SVM-based Audio Scene Classification

Hongchen JIANG, Junmei BAI, Shuwu ZHANG, Bo XU
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080
Email: { hcjiang, jmbai, swzhang, xubo }@hitic.ia.ac.cn

***Abstract*-Audio scene classification is very important in audio indexing, retrieval and video content analysis. In this paper we present our approach that uses support vector machine (SVM) for audio scene classification, which classifies audio clips into one of five classes: pure speech, non-pure speech, music, environment sound, and silence. Among of them, non-pure speech may further be divided into speech with music and speech with noise. We also describe two methods to select effective and robust audio feature sets. Based on these feature sets, we have evaluated and compared the performance of two kinds of classification frameworks on a testing database that is composed of about 4-hour audio data. The experimental results have shown that the SVM-based method yields high accuracy with high processing speed.**

I. INTRODUCTION

There are two key problems in audio scene classification: audio feature selection and classifier selection. The audio feature should be effective and robust under various audio circumstances, and the classifier should enable to separate the samples into different classes.

In recent years, there have been intensive studies on audio scene classification, and various audio feature and classification schemes have been proposed. For example, Scheirer et al.[1] introduced 13 features such as 4Hz modulation energy, zero-crossing rate, Spectral Flux and Spectral Centroid etc and Performed speech/music classification with different classifiers such as Gaussian Mixture Model (GMM) and K-nearest Neighbor (KNN). Zhang and Kuo [2] used a fixed threshold scheme to discriminate eight different audio classes including silence, speech, music, song, environment sound and their various combinations based on some simple features such as short-time energy, zero-crossing rate, the fundamental frequency and the spectral peak tracks. An accuracy of above 90% was reported. Kimber and Wilcox [3] employed Hidden Markov Model (HMM) to classify audio signals into music, speech or silence using cepstral features. Research by Zhu Liu et al. [4] aimed at the classification of TV programs into different categories, where cepstral coefficients were used as features and artificial neural network was used as the classifier. In the work by Lu et al. [5], Mel-frequency cepstral coefficients (MFCC), zero-crossing rate, spectral centroid, spectral spread, spectrum flux and band periodicity were introduced to hierarchically classify audio signals into five classes based on the SVM classifier. In this paper, we introduce our approach that uses support vector machine (SVM) for audio scene classification, which classifies audio clips into one of five classes: pure speech, non-pure speech, music, environment sound, and silence.

The rest of the paper is organized as follows: Section II discusses in detail how to select and construct effective and robust audio feature sets by two methods for different audio-pair types. Section III describes and compares two different SVM-based multi-classification frameworks. Experimental results are shown in Section IV. Finally, concluding remarks and future research work are given in Section V.

II. AUDIO FEATURE SELECTION

In our SVM-based audio classifiers, sixteen kinds of audio features are primarily taken into consideration, and the detailed descriptions of these features can be found in the references [1, 2, 3, 4, 5, 6]. These features are: Zero-Crossing Rate (ZCR), High ZCR Ratio (HZCRR), Short-Time Energy (STE), Low STE Ratio (LSTER), Root Mean Square (RMS), Silence Frame Ratio (SFR), Sub-band Energy Distribution (SED), Spectrum Flux (SF), Spectral Centriod (SC), Spectral Spread (SS), Spectral Rolloff Frequency (SRF), Sub-band Periodicity (BP), Noise Frame Ratio (NFR), Linear Spectrum Pair (LSP), Linear Predictive Cepstral Coefficients (LPCC) and Mel-frequency Cepstral Coefficients (MFCC).

Prior to feature extraction, the audio signal is firstly converted into a general format of 8-KHz, 16-bit, mono-channel. Then it is segmented into non-overlapping audio clips. Each clip is a basic test unit with 1s long. Next, the clip is further divided into non-overlapping 25 ms-long frames. Except HZCRR, LSTER, SFR and NFR, the other 12 features are extracted from each frame. The means and standard deviations in one audio clip are computed to get clip-based feature. In our method, the order of LPCC, LSP and MFCC is 10, 10 and 8 respectively, thus the total vector dimensions of the clip-based feature amount to 90.

