MIXED TYPE AUDIO CLASSIFICATION WITH SUPPORT VECTOR MACHINE

Lei Chen

Department of Computer Science
Hong Kong University of Sci. and Tech.
leichen@ust.hk

Şule Gündüz

Department of Computer Science
Istanbul Technical University
gunduz@cs.itu.edu.tr

M. Tamer Özsu

School of Computer Science
University of Waterloo
tozsu@uwaterloo.ca

ABSTRACT

Content-based classification of audio data is an important problem for various applications such as overall analysis of audio-visual streams, boundary detection of video story segment, extraction of speech segments from video, and content-based video retrieval. Though the classification of audio into single type such as music, speech, environmental sound and silence is well studied, classification of mixed type audio data, such as clips having speech with music as background, is still considered a difficult problem. In this paper, we present a mixed type audio classification system based on Support Vector Machine (SVM). In order to capture characteristics of different types of audio data, besides selecting audio features, we also design four different representation formats for each feature. Our SVM-based audio classifier can classify audio data into five types: music, speech, environment sound, speech mixed with music, and music mixed with environment sound. The experimental results show that our system outperforms other classification systems using k Nearest Neighbor (k-NN), Neural Network (NN), and Naive Bayes (NB).

1. INTRODUCTION

Audio classification can be used in many different application domains. For example, news information providers would like to label the huge amount of news audio data they collect everyday in a reliable and easy way, and video classification systems can use the audio information along with the video stream to achieve higher accuracy. Due the huge amount of audio data and the high expense of manual classification, an automatic audio classifier is need.

There are a lot of proposals for classifying single type audio data [1, 2, 3, 4, 5, 6], such as music and speech. However, the audio features that are used for differentiating single type audio data –such as Zero Crossing Rate (ZCR)– do not work for mixed type audio data. Figure 1 shows the characteristics of variance zero crossing rate[1] on mixed type audio data (speech with music background and music mixed with environment sound), which is the most often used audio feature to differentiate music from speech. From Figure 1, we can see that the ZCR values of speech with music background and of music mixed with environment sound are *non-linear separable*. Two data sets are non-linear separable if we can not find a hyper plan to separate two data sets. The challenge is how to classify mixed type audio data with the existing audio features.

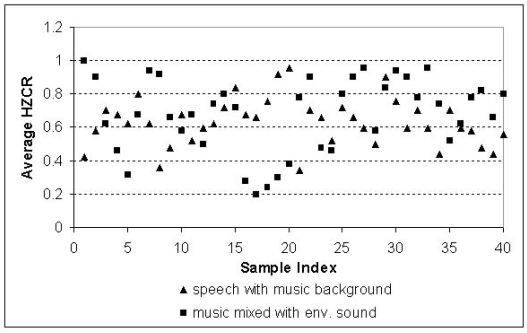

Fig. 1. Average HZCRR of speech mixture and music-env mixture

Support Vector Machine (SVM) has been successfully used in pattern recognition [7], such as speaker identification, face detection, and text recognition. Compared to other classifiers that separate the data in its original space, such as k Nearest Neighbor(k-NN), Neural Network (NN), and Naive Bayes (NB), SVM maps non-linear separable data to higher dimensional space and performs separation in that space. We exploit this characteristic and propose a SVM-based audio classifier to classify mixed type audio data. Besides selecting audio features, we also design four different representation formats for each feature in order to capture the characteristics of different types of audio data. In our work, audio clips are extracted from movies and manually marked into five types, music (mus), speech (spe), environment sound (env), speech mixed with music (spemus), and music mixed with environment sound (musenv). These are used as training and test data for the classifiers. In our experiment, we compare the performance of SVM compared to other three types of classification algorithms, k-NN, NN, NB. The result show that SVM performs better than other three classifiers confirming our initial expectations.

[1] We use HZCRR [6].

