Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 209814, 11 pages
http://dx.doi.org/10.1155/2015/209814

Research Article

Optimized Audio Classification and Segmentation Algorithm by Using Ensemble Methods

Saadia Zahid,[1] Fawad Hussain,[1] Muhammad Rashid,[2] Muhammad Haroon Yousaf,[1] and Hafiz Adnan Habib[1]

[1]*Department of Computer Engineering, University of Engineering and Technology Taxila, Taxila 47050, Pakistan*
[2]*Computer Engineering Department, College of Computer and Information Systems, Umm Al Qura University, Makkah 24382, Saudi Arabia*

Correspondence should be addressed to Fawad Hussain; fawad.hussain@uettaxila.edu.pk

Received 22 January 2015; Revised 16 April 2015; Accepted 16 April 2015

Academic Editor: Gen Qi Xu

Audio segmentation is a basis for multimedia content analysis which is the most important and widely used application nowadays. An optimized audio classification and segmentation algorithm is presented in this paper that segments a superimposed audio stream on the basis of its content into four main audio types: pure-speech, music, environment sound, and silence. An algorithm is proposed that preserves important audio content and reduces the misclassification rate without using large amount of training data, which handles noise and is suitable for use for real-time applications. Noise in an audio stream is segmented out as environment sound. A hybrid classification approach is used, bagged support vector machines (SVMs) with artificial neural networks (ANNs). Audio stream is classified, firstly, into speech and nonspeech segment by using bagged support vector machines; nonspeech segment is further classified into music and environment sound by using artificial neural networks and lastly, speech segment is classified into silence and pure-speech segments on the basis of rule-based classifier. Minimum data is used for training classifier; ensemble methods are used for minimizing misclassification rate and approximately 98% accurate segments are obtained. A fast and efficient algorithm is designed that can be used with real-time multimedia applications.

1. Introduction

The excessive rise in multimedia data over internet has created a major shift towards online services. In most multimedia applications, audio information is an important part. The most common and popular example of online information is music [1]. Audio analysis, video analysis, and content understanding can be achieved by segmenting and classifying an audio stream on the basis of its content [2]. For this purpose, an efficient and accurate method is required that segments out an audio stream. A technique, in which an audio stream is divided into homogenous (similar) regions, is called audio segmentation [1]. The advent of multimedia and network technology results in an emerging increase in digital data and this causes a growing interest in multimedia content-based information retrieval. For analyzing and understanding an audio signal, the fundamental step is to discriminate an audio signal on the basis of its content. Audio classification and segmentation are a pattern recognition problem. It comprises two main stages: feature extraction and then classification on the basis of these features (statistical information) extracted [3].

Applications of audio content analysis can be categorized in two parts. One part is to discriminate an audio stream into homogenous regions and the other part is to discriminate a speech stream into segments, of different speakers. Lu et al. [2, 4] discriminate an audio stream into different audio types. Classifier support vector machines [5–9] and K-nearest neighbor integrated with linear spectral pairs-vector quantization are used respectively. The training is done on 2-hour data.

Coz et al. [10] presented an audio indexing system that characterizes various content levels of a sound track by frequency tracking. The system does not require any prior

knowledge. A fuzzy approach is used by Kiranyaz et al. [11] in which hierarchic audio classification and segmentation algorithm based on automated audio analysis is proposed. An audio signal is divided into homogeneous regions by finding time boundaries also called change points detection. In audio segmentation, with the help of change detection a sound signal is segmented in homogenous and continuous temporal regions. The problem arises in defining the criteria of homogeneity. By computing exact generalized likelihood ratio statistics, the audio stream segmentation can be done without any prior knowledge of the classes. Mel-frequency cepstral coefficients are used as feature [12]. For calculating statistics large amount of training data is required.

Tasks like meeting transcription and automatic camera panning require the segmentation of a group meeting into different individual person's speech. Bayesian information criterion (BIC) is used for segmenting the feature vectors [13–15]. BIC requires a large amount of training data. Structured discriminative models use structures support vector machine (SSVM) in the mediums of large vocabulary speech recognition tasks. Hidden Markov models (HMMs) [16–21] are used to determine the features and Viterbi-like scheme is used [14].

Traditionally used audio retrieval systems are text based, whereas the human auditory systems principally rely on perception. As the text only elaborates the high level content, this is not sufficient to get any perceptual likeness between two acoustic audio clips. This problem can be solved easily by using Query by example technique. In this technique, only those audio samples are predicted from databases that sound similar to the example. Query by example is quite a different approach from audio classification. For modeling the continuous probability distribution of audio features, Gaussian mixture model (GMM) is used [22].

Janku and Hyniová [23] proposed that MMI-supervised tree-based vector quantizer and feedforward neural network [16, 17, 24, 25] can be used on a sound stream in order to detect environmental sounds and speech. Regularized kernel based method based on kernel Fisher discriminant can be used for unsupervised change detection [26, 27].

Speech is not only a mode of transmitting word messages; it also emphasizes emotions, personality, and so forth. Words contain vowel regions, which are of vital importance in many speech applications mainly in speech segmentation and verification of speaker. Vowel regions initiate when the vowel onset point occurs and ends when vowel offset point occurs. Audio segmentation is also possible, by dividing an audio stream into segments, on the basis of vowel regions [28].

Audio segmentation algorithms can be divided into three general categories. In the first category, classifiers are designed [29]. The features are extracted in time domain and frequency domain; then classifier is used to discriminate audio signals on the basis of its content. The second category of audio segmentation extracts features on statistics that is used by classifier for discrimination. These types of features are called posterior probability based features. Large amount of training data is required by the classifier to give accurate results. The third category of audio segmentation algorithm emphasizes setting up effective classifiers. The classifiers used in this category are Bayesian information criterion, Gaussian likelihood ratio, and a hidden Markov model (HMM) classifier. These classifiers also give good results when large training data is provided [29].

Audio segmentation and classification have many applications. Content-based audio classification and retrieval are mostly used in entertainment industry, audio archive management, commercial music usage, surveillance, and so forth. Nowadays, on the World Wide Web, millions of databases are present; for audio searching and indexing audio segmentation and classification are used. In monitoring broadcast news programs, audio classification is used, helping in efficient and accurate navigation through broadcast news archives [30].

