Multimedia Systems (2011) 17:101–112
DOI 10.1007/s00530-010-0205-x

ORIGINAL RESEARCH

Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news

Lei Xie · Zhong-Hua Fu · Wei Feng ·
Yong Luo

Received: 1 April 2010 / Accepted: 10 September 2010 / Published online: 28 September 2010

Abstract Audio classification is an essential task in multimedia content analysis, which is a prerequisite to a variety of tasks such as segmentation, indexing and retrieval. This paper describes our study on multi-class audio classification on broadcast news, a popular multimedia repository with rich audio types. Motivated by the tonal regulations of music, we propose two pitch-density-based features, namely average pitch-density (APD) and relative tonal power density (RTPD). We use an SVM binary tree (SVM-BT) to hierarchically classify an audio clip into five classes: pure speech, music, environment sound, speech with music and speech with environment sound. Since SVM is a binary classifier, we use the SVM-BT architecture to realize coarse-to-fine multi-class classification with high accuracy and efficiency. Experiments show that the proposed one-dimensional APD and RTPD features are able to achieve comparable accuracy with popular high-dimensional features in speech/music discrimination, and the SVM-BT approach demonstrates superior performance in multi-class audio classification. With the help of the pitch-density-based features, we can achieve a high average accuracy of 94.2% in the five-class audio classification task.

Communicated by T. Haenselmann.

L. Xie (✉) · Z.-H. Fu · Y. Luo
Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China
e-mail: lxie@nwpu.edu.cn

W. Feng
Media Computing Group, School of Creative Media, City University of Hong Kong, Kowloon, Hong Kong, China

Keywords Audio classification · Multimedia content analysis · Audio diarization · Support vector machines

1 Introduction

Multimedia contents are proliferating exponentially due to the increasing availability of media broadcasting and sharing, free digital storage, ubiquitous computing devices and constant Internet connectivity. Audio is an indispensable component of multimedia repositories, e.g., TV and radio broadcasting, music collections, historical archives, movies, home videos, lectures and meeting records. A fundamental step for semantic access to the media content is to automatically classify or divide an audio stream into homogenous segments, e.g., speech and music [1, 35]. Such a classification can facilitate effective searching and browsing of multimedia data [19].

This article describes our audio classification work in broadcast news, a major daily mass media with *rich* audio types in the media explosion era. Audio classification is not an easy task for audio streams such as broadcast news, containing not only single type classes (e.g., speech and music) but also mixed type of classes (e.g., speech with music, speech with environmental sound). Figure 1 shows a colored waveform of a broadcast news audio stream, where different audio classes are plotted with different colors. At least six audio (and mixed audio) types can be found in this audio segment:

- *music*: commercials, program titles, news about a concert, topic intervals and program closings.
- *pure speech*: anchor reports in a studio, post-production narratives, field speech from reporters in a clean environment.

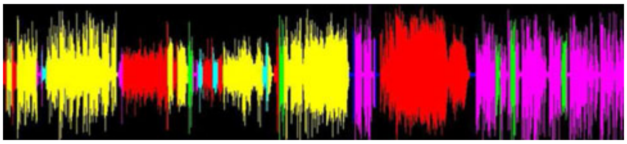

Fig. 1 Waveform of a broadcast news audio stream with various audio types (plotted in different *colors*)

- *speech with environment sound*: field speech or interviews in a noisy condition.
- *speech with music*: commercials, program titles and program closings.
- *environment sound*: field sounds from machines, crowds, birds, water, wind and applause.
- *silence*: speaker prosodic pauses, pause intervals between different speakers and news stories.

An effective audio classification scheme is quite useful for a diversity of subsequent tasks, such as:

- *Broadcast news transcription*. Filtering out non-speech segments and training acoustic models for pure speech and non-pure speech in order to improve speech recognition accuracy [36].
- *Multimedia fission*. Segmentation of continuous audio/video streams into self-coherent units or constituents. For example, speech/music shifts are effective acoustic cues for story/topic segmentation [33]; change of pure and non-pure speech is useful for shot boundary detection in a video stream.
- *Content-based indexing and retrieval*. Segmenting an audio stream into meaningful descriptions will support querying desired audio segments, e.g., music or speech clips [35].
- *Audio coding*. Different audio signals deserve different coding/compression schemes. By automatically determining the audio types, appropriate coding techniques can be applied. Also, identifying the silence segments can decrease the coding bit rate.

Audio classification has been an active research area in recent years. Plenty of previous studies address the audio classification task in two aspects, i.e., feature extraction and classification scheme. Most work focuses on the binary speech/music discrimination task and some researchers study on classification of mixed (or rich) types of audio. Features for classification can be time-domain descriptions or frequency-domain representations of audio signals. In general, features can be categorized into four groups:

- *Energy features*, such as root mean square (RMS) of short time energy [25], 4 Hz modulation energy [5, 27], percentage of "low-energy" frames [5, 27, 30], silence frame ratio (SFR) [16], noise frame ratio (NFR) [22] and subband energy distribution (SED) [23], etc.;
- *Statistical spectral features*, such as the roll-off or the centroid point of power spectrum [5, 21, 27], spectral flux [5, 21, 22, 23, 27, 29], spectral kurtosis [21] and zero-crossing rate and ratio [5, 21, 25, 27, 29,], etc.;
- *Spectral envelope features*, such as Mel frequency cepstral coefficients (MFCC) [5, 19, 21, 22, 23], mean of minimum cepstral distances (MMCD) [9], linear predictive cepstral coefficients (LPCC) [19, 21], linear spectral pairs (LSP) [19, 21] and power spectrum deviation (PSDev) [17], etc.;
- *Pitch features*, such as pitch and delta pitch [5, 17, 29], spectral peak duration [17, 29] and pitch tuning [37].

Discriminative abilities are quite different between various features and comparisons between a wide range of features can be found in [5, 24, 27]. Since features are often complementary, audio classification approaches usually make use of a combination of different features [18, 21, 23]. In our previous approach [21], we used a 94-dimensional feature vector including LPCC, LSP, spectral centroid, flux, rolloff, kurtosis and zero-crossing rate in classification of speech and music in Mandarin news broadcasts.

Despite of a variety of proposed features, most of them describe the speech/music difference solely from the speech point of view, e.g., limited bandwidth and alternative voiced/unvoiced fragments. As we know, music has well-defined systems, e.g., tones system, and each music composition should strictly follow the systematic regulations. As a result, these regulations reflect in the musical signals. With the help of the musical characteristics, we believe that the discrimination performance can be further pushed forward.