2. RELATED WORK

Based on the type of audio data, the work related to audio classification can be classified into two categories:

1. Single type audio classification
   Saunders [1] addresses the issue of single type audio classification for FM radio. Zero crossing rate (ZCR) and short time energy are used to classify input audio into two types: speech and music. Scheirer and Slaney [2] use thirteen features in time, frequency, spectrum and cepstrum domains and achieve better classification. Based on Scheirer's conclusion, Carey et al. [3] compare audio features for speech and music discrimination. They find that simple audio features, such as pitch and amplitude, have significant differences between music and speech. Since then, many approaches have been proposed to classify single type audio using different audio features and classifiers [4, 5, 8, 6, 9].
2. Mixed type audio classification.
   Srinivasan et al [10] propose a method to classify mixed type audio data, such as speech mixed with music. They use a fuzzy rule based model with empirically determined thresholds. Zhang and Kuo [11] propose a heuristic rule-based model for classification into speech, music, song, environmental sound, speech with background music, silence, etc. Empirically determined thresholds are used in this study as well. Different from previous approaches, our classifier does not need to set up classifying thresholds, the system can be trained with training data and can automatically classify the new data. Recently, Kiranyaz et al. [12] propose a generic frame work to classify audio into speech, music, fuzzy or silent. However, the fuzzy type can not tell whether the audio data is speech with music background or music mixed with environment sound. Our previous work [13] which employed audio cues to aid segmenting video clips mainly focus on the video segmentation accuracy.

3. SELECTED AUDIO FEATURES AND REPRESENTATIONS

In order to capture the characteristics of audio data, we select four audio features from two domains: variance of zero crossing rate and silence ratio from time domain, harmonic ratio and sub-band energy from frequency domain. We extract audio data from movies[2]. First, the extracted audio data are segmented into 5 second audio clips and each clip is further divided by a 1 second window into 5 segments. Then, the four audio features are computed for each audio segment. Finally, we propose four different representation formats as final features for each clip, which are $mean$, max, min and $(max + min)/2$, where $mean$, max, and min stand for mean, maximum, and minimum of segment-based feature values, respectively. In previous works [10, 11], only $mean$ is used to represent feature values. However, $mean$ blurs local characteristics of the data. For environment sounds that usually last a very short time, $mean$ is not a good representation. However, max, min, and $(max+min)/2$ can capture local characteristics of the data.

1. Variance of Zero Crossing Rate
   ZCR is defined as the number of zero crossing within an audio frame [1]. It has been widely used to differentiate speech from music. We compute the variance of ZCR for each audio segment.
2. Silence Ratio
   Silence Ratio (SR) is defined as the ratio of the amount of silence in an audio piece to the length of the piece. SR is an useful statistical feature for audio classification, it is usually used to differentiate music from speech [14]. Normally speech has higher SR than music. We divide a 1-second window into 50 frames. For each frame, the root mean square (RMS) is computed and compared to the RMS of the whole window. If the frame RMS is less than 50% of window RMS, we consider it as a silence frame.
3. Harmonic Ratio
   Spectrum analysis shows that music is more harmonic than speech, since speech contains a sequence of tonal (vowels) and noise (consonants) [10]. Harmonic sound is defined as one that contains a series of frequencies which are derived from a fundamental or original frequency as a multiple of that. For each 1-second window audio clip, we divide it into 10 frames. We compute the harmonic frequency of each frame using the algorithm in [15]. The harmonic ratio (HR) is defined as the ratio of the number of frames having a harmonic frequency to the total number of frames in the window.
4. Sub-band Energy
   The frequency of audio segment are segmented into four ranges based on the relevant frequencies on speech and music: R_1 (less than 1kHz), R_2 (1kHz-8kHz), R_3 (8kHz-16kHz), and R_4 (more than 16kHz) [10]. The sub-band energy for each audio segment is defined as the sum of the energy within each frequency range. We compute the variance of sub-band energy in R_1 (SBE1) as the feature for each segment. Because of the bandwidth limitation that speech is within 8kHZ and music can span over 16kHZ, the variance of SBE1 of speech is usually higher than that of music.

4. CLASSIFICATION ALGORITHMS

Some research that compare the performance of SVM with Naive Bayes, C4.5 and neural network [16, 17] show that SVM has a lower error rate. Since our main goal is to do a classification without assigning any threshold value for each

[2] 1. "Crouching Tiger and Hidden Dragon", 2000; 2. "Patch Adam", 1998

feature, we do not choose a rule based system as in [10] for comparing our results. Hence, we only compare our results with the three most popular learning classifiers, namely k Nearest Neighbor (k-NN), Neural Network (NN) and Naive Bayes (NB) in term of classification accuracy.

The k nearest neighbor classifier is an instance based classifier which stores all the entire training set in memory. To classify a new audio clip, the Euclidean distance is computed between the audio clip and each stored training audio clip. The new audio clip is assigned to the class that is most frequent among the nearest k training audio clips. Aha describes several space - efficient variations of nearest neighbor algorithms [18].