The analysis of superimposed speech is a complex problem and improved performance systems are required. In many audio processing applications, audio segmentation plays a vital role in preprocessing step. It also has a significant impact on speech recognition performance. That is why a fast and optimized audio classification and segmentation algorithm is proposed which can be used for real-time applications of multimedia. The audio input is classified and segmented into four basic audio types: pure-speech, music, environment sound, and silence. An algorithm is proposed that requires less training data and from which high accuracy can be achieved; that is, misclassification rate is minimum.

The organization of paper is as follows: Audio classification and segmentation algorithm (proposed), preclassification step, feature extraction step, hybrid classifier approach (bagged SVMs (support vector machines) with ANNs (artificial neural networks)), and steps used for discrimination are discussed. In Results and Discussion the experimental results are discussed.

2. Materials and Methods

2.1. Audio Classification and Segmentation Step. Hybrid classification scheme is proposed in order to classify an audio clip into basic data types. Before classification a preclassification step is done which analyzes each windowed frame of the audio clip separately. Then the feature extraction step is performed from which a normalized feature vector is obtained. After feature extraction the hybrid classifier approach is used. The first step classifies audio clips/frames into speech and nonspeech segments by using bagged SVM. As the silence frames are mostly present in speech signal so the speech segment is classified into silence and pure-speech segments on the basis of rule-based classifier. Finally, ANN classifier is used to further discriminate nonspeech segments into music and environment sound segments. This hybrid scheme is used to achieve high classification accuracy and can be used for different real-time applications of multimedia. Figure 1 illustrates the block diagram of the proposed algorithm. Audio stream is taken as an input, it is then downsampled to 8000 KHz, preclassification step is applied on this audio stream, features {zero-crossing rate, short-time energy, spectrum flux, Mel-frequency cepstral coefficients, and periodicity analysis} are extracted, and

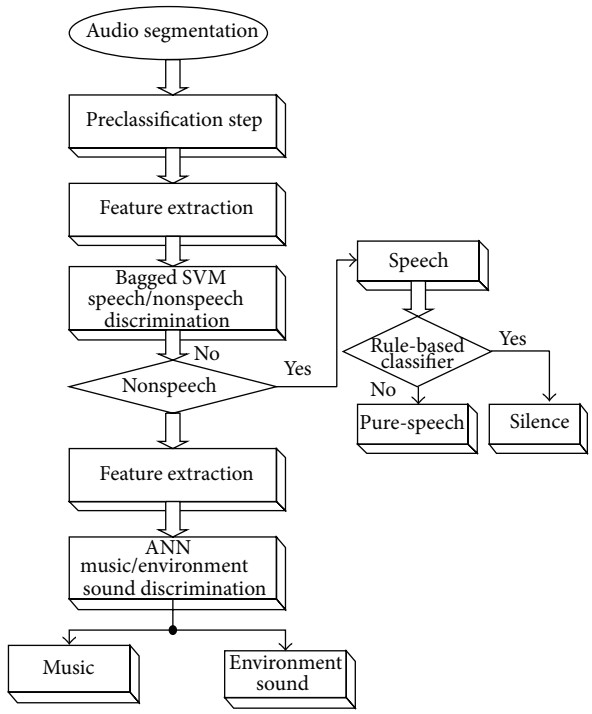

Figure 1: Proposed audio classification and segmentation algorithm.

hybrid classifier is used. Bagged SVM uses features {zero-crossing rate, short-time energy, spectrum flux, and Mel-frequency cepstral coefficients} and classifies audio clip into speech and nonspeech segments; features {spectrum flux, periodicity analysis, and Mel-frequency cepstral coefficients} are used and nonspeech segments are classified into music and environment sound using ANN. Rule-based classifier is used to discriminate silence and pure-speech segments.

In preprocessing step for audio segmentation all input signals are downsampled into 8 KHz sampling rate. Audio clips are subsequently segmented into 1-s frames. This 1-s frame is taken as the basic classifying unit. For feature extraction nonoverlapping frames are used. The features signify the characteristic information present within each 1-s audio clip.

2.2. Preclassification Step. Speech signal is superimposed (i.e., in mixed form) which means that a conversation is held at any place or party where there is music and lots of noise. This is also known as cocktail party effect. Separating the source or the desired segments within the independent component analysis framework is known as blind source separation [31–33]. Blind source is generally a method used to separate the mixed signal into independent sources (when the mixing process is not known) [34]. Most blind source separation techniques use higher order statistics. For higher order statistics these algorithms require iterative calculations [35]. Molgedey and Schuster method is used for separating the signals on the basis of second order statistics (correlation). This does not need higher order statistics and iterative calculations. The temporal structure of signals is analyzed and the separation is done on this basis.

The mixed signal is firstly converted to the time-frequency domain, also called spectrogram of signal, by applying Fourier transform at short-time intervals. Hamming window is used. In order to avoid mixing of spectrograms each spectrogram is dealt with separately. Correlation is performed on all these short intervals. The sphering and rotation step is then performed. Orthogonalizing source signals into an observing coordinate is called sphering. An observation is actually a projection of source signals in certain direction. Original observations are not orthogonal; by applying sphering these observations are arranged in such a way that they become orthogonal to each other. An ambiguity of rotation still remains, even after sphering. So the correct rotation can be examined by removing all the off-diagonal observations present in correlation matrix. Simultaneous diagonalization [36, 37] is applied at several time delays. Reconstruction step is performed on each separated signal's spectrogram. All the decomposed frequency components are then combined. At the end permutation step is performed for finding the relation between the separated signals shown in Figure 2. The decision is made by using classifier.

2.3. Feature Extraction Step. The process of converting an audio signal into a sequence of feature vectors is called feature extraction process. The feature vectors carry temporal as well as spectral characteristic information about the audio signal. Feature vectors are calculated on window basis. The feature selection has a great impact on the performance of audio segmentation systems. Three types of features are calculated in this proposed work: Mel-frequency cepstral coefficients (MFCCs), time-domain and frequency-domain features. To form a feature vector these normalized features are combined.

Initially the audio stream is converted to 16 bit chunks at a sampling rate of 8 kHz. Feature extraction step is performed on the separated signals obtained after preclassification step. These separated signals are divided into nonoverlapping frames. These frames are used as classification unit. On the basis of the classification results segmentation is performed.

As suggested by [38], 12 order Mel-frequency cepstral coefficients are used. Time-domain features are zero-crossing rate, short-time energy, and periodicity analysis. Frequency-domain feature is spectrum flux.

