Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 239892, 14 pages
doi:10.1155/2009/239892

Research Article

A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation

Yizhar Lavner[1] and Dima Ruinskiy[1,2]

[1] *Department of Computer Science, Tel-Hai College, Tel-Hai 12210, Israel*
[2] *Israeli Development Center, Intel Corporation, Haifa 31015, Israel*

Correspondence should be addressed to Yizhar Lavner, yizhar_l@kyiftah.org.il

Received 10 September 2008; Revised 5 January 2009; Accepted 27 February 2009

Recommended by Climent Nadeu

We present an efficient algorithm for segmentation of audio signals into speech or music. The central motivation to our study is consumer audio applications, where various real-time enhancements are often applied. The algorithm consists of a learning phase and a classification phase. In the learning phase, predefined training data is used for computing various time-domain and frequency-domain features, for speech and music signals separately, and estimating the optimal speech/music thresholds, based on the probability density functions of the features. An automatic procedure is employed to select the best features for separation. In the test phase, initial classification is performed for each segment of the audio signal, using a three-stage sieve-like approach, applying both Bayesian and rule-based methods. To avoid erroneous rapid alternations in the classification, a smoothing technique is applied, averaging the decision on each segment with past segment decisions. Extensive evaluation of the algorithm, on a database of more than 12 hours of speech and more than 22 hours of music showed correct identification rates of 99.4% and 97.8%, respectively, and quick adjustment to alternating speech/music sections. In addition to its accuracy and robustness, the algorithm can be easily adapted to different audio types, and is suitable for real-time operation.

1. Introduction

In the past decade a vast amount of multimedia data, such as text, images, video, and audio has become available. Efficient organization and manipulation of this data are required for many tasks, such as data classification for storage or navigation, differential processing according to content, searching for specific information, and many others.

A large portion of the data is audio, from resources such as broadcasting channels, databases, internet streams, and commercial CDs. To answer the fast-growing demands for handling the data, a new field of research, known as audio content analysis (ACA), or machine listening, has recently emerged, with the purpose of analyzing the audio data and extracting the content information directly from the acoustic signal [1] to the point of creating a "Table of Contents" [2].

Audio data (e.g., from broadcasting) often contains alternating sections of different types, such as speech and music. Thus, one of the fundamental tasks in manipulating such data is speech/music discrimination and segmentation, which is often the first step in processing the data. Such preprocessing is desirable for applications requiring accurate demarcation of speech, for instance automatic transcription of broadcast news, speech and speaker recognition, word or phrase spotting, and so forth. Similarly, it is useful in applications where attention is given to music, for example, genre-based or mood-based classification.

Speech/music classification is also important for applications that apply differential processing to audio data, such as content-based audio coding and compressing or automatic equalization of speech and music. Finally, it can also serve for indexing other data, for example, classification of video content through the accompanying audio.

One of the challenges in speech/music discrimination is characterization of the music signal. Speech is composed from a selection of fairly typical sounds and as such, can be represented well by relatively simple models. On the other hand, the assortment of sounds in music is much broader

and produced by a variety of instruments, often by many simultaneous sources. As such, construction of a model to accurately represent and encompass all kinds of music is very complicated. This is one of the reasons that most of the algorithmic solutions developed for speech/music discrimination are practically adapted to the specific application they serve. A single comprehensive solution that will work in all situations is difficult to achieve. The difficulty of the task is increased by the fact that on many occasions speech is superimposed on the music parts, or vice versa.

1.1. Former Studies. The topic of speech/music classification was studied by many researchers. Table 1 summarizes some of these studies. It can be seen from the table that, while the applications can be very different, many studies use similar sets of acoustic features, such as short time energy, zero-crossing rate, cepstrum coefficients, spectral rolloff, spectrum centroid and "loudness," alongside some unique features, such as "dynamism." However, the exact combinations of features used can vary greatly, as well as the size of the feature set. For instance, [3, 4] use few features, whereas [1, 2, 5, 6] use larger sets. Typically some long-term statistics, such as the mean or the variance, and not the features themselves, are used for the discrimination.

The major differences between the different studies lie in the exact classification algorithm, even though some popular classifiers (K-nearest neighbour, Gaussian multivariate, neural network) are often used as a basis. Finally, in each study, different databases are used for training and testing the algorithm. It is worth noting that in most of the studies, especially the early ones, these databases are fairly small [6, 7]. Only in a few works large databases are used [8, 9].