Another issue in audio classification is the selection of the classification scheme. The most popular classifiers include Gaussian likelihood ratio (GLR) [27], Bayesian information criterion (BIC) [24], Bayesian MAP classifier [30], Gaussian mixture model (GMM) [22, 27], hidden Markov model (HMM) [18], multi-layer perceptron (MLP) [18, 21] and other neutral networks [18], *K*-nearest neighbor (KNN) [21, 23] and support vector machines (SVM) [6, 16, 20, 21, 23]. Pikrakis et al. [26] proposed a multi-stage speech/music discriminator based on dynamic programming and Bayesian networks with a high accuracy for speech/music discrimination in radio recordings. Wu et al. [32] used a combination of data mining method with decision trees for speech/music classification. Evaluations on different audio classifiers can be found in [18, 21, 22].

SVMs have recently drawn much interest for audio classification due to their superior performances in various pattern classification tasks [28]. We studied music/speech classification in news broadcasts and discovered that SVM showed higher accuracy as compared with KNN and MLP

[21]. In IBM's approach [20], SVM was used for audio classification in instructional video analysis and the approach outperformed decision tree and threshold-based approaches. Lu et al. also employed SVMs in their work [23], which hierarchically classified audio signals into pure speech, non-pure speech, music and background sound.

Although the classification of audio data into single type, e.g., music, speech, environmental sound, is well studied, classification of mixed type audio such as clips having speech with musical background, is still considered as a challenging problem [6]. The use of SVM for multi-class, mixed-type audio classification deserves further investigation. The main reason is that SVM is originally designed for a binary classification problem and discriminating multiple classes is realized by either combining several binary classifiers or modification of a single SVM that considers all classes at once [7]. Therefore, an effective classification scheme is essential to the performance of SVM-based multi-class audio classification.

In this paper, we study multi-class audio classification in broadcast news from both the feature extraction and the classification points of view. The main contributions of this work are as follows. We propose two pitch-density-based features, namely average pitch-density (APD) and relative tonal power density (RTPD). Previous approaches mainly focus on speech characteristics while the two proposed features reflect the tonal regularities of music signals. We show that the two one-dimensional features are able to achieve comparable performance with high dimensional features such as MFCC and even outperform other competitive features in classification accuracy of music.

We use an SVM binary tree (SVM-BT) approach to solve the multi-class audio classification problem, which employs a coarse-to-fine stepwise classification strategy that allows us to first discriminate on broad audio classes easy to differentiate. Lu et al. [23] also used an SVM tree to hierarchically classify audio into four classes. In contrast, five classes are considered in our study where non-pure speech are further classified into two mixed type audio classes, i.e., speech with music and speech with environment sound. This fine-gained classification is useful to broadcast news content analysis, where speech with music clips are mainly from commercials and speech with environment sound clips are generally field speech in news stories. Specifically, considering the feature discrepancy in discriminative ability, we use a feature selection process to select an optimal feature subset from a general feature pool for each binary SVM on the SVM binary tree. We also evaluate two SVM-BT schemes, i.e., discriminating with-speech and without-speech and discriminating with-music and without-music at the top level of the SVM-BTs. Experiments show the superior performance of the SVM-BT approach in multi-class audio classification.

The rest of this article is organized as follows. Section 2 describes the new pitch-density-based features. Section 3 presents the SVM binary tree approach to multi-class audio classification. In Sect. 4, experimental evaluations are reported. Finally, we summarize the article in Sect. 5.

2 Pitch-density-based features

Conventional speech/music classification approaches mainly focus on speech characteristics, e.g., limited bandwidth and alternative voiced/unvoiced fragments, etc. As we know, music is a human art that has a more precise theoretical foundation, i.e., the *tones system*. Therefore, we propose two tone-related features motivated by the fact that music signals have distinct pitch characteristics that can discriminate them from speech signals.

2.1 Pitch characteristics in music and speech

Music types or genres are extremely manifold, e.g., classical, rap, jazz, blues and hip hop, etc. However, diverse music signals can be roughly categorized into *tone-like* and *noise-like*, according to melody and rhythm. Notes are the atoms of tone-like music, which correspond to musical tones of audio. Tone-like music is played by musical instruments, e.g., piano, violin, guitar and organ, etc. The tones, or pitches, have precise definition in the musical system. For example, standard pitch or concert pitch is defined as $A = 440$ Hz ($A440$) for musical note central *la*. Equal temperament is a system of tuning in which every pair of adjacent notes has an identical frequency ratio, usually the octave. Each *octave* contains 12 semitones, including 7 white keys and 5 black ones on a piano keyboard. Each transition to go up one octave corresponds to twice the frequency. For instance, $A5$ is 880 Hz (81 semitones), which is twice of $A4$ (central *la*, 440 Hz). Tone-like music should follow these tonal regulations. This tone pattern is consistent regardless of types of music or instruments.

Another broad music category, namely noise-like music, is usually played by percussion instruments, e.g., drum, cymbal and maracas, etc. These instruments perform noise-like sounds in short time slice and cannot play clear notes. Therefore, noise-like music signals do not have salient pitch patterns. However, percussion instruments present energy pulses with strong power and thus are often used for a strong rhythm in music performance.

Due to the regularities of the tones system, the pitches of music signals usually remain relative constant for a moment and can only jump between discrete frequencies, except for vibratos or glissandi. But this phenomenon seldom occurs in human speech. The pitches (i.e., F_0) of speech signals change continuously and will not keep on a fixed frequency for a long time. Instead, speech exhibits an alternating sequence of tonal and noise-like segments. On the other hand, noise-like musical signals played by percussion instruments usually do not contain distinct pitches. Instead, they usually exhibit stronger powers. In speech signals, the strong power components are generally voiced speech, which conversely consist of distinct pitches. These facts in pitch have motivated us to propose new features for speech/music discrimination.

2.2 Real cepstrum

Motivated by the distinct pitch characteristics between music and speech signals, we propose two pitch-density-based features. However, we do not detect pitches directly because current pitch extraction methods still do not reach a desired level of accuracy and robustness [12]. Instead, we use real cepstrum to analyze the pitch information. A *cepstrum* (named by reversing the first four letters of spectrum) is the result of taking the Fourier transform (FT) of the log spectrum, which can been seen as the rate of change in the different spectrum bands [8]. Cepstrum is a powerful tool to show the detail of spectrum by separating the pitch information from the spectral envelope. As a power representation, the real cepstrum is defined as

$$\mathbf{rc}(n) \triangleq \mathrm{real}\left(\frac{1}{2}\int_{-\pi}^{\pi} \log(|\mathbf{X}_n(j\omega)|)e^{j\omega n}\mathrm{d}\omega\right), \qquad (1)$$

where $\mathbf{X}_n(j\omega)$ is the short-time Fourier transform of the nth windowed audio frame and real($\cdot$) denotes the real part of the complex cepstrum. Note that $\mathbf{rc}_x(n)$ is a vector that contains all real cepstral coefficients of the nth audio frame. The lower order coefficients of $\mathbf{rc}(n)$ refer to the large scale information of the spectrum like formants, and the higher order coefficients exhibit the detail information like pitches. Hence we use the higher order coefficients to discriminate music from speech. Figure 2 shows the difference between music and speech by means of the high-order coefficients of real cepstrum. We can clearly observe that the pitches of speech change continuously, as shown in Fig. 2b, and the pitches of music jump discretely and keep on certain frequencies for a moment, as shown in Fig. 2d. Figure 2f shows a segment of drum music, where the pitch curve is blurred especially when strong ictuses (i.e. beats) occur.