2.3.1. Zero-Crossing Rate (ZCR). Zero-crossing is a measure of signal changes that occurs from positive to negative or vice versa as shown in Figure 3. General definition is the amount of zero-crossing within a frame. Zero-crossing rate discriminates speech and music effectively, as the speech contains more silent regions as compared to music so the zero-crossing rate for speech is greater than music [4, 30].

The expression for the zero-crossing rate is given by

$$\mathrm{ZCR} = \frac{1}{2\,(M-1)} \sum_{n=1}^{M-1} \left|\mathrm{sgn}\left[x\,(n+1)\right] - \mathrm{sgn}\left[x\,(n)\right]\right|, \quad (1)$$

where $x(n)$ represents the discrete signal that is in the range of $n = 1, \ldots, M$. sgn$[\cdot]$ is known as sign function.

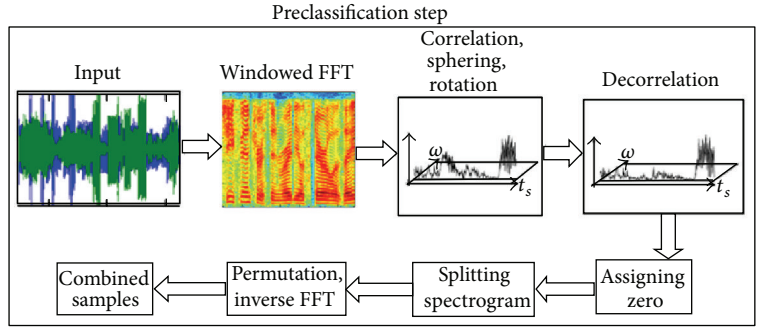

Figure 2: An overview of the preclassification step.

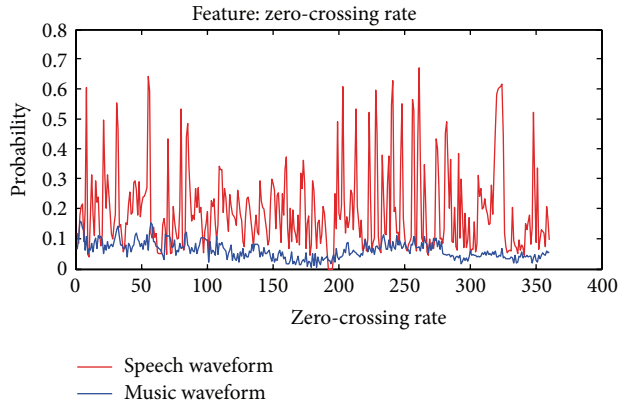

Figure 3: Zero-crossing rate (calculated for multiple frames of speech and music waveforms).

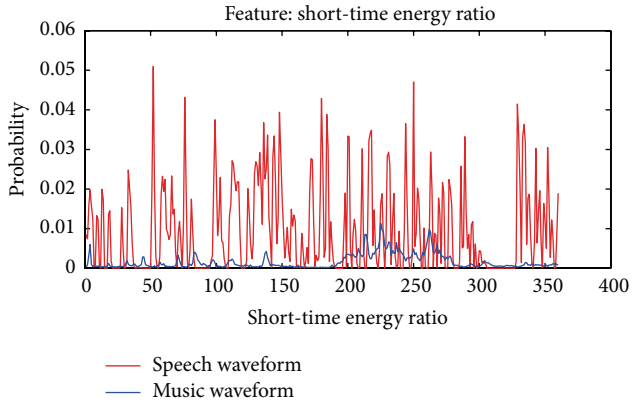

Figure 4: Short-time energy (calculated for multiple frames of speech and music waveforms).

2.3.2. Short-Time Energy (STE). Short-time Energy is a measure of the total energy power of a frame. STE is a useful feature for distinguishing speech and music segments. STE measure for speech signals has large variations as compared to music signals [4, 30], because the frequency characteristics of human voice are extremely different from music apparatus. Figure 4 shows the short-time energy calculated from the multiple frames of speech and music.

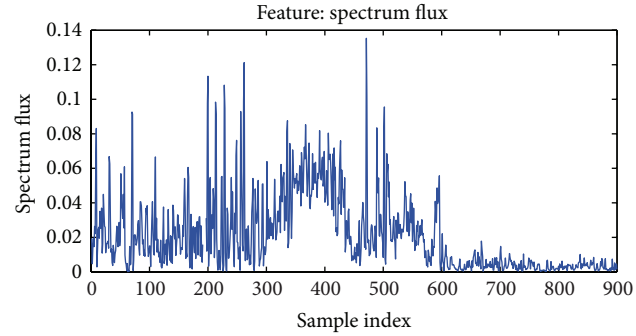

Figure 5: Spectrum flux plot (calculated for multiple frames of speech, environment sound, and music waveforms; from the plot it can be easily observed that spectrum flux of speech which is from 0 to 280 s is greater than music spectrum flux which is from 561 to 850 s. And the spectrum flux for environment sound which is 281 to 560 s is highest).

Mathematically short-time energy feature is expressed as

$$E_m = \sum_{n=-\infty}^{\infty} [x(n)\, w(m-n)]^2, \tag{2}$$

where $x(n)$ is the input discrete signal, m is the number of frames, and $w(n)$ is the window used for analysis.

2.3.3. Spectrum Flux (SF). Spectrum flux is a measure of the changeable power spectrum of an audio signal. It is calculated by computing distance between current frame and the previous frame. Precisely, spectrum flux can be obtained by calculating the Euclidean distance between two normalized spectra. Spectrum flux helps in discriminating speech, music, and environment sound. Speech signals have higher spectral variations as compared to music. However, for environment sound the spectral variation is the highest as compared to music and speech [2, 4, 30]. Figure 5 shows the spectrum flux plot. Consider

$$\mathrm{SF} = \frac{1}{(M-1)(N-1)} \times \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} [\log(X(m,n)+\Delta) - \log(X(m-1,n)+\Delta)]^2, \tag{3}$$

where

$$X(m,n) = \left| \sum_{k=-\infty}^{\infty} x(k)\, w(mL-k)\, e^{-j(2\pi/L)Nk} \right|. \quad (4)$$

Equation (3) illustrates the formula that is used to calculate spectrum flux. $x(k)$ is the input discrete audio signal. Window function $w(k)$ of length L is used; the order of DFT is N. The total number of frames is M and a small value Δ is introduced in order to avoid calculation overflow, whereas $X(m,n)$ is the Fourier transform of $x(k)$.