Theoretically, each dimension of the above clip-based feature vector can be used the classification feature, but for different audio classes, the effectiveness and robustness of these features are not identical. In other words, it is not reasonable that we select a uniform feature set for discriminating any pair of audio classes. For instance, if we only use SFR to classify music and noise, the result will not be satisfied. The reason is that the SFR statistics value of music and noise are both very low. In the other hand, if we only use SFR to classify music and speech, the result will be very good because the SFR statistics value of speech is greater than that of music. Therefore, in order to improve the accuracy of classification, we will analysis these features by two methods, thus select effective and robust feature combinations for discriminating any pair of audio classes.

0-7803-9361-9/05/$20.00 ©2005 IEEE

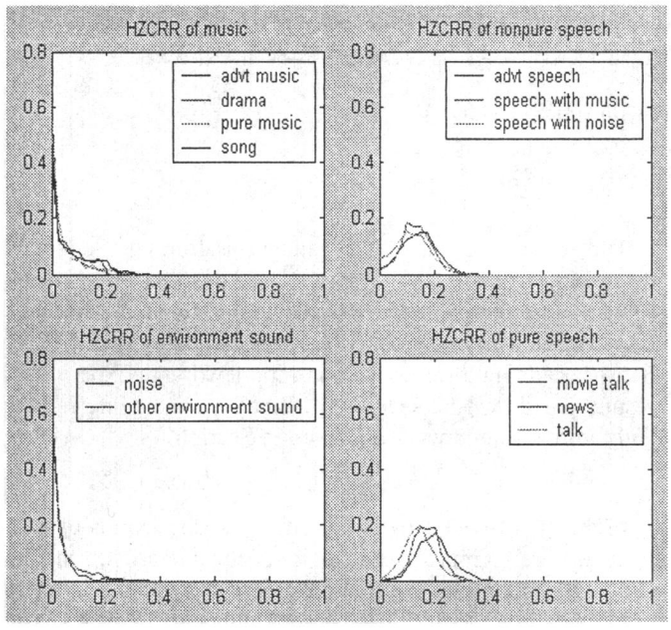

Fig.1. Probability distribution curves of HZCRR

The first method is that we draw the probability distribution curves of statistics feature value computed from the audio database using 1s windows, then observe 1) whether the curves of different audio classes are similar or not and 2) whether the curves of different sources of each audio class are similar or not. For one feature, if two audio classes have similar feature curve shape, this feature is not effective on discriminating between these two classes, and if the different signal sources of each audio class have dissimilar feature curve shape, this feature is not robust to represent the feature of this audio class.

The second method is that we first implement a baseline classifier employing some commonly used feature, then add each new feature to construct a new classifier, and then compare the classification results between using the new classifier and the baseline classifier to observe the effectiveness of each new feature. If the classification accuracy of the new classifier improves quite a lot, the new added feature is considered to be effective.

For the first method, we take the probability distribution curves of HZCRR as the example to illustrate the effectiveness and robustness of HZCRR feature for classifying audio classes, as shown in Fig. 1.

It can be seen from Fig. 1 that HZCRR values of music and environment sound clips mostly fall below 0.2, while HZCRR values of pure speech and non-pure speech centralize around 0.15. Therefore, it is difficult to discriminate between music and environment sound if we only use HZCRR feature, so does between pure speech and non-pure speech. However, if music and environment sound constitute the non-speech class, and pure speech and non-pure speech constitute the speech class, it is not hard to see that HZCRR feature can well discriminate speech from non-speech.

TABLE I

EFFECTIVENESS OF NEW ADDED FEATURES ON MUSIC/ENVIRONMENT SOUND CLASSIFICATION

Baseline	Baseline +ZCR	Baseline +BP	Baseline +SF	Baseline +SC
98.32%	97.85% (-0.47%)	97.85% (-0.47%)	98.52% (+0.20%)	97.86% (-0.46%)
Baseline +SS	Baseline +SRF	Baseline +LPCC	Baseline +LSP	Baseline +MFCC
97.72% (-0.60%)	98.46% (+0.14%)	98.52% (+0.20%)	98.70% (+0.38%)	98.48% (+0.16%)

TABLE II

THREE GROUPS OF FEATURE SETS

Pair of Audio types	Feature Sets
Speech/ non-speech	HZCRR、ZCR、LSTER、RMS、SC、SS、BP、NFR、SF、LPCC、LSP、MFCC
Pure speech/ Non-pure speech	SFR、ZCR、RMS、SC、SS、SF、LPCC、LSP、MFCC
Music/ Environment sound	NFR、STE、SBE、SF、LPCC、LSP

From Fig. 1, we can also see that each of four audio classes is composed of different audio signals from different TV programs, but HZCRR probability distribution curves of signals of different TV programs of the same audio type are similar. For example, for pure speech type, the signals from movie/sitcom (movie talk), news report (news) and dialog/talk/interview (talk) all have the resemble curve shape. The fact shows that HZCRR is robust to each audio type.