1.2. The Algorithm. In this paper we present an efficient algorithm for segmentation of audio signals into speech or music. The central motivation to our study is consumer audio applications, where various real-time enhancements are often applied to music. These include differential frequency gain (equalizers) or spatial effects (such as simulation of surround and reverberation). While these manipulations can improve the perceptive quality of music, applying them to speech can cause distortions (for instance, bass amplification can cause an unpleasant booming effect).

As many audio sources, such as radio broadcasting streams, live performances, or movies, often contain sections of pure speech mixed between musical segments, an automatic real-time speech/music discrimination system may be used to allow the enhancement of music without introducing distortions to the speech.

Considering the application at hand, our algorithm aims to achieve the following:

(i) Pure speech must be identified correctly with very high accuracy, to avoid distortions when enhancements are applied.

(ii) Songs that contain a strong instrumental component together with voice should be classified as music, just like purely instrumental tracks.

(iii) Audio that is neither speech, nor music (noise, environmental sounds, silence, and so forth) can be ignored by the classifier, as it is not important for the application of the manipulations. We can therefore assume that a priori the audio belongs to one of the two classes.

(iv) The algorithm must be able to operate in real time with a low computational cost and a short delay.

The algorithm proposed here answers all these requirements: on one hand, it is highly accurate and robust, and on the other hand, simple, efficient, and adequate for real-time implementation. It achieves excellent results in minimizing misdetection of speech, due to a combination of the feature choice and the decision tree. The percentage of correct detection of music is also very high. Overall the results we obtained were comparable to the best of the published studies, with a confidence level higher than most, due to the large size of test database used.

The algorithm uses various time domain parameters of the audio signal, such as the energy, zero-crossing rate, and autocorrelation as well as frequency domain parameters (spectral energy, MFCC, and others).

The algorithm consists of two stages. The first stage is a supervised learning phase, based on a statistical approach. In this phase training data is collected from speech and music signals separately, and after processing and feature extraction, optimal separation thresholds between speech and music are set for each analyzed feature separately.

In the second stage, the processing phase, an input audio signal is divided into short-time segments and feature extraction is performed for each segment. The features are then compared to their corresponding thresholds, which were set in the learning phase, and initial classification of the segment as speech or music is carried out. Various post-decision techniques are applied to improve the robustness of the classification.

Our test database consisted of 12+ hours of speech and 20+ hours of music. This database is significantly larger than those used for testing in the majority of the aforementioned studies. Tested on this database, the algorithm proved to be highly accurate both in the correctness of the classification and the segmentation accuracy. The processing phase can also be applied in a real-time environment, due to low computation load of the process, and the fact that the classification is localized (i.e., a segment is classified as speech or music independently of other segments). A commercial product based on the proposed algorithm is currently being developed by Waves Audio, and a provisional patent has been filed.

The rest of the paper is arranged as follows: in Section 2 we describe the learning procedure, during which the algorithm is "trained" to distinguish between speech and music, as well as the features used for the distinction. Next, in Section 3, the processing phase and the classification algorithm are described. Section 4 provides evaluation of the algorithm in terms of classification success and comparison to other approaches, and is followed by a conclusion (Section 5).

Table 1: Summary of Former studies.