2.3.4. Mel-Frequency Cepstral Coefficients (MFCCs). Mel-frequency cepstral coefficients are the logarithmic measure of the Mel magnitude spectrum, which is calculated by triangular band-pass filter. These values are decorrelated using discrete cosine transform. MFCC is a real-valued implementation of complex cepstrum; that is why it is calculated by taking FFT. The steps for calculating MFCC are as follows:

(1) Divide the audio signal into short frames.

(2) Take Fourier transform of each frame and calculate periodogram-based power spectral estimate for each frame.

(3) Take log of all filter bank energies.

(4) Take discrete cosine transform of each Mel log power.

(5) The amplitudes of resulting spectrum are MFCCs.

MFCCs have good discriminating capability. That is why most of the speech recognition systems use them as a strong feature [4, 27, 30, 39] (see Figure 6). Consider

$$s_m = \sqrt{\frac{2}{K}} \sum_{i=1}^{K} (\log S_K) \cos\left[\frac{m(i-0.5)\pi}{K}\right], \quad (5)$$

where $m = 1, 2, \ldots, L$, K represents the total band-pass filters and the order of cepstrum is L. 12 order MFCCs are used. After passing ith triangular band-pass filter, the resulting Mel-weighted spectrum is S_K. s_m is the transformed Mel-weighted spectrum to MFCCs.

2.3.5. Periodicity Analysis. Periodicity analysis can be calculated by estimating the periodicity of each frame and periodicity is obtained by correlation. Periodicity analysis for music is higher than environment sound because music signals are more periodic in nature as compared to environment sound signals. Periodicity is useful feature for discriminating music and environment sound [2] (see Figure 7). Consider

$$r = \frac{\sum_{n=0}^{N-1} X_{i(n-k)} X_i(n)}{\sqrt{\sum_{n=0}^{N-1} X_i(n-k)} \sqrt{\sum_{n=0}^{N-1} X_i(n)}}. \quad (6)$$

Equation (6) illustrates the periodicity calculation for each frame. i is the frame index, N is the total frame number, and r is the normalized correlation function calculated from current sample $X_i(n)$ and previous sample $X_i(n-k)$.

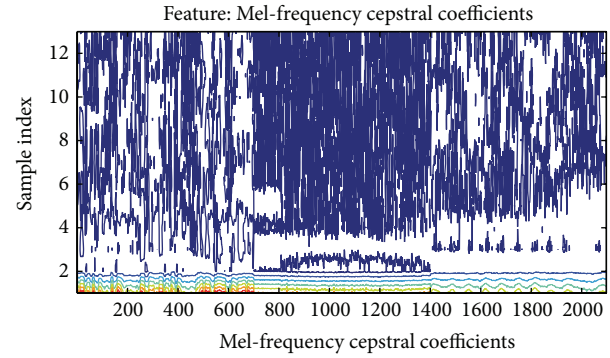

Figure 6: Mel-frequency cepstral plot (calculated for multiple frames of speech, music, and environment sound; from the plot the cepstral behavior for three different audio types can be observed; from 0 to 700 s is speech, from 701 to 1400 is environment sound, and from 1401 to 2100 s is music).

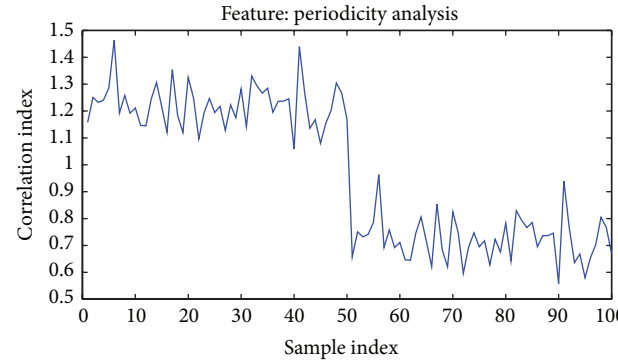

Figure 7: Periodicity analysis plot (calculated for multiple frames of music and environment sound; from the plot it can be easily observed that periodicity of music which is from 0 to 50 s is greater than environment sound which is from 51 to 80 s).

These features are concatenated to form a feature vector. All the features have different characteristics; that is, combining these features to create a feature vector is not appropriate. That is why each feature component is normalized which makes their scale comparable. Normalization is computed as

$$m(x_i) = \frac{(x_i - \min(x))}{(\max(x) - \min(x))}, \quad (7)$$

where x_i represents the ith feature value. This normalized feature vector is further used for classification.

3. Hybrid Approach

Bagged SVM is combined with ANN classifier to provide a hybrid approach. Only using a single classifier most of the environment data is misclassified as speech and music, which is not a good approach. So in order to avoid this misclassification hybrid approach is used in which firstly speech and nonspeech are discriminated and then further nonspeech is discriminated into music and environment sound.

3.1. Support Vector Machine (SVM). Support vector machine [40–42] uses a given set containing positive and negative examples in order to learn an optimized separating hyperplane. For a fixed data having unknown probability distribution, SVM minimizes the probability of misclassifying unseen patterns. SVM minimizes the structural risk; that is why optimized performance on training data is obtained. This property of SVM makes it more optimal when compared with traditional pattern recognition techniques. The support vector machines are of two types; linear and nonlinear (kernel based). In an audio data the feature distribution is so complicated. Different classes of an audio data may have overlapping areas and cannot be separated linearly; such situation can be handled by kernel support vector machine. In this section, some concept of kernel based SVM with bagging (bagged SVM) is introduced.

3.2. Kernel Support Vector Machines. Considering the case in which the vectors are linearly nonseparable but are nonlinearly separable, SVM uses a kernel function $K(x, y)$. SVM uses the kernel to create an optimal separating hyperplane [43–45]. The curse of dimensionality can be addressed in such a way that the input vectors can be mapped implicitly by the kernel function to a high dimensionality feature space; in this feature space the mapped data is linearly separable. Most commonly used kernel functions are polynomial, Gaussian radial basis function, and multilayer perception. It was empirically observed that the Gaussian radial basis kernel performs better than the other two; that is why in the proposed method Gaussian radial basis kernel is used. Consider

$$K(x, y) = \exp\left(-\frac{x - y^2}{2\sigma^2}\right), \tag{8}$$

where σ is the Gaussian function's width.