As for the second method, we take the experimental results shown in table I as an example to illustrate the effectiveness of new added features.

From table I, we can see that the baseline classifier performs well for music/environment sound classification. However, after adding the new feature of ZCR, BP, SC and SS, the classification accuracy is reduced by 0.47%, 0.47%, 0.46% and 0.60% respectively, while after adding the new feature of SRF, SF, LPCC, LSP and MFCC, the accuracy is increased by 0.14%, 0.20%, 0.20%, 0.38% and 0.16% respectively. These facts show that ZCR, BP, SC and SS are not effective on music/environment sound classification while SRF, SF, LPCC, LSP and MFCC are the very reverse.

Based on these experiments and analyses, in this paper we construct three groups of feature sets for speech/non-speech, pure speech/non-pure speech, and music/environment sound classification respectively, as shown in table II.

III. TWO FRAMEWORKS OF SVM-BASED AUDIO SCENE CLASSIFICATION

Support vector machine (SVM), as a very efficient classifier, has been extensively used in many fields of pattern recognition such as image analysis, text classification, speech

recognition and speaker identification. However fewer works have tried to apply SVM to audio scene classification. Compared to GMM, HMM, KNN and ANN, SVM uses a nonlinear kernel function to map samples of two classes, which can't be separated in low dimensional feature space, into a high dimensional feature space in which SVM learns to find the optimal separating hyper-plane, thus maximizes the margin of two classes. In our work, we classify a 1-s audio clip into five classes of pure speech, non-pure speech, music, environment sound and silence. Based on SVM, we implement two kinds of classification frameworks using two strategies.

The first framework [5] is hierarchical structure. The input audio is first classified into silence and non-silence clip by SFR. It will be marked as silence if the SFR value is lower than a fixed threshold. Then the non-silence clips are classified into speech and non-speech by the first SVM classifier. Next, non-speech signals are further classified into music and environment sound by the second SVM classifier, while speech signals are classified into pure speech and non-pure speech by the third SVM classifier respectively. This kind of framework is diagramed in the following Fig. 2.

The second framework [7] is parallel structure. As the first framework, the input audio is first classified into silence clip and non-silence clip by SFR. Silence will be marked if the SFR value is lower than a fixed threshold. Then unlike framework I , for the pure speech, non-pure speech, music and environment sound 4 classes, we construct $4\times(4-1)/2=6$ SVM classifiers where each one is trained on data from two classes. When constructing the SVM classifier for discriminating between the *i*th class and the *j*th class, we use the data from the *i*th class and the *j*th class for training. We label the *i*th class with +1 while label the *j*th class with −1. During testing, test data is scored by each SVM classifier, and then we vote for each class according to these scores. If the score of the *i*th class is larger than that of the *j*th class for one SVM classifier, then the vote for the *i*th class is added by one, otherwise, the *j*th class is increased by one. Finally we predict the test sample in the class with largest vote. This voting method described above is also called the "Max Wins" strategy. This kind of framework is diagramed in the following Fig. 3.

Compared Fig.2 with Fig.3, we can see that each of the two frameworks has some advantages and disadvantages: 1) Framework II has three SVM classifier more than framework I , which decides that the training time of framework II is much longer than framework I . 2) Framework I has a drawback that if the signal is misclassified by the first SVM classifier, it will be never reach the correct type. 3) Framework II has the case that two classes have identical votes so that difficult to distinguish from each other. 4) if we want to detect two new classes of speech with music and speech with noise, framework I needs only another SVM classifier for distinguishing the two new classes in non-pure speech while framework II needs to add four SVM classifiers. However, if we distinguish more complicated audio classes,