Paper	Main Applications	Features	Classification method	Audio material	Results
Saunders, 1996 [4]	Automatic real-time FM radio monitoring	Short-time energy, statistical parameters of the ZCR	Multivariate Gaussian classifier	Talk, commercials, music (different types)	95%–96%
Scheirer and Slaney, 1997 [6]	Speech/music discrimination for automatic speech recognition	13 temporal, spectral and cepstral features (e.g., 4 Hz modulation energy, % of low energy frames, spectral rolloff, spectral centroid, spectral flux, ZCR, cepstrum-based feature, "rhythmicness"), variance of features across 1 sec.	Gaussian mixture model (GMM), K nearest neighbour (KNN), K-D trees, multidimensional Gaussian MAP estimator	FM radio (40 min): male and female speech, various conditions, different genres of music (training: 36 min, testing: 4 min)	94.2% (frame-by-frame), 98.6% (2.4 sec segments)
Foote, 1997 [10]	Retrieving audio documents by acoustic similarity	12 MFCC, Short-time energy	Template matching of histograms, created using a tree-based vector quantizer, trained to maximize mutual information	409 sounds and 255 (7 sec long) clips of music.	No specific accuracy rates are provided. High rate of success in retrieving simple sounds.
Liu et al., 1997 [5]	Analysis of audio for scene classification of TV programs	Silence ratio, volume std, volume dynamic range, 4 Hz freq, mean and std of pitch difference, speech, noise ratios, freq. centroid, bandwidth, energy in 4 sub-bands	A neural network using the one-class-in-one-network (OCON) structure	70 audio clips from TV programs (1 sec. long) for each scene class (training: 50, testing: 20)	Recognition of some of the classes is successful
Zhang and Kuo, 1999 [11]	Audio segmentation/retrieval for video scene classification, indexing of raw audio visual recordings, database browsing	Features based on short-time energy, average ZCR, short-time fundamental frequency	A rule-based heuristic procedure for the coarse stage, HMM for the second stage	Coarse stage: speech, music, env. sounds and silence. Second stage: fine-class classification of env. sounds.	>90% (coarse stage)
Williams and Ellis, 1999 [12]	Segmentation of speech versus nonspeech in automatic speech recognition tasks	Mean per-frame entropy and average probability "dynamism", background-label energy ratio, phone distribution match—all derived from posterior probabilities of phones in hybrid connectionist-HMM framework	Gaussian likelihood ratio test	Radio recordings, speech (80 segments, 15 sec. each) and music (80, 15), respectively. Training: 75%, testing: 25%.	100% accuracy with 15 seconds long segments 98.7% accuracy with 2.5-seconds long segments
El-Maleh et al., 2000 [13]	Automatic coding and content-based audio/video retrieval	LSF, differential LSF, measures based on the ZCR of high-pass filtered signal	KNN classifier and quadratic Gaussian classifier (QCG)	Several speakers, different genres of music (training: 9.3 min. and 10.7 min., resp.)	Frame level (20 ms): music 72.7% (QGC), 79.2% (KNN). Speech 74.3% (QGC), 82.5% (KNN). Segment level (1 sec.), music 94%–100%, speech 80%–94%.

Table 1: Continued.

Paper	Main Applications	Features	Classification method	Audio material	Results
Buggati et al., 2002 [2]	"Table of Content description" of a multi-media document	ZCR-based features, spectral flux, short-time energy, cepstrum coefficients, spectral centroids, ratio of the high-frequency power spectrum, a measure based on syllabic frequency	Multivariate Gaussian classifier, neural network (MLP)	30 minutes of alternating sections of music and speech (5 min each)	95%–96% (NN). Total error rate: 17.7% (Bayesian classifier), 6.0% (NN).
Lu, Zhang, and Jiang, 2002 [9]	Audio content analysis in video parsing	High zero-crossing rate ratio (HZCRR), low short-time energy ratio (LSTER), linear spectral pairs, band periodicity, noise-frame ratio (NFR)	3-step classification: 1. KNN and linear spectral pairs-vector quantization (LSP-VQ) for speech/nonspeech discrimination. 2. Heuristic rules for nonspeech classification into music/background noise/silence. 3. Speaker segmentation	MPEG-7 test data set, TV news, movie/audio clips. Speech: studio recordings, 4 kHz and 8 kHz bandwidths, music: songs, pop (training: 2 hours, testing: 4 hours).	Speech 97.5%, music 93.0%, env. sound 84.4%. Results of only speech/music discrimination: 98.0%
Ajmera et al., 2003 [14]	Automatic transcription of broadcast news	Averaged entropy measure and "dynamism" estimated at the output of a multilayer perceptron (MLP) trained to emit posterior probabilities of phones. MLP input: 13 first cepstra of a 12th-order perceptual linear prediction filter.	2-state HMM with minimum duration constraints (threshold-free, unsupervised, no training).	4 files (10 min. each): alternate segments of speech and music, speech/music interleaved	GMM: Speech 98.8%, Music 93.9%. Alternating, variable length segments (MLP): Speech 98.6%, Music 94.6%.
Burred and Lerch, 2004 [1]	Audio classification (speech/ music/back-ground noise), music classification into genres	Statistical measures of short-time frame features: ZCR, spectral centroid/rolloff/flux, first 5 MFCCs, audio spectrum centroid/flatness, harmonic ratio, beat strength, rhythmic regularity, RMS energy, time envelope, low energy rate, loudness, others	KNN classifier, 3-component GMM classifier	3 classes of speech, 13 genres of music and background noise: 50 examples for each class (30 sec each), from CDs, MP3, and radio.	94.6% /96.3% (hierarchical approach and direct approach, resp.)
Barbedo and Lopes, 2006 [15]	Automatic segmentation for real-time applications	Features based on ZCR, spectral rolloff, loudness and fundamental frequencies	KNN, self-organizing maps, MLP neural networks, linear combinations	Speech (5 different conditions) and music (various genres)more than 20 hours of audio data, from CDs, Internet radio streams, radio broadcasting, and coded files.	Noisy speech 99.4%, Clean speech 100%, Music 98.8%, Music without rap 99.2%. Rapid alternations: speech 94.5%, music 93.2%.
Muñoz-Expósito et al., 2006 [3]	Intelligent audio coding system	Warped LPC-based spectral centroid	3-component GMM, with or without fuzzy rules-based system	Speech (radio and TV news, movie dialogs, different conditions); music (various genres, different instruments/singers) -1 hour for each class.	GMM: speech 95.1%, music 80.3%. GMM with fuzzy system: speech 94.2%, music 93.1%.