3.3. Bagging Approach. Multiple classifier system, also known as ensemble learning, includes training of different classifiers and combining their predictions in order to obtain improved classification accuracy. Using multiple classifiers in a system outperforms the single classifier results. Ensemble method tries to combine a set of learners in contrast to the ordinary learning approaches that make their predictions on the basis of a single learner [46, 47]. Bagging gives best results for unstable classifiers [48]. Bagging approach is used to improve the accuracy of certain classifiers while dealing with artificial and real-world datasets. The general concept of bagging is to generate multiple training subsets through bootstrapping. In bootstrapping random samples with replacement are picked [47, 49]. The training is performed on each subset and their output is aggregated via majority voting [50, 51].

An audio segmentation technique is presented that reduced the misclassification rate of the classifier in order to achieve high accuracy and to fully preserve the information inside the audio stream. That is why support vector machine (SVM) classifier is bagged. SVM gives good results when used separately as compared to artificial neural networks (ANNs) and k-nearest neighbors (KNN) [38]. Different subsets for bagging approach are randomly selected; on each subset training and testing are performed. The predictions of each SVM are aggregated through majority voting.

3.4. Artificial Neural Network (ANN). A computational or mathematical model that is inspired from the structural and functional characteristics of human nervous system is called artificial neural network (ANN) or neural networks (NNs) [52]. A neural network is composed of multiple interconnected groups of artificial neurons. These interconnected neurons use a connectionist approach for computing any information. ANNs have adaptive nature, which means that they change their structure on the basis of the information (either internal or external) which passes through network. Artificial neural networks are composed of simple elements working in parallel, known as nodes. Neural networks are trained by adjusting the connection values between nodes. Training is performed until a specific output appears for a corresponding input. Network is adjusted on the basis of the target and output difference and it stops when the difference between targets and output is zero or minimum; that is, output matches input [53].

ANN training process is called supervised learning. ANNs are trained in such a way that an input is given to the nodes, the nodes calculate the output, and the predetermined targets are compared with the output. If the targets and output do not match then it is given back to the node and the weights readjust. This process continues until the output and targets have maximum matches. Due to the knowledge storing ability and decision making ANNs are used extensively in pattern recognition tasks. ANNs are of two types: single layer and multiple layer perceptron (MLPs) [54, 55]. Single layer perceptron uses single layer of weights; that is, input is directly connected to the output. Single layer perceptron only handles linearly separable problems. Multiple layer perceptron (MLP) uses multiple layers of weights. It consists of input layer, hidden nodes, and output layers. The proposed algorithm uses multiple layer perceptron ANN. Back propagation algorithm is used for training ANN classifier.

3.5. Discrimination Steps. The steps used for discriminating an audio data into different audio types are discussed in detail.

3.5.1. Speech and Nonspeech Discrimination. Speech and nonspeech frames are discriminated on the basis of bagged SVM classifier. On the processed audio clip bagged SVM classifier is applied based on spectrum flux, Mel-frequency cepstral coefficients, zero-crossing rate, and short-time energy. Speech and nonspeech codebooks are generated by training databases.

3.5.2. Silence and Pure-Speech Discrimination. Silence is detected on the basis of features {short-time energy, zero-crossing rate} by using 1-s window. The classification is done by rule-based classifier; a threshold value is set. If {short-time energy, zero-crossing rate} are less than the predefined threshold then it is a silence frame; otherwise it is classified as pure-speech frame. This is a simple approach used for distinguishing silence and pure-speech frames (see Figure 8).

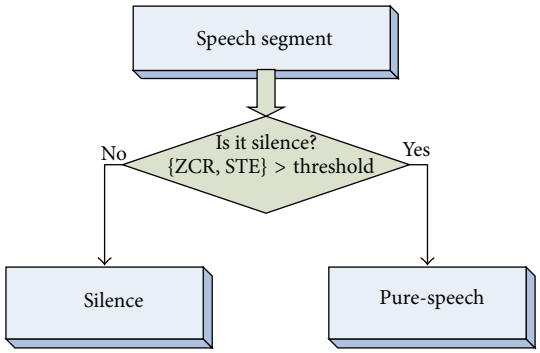

Figure 8: Block diagram for silence and pure-speech discrimination (silence is separated from speech segment by using a rule-based classifier; if the value of features {ZCR, STE} is greater than a predefined threshold value then it is a pure-speech segment; otherwise it is a silence segment.).

The mixed audio stream is never fully silent; it always contains some kind of sound. That is why silence is only present in the pauses in the speech and it can be detected.

3.5.3. Discriminating Music and Environment Sound. Nonspeech segment is used for discriminating music and environment sound segments. Spectrum flux, Mel-frequency cepstral coefficients, and periodicity analysis are used to discriminate music from environment sound. Music signals are more periodic as compared to environment sound signals. But discriminating the signals alone on the basis periodicity analysis is not a good choice. For more accurate and precise results spectrum flux and Mel-frequency cepstral coefficients are also incorporated with periodicity analysis feature. Spectrum flux of environment sounds in most cases is greater than music. Using ANN classifier the music and environment sound segments are discriminated.

3.6. Postprocessing Step. Audio stream is continuous, so it does not change frequently or abruptly. A few smoothing steps are performed at the end on speech segment. Between three 1-s frames if first frame and last frame are speech segments then most probably the mid frame is also a speech frame. Similarly, in three 1-s frames if first, second, and last frames are not same then the mid frame may be a silence frame and first and last frames are speech frames. This postprocessing step refines the results and increases the accuracy measure for pure-speech and silence discrimination.

4. Results and Discussions

The audio dataset used for evaluation of proposed algorithm is speech/music collection of GTZAN. It consists of personal CDs, microphone, and radio recordings in order to provide different varieties of recording conditions. The dataset contains 120 audio files that are 30-s long and are downsampled to 8 KHz before processing. 1000-s environment sounds are also included in the dataset. The audio stream used is in mixed form; that is, speech is superimposed with music and noise.

Table 1: The accuracies for randomly selected five datasets.

Sets	Accuracy %	Misclassification rate %
1	97.2	2.8
2	97.5	2.5
3	97.1	2.9
4	97.0	3
5	96.6	3.4
Aggregate labels	**98.2**	**1.8**

Table 2: Results obtained for different classifying types.

Classifying type	Accuracy %	Misclassification rate %
Speech/nonspeech	98.2	1.8
Music/environment sound	97.6	2.4
Silence/pure-speech	98.5	1.5

Half-hour data is used for training and approximately two-hour data for testing. Noise is segmented out as environment sound. 1/3 of the dataset is used for training classifier and 2/3 of the dataset is used for testing the classifier.