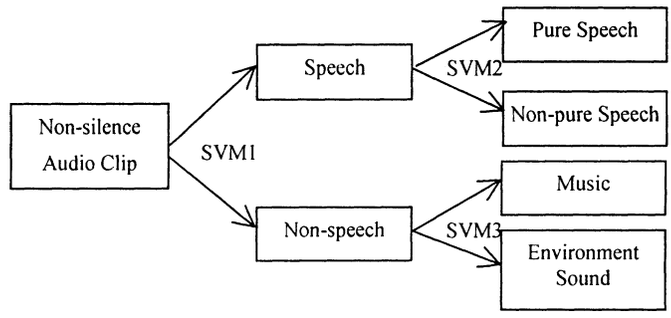

Fig.2. Framework I of audio scene classification

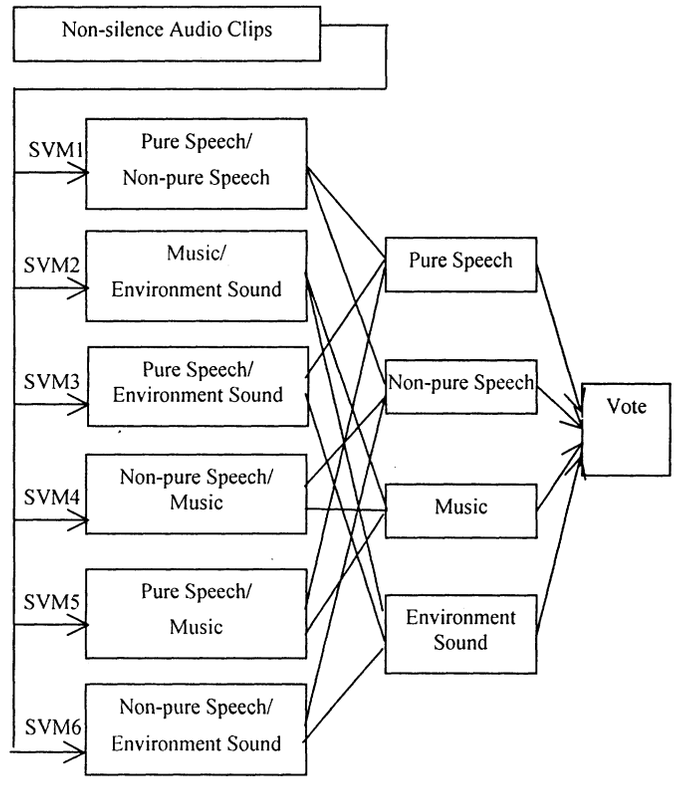

Fig.3. Framework II of audio scene classification

for example, music genre classification, obviously framework I will be invalid.

IV. Experimental Results

The data used in our experiment is collected from real TV programs, which is about 343 minutes in total. 94 minutes of data is used for training, and 249 minutes of data is used for testing. The training set consists of 25 minutes of pure speech, 25 minutes of non-pure speech, 25 minutes of music and 19 minutes of environment sound. The test set includes 109 minutes of pure speech, 103 minutes of music, 25 minutes of non-pure speech and 12 minutes of environment sound. Pure speech is mainly selected from the program of news report, movie/sitcom, or dialog/talk/interview.. Non-pure speech mainly includes speech with music and speech with noise, and the music background in advertisement speech is a

little stronger. Music is mainly consisted of three components: pure instrumental music produced by different musical instruments, songs sung by male, female, or children, and some drama. Environment sound is composed of the sounds of applause, animal, footstep, explosion, vehicles, laugh, crowds, and so on. Pure speech and non-pure speech can be combined into speech class, while music and environment sound can be combined into non-speech class.

The training and testing data are all converted into the uniform format of 8-KHz, 16-bit, mono-channel. In our experiments, we set 1s as a test unit. If there are two audio types in a 1s audio clip, we will classify it as the time-dominant audio type.

The three SVM models of SVM1, SVM2 and SVM3 in framework I are trained using the first, the second and the third feature set respectively as shown in table II. In framework II, the SVM1 and SVM2 are modeled using the second and the third feature set respectively; while SVM3, SVM4, SVM5 and SVM6 are modeled using the first feature set. The features in one of feature sets constitute a clip-based feature vector after scaling to [−1, +1].

A. Training models

Framework I needs to train three SVM kernel models and framework II needs to train six kernel SVM models. We select the Radial Basis kernel function (RBF) to train these models in our experiments. There are two parameters while using RBF kernels: C and σ. It is not known beforehand which C and σ are the best for one model. Consequently in order to get the optimal parameters, we use cross-validation and grid-search technology [8] in our work.