Table 1: Continued.

Paper	Main Applications	Features	Classification method	Audio material	Results
Alexandre et al, 2006 [16]	Speech/music classification for musical genre classification	Spectral centroid/rolloff, ZCR, short-time energy, low short-time energy ratio (LSTER), MFCC, voice-to-white	Fisher linear discriminant, K nearest-neighbour	Speech (without background music), and music without vocals. (training: 45 min, testing: 15 min)	Music 99.1%, speech 96.6%. Individual features: 95.9% (MFCC), 95.1% (voice to white).

2. The Learning Phase

2.1. Music and Speech Material. The music material for the training phase was derived mostly from CDs or from databases, using high bitrate signals with a total duration of 60 minutes. The material contained different genres and types of music such as classical music, rock and pop songs, folk music, etc.

The speech material was collected from free internet speech databases, also containing a total of 60 minutes. Both high and low bitrate signals were used.

2.2. General Algorithm. A block diagram of the main algorithm of the learning phase is depicted in Figure 1. The training data is processed separately for speech and for music, and for each a set of candidate features for discrimination is computed. A probability density function (PDF) is then estimated for each feature and for each class (Figure 1(a)). Consequently, thresholds for discrimination are set for each feature, along with various parameters that characterize the distribution relative to the thresholds, as described in Section 2.5. A feature ranking and selection procedure is then applied to select the best set of features for the test phase, according to predefined criteria (Figure 1(b)). A detailed description of this procedure is given in Section 2.6.

2.3. Computation of Features. Each of the speech signals and music signals in the learning phase is divided into consecutive analysis frames of length N with hop size h_f, where N and h_f are in samples, corresponding to 40 milliseconds and 20 milliseconds, respectively. For each frame, the following features are computed:

Short-Time Energy. The short-time energy of a frame is defined as the sum of squares of the signal samples normalized by the frame length and converted to decibels.

$$E = 10\log_{10}\left(\frac{1}{\mathrm{N}}\sum_{n=0}^{N-1}x^2[n]\right). \quad (1)$$

Zero-Crossing Rate. The zero-crossing rate of a frame is defined as the number of times the audio waveform changes its sign in the duration of the frame:

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

Band Energy Ratio. The band energy ratio captures the distribution of the spectral energy in different frequency bands. The spectral energy in a given band is defined as follows: Let $x[n]$ denote one frame of the audio signal ($n = 0, 1, \ldots, \mathrm{N}-1$), and let $X(k)$ denote the Discrete Fourier Transform (DFT) of $x[n]$. The values of $X(k)$ for $k = 0, 1, \ldots, \lfloor K/2 \rfloor - 1$ correspond to discrete frequency bins from 0 to π, with π indicating half of the sampling rate F_s. Let f denote the frequency in Hz. The DFT bin number corresponding to f is given by

$$\hat{f} = \left\lfloor \frac{f}{F_s}\cdot K \right\rfloor. \quad (3)$$

For a given frequency band $[f_L, f_H]$ the total spectral energy in the band is given by

$$E_{f_L,f_H} = \sum_{k=\widehat{f_L}}^{\widehat{f_H}} |X(k)|^2. \quad (4)$$

Finally, if the spectral energies of the two bands $B_1 = [f_L^1, f_H^1]$ and $B_2 = [f_L^2, f_H^2]$ are denoted E_{B_1} and E_{B_2}, respectively, the ratio is computed on a logarithmic scale, as follows:

$$E_{\mathrm{ratio}} = 10\log_{10}\left(\frac{E_{B_1}}{E_{B_2}}\right). \quad (5)$$

We used two features based on band energy ratio—the *low energy ratio*, defined as the ratio between the spectral energy below 70 Hz and the total energy, and the *high energy ratio*, defined as the ratio between the energy above 11 KHz and the total energy, where the sampling frequency is 44 KHz.

Autocorrelation Coefficient. The autocorrelation coefficient is defined as the highest peak in the short-time autocorrelation sequence and is used to evaluate how close the audio signal is to a periodic one. First, the normalized autocorrelation sequence of the frame is computed:

$$\hat{A}(m) = \frac{A(m)}{A(0)} = \frac{\sum_{n=0}^{N-m-1}x[n]\,x[n+m]}{\sum_{n=0}^{N-1}(x[n])^2}. \quad (6)$$

Next, the highest peak of the autocorrelation sequence between m_1 and m_2 is located, where $m_1 = \lfloor 3\cdot F_s/1000 \rfloor$ and $m_2 = \lfloor 16\cdot F_s/1000 \rfloor$ correspond to periods between 3 milliseconds and 16 milliseconds (which is the

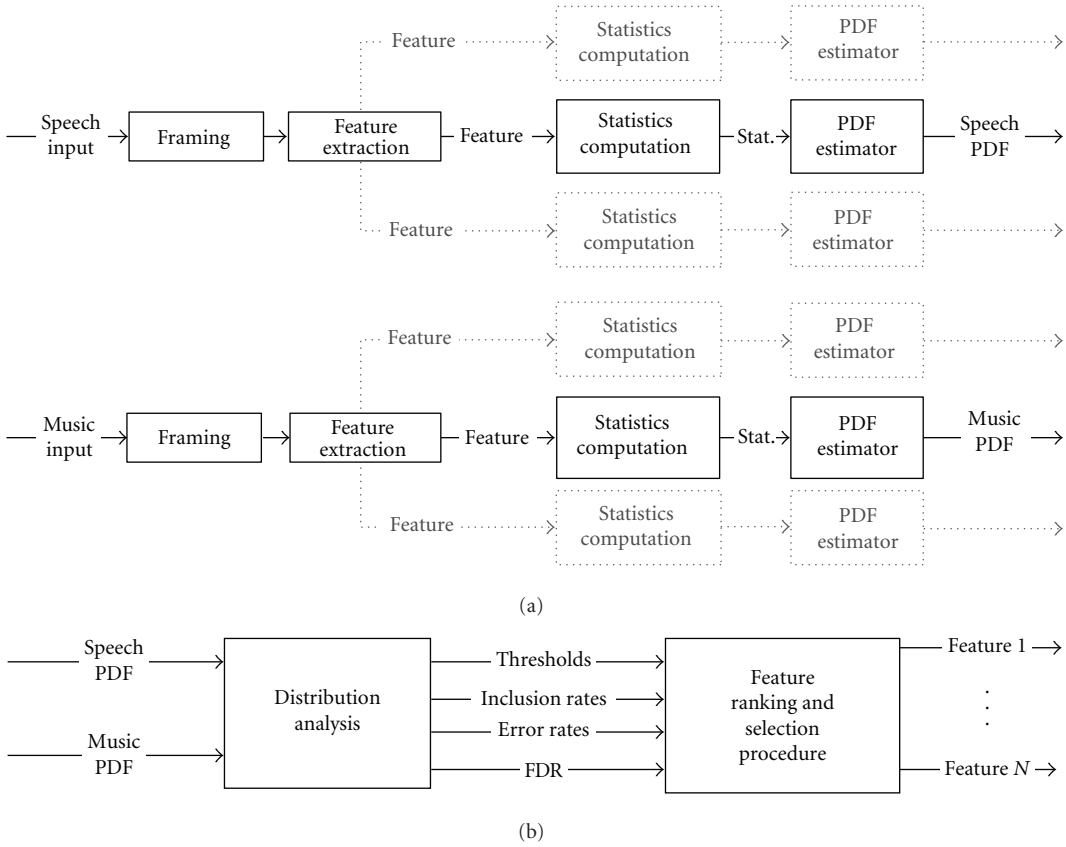

Figure 1: A block diagram of the training phase. (a) Feature extraction and computation of probability density functions for each feature. (b) Analysis of the distributions, setting of optimal thresholds, and selection of the best features for discrimination.