After preprocessing, the audio stream is divided into short segments by applying hamming window, which is a moderate window. Each segment is processed independently. Hamming window is mostly used in narrowband applications (e.g., spectrum of telephone signal). Fourier transform is computed for each segment. Correlation of these short segments is taken, sphering and rotation step is performed in order to estimate the direction. Decorrelation is done to reconstruct these short segments. The cross correlating values are eliminated. Permutation is applied to find the relation between these independent segments. On the relation basis these segments are combined.

The combined audio stream is divided into nonoverlapping 1-s frames with the help of a moving window. On these 1-s frames classification is performed. For classification the features are extracted from each frame. These features are combined to form a feature vector.

The speech and nonspeech frames are discriminated by using bagged SVM. Randomly five different training and testing sets with replacement are selected. The results of these classifiers are majority voted and the final aggregate labels are obtained. The accuracies for these five randomly selected datasets are shown in Table 1. The output labels for each set are compared with each other; if the maximum number of output labels for a single frame is 1 then it is labeled as 1; otherwise it is labeled as 0. This method is known as majority voting.

Bagged SVM classifier is based on features {ZCR, STE, SF, and MFCC}. This baseline model gives good results for speech and nonspeech discrimination. Bagged SVM results are compared with the simple SVM classifier. The results obtained for different classifying types are shown in Table 2. Bagged SVM gives 98.2% accurate results and 1.8% reduced misclassification rate whereas simple SVM gives 92% accuracy and 8% misclassification as evident in Figure 9. The

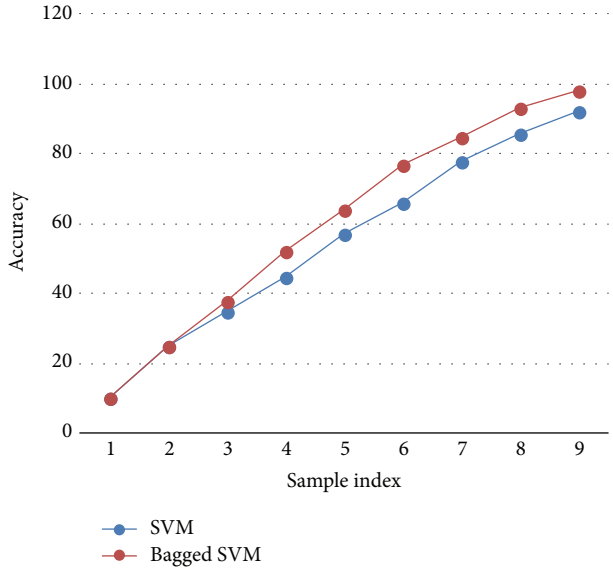

Figure 9: Bagged SVM versus SVM (the accuracy results for (A) SVM and (B) bagged SVM are compared; bagged SVM achieves high accuracy as compared to SVM).

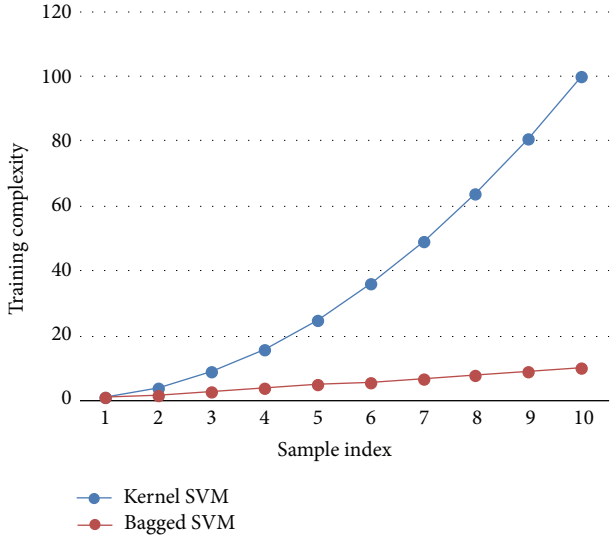

Figure 10: Training complexity of kernel SVM versus bagged SVM.

results of different SVMs applied on five different randomly selected sets with replacement are boosted by using ensemble methods technique called bagging.

The use of ensemble approaches reduces training complexity drastically, especially when high predictive accuracy is maintained. In bagged SVM, different SVM models are aggregated; in this case five SVM models are aggregated. Each SVM randomly selects training set with replacement; that is, training is performed on small samples of training set. Because of this subdivision, total training time decreases. The computation complexity of kernel based SVM is $\Omega(n^2)$, but when m classifiers are used on subsamples of size n/m, then the computational complexity is approximately $\Omega(n^2/m)$

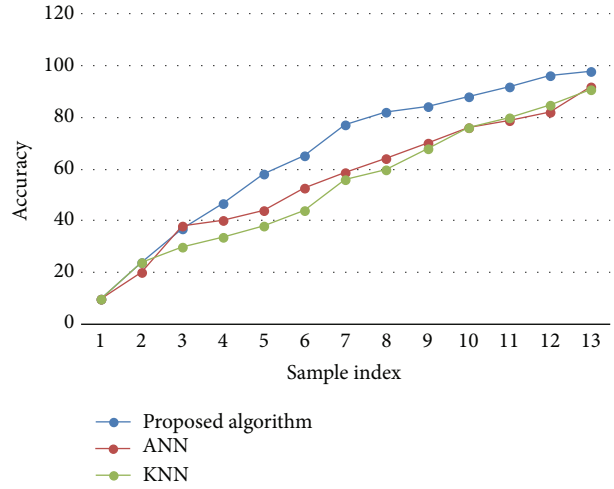

Figure 11: Comparison of proposed algorithm versus ANN and KNN (the accuracy results for (A) proposed algorithm, (B) ANN, and (C) KNN are compared; from the plot it can be observed that the proposed algorithm achieves high accuracy as compared to ANN and KNN).

[56]. Because of this reduced complexity larger dataset and nonlinear kernels can be handled easily as shown in Figure 10.

Nonspeech segments are further discriminated into music and environment sound by using ANN. The features {periodicity analysis, Mel-frequency cepstral coefficients, and spectrum flux} are used. The speech segment can further be segmented into silence and pure-speech segment by using rule-based classifier. The overall classification results are shown in Table 3. 98% classification results are obtained with 1.9% misclassification rate.