The neural network uses backpropogation to classify instances. Backpropogation (BP) is an algorithm for modifying the weights of a Multilayered Perception based on incremental gradient descent of mean-square error.

The naive Bayes algorithm computes a discriminant function for each n possible classes. It assumes that each feature of an audio clip is drawn independently from a normal distribution and classifies according to the Bayes optimal decision rule. We choose NB and NN from Weka data mining tool for running the experiments [19].

The Support Vector Machine is a classifier, originally proposed by Vapnik, that finds a maximal margin separating hyperplane between two classes of data [20]. There are non-linear extensions to the SVM that use kernel function to map the input points to a high dimensional space. For more information, see Burges' tutorial [7]. Since SVM is based on two-class classification problems, several solutions have been proposed to handle a n-class problem. A more general solution is to convert a n-class problem into n two-class problems and for the ith two-class problem, class i is separated from the remaining classes, which is defined as *one-against-all* [21]. Another approach is to convert a n-class problem into $n(n-1)/2$ two-class problems which cover all pairs of classes. This method is called *pairwise classification*. There is no theoretical analysis of the two strategies with respect to classification performance. However, regarding the training effort, the one-against-all approach is preferable since only n SVMs have to be trained compared to $n(n-1)/2$ SVMs in the pairwise approach. In our study, we use the SvmFu package to run the experiments [22].

5. EXPERIMENTS

In our experiments, one hour audio data are extracted from each movie as mentioned in Section 3, and they are segmented into 5 second audio clips. We manually marked each audio clip with one of the five class labels, mus, spe, env, spemus, and musenv. In our study, the environment sound is defined as the sound that is not music or speech, such as the sound of door close and opening, the sound of footsteps, and the sound of a bird singing, etc. In movies, there are no single type sound, such as speech, because people always talk in a real environment and some environment sound is mixed with speech. Therefore, we label all the clips in which people talk with environment sound background as speech. Approximately 30% of these extracted audio clips are randomly selected as the test set, and the remaining part as the training set.

We first test the classification accuracy of our classifier using four different representation formats, $mean$, max, min, and $max + min/2$. The classification accuracies are: 71.61%, 78.05%, 62.24%, and 69.49%, respectively. By checking the feature values of environment sound audio clips, we find that the environment sound within the clips always produces the local maximum values of the features. This phenomena explains that maximum representation outperforms the other three. We also find that min and $max + min/2$ do not perform better than $mean$. This is because min is close to 0 in some audio clips, which cause it not useful in separating data. In the rest of the experiment, we only report the results using max as the representation of feature values in each clip.

Our SVM-based mixed type audio classifier is then tested together with other three classification methods, k-NN, NN, NB. The experiments are repeated for different k values for k-NN, and different kernel functions, kernel parameters and multi-class approaches for SVM. Despite the fact that polynomial kernel provides better performance on different data sets than the other kernels as pointed in [16], our results show that the Gaussian kernel performs better with the SVM in audio classification. The one-against-all approach performs better than the pairwise classification which shows that the first one is a considerable technique in audio classification for the multi-class problem. The SVM has a user defined C parameter which is the cost of the penalty of the errors. Table 1 shows multi-class accuracy for a variety of conditions of SVM with one-against-all approach. We have 75.61% accuracy with 5-NN, which is the best result among other k values. The accuracies for NN and NB are 74.69% and 73.78% respectively. The SVM achieves $8-15\%$ lower error than the other three classifiers. It is observed that the selection of user defined kernel parameter and the C parameter is not an easy task and has a significant effect on the performance of the classifier. However, in general SVM is the best choice for classifying multi-class audio data.

Finally, to evaluate the effect of the environment sound on the accuracy we repeat the experiments with three types, namely speech, music and speech mixed with music, on the same training and test sets. The accuracies for 5-NN, NN, NB and SVM are 89.69%, 90.07%, 89.93% and 90.84% respectively. The experiments are repeated on the other three classes that consist of environment sound, music, and music mixed with environment sound. The accuracies for 5-NN, NN, NB and SVM are 60%, 63.81%, 61.9% and 64.76% respectively. From these results, we find that environment

	C=1	C=10	C=50	C=100
Linear SVM	72.561	73.171	73.171	73.476
Gaussian Kernel	74.390	77.439	75	76.524
Gaussian Kernel sigma =0.1	75.915	74.695	73.171	72.256
Gaussian Kernel sigma = 0.5	75.610	75.915	75.610	75.610
Gaussian Kernel sigma = 0.8	75	77.744	76.524	75.915
Gaussian Kernel sigma=0.9	74.39	**78.049**	76.22	75.915

