Audio Content Classification Method Research Based on Two-step Strategy

Sumei Liang
Department of Computer Science and Technology
Chongqing University of Posts and Telecommunication
Chongqing, China

Xinhua Fan
Department of Computer Science and Technology
Chongqing University of Posts and Telecommunication
Chongqing, China

***Abstract*—Audio content classification is an interesting and significant issue. Audio classification technique has two basic parts: audio feature extraction and classifier. In general the audio content classification method is firstly to identify the original audio into text, then use the identified text to classify. But the text recognition rate is not high, some words that good for classification are identified by mistake causing that the classification effect is not ideal. In order to solve these problems above, this paper proposes a new effective audio classification method based on two-step strategy. In the first step the features are extracted by using the improved mutual information and classified with Naïve Bayes classifier. After classification of the first step, an unreliable area is determined, and samples with features in this area go on to be classified with the second step. In the second step, textual features extracted with CHI statistic method are used to build a text feature space model. Then audio features containing MFCC and frame energy are combined together with the text features to build a new feature vector space model. Finally, the new feature vector space model is classified using Support Vector Machine (SVM) classifier. The experiments show that the two-step strategy classification method for audio classification achieves great classification performance with the accuracy rate of 97.2%.**

***Keywords*—Two-step Strategy; Audio classification; MFCC; Frame energy; Naive Byes; Support vector machine (SVM)**

I. INTRODUCTION

With the high-speed development of information industry, the digital information grows rapidly. People have urgent demand on the process of digital information. Images, video and audio are the main forms of media in the field of information processing, and audio occupies very important position. How to quickly grasp the most effective information is an important problem people have to be faced with. Because the audio classification can solve the problems of information clutter to a certain extent and it is convenient for users to accurately locate the required information, it has become a key practical technology. For example, in the telephone booking and mobile service hotline, proprietor can evaluate employee's job performance according to the employee's contents, attitude, tone in the phone etc. At the same time it plays an important role in speech retrieval and the depth of the speech information processing with a broad application prospect. Typical audio classifiers used in the related papers contain Minimum Distance method, Support Vector Machine (SVM), neural network, Decision Tree method and the Hidden Markov Model[1][2][3][4] etc.

Currently audio content classification research mainly has two directions: one method is firstly to recognize the audio into text, and then classify the text after the identification; the other method is directly to use audio features, such as MFCC, frame energy, and pitch frequency and so on to classify audio. However using the text recognized form the audio information to classify in the first method has some problems in the following.

- The recognition rate of the first method is not high.
- Some words that contribution to the classification are recognized by mistake.
- The classification method is usually in single step classification strategy.

These problems cause that the classification effect is not ideal. Fan XingHua[5] etc proposed a new highly effective two-step Chinese text classification method; it achieved good effect in Chinese text classification research. This paper considers importing this two-step classification strategy into audio classification, and verifying the actual classification effect through experiment.

It is focused on applying the two-step classification strategy in audio classification in this paper. There are several problems needed to be studied.

- Whether the texts after identification are the same as plain texts with the phenomena that most misclassified ones are in a fuzzy region constituted in the two-dimensional space structure.
- Whether audio features such as MFCC, frame power could be effective for audio content classification.
- Text features of texts identified from audio are fewer than features of the plain texts, and especially some words with excellent contribution to the classification are recognized by mistake. These problems cause the classification effect in the first step is not ideal. Whether the combine of text features and audio features such as MFCC and frame energy could enhance the accuracy of audio content classification in the second step.
- Whether this two-step classification strategy for audio classification is feasible.

To aim of solving the above problems, we proposed a new audio classification method based on two-step strategy

combining text features and audio features. At last the method is studied with Naive Bayes and Support Vector Machine classifiers. The basic fundamental is as follows: in the first step improved mutual information is used to extract the characteristics, and the Naïve Bayes classifier is used for classifying. If the classification result of the first step is reliable the classification decision will be given, otherwise the samples will go to the second step. In the second step it combines audio features MFCC, frame energy with text features selected with CHI statistic formula as the total classify features, then uses Support Vector Machine classification method to classify. In the end, the final classification judgment is given according to the results of the two steps.

II. Characteristic Analysis In The Audio Content Classification

According to the analysis of sports audio and news audio, it is found that usually there are some special unsteady sounds in the sports audio such as whistle, sonorous voice of narrator and cheers etc. Oppositely the news audio usually is relatively stable, and the kinds of sounds above don't exist. These voices contain rich semantic information and can be very useful to distinguish these two types of audio. Through the experiments, it is proved that audio features MFCC, frame energy under different audio category have obvious distinguish ability.

A. Frame Energy

Short-time Energy[6][7] is the energy focused by the sample signal in a short audio frame. Sequence of Short-time energy reflects the detailed change rule of the voice amplitude or energy over time.

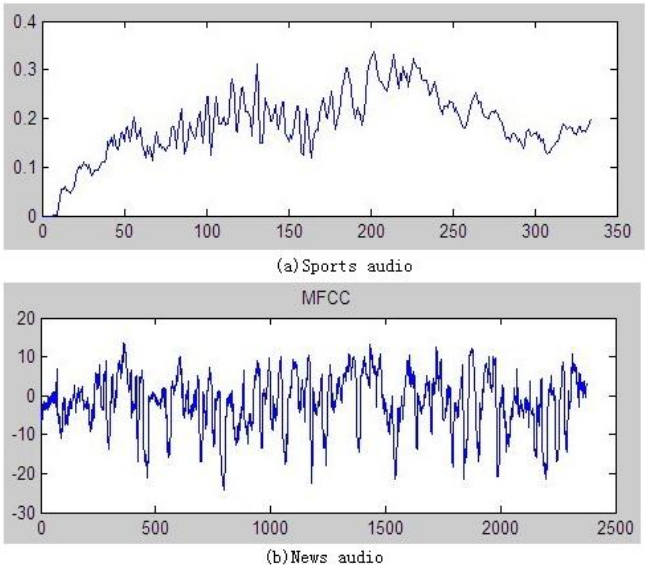

Fig. 1. the distribution curve of frame energy

Figure1 (a) and (b) respectively describe the short-time energy distribution curve of these two types of audio. The Energy of sports audio generally is high, and the amplitude usually changes slowly. On the other hand the energy distribution of the news audio is relatively concentrated, and amplitude changes quickly. Actually there are a lot of sounds such as whistle, sonorous voice of narrator and cheers in the sports audio, but news audio rarely appears this kind of sounds through analysis. Because of these differences theirs energy distribution curves are obviously different between with each other. According to the above analysis Short-time Energy can be considered as the feature used to classify the audio.

B. Mel Frequency Cepstrum Coefficient

Mel frequency cepstrum coefficient (MFCC) [8][9] is the cepstrum parameters extracted from the Mel scale frequency domain, and also a kind of perception frequency cepstrum parameters. In order to accord with the human's auditory characteristics, MFCC generally adopts the triangular filter group to filter the energy coefficient of Fourier Transform, and do the Mel scale transformation for frequency domain. MFCC coefficient is firstly used for speech or speaker recognition, but the results of literature [10][11] show that MFCC coefficient can improve the accuracy of audio content classification

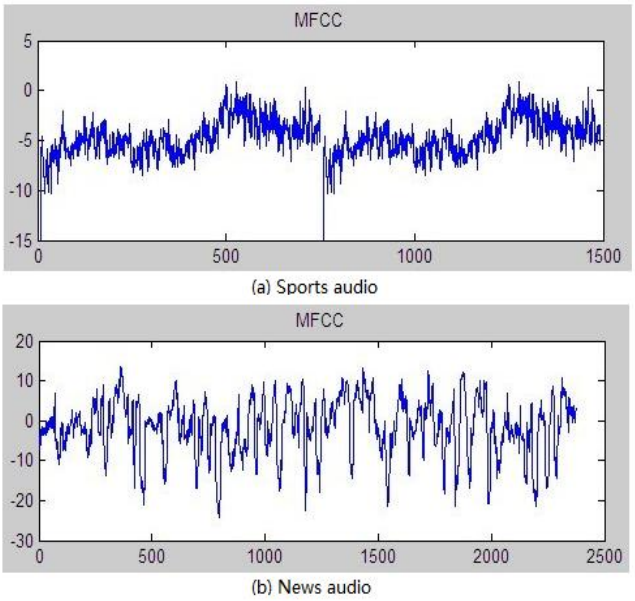

Fig. 2. one-dimensional map of MFCC feature

Lu Jian [12] puts forward a kind of audio classification method based on Hidden Markov Model. That paper pointed out that the difference of MFCC coefficient Δ MFCC could well reflect the dynamic change characteristics of audio signal with the calculation and analysis of multi-stage MFCC and the difference coefficient ΔMFCC, so they can be used to revealed the time statistical properties of different types of audio.

Figure2 (a) and (b) respectively represent the short MFCC feature of sports audio and news audio in one dimensional mapping. Through the compare of these two pictures it can be found that the MFCC feature mapping values of sports audio jump densely in a short local area and amplitude is small. Oppositely MFCC feature mapping values of news audio jump sparsely densely in a short local and the amplitude is much larger. These differences prove that the MFCC features can reflect the rich semantic characteristics, and the most important of all they have good distinction between two types of audio category.

III. Audio Classification Method Based On Two-Step Strategy

A. The Rewriting for two types of Naïve Bayes Classifier

Given a binary text vector d=$(w_1, w_2, \cdots, w_D)$, $w_i = 0$ or 1. If the ith feature appears in the text w_i =1, otherwise $w_i = 0$, $P_{ki} = P(w_k = 1 | c_i)$, $P(\cdot)$ means the probability of event $(\cdot)$. The discrimination function for two types of Naïve Bayes classifier can be expressed as follows:

$$f(d) = log\frac{P(c_1|d)}{P(c_2|d)} = log\frac{P(c_1)}{P(c_2)} + \sum_{k=1}^{|D|} log\frac{1-P_{k1}}{1-P_{k2}} + \sum_{k=1}^{|D|} W_K log\frac{P_{k1}}{1-P_{k1}} - \sum_{k=1}^{|D|} W_K log\frac{P_{k2}}{1-P_{k2}} \quad (1)$$

When $f(d) \geq 0$, text d belongs to type c_1. Otherwise it belongs to type c_2.

B. The design of the Support Vector Machine SVM Classifier

The principle of Support Vector Machine (Support Vector Machine) [13] can be simply described as follows: it hopes to seek a hyper plane which can separate positive samples from negative samples in the training set with the largest blank space on either side. It is given a set of training samples as follows.

$$T = \{(x_i, y_i)\} \in (R^n \times Y)^l$$

s.t $x_i \in R^n$, $i = 1, ...l$

If $y_i \in Y = \{1, -1\}$, SVM becomes the process of constructing hyperplane $(w.x) + b = 0$ which separates the two types of sample points. Among them, the distance from the nearest point in the samples to the hyperplane is called interval, as shown in the Figure 3.

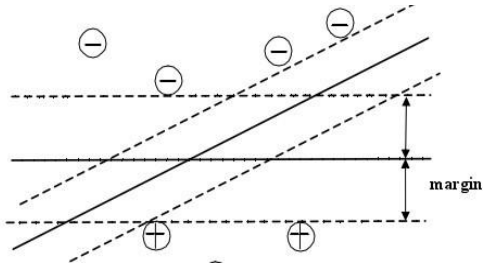

Fig. 3. SVM schematic diagram

C. The observations of misclassification samples in the first step

In this two-step strategy classification method, in the first the Hidden Markov Model is used for speech recognition in which the audio samples are recognized to the texts. The improved mutual information formula is used for feature selection of identified texts in the first step in order to get the Binary text vector, then the Bayes classifier in part A is used to classify. In order to study and analysis the result of classification in the first step, the formula (1) is took apart and two posterior probability parameters representing the probability where one sample belongs to one of two types are respectively fetched out as follows.

$$X = \sum_{k=1}^{|D|} W_K log\frac{P_{k1}}{1-P_{k1}} \quad (2)$$

$$Y = \sum_{k=1}^{|D|} W_K log\frac{P_{k2}}{1-P_{k2}} \quad (3)$$

$$con = log\frac{P(c_1)}{P(c_2)} + \sum_{k=1}^{|D|} log\frac{1-P_{k1}}{1-P_{k2}} \quad (4)$$

X represents posterior probability where the text d belongs to type c_1, and Y represents posterior probability where the text d belongs to type c_2. Con is a constant only related to the training samples set, and would not be changed by text d. So the formula (1) can be rewritten as follows.

$$f(d) = X - Y + con \quad (5)$$

Formula (5) represents that the two-step Naïve Bayes classifier can be viewed as a process of seeking a straight line $f(d) = 0$ in two-dimensional space constituted by X and Y. In this way, sample text can be expressed as a single point (x, y) in the two-dimensional space determined by formula (2) and formula (3), the distance *Dist* from this point to dividing line $f(d) = 0$ is as follows.

$$Dist = \frac{1}{\sqrt{2}}(x - y + con) \quad (6)$$

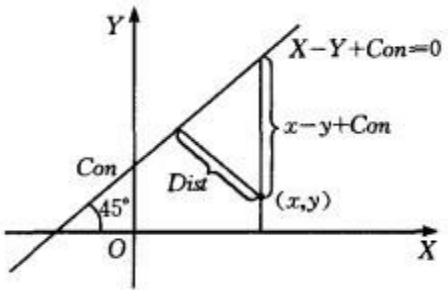

Fig. 4. the distance from the text point to the dividing line.

As shown in Figure 4, sample text d belongs to type c_1 when $Dist \geq 0$, sample text d belongs to type c_2 when $Dist < 0$.

The purposes that the formula (1) is changed to the formula (5), and then evolved into formula (1) are as follows.

- Taking advantage of formula (6) can easily investigate and analyze text classification error, and discuss the relationship between the distance *Dist* and the

classification error in the condition of a given classification method and textual features set in the two-dimensional space made up of X and Y.

- It is convenient to assess the relationship of reliability of classification and the value of the distance *Dist* and determine the unreliable part of the classification results in the first step by taking advantage of formula (6).

In this paper C_1 and C_2 respectively represent for sports category and news category. Using the corpus in the experimental section as samples set, distribution of the text after identification can be calculated in the two-dimensional space with X as the abscissa value and Y as the ordinate value as shown in Figure 5. Figure 5 (a) and Figure 5 (b) respectively corresponding to the distribution of sports audio and the news audio. It can be seen from the figure that two types of audio are distributed on two sides of the dividing line in two-dimensional space. The texts after identification classified by mistake are located in the above area of dividing line in Figure 5 (a) and in the below area of dividing line in Figure 5 (b). Through observing of these texts it is clear that the samples classified by mistake mainly concentrate in the area very close to the dividing line.

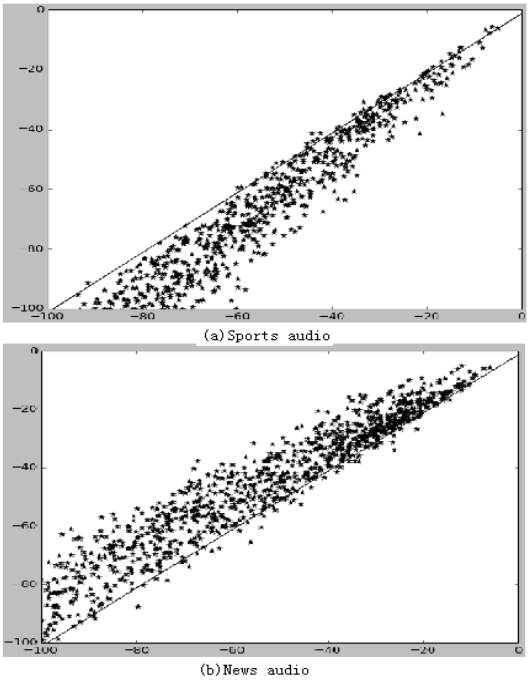

Fig. 5. the distribution of identified text

Fan XingHua [5] puts forward most of the errors are in a narrow area close to the dividing line in the plain text classification study. That is to say, if the texts whose distance are very close to the dividing line are got rid of, then the test performance of the classifier will be improved. This assumption has been proven by experiments. This paper imports this idea to the first step of classification process in which the text after speech recognition is classified. According to the observation of the Figure 5, it is assumed that the performance of the classifier is relate to the distance *Dist* calculated by formula (6), and major errors happen in the narrow area close to the dividing line.

D. The method of determining the unreliable area after classification of the first step

Through the observation of classified result in the first step of classification process it can be seen that the samples are classified by mistake mainly concentrated on the area close to the dividing line. A range where *Dist* values are near to zero can be determined as the unreliable area in the first step of classification process, and then decision whether reliable or unreliable can be made. Formula (7) is the discriminant formula for classification.

$$\begin{cases} Dist1 \le dist \le Dist2 & \text{the classification is unreliable} \\ dist \text{ for other values} & \text{the classification is reliable} \end{cases} \quad (7)$$

In order to get the most optimal boundary constant $Dist1$, $Dist2$, two evaluation indexes are introduced: error rate and area percentage.

Error Rate:

$$ER(dist1, dist2) = \frac{EC(dist_1, dist_2)}{EC\_CON} \times 100\%$$

Area Percentage:

$$RP(dist1, dist2) = \frac{T(dist_1, dist_2)}{T\_CON} \times 100\%$$

$EC(dist_1, dist_2)$ is the count of the samples that are classified by mistake with the Dist value in the range of $[dist1, dist2]$. EC_CON is the total number of samples that are classified by mistake. $T(dist_1, dist_2)$ is the count of samples with the *Dist* value in the range of $[dist1, dist2]$. T_CON is the total number of all test samples.

In order to conveniently draw the curve lines that show changes of ER and RP following *Dist* range, one of two endpoints could be fixed to a constant value. Using the corpus in the experimental section as samples, Figure 6 shows the curves of ER and RP with the fixed constant value of zero. The curve lines on the right side of Y axis reflect the changes of ER and RP following $dist2$ value when the value of $dist1$ is zero, and similarly curve lines in the other side Y axis reflect the changes of ER and RP following $dist1$ value when the value of $dist2$ is zero. Through observation of Figure 6 can be seen that the ER grows rapidly when the value of $dist2$ is small, and then gradually stabilizes form the inflection point. The situation in the left of Y axis is the same as the right of Y axis. The value of ER in position of right inflection point with the $dist2$ value of 5 is 49% meaning 49% samples(sports samples) classified by mistake are concentrated on the range of [0,5], and RP is 15% meaning the sports samples whose *Dist* values are in the range of [0,5] account for 15% of the total test samples. The inflection point on the other side is the position with $dist2$ value of - 4, where the ER and RP values are 47% and 19%

respectively. It means that 47% samples (news samples) classified by mistake are concentrated on the range of [-4, 0] where samples account for 19% of the total test sample. To sum up, the 97% part of samples wrongly classified distribute on the scope of the [-4, 5], and at the same time samples whose *Dist* value are in this area of only account for 34% of all samples. Therefore, the two inflection points' values are fairly appropriate to be as the endpoint value of unreliable area for formula (7).

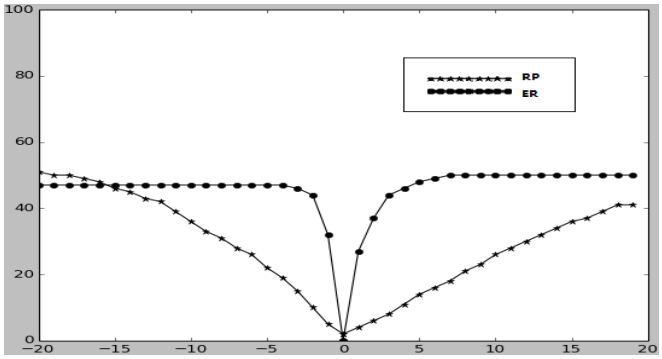

Fig. 6. ER and RP curve changed with distance Dist

E. The characteristics of the two-step classification method

The most significant value of two-step classification method put forward in this paper is able to mix text and audio feathers together for classifying. Generally the traditional method firstly recognizes the audio into text, does feature extraction and chooses an appropriate classifier to classify. But the recognition rate is not high and some words that well contribute to the classification are recognized by mistake or recognized to other homophones, these problems cause that the classification effect is not ideal. According to the analysis of two kinds of audio, it is found that there are usually some special unsteady sounds in the sports audio such as whistle, sonorous voice of narrator and cheers etc. Oppositely the news audio is usually relatively stable, and these kinds of sounds don't exist in general. Through the experiment, it is proved that the audio features containing MFCC and frame energy have good ability to distinguish raw audio differences above. Two-step strategy classification method in this paper effectively combines audio and text features with each other in order to get a high performance classification.

IV. EXPERIMENT

A. The Experiment Database

The audio corpuses used in the experiment are all from the real environment context. Sports corpus is mainly from CCTV live program basketball and soccer sports events and news corpus are from CCTV news broadcast program. There are 1050 sports audio samples and 1200 news audio samples. The proportion of training samples and testing samples is 3:1 on the experiment. In order to ensure the uniformity of training samples and testing samples, the training sample and test samples are obtained in a cross way. Sampling frequency is 16 KHZ, and capacity of data is 491 MB.

B. Evaluation index

The accuracy performance of classification can be measured in the following index: the classification accuracy of one category and the average classification accuracy. The definitions are as follows:

$$accuracy\ of\ C_i = \frac{count\ of\ C_i\ samples\ predicted\ to\ be\ C_i}{count\ of\ samples\ predicted\ to\ be\ Ci}$$

$$average\ accuracy = \frac{\sum count\ of\ C_i\ samples\ predicted\ to\ be\ C_i}{the\ total\ number\ of\ test\ samples}$$

C. Experiment steps of Two-step classification method

- Design and implement the continuous Chinese speech recognition system based on HTK. The speech recognition rate is 78.04% in this paper's experiment.
- Complete feature extraction using the enhanced mutual information formula for the text after identification.
- Use the improved Naïve Bayes classifier for the first step.
- Analyze and observe the experiment result in the first step, and get a good boundary of unreliable area.
- According to the boundary of the unreliable area, determine a fuzzy region.
- Use the SVM classifier to classify the samples in the fuzzy region in the second step.
- Make the final classification decision.

D. Analysis of experimental results

This paper has done the classification experiments by five kinds of methods:

- Method 1: extract text features using the improved mutual information formula, and use Naïve Bayes classifier to classify the audio.
- Method 2: use audio features containing MFCC and frame energy to construct vector space model, and use the SVM classifier to classify the audio.
- Method 3: use CHI statistic formula to extract text features, and combine two kinds of features including audio features containing MFCC、frame energy and text features to construct vector space model. At last use the SVM classifier to classify the audio.
- Method 4: Use Two-step strategy method combining method 1 with method 2 to classify the audio.
- Method 5: Use Two-step strategy method combining method 1 with method 3 to classify the audio.

TABLE I. THE CLASSIFICATION RESULTS

Classification Method	Sports	News	Average accuracy rate
Method 1	90.77%	88.78%	89.72%
Method 2	81.15%	85.96%	83.84%
Method 3	95.65%	93.23%	94.16%
Method 4	96.75%	95.32%	95.79%
Method 5	97.76%	96.62%	97.2%

The table 1 shows the results of five kinds of classification methods. Through analysis of method 2's experiment results it can be seen that the performance of the method using only audio features to classify is not good. The contrast of results of method 2 and method 3 shows that combining two kinds of features including text features and audio features with each other simply can well improve the classification effect.

According to the compare of classification effects of method 4 with method 1 and method 2, it can be seen that the addition of audio features improves the problem that the accuracy rate is not high because of the inaccuracy in text information identification. Obviously it comes to a very import conclusion that the classification accuracy rate of Method 4 based on two-step strategy has great classification accuracy promotion. This conclusion applies to the contrast method 5 with method 1 and method 3 at the same time. What's more, the difference between Method 5 and Method 4 is that Method 5 imports text features again in the second step of classification process in order to get a larger number of classification features in the second step. The contrast of results of method 4 and method 5 shows that importing the text features for samples in the unreliable area of the first step can achieve a better correction effect.

V. CONCLUSION

As a main form of media audio plays an import role in the field of information processing. Audio classification has become a hot practical technology with a wide application prospect in the fields of speech retrieval, deep voice information processing. In general the audio content classification method is firstly to identify the original audio into text, then use the identified text to classify. But the text recognition rate is not high, some words that are good for classification are identified by mistake causing that the classification effect is not ideal. This paper provides a new effective audio content classification method based on two-step strategy.

The basic fundamental is as follows: in the first step improved mutual information is used to extract the characteristics, and the Naïve Bayes classifier is used for classifying. If the classification result of the first step is reliable the classification decision will be given, otherwise the samples go to the second step. In the second step it combines audio features MFCC, frame energy with text features selected with CHI statistic formula as the total classify features, then uses Support Vector Machine classification method to classify. Through the experiments, it comes to a conclusion that the audio content classification method based on two-step strategy in this paper is effective in enhancing the performance of audio content classifying, and it can achieve the great classification performance with the classification accuracy rate of 97.2%.

REFERENCES

[1] K.Subashini,S.Palanivel,V.Ramaligam,"Audio-video based segmentation and classification using SVM," 2012 Third International Conference on Computing Communication & Networking Technologies (ICCCNT),vol., no., pp.1-6, 26-28 July 2012

[2] T.Giannakopoulos, D.I.Kosmopoulos, A.Aristidou,and S. Theodoridis, "A multi-class audio classification method with respect to violent content in movies using bayesian networks," in IEEE Workshop on MSP, pp.90-90, 2007.

[3] Kiranyaz.S, Ahmad Farooq Qureshi, Gabbouj.M, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," IEEE Transactions on Audio, Speech and Language Processing, vol.14(3), pp.1062-1081, 2006.

[4] B.Liang, SY.Lao, HX.Liao, JY. Chen, "Audio Classification and Segmentation for Sports Video Structure Extraction using Support Vector Machine," 2006 International Conference on Machine Learning and Cybernetics, pp.3303-3307, 2006.

[5] XH. Fan, and MS. Sun, "A high performance two-class chinese text categorization method," Journal of Computers, China, vol. 29(1), pp. 124-131, 2006.

[6] Umapathy,K, Krishnan,S, Rao,R.k. "Audio Signal Features Extraction and Classification Using Local Discriminant Bases," IEEE transactions on audio, speech and language processing, vol.15(4) , pp.1236-1246, 2007.

[7] U.Shrawankar, V.Thakare, "Feature Extraction for a Speech Recognition System in Noisy Environment: A Study", ICCEA, pp.358-361, 2010.

[8] S. Jothilaskmi, S. Palanivel, V. Ramalingam, "Unsupervised speaker segmentation with residual phase and MFCC features," Expert System With Applications, India, pp. 9799-9804, 2009

[9] H. Cao, C. Xu, X. Zhao, SJ. Wu. "The Mel-frequency cepstral coefficients in speaker recognition," XiAn: Journal of northwest university (natural science edition), vol.43, , pp.203-208, 2013

[10] SahamiM, Dumais S, Hecheman D, Horvitz E. "A Bayesian approach to filtering junk E-mail," Madison Wisconsin AAAI Technical Report WS-98-05, pp.55-62. 1998

[11] Li, S.Z. "Content- Based classification and retrieval of audio using the nearest feature line method," IEEE Transactions on Speech and Audio Processing, vol.8(5), pp.619~625. 2000

[12] J. Lu, YS. Chen, ZS. Sun, FY. Zhang. "Automatic audio classification based on Hidden Markov Model," BeiJing: Journal of software, vol.13(8), pp.1593-1597, 2002.

[13] Subashini. K, Palanivel.S, Ramaligam.V, "Audio-video based segmentation and classification using SVM," 2012 Third International Conference onComputing Communication & Networking Technologies(ICCCNT), pp.1-6,26-28, .2012.