The performance of proposed algorithm is also tested with K-nearest neighbor (KNN) and artificial neural networks (ANNs) as shown in Figure 11. Both these classifiers are also used for audio classifications. The Bayesian information criterion (BIC), Gaussian likelihood ratio, and a hidden Markov model (HMM) use large training data for good results [29]; that is why proposed algorithm is only compared with the conventional classifiers SVM and ANN, as the proposed algorithm uses less training data. ANN performs better than KNN but still there is some misclassification. With the hybrid approach proposed minimum data is misclassified and maximum information within the audio stream is preserved.

In addition to the above performance analysis, a comparison is done between the algorithm used and the already existing audio segmentation and classification techniques. Audio classification and segmentation was presented in [57], in which audio stream is segmented into speech, music and silence. The respective algorithm uses general mixture model (GMM) and K-nearest neighbor (KNN). The algorithm achieves 95% accuracy for discrete audio signals.

Audio stream is segmented into music, speech, environment sound, and silence [2]. The respective algorithm

Table 3: Overall classification results.

Parameters	Bagged SVM %	ANN %	Rule-based classifier %	Final results (average) %
Accuracy	98.2	97.6	98.5	**98.1**
Misclassification rate	1.8	2.4	1.5	**1.9**

Table 4: Statistical comparison of proposed technique with the existing ones.

	Audio segmentation techniques					
Accuracy	Proposed optimized audio classification and segmentation algorithm by using ensemble methods	Audio segmentation and classification using GMM and KNN [57]	Content analysis for audio classification and segmentation using KNN and LSP-VQ [2]	Sports audio segmentation and classification using BIC and GMM [15]	Classification of audio signals using AANN [58]	Classification of audio signals using GMM [58]
	98%	95%	96%	87.3%	93.1%	92.9%

uses K-nearest neighbor (KNN) and linear spectral pairs-vector quantization (LSP-VQ). This algorithm achieves 96% accuracy.

Sports audio stream is segmented and classified into speech and nonspeech [15]. For segmentation Bayesian information criterion (BIC) is used. Clusters are formed and Gaussian mixture model (GMM) is used for classification. 87.3% accurate segmentation and classification results are achieved.

Audio stream is classified into music, sports, advertisement, cartoon, and movie [58]. In order to capture feature vectors auto associative neural network model (AANN) is used. For training GMM is used.

Table 4 shows the comparison of proposed audio segmentation technique with the existing audio segmentation techniques.

5. Conclusion and Future Work

An efficient and fast audio classification and segmentation approach has been discussed that does not require large amount of training data yet gives good discrimination results. In this work, an audio stream is discriminated into homogenous regions and classified into basic audio types such as pure-speech, music, environment sound, and silence. Main goal is to design an audio segmentation algorithm which can be incorporated with multimedia content analysis applications and audio recognition systems.

Hybrid approach has been used for audio classification and segmentation. Firstly, audio clips are discriminated into speech and nonspeech segments by using bagged SVM classifier. Nonspeech segments are further classified into environment sound and music by using ANN classifier. Speech segment is discriminated into silence and pure-speech segments by using rule-based classifier. Experiments have showed that the algorithm is very efficient for real-time multimedia applications.

In future work, this algorithm can be used as a preprocessing step in automatic speech recognition, video conferencing, human-computer interaction systems (for identifying human activities involving speech), and speaker tracking. This algorithm can be used in video content analysis, audio retrieval, and indexing, for attaining useful information.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was fully supported by the Directorate of Advanced Research and Technological Studies at University of Engineering and Technology Taxila, Pakistan.

References

[1] A. Lerch, *An Introduction to Audio Content Analysis*, Wiley-IEEE Press, 2012.

[2] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[3] I. McLoughlin, *Applied Speech and Audio Processing: With MATLAB Examples*, Nanyang Technological University, Cambridge University Press, 2009.

[4] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 2003.

[5] H. Tang, C.-H. Meng, and L.-S. Lee, "An initial attempt for phoneme recognition using structured support vector machine (SVM)," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 4926–4929, IEEE, Dallas, Tex, USA, March 2010.

[6] S. E. Krüger, M. Schafföner, M. Katz, E. Andelic, and A. Wendemuth, "Speech recognition with support vector machines in a hybrid system," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, Lisbon, Portugal, September 2005.

[7] M. Davy, A. Gretton, A. Doucet, and P. J. W. Rayner, "Optimized support vector machines for nonstationary signal classification," *IEEE Signal Processing Letters*, vol. 9, no. 12, pp. 442–445, 2002.

[8] R. Solera-Ureña, J. Padrell-Sendra, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de-María, "SVMs for automatic speech recognition: a survey," in *Progress in Nonlinear Speech Processing*, pp. 190–216, Springer, Berlin, Germany, 2007.

[9] S. Haykin, *Neural Networks: A Comprehensive Foundation*, vol. 2, 2004.

[10] M. L. Coz, J. Pinquier, R. Andre-Obrecht, and J. Mauclair, "Audio indexing including frequency tracking of simultaneous multiple sources in speech and music," in *Proceedings of the 11th International Workshop on Content-Based Multimedia Indexing (CBMI '13)*, pp. 23–28, IEEE, Veszprem, Hungary, June 2013.

[11] S. Kiranyaz, A. F. Qureshi, and M. Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1062–1081, 2006.

[12] A. Dessein and A. Cont, "An information-geometric approach to real-time audio segmentation," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 331–334, 2013.

[13] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 2, pp. 1150–1153, IEEE, Hong Kong, August 2006.

[14] S.-X. Zhang and M. J. F. Gales, "Structured SVMs for automatic speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 3, pp. 544–555, 2013.

[15] J. Huang, Y. Dong, J. Liu, C. Dong, and H. Wang, "Sports audio segmentation and classification," in *Proceedings of the IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC '09)*, pp. 379–383, IEEE, Beijing, China, November 2009.

[16] J. Hennebert, M. Hasler, and H. Dedieu, "Neural networks in speech recognition," in *Proceedings of the 6th Microcomputer School of Neural Networks, Theory and Applications (Micro-Computer '94)*, pp. 23–40, Prague, Czech Republic, 1994.

[17] D. O'Shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1272–1305, 2003.

[18] L. D. Paulson, "Speech recognition moves from software to hardware," *Computer*, vol. 39, no. 11, pp. 15–18, 2006.

[19] Ø. Birkenes, T. Matsui, K. Tanabe, S. M. Siniscalchi, T. A. Myrvoll, and M. H. Johnsen, "Penalized logistic regression with HMM log-likelihood regressors for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1440–1454, 2010.

[20] B.-H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

[21] P. Nguyen, G. Heigold, and G. Zweig, "Speech recognition with flat direct models," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 994–1006, 2010.