Table 1. Accuracy % of the multi-class classification with SVM

sound has a negative effect on the accuracy of classification. This may be due to the fact that the definition of environment sound is too board. The environment sound can include sound of nature (e.g. sound of sea), sound of animal (e.g. singing of a bird), some man-made sound (e.g. sound of footsteps or closing a door), etc. Therefore, the characteristics displayed by different environment sounds may be quite different, which results in lower accuracy compared with that of the experiments on types without environment sound. However, in all cases, SVM performs better than the other three classifiers. This is not surprising, because SVM has a good performance on non-linear separable classes. The results also confirm another advantage of SVM that it performs better compared to the other three classifiers if the number of training examples are few.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we examined the suitability of SVM on mixed type audio classifier and proposed four different representation formats for extracted feature values of audio clips. Our comparison experiments show that the maximum of feature values in each audio clip can capture the characteristic of mixed type audio data and SVM-based classifier outperforms other popular classifier such as k-NN, NB, and NN. Once kernel type is fixed, SVM has only two user defined parameters (the error penalty, C, and kernel parameter) but the best choice of kernel for a given problem is still a research issue. Based on our experiments, Gaussian kernel is the best choice for mixed type audio classification.

Our experiments also show that the fuzzy definition of environment sound has a negative effect on the accuracy of SVM (and other classifiers). In future, we will divide environment sound into finer subclasses, such as nature environment sound, animal sound and man-made sound and investigate common characteristics of each subclasses. Some new audio features will be introduced to characterize these subclasses in order to reduce the error rate in the classification of classes that have environment sound. Since manually labelling of audio data is expensive we are extending our work to do a classification using both labelled and unlabelled data as training set for SVM.

7. REFERENCES

[1] J. Saunders, "Real-time discrimination of broadcast speech/music," in *ICASSP*, 1996.

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *ICASSP*, 1997.

[3] M.J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *ICASSP*, April 1999.

[4] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia applications," in *ICASSP*, June 2000.

[5] S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method," *IEEE Trans. on Speech and Audio Processing*, 2000.

[6] L. Lu, H. Jiang, and H. J. Zhang, "A robust audio classification and segmentation method," in *Proc. 9th ACM Int. Conf. on Multimedia*, 2001.

[7] C. J. C. Burges., "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 1998.

[8] S.-Z. Li and G. Guo, "Content-based audio classification and retrieval by support vector machines," in *PRCM (invited talk)*, 2000.

[9] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, 1996.

[10] S. Srinivasan, D. Petkovic, and D. Ponceleon, "Towards robust features for classifying audio in the CueVideo system," in *Proc. 7th ACM Int. Conf. on Multimedia*, 1999.

[11] Tong Zhang and C. C. Jay Kuo, "Audio content analysis for online audiovisual data," *IEEE Trans. on Speech and Audio Processing*, 2001.

[12] A. F. Qureshi S. Kiranyaz and M. Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," *IEEE Trans. on Speech and Audio Processing*, 2006, in print.

[13] L. Chen, S. Rizvi, and M. T. Özsu, "Incorporating audio cues into dialog and action scene extraction," in *SPIE Storage and Retrieval for Media Databases*, 2003.

[14] G. J. Lu and T. Hankinson, "A technique towards automatic audio classification and retrieval," in *ICIP*, 1998.

[15] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," in *Proc. 4th ACM Int. Conf. on Multimedia*, 1996.

[16] A. S. Shawkat and A. Abraham, "An empirical comparision of kernel selection for supoort vector machines," in *Second International Conference on Yhbrid Intelligent Systems*, 2002.

[17] J.D.M. Rennie and R.Rifkin, "Improving multiclass text calssification with the support vector machine," Tech. Rep., Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 2001.

[18] D.W.Aha, "A study of instance-based algorithms for supervised learning tasks," Tech. Rep., University Of California, Irvine., 1990.

[19] Weka3, "The machine learning software in java," http://www.cs.waikato.ac.nz/ml/weka/index.html.

[20] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 2001.

[21] E.L. Allwein, R.E.Schapire, and Y.Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *Journal of Machine Learning*, 2002.

[22] R. Rifkin, "Svmfu," http://five-percent-nation.mit.edu/SvmFu/.