We first divide the training data set into five subsets of equal size. Sequentially one subset is tested using the classifier model trained on the remaining four subsets. Thus, each sample of the whole training set is predicted once so the cross-validation accuracy is the percentage of data that are correctly classified. Meanwhile, we use grid-search on C and σ using cross-validation. We try exponentially growing sequences for pairs of (C, σ) (for example, $C=2^{-1}$, $2^{0},\cdots,2^{13}$, $\sigma=2^{-4}$, $2^{-3},\cdots,2^{3}$) and the one with the best cross-validation accuracy is picked as the optimal parameters.

Under the optimal parameters C and σ, the cross-validation accuracy of framework I and framework II is listed in table III and table IV respectively. Where, nSample means the total number of samples in a training set, and nSV means the number of support vectors obtained from the training set.

It is apparent from table III and table IV that the cross-validation accuracy of each SVM classifier in framework I and framework II is beyond 97% and differs little with the training set. The average accuracy of framework I is up to 98.18%, while the average accuracy of framework II reaches 98.81%. This reveals that SVM-based method is very effective on audio scene classification.

TABLE III

CROSS-VALIDATION ACCURACY OF FRAMEWORK I

	C	σ	NSample	nSV	Accuracy
SVM1	588.13	0.44	5546	1427	98.43%
SVM2	14263.10	0.38	2918	1000	97.81%
SVM3	9.19	0.25	2628	435	98.29%

TABLE IV

CROSS-VALIDATION ACCURACY OF FRAMEWORK II

	C	σ	nSample	nSV	Accuracy
SVM1	14263.10	0.38	2918	1000	97.81%
SVM2	9.19	0.25	2628	435	98.29%
SVM3	10.56	0.16	2566	207	99.73%
SVM4	16.00	0.15	2983	516	97.85%
SVM5	73.52	0.22	3032	322	99.87%
SVM6	1.52	0.14	2517	229	99.28%

B. Comparisons of different classification frameworks

The table V and table VI show the classification results of the four types in the form of a confusion matrix (The number of in the tables is in units of second). Accuracy and average accuracy are used to evaluate the classification performance of each class and the total classes respectively.

Suppose n_{ij} is the number of clips in ith class classified into jth class, the accuracy of ith class and the average accuracy of the total classes can be defined as the following formulae (1) and (2).

$$P_i = \frac{n_{ii}}{\sum_{j=1}^{4} n_{ij}} \tag{1}$$

$$Q = \frac{\sum_{i=1}^{4} n_{ii}}{\sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij}} \tag{2}$$

It is noted that we have excluded the silence clips when calculating the accuracy because the four audio types all have more or less silence clips but we have not built special silence training and testing data set.

TABLE V

SVM-BASED CLASSIFICATION RESULTS OF FRAMEWORK I

	Pure Speech	Non-pure Speech	Music	Environment Sound
Pure Speech	6054	38	22	0
Non-pure Speech	49	1345	172	0
Music	16	90	5930	24
Environment Sound	6	0	75	674
Silence	457	0	5	2
Accuracy	98.84%	91.31%	95.66%	96.56%

TABLE VI

SVM-BASED CLASSIFICATION RESULTS OF FRAMEWORK II

	Pure Speech	Non-pure Speech	Music	Environment Sound
Pure Speech	6075	40	16	0
Non-pure Speech	40	1375	197	0
Music	5	57	5926	29
Environment Sound	5	1	60	669
Silence	457	0	5	2
Accuracy	99.18%	93.35%	95.60%	95.85%

It can be calculated that the average accuracy of the total classes of framework I is about 96.61% and that of framework II is about 96.90%. Therefore, the performance of framework II is a little better than that of framework I. The main reason is that there is still a drawback in framework I as said in Section III, though we apply some smooth rules after the first classification by SVM1. Using the smooth rules is based on the empirical facts that a continuous audio stream does not have abruptly and frequently changed audio types. However, because framework II applies the voting strategy to predict the test sample, it is unavoidable that two classes have the same votes. If this happens, we simply select the one with the smaller index.