[22] P. Hu, W. Liu, W. Jiang, and Z. Yang, "Latent topic model for audio retrieval," *Pattern Recognition*, vol. 47, no. 3, pp. 1138–1143, 2014.

[23] L. S. Janku and K. Hyniová, "Application of feed-forward neural network and MMI-supervised vector quantizer to the task of content based audio segmentation by co-operative unmanned flying robots," in *Proceedings of the 1st International Conference on Intelligent Systems, Modelling and Simulation (ISMS '10)*, pp. 111–115, IEEE, Liverpool, UK, January 2010.

[24] Z. Ping, T. Li-Zhen, and X. Dong-Feng, "Speech recognition algorithm of parallel subband HMM based on wavelet analysis and neural network," *Information Technology Journal*, vol. 8, no. 5, pp. 796–800, 2009.

[25] V. R. V. Krishnan and P. BabuAnto, "Features of wavelet packet decomposition and discrete wavelet transform for malayalam speech recognition," *International Journal of Recent Trends in Engineering*, vol. 1, no. 2, pp. 93–96, 2009.

[26] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappé, "A regularized kernel-based approach to unsupervised audio segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 1665–1668, IEEE, Taipei, Taiwan, April 2009.

[27] T. Giannakopoulos and S. Petridis, "Detection and clustering of musical audio parts using Fisher linear semi-discriminant analysis," in *Proceedings of the IEEE 20th European Signal Processing Conference (EUSIPCO '12)*, pp. 1289–1293, IEEE, Bucharest, Romania, August 2012.

[28] J. Yadav, *Detection of Vowel Offset Point from Speech Signal*, 2013.

[29] C.-C. Huang, J.-F. Wang, and D.-J. Wu, "Audio signal segmentation algorithm," U.S. Patent No. 7,774,203, 2010.

[30] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[31] M. Yanai, F. Sasaki, O. Tanaka, and M. Yasuoka, "Blind source separation and two-signal localization in time-frequency domain considering time lag information: application to the case where one signal includes a reflected signal," *Acoustical Science and Technology*, vol. 35, no. 1, pp. 55–58, 2014.

[32] Y. Liang, J. Harris, S. M. Naqvi, G. Chen, and J. A. Chambers, "Independent vector analysis with a generalized multivariate Gaussian source prior for frequency domain blind source separation," *Signal Processing*, vol. 105, pp. 175–184, 2014.

[33] J. Ye, H. Jin, and Q. Zhang, "Adaptive weighted orthogonal constrained algorithm for blind source separation," *Digital Signal Processing*, vol. 23, no. 2, pp. 514–521, 2013.

[34] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.

[35] G. Lio and P. Boulinguez, "Greater robustness of second order statistics than higher order statistics algorithms to distortions of the mixing matrix in blind source separation of human EEG: implications for single-subject and group analyses," *NeuroImage*, vol. 67, pp. 137–152, 2013.

[36] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 1, pp. 161–164, 1996.

[37] K. Abed-Meraim and A. Belouchrani, *Algorithms for Joint Block Diagonalization*, 2004.

[38] M. Cutajar, J. Micallef, O. Casha, I. Grech, and E. Gatt, "Comparative study of automatic speech recognition techniques," *IET Signal Processing*, vol. 7, no. 1, pp. 25–46, 2013.

[39] S. Zubair, F. Yan, and W. Wang, "Dictionary learning based sparse coefficients for audio classification with max and average pooling," *Digital Signal Processing*, vol. 23, no. 3, pp. 960–970, 2013.

[40] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN '99)*, vol. 99, 1999.

[41] V. Franc and V. Hlavác, "Multi-class support vector machine," in *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 2, IEEE, 2002.

[42] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[43] K.-B. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," in *Multiple Classifier Systems*, vol. 3541 of *Lecture Notes in Computer Science*, pp. 278–285, Springer, Berlin, Germany, 2005.

[44] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *The Annals of Statistics*, vol. 26, no. 2, pp. 451–471, 1998.

[45] P. Clarkson and P. J. Moreno, "On the use of support vector machines for phonetic classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 2, pp. 585–588, IEEE, Phoenix, Ariz, USA, March 1999.

[46] S. Dupont and T. Ravet, "Improved audio classification using a novel non-linear dimensionality reduction ensemble approach," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR '13)*, November 2013.

[47] E. Bauer and R. Kohavi, "Empirical comparison of voting classification algorithms: bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1, pp. 105–139, 1999.

[48] Y. Zhang, D. J. Lv, and H. S. Wang, "The application of multiple classifier system for environmental audio classification," *Applied Mechanics and Materials*, vol. 462-463, pp. 225–229, 2014.

[49] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[50] Y. Zhang and Y. Lin, "The classification of environmental audio with ensemble learning," in *Proceedings of the International Conference on Advanced Computer Science and Electronics Information (ICACSEI '13)*, Atlantis Press, Beijing, China, July 2013.

[51] C. C. M. Yeh, J. Wang, Y. Yang, and H. Wang, "Improving music auto-tagging by intra-song instance bagging," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '14)*, pp. 2139–2143, Florence, Italy, May 2014.

[52] B. Yegnanarayana, *Artificial Neural Networks*, Phi Learning Private Limited, New Delhi, India, 2009.

[53] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, vol. 1857 of *Lecture Notes in Computer Science*, pp. 1–15, Springer, Berlin, Germany, 2000.

[54] N. García-Pedrajas, C. Hervás-Martínez, and D. Ortiz-Boyer, "Cooperative coevolution of artificial neural network ensembles for pattern classification," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 3, pp. 271–302, 2005.

[55] J. S. Yadav, M. Yadav, and A. Jain, "Artificial neural network," *International Journal of Scientific Research and Education*, vol. 1, no. 6, pp. 108–117, 2014.

[56] M. Claesen, F. de Smet, J. A. K. Suykens, and B. de Moor, "EnsembleSVM: a library for ensemble learning using support vector machines," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 141–145, 2014.

[57] M. Kos, Z. Kačič, and D. Vlaj, "Acoustic classification and segmentation using modified spectral roll-off and variance-based features," *Digital Signal Processing*, vol. 23, no. 2, pp. 659–674, 2013.

[58] V. P. Minotto, C. B. O. Lopes, J. Scharcanski, C. R. Jung, and B. Lee, "Audiovisual voice activity detection based on microphone arrays and color information," *IEEE Journal on Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 147–156, 2013.