2.3 Pitch-density

In real-world audio, such as broadcast news, pitch extraction is often sensitive to noises, resulting in inaccurate pitch values and their holding lengths. Therefore, we use pitch-density (PD), which is based on real cepstrum, to

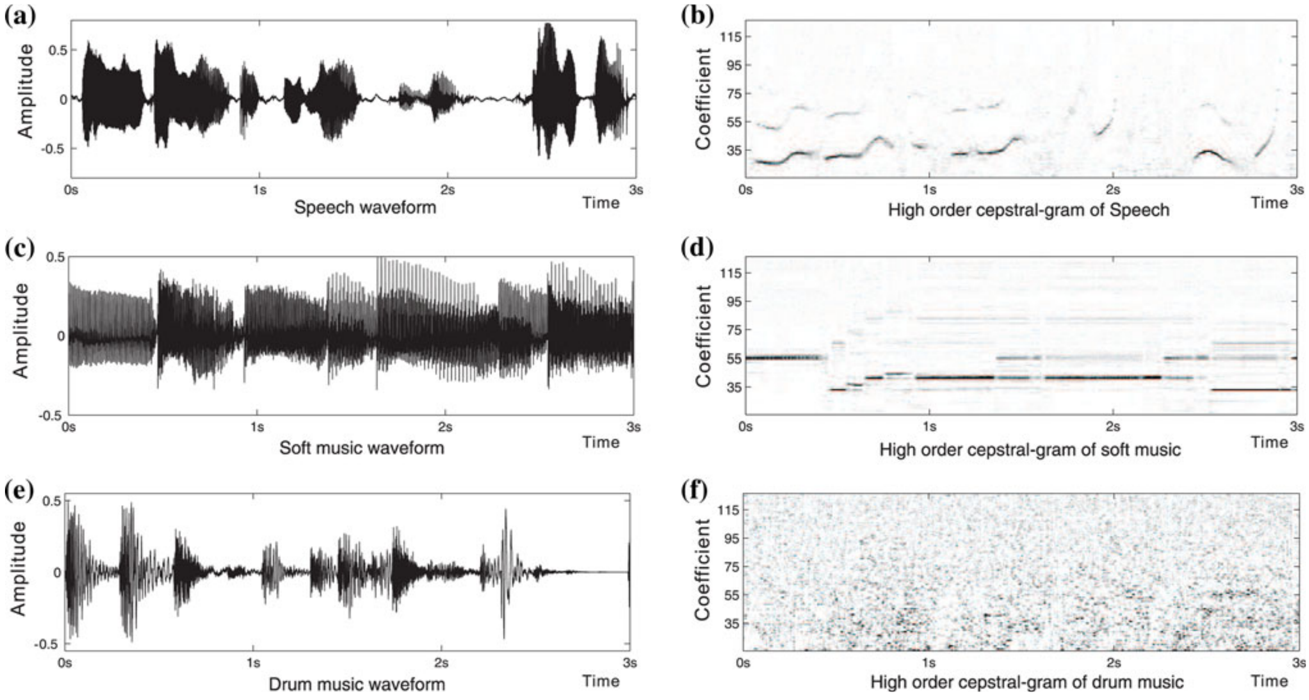

Fig. 2 Audio signals (*left column*) and their high-order cepstral-grams (*right column*). From *top* to *bottom*, the signals are speech, soft music and drum music, respectively

roughly characterize the pitch properties of music and speech signals. The PD is defined as

$$\mathrm{PD}(n) = \frac{1}{L}\sum_{m=l_1}^{l_2} |\mathbf{rc}(n,m)| \qquad (2)$$

with

$$L = l_2 - l_1 + 1, \qquad (3)$$

where $\mathbf{rc}_x(n, m)$ is the mth coefficient of $\mathbf{rc}(n)$. PD(n) represents the mean of absolute values of high-order real cepstral coefficients within the range of [l_1, l_2]. Our empirical study shows that simply averaging the overall high-order cepstrum content is effective in speech/music discrimination. For music signals, due to the characteristics of musical instruments and the existence of polyphony, the PD is tend to be higher than that of speech signals.

For an audio clip, average pitch-density (APD) can be calculated, which is defined as

$$\mathrm{APD}(k) = \frac{1}{N}\sum_{n=k\beta N+1}^{k\beta N+N} \mathrm{PD}(n) \qquad (4)$$

where N is the number of frames in an audio clip k, and β is the overlapping factor of each clip.

Figure 3 shows the relative frequency histograms of APD for music and speech. We can clearly see that music can be largely discriminated from speech by the 1-dimensional APD feature.

2.4 Relative tonal power density

For noise-like music, we consider its pitch and energy characteristics different from speech, as described in Sect. 2.1. First, for an audio clip, we mark each frame as a tonal-frame or a non-tonal-frame according to the following rule:

$$\max_{m\in[l_1,l_2]}(\mathbf{rc}(n,m))\begin{cases} > \theta & \text{tonal-frame;} \\ \le \theta & \text{non-tonal-frame.}\end{cases} \qquad (5)$$

That is, if the maximum value of high order real cepstrum is bigger than a predefined threshold, which indicates a salient peak in the high-order part, the frame is marked as a tonal-frame. Second, we compute the relative tonal power density (RTPD) between the tonal-frames and all frames in the audio clip, i.e.,

$$\mathrm{RTPD}(k) = \frac{\frac{1}{|\Theta_k|}\sum_{n\in\Theta_k}\mathrm{RMS}(n)}{\frac{1}{|\Psi_k|}\sum_{n\in\Psi_k}\mathrm{RMS}(n)} \qquad (6)$$

where Θ_k denotes the tonal frame set of the kth clip with the number of tonal frames $|\Theta_k|$, and Ψ_k denotes the overall frame set of the kth clip with the number of frames $|\Psi_k|$. RMS(n) is defined as the root mean square energy of the nth audio frame:

$$\mathrm{RMS}(n) = \sqrt{\sum_i x^2(i)}. \qquad (7)$$

Voiced speech usually has stronger energy than unvoiced speech and background noise. Therefore, a small RTPD may be indicative of a noise-like music clip, such as rock music. Figure 4 shows the relative frequency histograms of RTPD value for music and speech. We can clearly see that RTPD is quite effective in discriminating music from speech.