From the two tables, we can also see that the accuracy of pure speech is much higher than the other three classes. However the non-pure speech has the lowest classification accuracy. The non-pure speech and music cannot be well discriminated from each other. This is because the advertisement speech in non-pure speech class has so strong music background that it is easily misclassified into music; while some songs in music class has weaker music background thus it is easily misclassified into non-pure speech. But generally speaking, the classification performance by SVM-base method is acceptable.

In addition, the processing time for these two frameworks on a Pentium 4 computer with 3GHz CPU and 512M memory is about 0.07 times real-time. In other words, classifying the total 249 minutes test data only takes 18 minutes, and this is enough to meet the need of audio scene real-time classification.

TABLE VII

CLASSIFICATION RESULTS OF FRAMEWORKS I USING A UNIFORM FEATURE SET

	Pure Speech	Non-pure Speech	Music	Environment Sound
Pure Speech	6030	28	23	0
Non-pure Speech	56	1336	225	0
Music	21	107	5878	37
Environment Sound	14	2	73	661
Silence	457	0	5	2
Accuracy	98.45%	90.70%	94.82%	94.70%

C. *Comparisons of using a uniform feature set and different feature sets*

Before this experiment, we first construct a uniform feature set, which includes all of the sixteen features described in Section II. These features are extracted to form a clip-based feature vector after scaling to [-1, +1]. Using the uniform feature set and framework I, we perform the classification for the three kinds of audio-pair types, which are speech/non-speech, pure speech/non-pure speech and music/environment sound classification. The final classification results are shown in table VII.

Compared with the classification results shown in table V, which are achieved by using different feature sets, we can find that for each audio type, the classification accuracy is reduced by 0.39% — 1.86%, and the total average accuracy is decreased to 95.93%. This experiment results indicate that using different feature sets for two different audio types is more effective and reasonable than that only using a uniform feature set.

D. *Classification of speech with music and speech with noise*

On the basis of framework I, non-pure speech is further classified into speech with music and speech with noise. Using the feature set including NFR, ZCR, RMS, SF, SC, SS, BP, SBE, LPCC, LSP and MFCC, we perform the training of a new SVM model. The final classification accuracy is up to 92.19%. The accuracy is relatively lower. This is because most of statistics feature of these two classes are so similar. In spite of this, this result proves the fact that non-pure speech is further classified into speech with music and speech with noise is feasible.

V. CONCLUSION AND FUTURE WORK

In this paper, we have employed two kinds of SVM-based classification frameworks to classify audio signals into five classes, which are pure speech, non-pure speech, music, environment sound and silence. Experiments have achieved

the average 96.61% and 96.90% classification accuracy respectively. With the first kind of SVM-based classification framework, we have introduced our efforts on how to select and construct optimal feature combination for classifying different pairs of audio types. The experimental results have showed that our method can achieve an average 96.61% of accuracy rate by using different feature combination for classifying different pairs of audio types. It relatively raised about 17% of accuracy rate compared to the method using a uniform feature set. In the mean time, we have also explored the feasibility that non-pure speech is further classified into speech with music and speech with noise.

Based on the audio scene classification results, our future work will mainly focus on developing an effective scheme to detect all change points in an audio signal so as to segment the audio stream.

ACKNOWLEDGEMENT

This research work in this paper is partially supported by the National Natural Science Foundation of China under grant No. 60475014 and the National Hi-tech Research Plan under grant No. 2003AA115520 & 2005AA114130.

REFERENCES

[1] Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature music/speech discriminator", Proc. of ICASSP97, vol. II, pp.1331-1334, April 1997.

[2] T. Zhang, C.-C J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification", IEEE Transactions on Speech and Audio Processing, 3(4), 2001.

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

[4] Z. Liu, J. Huang, Y. Wang and T. Chen, "Audio feature extraction and analysis for scene classification", IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing, 1997.

[5] Lie Lu, Hong-Jiang Zhang, Stan Li, "Content-based audio classification and segmentation by using support vector machines", ACM Multimedia Systems Journal 8 (6), pp. 482-492, March, 2003.

[6] Tobias Andersson, "Audio classification and content description", Audio Processing & Transport Multimedia Technologies Ericsson Research, Corporate Unit Lulea, Sweden, March, 2004.

[7] HSU C-W, LIN C-J, "A comparison of methods for multiclass support vector machines", IEEE Trans on Neural Networks, 2002

[8] Chang, C.-C and C.-J. Lin, "LIBSVM: a library for support vector machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm