by thresholding its values of the following four features: short-time energy, zero-crossing rate, fundamental frequency and spectral peak tracks. These features were proved to be very effective in classifying generic audio signals [4]. The thresholds used for this purpose were empirically determined from experimental results. Moreover, to achieve the best performance, we have tried to find the optimal thresholds for every test video.

Table 2. Overall classification accuracies of seven test videos.

Test Video	Length (min)	Classif. Accuracy
Train1	15:13	95%
Train2	19:19	98%
Train3	21:28	96%
Train4	13:23	99.5%
Education1	20:12	98.9%
Education2	40:12	99%
Education3	55:50	99%

The test set in this case contains 214-minute audio data, which only contains pure sound types and are collected from both training and education videos. Note that when we determine the audio type from SVM classifier outputs, we will only choose the class with the highest decision value as we only consider pure audio types in this comparison.

Table 3 compares the performance of the three classification schemes in terms of audio classification accuracy. Clearly, the SVM-based approach has outperformed the other two for every audio class. But surprisingly, the DT-based approach does not give the performance as we expected. Although its classification accuracy for speech is fairly good, the accuracies for the other three are rather low, especially for the environmental sound. Nevertheless, its resulting decision rules are quite readable, which provides us certain knowledge about the effectiveness of the employed audio features. For instance, from the output tree, we see that features such as the first subband energy mean, variances of the third and fourth subband energy ratio, energy and ZCR means, low ST-energy ratio, spectrum flux and harmony, are more effective than the others.

The performance of the threshold-based approach is acceptable, which however, is achieved based on tedious fine-tunings of various thresholds for every test video. In fact, we have attempted to fix the thresholds once we derived their optimal values from one test video. The classification accuracies in that case, however, varied significantly across different videos. For instance, we achieved only 15% classification accuracy for one test video where most of its speech signals were mis-classified as silence because of the extremely low voice volumes and fixed thresholds. This approach is thus impractical when a fully automated system is needed for very large video archives.

3.4. MFCC Feature Examination for Audio Classification

Mel-frequency cepstral coefficient (MFCC) is the feature popularly used in various speech processing and recognition applications. It was also employed in [5, 7] for the audio classification purpose. As an attempt to verify the effectiveness of the MFCC, we have conducted another two experiments: a) only use MFCC feature for SVM training and testing. In this case, each clip will be represented by 28 features with 14 MFCC mean and 14 MFCC variance values; b) use MFCC as the extra feature to the original feature set as described in Section 2.1. Thus in this case, each fea-

Table 1. SVM-based classification results of seven sound types where each number is in units of second.

Sound Type	Speech	Silence	Music	Environmental Sound	Speech w/ music	Speech w/ environ.	Environ. w/ music
Speech	**9160**	0	2	28	1	19	0
Silence	0	**636**	2	0	0	0	0
Music	2	0	**277**	0	0	0	0
Environ. Sound	30	0	6	**268**	0	4	6
Speech w/ music	4	0	12	0	**28**	0	4
Speech w/ environ.	23	0	20	26	0	**427**	0
Environ. w/ music	0	0	0	0	0	0	**83**
Classification Accuracy	**99.4%**	**100%**	**87%**	**84%**	**96.5%**	**95.7%**	**89%**

Table 3. Classification accuracy comparisons between the SVM-based, DT-based and threshold-based classification schemes.

Sound type	SVM-based	DT-based	Threshold-based
Speech	99.6%	97.9%	95.4%
Silence	100%	75.5%	93.7%
Music	93.1%	71.9%	77.7%
Environ.	90.3%	66.8%	75.4%
Average	95.8%	78%	85.6%

ture vector will contain 54 features with 26 from the original set and 28 from the MFCC feature. The original training and test sets are used for these two experiments.

The test results are summarized as follows: 1) experiment (a) produces comparable classification accuracy for speech, while it performs rather poorly for silence, music and environmental sound whose classification accuracies have dropped 10% on average; 2) experiment (b) has achieved a slightly better performance for speech, while the classification accuracy drops for silence, music and environmental sound. As this is not the result we expected, we suspect that the performance degradation in case (b) was mainly caused by the insufficiency of training data since the feature dimensionality has now doubled. We have thus collected more training data from various sources for silence, music and environmental sound, and repeated experiment (b). This time, no performance degradation was observed, yet the performance increment was also slight, only around 0.7%.

The conclusion we draw from these two experiments is, MFCC is an effective feature for recognizing speech signals, but in order to achieve good audio classification accuracies across various sound types, we must combine it with other perceptual features. Moreover, when the training data is insufficient, we may exclude the MFCC feature without sacrificing the system performance.

4. CONCLUSION

This paper presents our efforts on applying the Support Vector Machine (SVM) technology to the audio classification task, which classifies audio tracks of instructional videos into homogeneous audio segments with each having a distinct audio class label. Experiments performed on three education and four training videos have yielded an average 97.9% classification accuracy. The performance comparison between the proposed SVM-based, the decision tree-based and the threshold-based approaches also demonstrated the superiority of the proposed classification scheme. Finally, we examined the MFCC feature for the classification task, and concluded from our experiments that while it is effective in identifying speech signals, it should be combined with other features for a better system performance.

5. REFERENCES

[1] D. Kimber and L. Wilcox, "Acoustic segmentation for audio browsers," *Proc. of Interface Conference, Sydney, Australia*, July 1996.

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discrimination," *ICASSP'97*, vol. 4, Munich, Germany, 1997.

[3] S. Srinivasan, D. Petkovic, and D. Ponceleon, "Towards robust features for classifying audio in the CueVideo system," *ACM Multimedia'99*, 1999.

[4] T. Zhang and C.-C. Kuo, "Audio content analysis for online audiovisual data segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[5] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang, "Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification," *ICME'03*, 2003.

[6] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 1–47, 1998.

[7] L. Lu, S. Z. Li, and H. J. Zhang, "Content-based audio segmentation using support vector machine," *ICME'01*, 2001.

[8] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," *Proc. of the IEEE Signal Processing Society Workshop on Neural Networks*, vol. 2, pp. 775–784, 2000.

[9] T. Joachims, *Making large-scale SVM learning practical*, MIT Press, In Advances in Kernel Methods - Support Vector Learning, 1999.

[10] J. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, San Mateo, CA, 1993.