3 SVM binary tree approach for multi-class audio classification

We classify a broadcast news audio clip into five classes, i.e., music, pure speech, speech with environment sound

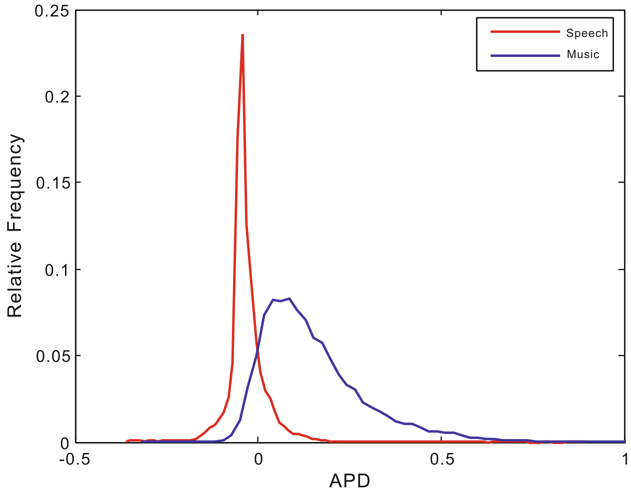

Fig. 3 Relative frequency histogram of APD for speech and tone-like music

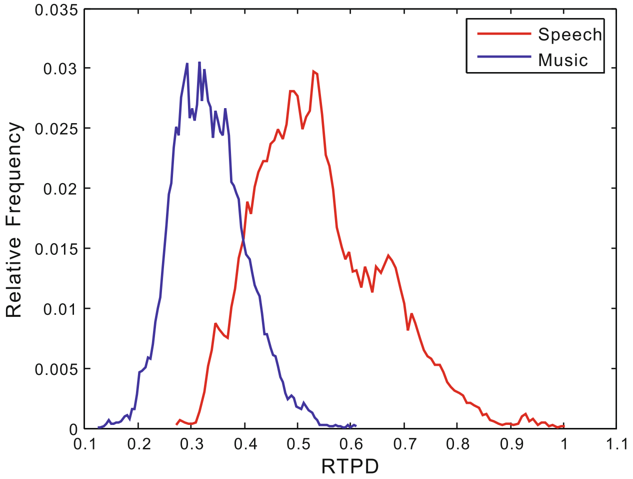

Fig. 4 Relative frequency histogram of RTPD for speech and noise-like music

(speech + envsnd), speech with music (speech + music), environment sound (envsnd). We adopt an SVM binary tree approach to classify an audio clip into the above five classes in a step-by-step manner.

3.1 Support vector machines

Plenty of previous work point out that support vector machines (SVM) [10, 28] show superior performance in pattern classification. Recently, this discriminative model has drawn much attention in audio classification. The SVM classifier outperforms other classifiers such as Gaussian mixture model (GMM), K-Nearest Neighbor (KNN) and multi-layer perceptron (MLP) [21].

Support vector machines (SVMs) [10] are a set of supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM constructs a separating hyperplane in that space, which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier.

Traditional classification techniques optimize the performance on the training data with the criterion of minimizing the empirical risk [10], which leads to high variance and poor generalization (overfitting). SVM tries to minimize the structural risk [10] that prevents overfitting by incorporating a regularization penalty into the optimization. It is equivalent to minimize an upper bound on the classification error. In this sense, SVM is a kind of discriminative models that do not model the whole distribution, but the most discriminative regions of the distribution, e.g., class boundary or margin. Therefore we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is clearly of interest and is known as the maximum-margin hyperplane and such a linear classifier is known as a maximum margin classifier.

With a set of labeled training data for two separate classes, i.e., $\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n)\}$, where $\mathbf{x}_j$ is the feature vector ($\mathbf{X}_j \in \mathbf{R}^d$) with class label $c_j \in \{1, -1\}$, the hyperplane classifier can be represented by

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{j=1}^{l} \alpha_j^* c_j K(\mathbf{x}_j \cdot \mathbf{x}) + b^*\right), \tag{8}$$

where α^* and b^* are classifier parameters, and $\mathbf{x}_j$ is called support vector when α_j^* is not zero. $K(\mathbf{x}_j \cdot \mathbf{x})$ is called the kernel function. An SVM can be linear or non-linear according to the different kernel functions, for example,

- Linear kernel: $K(\mathbf{x}_j \cdot \mathbf{x}) = (\mathbf{x}_j, \mathbf{x})$;
- Polynomial kernel: $K(\mathbf{x}_j \cdot \mathbf{x}) = (s(\mathbf{x}_j, \mathbf{x}) + c)^d$, where s, c and d are parameters;
- RBF kernel: $K(\mathbf{x}_j \cdot \mathbf{x}) = \exp(-\gamma|\mathbf{x} - \mathbf{x}_j|^2)$, where γ is a parameter.

SVM is also good at classifying data with small number of training samples [10]. In the audio classification of broadcast news, the labeled data is very expensive. Due to the proliferation of the broadcast news documents, we prefer very much to use limited number of training samples to conduct the classification. For this purpose, SVM is more suitable.

As a discriminative model, SVM has no prior assumption to distribution of a class in the feature space [10]. In contrast, it is well-known that many conventional methods, e.g., K-means, intrinsically assume that the samples in a class form a multi-dimensional ball in the feature space, or obey an ellipsis distribution when distance metric learning is used [13, 15]. In the problem of broadcast news classification, it is very difficult to find a proper and general prior distribution for a class due to the diversity of the news contents, speakers, and environmental noises. Hence, SVM is a preferable choice for multi-class audio classification in broadcast news.

3.2 SVM binary tree

SVMs are originally developed for a binary classification problem [10, 28]. In real-word applications, such as audio classification in broadcast news, we often need to distinguish between multiple classes. How to effectively extend SVMs for multi-class classification is still an on-going research issue. Some methods directly solve the multi-class problem by considering an SVM target function for all classes at once. Weston et al. [31] extended the SVM target function that integrated multiple decision rules. However, the multi-class target function is quite complicated and the training and testing processes are time-consuming. Therefore, the dominating approaches are to decompose the multi-class problem into multiple binary classification problems and incorporate binary-class SVMs. The one-against-rest method constructs an SVM for each class by discriminating that class against the remaining $K - 1$ classes, in which N SVMs are needed. The one-against-one strategy constructs an SVM for each pair of classes and the number of SVMs needed are $K(K - 1)/2$.

In this study, we use an SVM binary tree (SVM-BT) approach [7] to model the multi-class audio classification task. An SVM-BT represents the multi-class problem by a

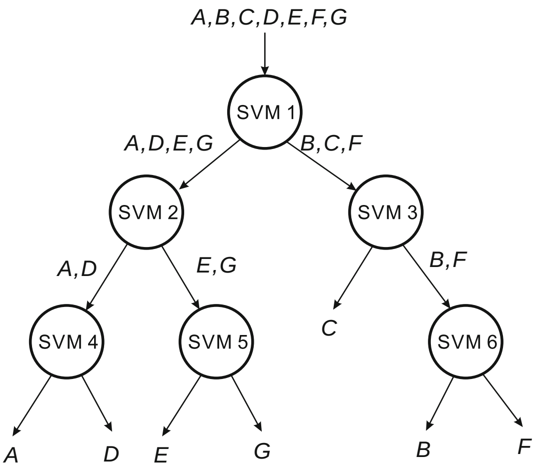

Fig. 5 SVM binary tree architecture for multi-class classification

hierarchical binary tree, where each node makes a binary-class decision by a common SVM, as shown in Fig. 5. The SVM-BT is a preferable solution to the multi-class audio classification task in terms of efficiency and strategy.

The general K-class labeling is NP-hard [2, 11]. If we want to label N data samples into K classes, the size of the searching space is K^N. It is clear that the complexity of the problem is highly related to the number of classes K. It is desirable if we could somehow use smaller K, e.g. $K = 2$, to obtain a reasonable multi-class labeling, since this may lead to a more effective solution. In this paper, we show this is possible in the audio classification task of broadcast news by our SVM binary tree approach. The tree approach only needs to train maximally $(K - 1)$ SVMs for a K-class problem. Moreover, the binary tree architecture ensures that only $\lceil \log_2 K \rceil$ SVMs are subjected to operation in the classification of each sample. Therefore, it is more efficient in computation as compared with the one-against-rest and one-against-one methods. Similar tree-structured extensions to binary labeling methods to handle multi-label clustering problems were also successfully used in computer vision [2, 11] and other fields. As a result, we believe the proposed binary SVM tree approach is a general solution and can be used in other multi-class labeling problems.

Our problem, in nature, involves the classification of mixed type broadcast news audio. We need to discriminate speech + music with pure speech and pure music, and speech + envsnd with pure speech and environment sound. The one-against-rest scheme is apparently unappropriate because the mixed types complicate the decision making. For example, it is quite difficult to distinguish pure speech from the combination of the rest classes (which also contains the speech component). The SVM-BT is able to employ a *coarse-to-fine* strategy that allows us to first make a discrimination on coarse classes easy to differentiate. For example, we can first make classification with a high accuracy on two coarse classes, i.e., with-speech and without-speech, and then go through next level of the classification on fine-gained classes.

Another consideration for a stepwise classification approach is that the effectiveness of various features are not identical for discriminating between all the classes. A universal feature set is not appropriate to discriminate each pair of classes. For multi-class classification, it is often the case that features are only effective to distinguish between some classes. For example, silence frame ratio (SFR) is quite effective to discriminate speech from music because music shows a lower SFR while speech show a higher SFR. However, it cannot demonstrate superior performance when categorizing noise and music since the two classes both tends to show a lower value in SFR.

3.3 Multi-class audio classification approach

3.3.1 System diagram

Our broadcast news audio classification approach is composed of three steps, i.e., front-end processing, feature extraction and classification. In front-end processing, a broadcast news audio stream is segmented into short clips with fixed length. A short clip is used as the classification unit. Each clip is further divided into non-overlapping frames for feature extraction. Before feature extraction, we pre-emphasize audio frames with an FIR filter ($H(z) = 1 - az^{-1}$, $a = 0.97$), and weight them with a Hamming window to avoid spectral distortions. We filter out silence clips using an empirical threshold on silence frame ratio (SFR).

In the classification step, the non-silence clips are classified into one of the five pre-defined classes based on the extracted features using the SVM-BT approach. We investigate two SVM-BT architectures, i.e., with-speech and without-speech classification at the top level (SVM-BT-S) and with-music and without-music classification at the top level (SVM-BT-M), as shown in Fig. 6a, b, respectively. Since the misclassification at the upper level may propagate to the lower level of the tree, we follow the coarse-to-fine strategy. That is, given a non-silence clip, starting at the root node, a pairwise SVM decision is made on broad classes (with-speech versus without-speech in Fig. 6a or with-music versus without-music Fig. 6b). Then it moves to either left or right of the tree depending on the result, and continues lower rounds of classification until reaching to one of leaves which represent the fine-gained class. Specifically, our multi-class audio classification needs to train four SVMs and maximally three rounds of binary classification for each audio clip. The SVM-BT

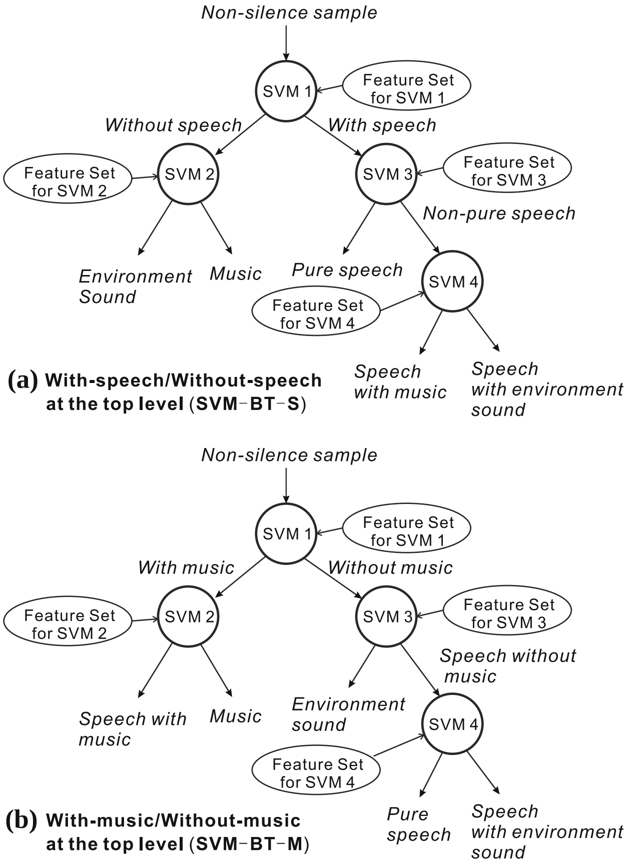

Fig. 6 Two SVM-BT architectures for multi-class audio classification

structure allows the use of different sets of features at each SVM node.

3.3.2 Feature pool

We use a broad feature pool of 18 types of audio features for multi-class audio classification, which includes 6 types of frequently-used time-domain features, 10 types of popular frequency-domain features and two new pitch-density-based features, as listed in Table 1. Some features are extracted at clip level and others are frame-based. A feature selection step is used to select an effective feature subset from the broad pool for each SVM.

3.3.3 Feature selection

The various features in the feature pool may be irrelevant and redundant for each specific SVM classifier in the binary tree. For more interpretability, robustness and efficiency, we adopt a wrapper-based feature selection method [14], which selected an optimal feature subset according to their usefulness to the classifier. A greedy heuristic search algorithm, i.e., backward elimination, is used to seek the optimal subset by iteratively eliminating the features whose absence do not decrease the classification accuracy, as shown in Fig. 7. In each iteration of the elimination procedure, we use the training set for SVM training, and a validation set for performance validation. After the feature selection process, an optimal feature subset is selected for each SVM on the SVM tree.

4 Experiments

4.1 Corpus and experimental setup

Our experimental data is derived from CCTV broadcast news,[1] which is about 8 h in total with 16 0.5-h-long news episodes. The broadcast news corpus contains rich audio types. The speech data is composed of studio speech, telephone interview and field speech with different ambient noises. The music data comes from both non-vocal music (e.g. orchestral and percussion) and vocal music (e.g. pop songs). The environment sounds originate from various audio scenes, such as machines, engines, winds, crowds, birds, cars and applause, etc. The audio streams are recorded in 16 kHz, mono with 16 bit per sample, and manually annotated in terms of the pre-defined five classes, i.e., pure speech, speech with environment sound (speech + envsnd), speech with music (speech + music), music and environment sound (envsnd). Three hours of audio data were used for training, one hour for validation of feature selection, and the rest four hours were used for testing.

In the experiments, we divided each audio stream into non-overlapping clips with length of 0.96 s by a sliding window and then various features are extracted for each clip. A clip was further divided into 30 32 ms-long audio frames and frame-based features were calculated for each frame. Since classification was performed at clip level, we used mean and variance of frame-based features as their clip level features. After silence removal by SFR thresholding, classification experiments were performed on these short clips and evaluated in terms of classification accuracy. It is defined as the ratio of correctly classified samples over all predicted samples of each audio class. We used the open-source Marsyas[2] toolkit for feature extraction, SVM training, SVM-BT implementation and classification experiments.

4.2 Experimental results

4.2.1 Performance of pitch-density-based features

We first carried out experiments to evaluate the proposed pitch-density-based features. We compared the two new

[1] http://www.cctv.com.

[2] http://marsyas.sness.net/.

Table 1 Feature candidates for multi-class audio classification in this study

Type	Feature	Abbreviation	Frame/clip
Time domain	Root mean square	RMS	Frame
	Zero-crossing rate	ZCR	Frame
	High zero-crossing rate ratio	HZCCR	Clip
	Low short time energy ratio	LSTER	Clip
	Noise frame ratio	NFR	Clip
	Silence frame ratio	SFR	Clip
Frequency domain	Spectral centroid	SC	Frame
	Spectral spread	SS	Frame
	Spectral flux	SF	Clip
	Spectral kurtosis	SK	Frame
	Spectral roll-off frequency	SRF	Frame
	Band period	BP	Clip
	Subband energy distribution	SED	Frame
	Mel frequency cepstral coefficients	MFCC	Frame
	Linear predictive cepstral coefficients	LPCC	Frame
	Linear spectral pairs	LSP	Frame
New pitch-density feature	Average pitch-density	APD	Clip
	Relative tonal power density	RTPD	Clip

```
Backward Elimination()
Input   : $\mathcal{P}^0 = \phi, \mathcal{P}^1 =$ {all features}, $\mathcal{R} = \phi, i = 1$
Output  : Best subset $\mathcal{P}^i$
begin
    while $\mathcal{P}^i \neq \mathcal{P}^{i-1}$ do
        for each $v \in \mathcal{P}^i$ do
            set $\mathcal{P}^{'} \leftarrow \mathcal{P}^i \setminus v$;
            train the SVM with $\mathcal{P}^{'}$ and get the
            validation performance $F(\mathcal{P}^{'})$;
            if $F(\mathcal{P}^{'}) \geq F(\mathcal{P}^i)$ then
                $\mathcal{R} \leftarrow \mathcal{R} \cup \{v\}$;
            end
        end
        $\mathcal{P}^{i+1} \leftarrow \mathcal{P}^i \setminus \mathcal{R}$;
        $i++$;
        $\mathcal{R} = \phi$;
    end
    return $\mathcal{P}^i$;
end
```

Fig. 7 Backward elimination algorithm for feature selection

Table 2 Speech/music discrimination accuracy for individual features

Feature	Dimension	Accuracy (%)		
		Speech	Music	Average
HZCCR	1	87.7	87.1	87.4
LSTER	1	88.0	88.1	88.1
SFR	1	86.4	85.9	86.2
SC	2	86.2	84.9	86.6
SS	2	85.7	85.8	85.8
SF	1	85.2	83.8	84.5
SK	2	86.2	86.9	86.6
SRF	2	87.7	87.4	87.6
BP	4	88.0	88.0	88.0
SED	2	89.2	88.7	89.0
MFCC	24	94.3	88.8	91.6
LPCC	24	91.7	88.2	90.0
LSP	36	92.2	87.9	90.1
APD	1	90.6	89.6	90.1
RTPD	1	90.0	89.1	89.6

features with several popular features (as listed in Table 1) in the task of speech/music discrimination. We grouped the categories of pure speech and speech with environment sound together into a broad "speech" category in the discrimination experiments. An SVM with RBF kernel was trained for speech/music discrimination using the training set. Experimental results on the test set are summarized in Table 2.

From Table 2, we can clearly see that the MFCC feature obtains the best classification performance, followed by LSP and LPCC. Interestingly, the proposed one-dimensional APD and RTPD features achieve comparable performance with high-dimensional LSP and LPCC features. Especially, APD and RTPD achieve superior performance in music classification with higher accuracy as compared with other features. The promising performance is due to the specific consideration of music characteristics of the pitch-density-based features, as studied in Sect. 2.

Table 3 Optimal feature subsets achieved by feature selection for various schemes

Scheme	SVM#	Discrimination	Optimal feature subset	
			Pool without APD and RTPD	Pool with APD and RTPD (-N)
SVM-BT-S	SVM1	With speech versus without speech	HZCRR, LSTER, RMS, SC, SS, BP, NFR, SF, SFR, SK, LPCC, LSP, MFCC	HZCCR, LTSER, RMS, SFR, SK, BP, NFR, SF, LPCC, LSP, MFCC, APD, RTPD
	SVM2	Environment sound versus music	NFR, STE, SED, SF, LPCC, LSP	NFR, STE, SED, SF, LPCC, LSP, APD
	SVM3	Pure speech versus non-pure speech	SFR, ZCR, SC, SSSF, LPCC, LSP, MFCC	SFR, HZCCR, SC, SS SF, LPCC, MFCC, APD, RTPD
	SVM4	Speech with music versus speech with environment sound	NFR, STE, SED, SF, LPCC, LSP	NFR, STE, SED, SF, LPCC, LSP, SK, APD, RTPD
SVM-BT-M	SVM1	With music versus without music	HZCRR, LSTER, RMS, BP, SED, MFCC, LSP	HZCRR, LSTER, BP, SED, MFCC, LPCC, APD, RTPD
	SVM2	Music versus speech with music	HZCRR, ZCR, LSTER, RMS, SC, SS, BP, NFR, SF, LPCC, LSP, MFCC	ZCR, LSTER, RMS, SC, SS, BP, NFR, SED, LPCC, MFCC, RTPD
	SVM3	Speech without music versus environment sound	HZCRR, ZCR, LSTER, RMS, SC, SS, SK, SFR, BP, SF, LPCC, LSP, MFCC	HZCRR, ZCR, LSTER, RMS, SC, SS, SK, SFR, BP, NFR, SF, LPCC, LSP, MFCC
	SVM4	Pure speech versus speech with environment sound	SFR, ZCR, RMS, SC, SSSF, SK, LPCC, MFCC	SFR, HZCCR, RMS, SC, SS, SF, SK, LPCC, LSP, MFCC, RTPD
Weston	Single SVM	All five classes	RMS, HZCRR, LSTER, NFR, SFR, SC, SF, SK, SRF, BP, LPCC, LSP, MFCC	RMS, HZCRR, LSTER, NFR, SFR, SC, SF, SRF, BP, SED, LPCC, LSP, MFCC, APD

SVM-BT-S with-speech/without-speech discrimination at the top level of the SVM-BT; *SVM-BT-M* with-music/without-music discrimination at the top level of the SVM-BT; *Weston* Weston's multi-class SVM

Table 4 Multi-class audio classification results for Weston's multi-class SVM and SVM-BT approaches in terms of classification accuracy (%)

Method	Pure speech	Music	Speech + music	Speech + envsnd	envsnd	Average
Baseline	89.7	88.7	86.4	84.3	80.0	85.8
Weston	93.7	92.4	88.5	87.8	86.6	89.8
Weston-N	93.6	93.0	88.1	87.6	86.6	89.8
SVM-BT-S	97.3	95.3	90.0	93.6	89.7	93.2
SVM-BT-S-N	97.3	96.5	93.9	93.7	89.7	94.2
SVM-BT-M	96.7	95.1	93.7	93.0	88.4	93.4
SVM-BT-M-N	96.9	95.1	96.2	93.1	88.1	93.9
Average	95.9	94.6	91.7	91.5	88.2	

Baseline MFCC/GMM-SVM supervector, *-N* with the pitch-density-based features

4.2.2 Multi-class audio classification results

We carried out multi-class audio classification experiments to evaluate the performances of Weston's multi-class SVM and the SVM-BT approaches. The two SVM-BT architectures, as drawn in Fig. 2, were both evaluated in the experiments. We performed feature selection (see Sect. 3.3.3) for each binary SVM on the SVM-BT to achieve the best feature subset from the general feature pool which contains 18 types of features (Table 1). To measure the usefulness of the pitch-density-based features, feature selection was initiated from both a feature pool that contains all 18 features and a feature pool without the two pitch-density-based features (16 features). The optimal feature subsets achieved by feature selection are summarized in Table 3. We observe that SVM classifiers at different levels of the SVM tree chose different sets of features from the feature pool. This accords with our argument that the effectiveness of different features are not identical for discriminating between all the classes and a feature selection procedure is essential to the performance. From Table 3, we also see that APD and RTPD are frequently selected from the feature pool. This shows the effectiveness of the pitch-density-based features in multi-class audio classification.

Experimental results on the test set for multi-class audio classification are summarized in Table 4. Note that the results for an MFCC/GMM-SVM supervector baseline

[3, 4] were also included. In this baseline approach, the 24-dimension MFCC feature vector was mapped to a supervector (i.e., concatenation of adapted GMM means) via MAP adaption from a GMM with 256 mixtures [3]. The GMM was trained using all the data from the training set. The output supervector was served as the feature vector and was classified by a Weston's multi-class SVM. From the results, we can see that all the tested approaches achieve higher accuracy as compared with the baseline approach. In general, SVM-BTs outperform Weston's multi-class SVM in terms of classification accuracy. The four SVM-BT approaches can improve average accuracy by 3.3–4.4 points as compared to Weston's multi-class SVM approaches. The SVM-BT-S-N approach achieves the best average classification accuracy which is as high as 94.2%. We observe that the pitch-density-based features can generally improve the multi-class audio classification performance. Especially, the addition of the two features has increased the accuracy rate for the speech + music class by 3.9 points when we migrate from the SVM-BT-S scheme to the SVM-BT-S-N scheme.

When comparing the speech-dominated with music-dominated schemes (SVM-BT-S vs. SVM-BT-M and SVM-BT-S-N vs. SVM-BT-M-N), we observe that the speech-dominated schemes achieve higher accuracy for most classes except for the speech + music class. However, discriminating between with-music and without-music at the top level of the SVM tree can improve the classification accuracy for the mixed type of speech with music. The SVM-BT-M-N scheme obtains a high accuracy of 96.2% for the speech + music class.

When we compare different classes, we see that the classification accuracy rates for pure speech, music, speech + music and speech + envsnd are all over 92% except for the envsnd class. The classification accuracy for environment sound remains the lowest in the five testing audio classes. This is probably because the environment sound class is composed of various types of sounds and they have quite different acoustic characteristics. For example, the crowd sound may be mis-classified as speech and the bird tweets have similar characteristics with music. The classification scheme may be further improved if we discriminate more fine-gained audio classes.

5 Summary

In this paper, we have studied multi-class audio classification in broadcast news. After silence removal, we classify an audio clip into one of the five classes with mixed audio types: pure speech, music, environment sound, speech with music and speech with environment sound. This fine-gained classification is an important precursor to many broadcast news management tasks, such as segmentation, browsing and retrieval, etc.

We have proposed two pitch-density-based features, namely APD and RTPD. APD is motivated by the tonal regularities of music signals and RTPD represents the characteristics of percussion instruments. We adopt an SVM binary tree (SVM-BT) approach to solve the multi-class audio classification problem, which employs a coarse-to-fine hierarchical classification strategy. SVM-BT allows us to first discriminate on broad audio classes easy to differentiate and employ different feature set for each individual SVM on the tree. Specifically, we use a feature selection process to select an optimal feature subset from a pool of 18 types of features for each SVM. Experiments show that the proposed 1-dimensional APD and RTPD features are able to achieve comparable accuracy with high-dimensional features in speech/music discrimination, and the SVM-BT approach demonstrates superior performance in multi-class audio classification.

Future work will focus on the integration of multi-class audio classification with speaker annotation [34] to realize an automatic broadcast news diarization system.

Acknowledgments This work was supported by the National Natural Science Foundation of China (60802085), the Program for New Century Excellent Talents in University (2008) supported by the Ministry of Education (MOE) of China, the Research Fund for the Doctoral Program of Higher Education in China (20070699015), the Natural Science Basic Research Plan of Shaanxi Province (2007F15) and the NPU Foundation for Fundamental Research (W018103).

References

1. Androutsos, D., Guan, L., Venetsanopoulos, A.N.: Semantic retrieval of multimedia. IEEE Signal Process. Mag. **14**, 237–253 (2006)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1222–1239 (2001)
3. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
4. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: The GMM-SVM supervector approach for the recognition of the emotional status from speech. LNCS, vol. 5768, pp. 894–C903 (2009)
5. Carey, M.J., Parris, E.S., Lloyd-Thomas, H.: A comparison of features for speech, music discrimination. In: ICASSP, vol. 1, pp. 149–152. Phoenix, USA (1999)
6. Chen, L., Gunduz, S., Ozsu, M.T.: Mixed type audio classification with support vector machine. In: International Conference on Multimedia and Expo, pp. 781–784. Toronto, Canada (2006)
7. Cheong, S., Oh, S.H., Lee, S.Y.: Support vector machines with binary tree architecture for multi-class classification. Neural Inf. Process. **2**(3), 47–51 (2004)
8. Childers, D.G., Skinner, D.P., Kemerait, R.C.: The cepstrum: a guide to processing. Proc. IEEE **65**(10), 1428–1443 (1977)
9. Choi, M.Y., Song, H.J., Kim, H.S.: Discrimination for robust speech recognition in robots. In: International Symposium on

Robot and Human Interactive Communication, vol. 1, pp. 118–121. Jeju, Korea (2007)

10. Cortes, C., Vapnik, V.: Support network vectors. Mach. Learn. **20**, 273–297 (1995)
11. Feng, W., Jia, J., Liu, Z.Q.: Self-validated labeling of Markov random fields for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (2010)
12. Gerhard, D.: Pitch extraction and fundamental frequency: History and current techniques. Tech. rep., University of Regina (2003)
13. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighborhood component analysis. Adv. Neural Inf. Process. Syst. **17**, 513–520 (2005)
14. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
15. Hastie, T., Tibshirani, R.: Discriminant adaptive nearest neighbor classification. IEEE Trans. Pattern Anal. Mach. Intell. **18**(6), 607–616 (1996)
16. Jiang, H., Bai, J., Zhang, S., Xu, B.: Svm-based audio scene classification. In: NLP-KE, vol. 131–136, pp. 897–900 (2005)
17. Keum, J.S., Lee, H.S.: Speech/music discrimination using spectral peak feature for speaker indexing. In: International Symposium on Intelligent Signal Processing and Communication Systems, pp. 323–326 (2006)
18. Khan, M.K.S., Al-Khatib, W.G.: Machine-learning based classification of speech and music. Multimedia Syst. **12**(1), 55–67 (2006)
19. Li, D., Sethi, I.K., Dimitrova, N., McGee, T.: Classification of general audio data for content-based retrieval. Pattern Recognit. Lett. **22**, 533–544 (2001)
20. Li, Y., Dorai, C.: Svm-based audio classification for instructional video analysis. In: ICASSP, vol. 5, pp. 897–900. Toronto, Canada (2004)
21. Liu, C., Xie, L., Meng, H.: Classification of music and speech in mandarin news broadcasts. In: National Conference on Man–Machine Speech Communication. Huangshan, China (2007)
22. Lu, L., Zhang, H.J.: Content analysis for audio classification and segmentation. IEEE Trans. Speech Audio Process. **10**(7), 504–516 (2002)
23. Lu, L., Zhang, H.J., Li, Z.: Content-based audio classification and segmentation by using support vector machines. Multimedia Syst. **8**, 482–491 (2003)
24. Mckinney, M., Breebaart, J.: Features for audio and music classification. In: Proceedings of the International Symposium on Music Information Retrieval, pp. 151–158 (2003)
25. Panagiotakis, C., Tziritaz, G.: A speech/music discriminator based on rms and zero-crossings. IEEE Trans. Multimedia **7**(1), 155–166 (2005)
26. Pikrakis, A., Giannakopoulos, T., Theodoridis, S.: A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks. IEEE Trans. Multimedia **10**(5), 846–857 (2008)
27. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: ICASSP, vol. 2, pp. 1331–1334 (1997)
28. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
29. Wang, J., Wu, Q., Deng, H., Yan, Q.: Real-time speech/music classification with a hierarchical oblique decision tree. In: ICASSP, pp. 2033–2036 (2008)
30. Wang, W.Q., Gao, W., Ying, D.W.: A fast and robust speech/music discrimination approach. Inf. Commun. Signal Process. **3**, 1325–1329 (2003)
31. Weston, J., Watkins, C.: Multi-class support vector machines. Tech. Rep. CSD-TR-98-04, University of London, Egham, UK (1998)
32. Wu, Q., Yan, Q., Deng, H., Wang, J.: A combination of data mining method with decision trees building for speech/music discrimination. Comput. Speech Lang. **24**(7), 257–272 (2010)
33. Xie, L.: Discovering salient prosodic cues and their interactions for automatic story segmenation in Mandarin broadcast news. Multimedia Syst. **14**, 237–253 (2008)
34. Xie, L., Wang, G.: A two-stage multi-feature integration approach to unsupervised speaker change detection in real-time news broadcasting. In: International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 350–353 (2008)
35. Zhang, T., Jay Kuo, C.C.: Audio content analysis for online audiovisual data segmentation and classification. IEEE Trans. Speech Audio Process. **9**(4), 441–457 (2001)
36. Zheng, L., Xie, L., Wang, X., Lu, M., Yang, Y., Zhang, Y.: An antomatic caption generator for mandarin broadcast news. In: 5th Joint Conference on Harmonious Human Machine Environment. Xi'an, China (2009)
37. Zhu, Y., Sun, Q., Rahardja, S.: Detecting musical sounds in broadcast audio based on pitch tuning analysis. In: International Conference on Multimedia and Expo, pp. 13–16. Toronto, Canada (2006)