expected fundamental frequency range in voiced speech). The autocorrelation coefficient is defined as the value of this peak:

$$AC = \max_{m=m_1,\ldots,m_2} \left\{\widehat{A}(m)\right\}. \tag{7}$$

Mel Frequency Cepstrum Coefficients. The mel frequency cepstrum coefficients (MFCCs) are known to be a compact and efficient representation of speech data [17, 18]. The MFCC computation starts by taking the DFT of the frame $X(k)$ and multiplying it by a series of triangularly shaped ideal band-pass filters $V_i(k)$, where the central frequencies and widths of the filters are arranged according to the mel scale [19]. Next, the total spectral energy contained in each filter is computed:

$$E(i) = \frac{1}{S_i} \sum_{k=L_i}^{U_i} \left(|X(k)| \cdot V_i(k)\right)^2, \tag{8}$$

where L_i and U_i are the lower and upper bounds of the filter and S_i is a normalization coefficient to compensate for the variable bandwidth of the filters:

$$S_i = \sum_{k=L_i}^{U_i} \left(V_i(k)\right)^2. \tag{9}$$

Finally, the MFCC sequence is obtained by computing the Discrete Cosine Transform (DCT) of the logarithm of the energy sequence $E(i)$:

$$\mathrm{MFCC}(l) = \frac{1}{N} \sum_{i=0}^{N-1} \log\left(E(i)\right) \cdot \cos\left(\frac{2 \cdot \pi}{N}\left(i + \frac{1}{2}\right) \cdot l\right). \tag{10}$$

We computed the first 10 MFC coefficients for each frame. Each individual MFC coefficient is considered a feature. In addition, the MFCC difference vector between neighboring frames is computed, and the Euclidean norm of that vector is used as an additional feature:

$$\Delta\mathrm{MFCC}(i, i-1) = \sqrt{\sum_{l=1}^{10} |\mathrm{MFCC}_i(l) - \mathrm{MFCC}_{i-1}(l)|^2}, \tag{11}$$

where i represents the index of the frame.

Spectrum Rolloff Point. The spectrum rolloff point [6] is defined as the boundary frequency f_r, such that a certain percent p of the spectral energy for a given audio frame is concentrated below f_r:

$$\sum_{k=0}^{f_r} |X(k)| = p \cdot \sum_{k=0}^{K-1} |X(k)|. \tag{12}$$

In our study $p = 85\%$ is used.

Spectrum Centroid. The spectrum centroid is defined as the center of gravity (COG) of the spectrum for a given audio frame and is computed as

$$S_c = \frac{k \cdot \sum_{k=0}^{K-1} |X(k)|}{\sum_{k=0}^{K-1} |X(k)|}. \quad (13)$$

Spectral Flux. The spectral flux measures the spectrum fluctuations between two consecutive audio frames. It is defined as

$$S_f = \sum_{k=0}^{K-1} (|X_m(k)| - |X_{m-1}(k)|)^2, \quad (14)$$

namely, the sum of the squared frame-to-frame difference of the DFT magnitudes [6], where $m-1$ and m are the frame indices.

Spectrum Spread. The spectrum spread [1] is a measure that computes how the spectrum is concentrated around the perceptually adapted audio spectrum centroid, and calculated according to the following:

$$S_{\text{sp}} = \sqrt{\frac{\sum_{k=0}^{K-1} \left(\left[\log_2 (f(k)/1000) - \text{ASC} \right]^2 \cdot |X(k)|^2 \right)}{\sum_{k=0}^{K-1} |X(k)|^2}}, \quad (15)$$

where $f(k)$ is the frequency associated with each frequency bin, and ASC is the perceptually adapted audio spectral centroid, as in [1], which is defined as

$$\text{ASC} = \frac{\sum_{k=0}^{K-1} \log_2 (f(k)/1000) \cdot |X(k)|^2}{\sum_{k=0}^{K-1} |X(k)|^2}. \quad (16)$$

2.4. Computation of Feature Statistics. Each of the above features is computed on frames of duration N, where N is in samples, typically corresponding to 20–40 milliseconds of audio. In order to extract more data to aid the classification, the feature information is collected over longer segments of length S (2–6 seconds). For each such segment and each feature the following statistical parameters are computed:

(i) Mean value and standard deviation of the feature across the segment.

(ii) Mean value and standard deviation of the difference magnitude between consecutive analysis points.

In addition to that, for the zero-crossing rate, the skewness (third central moment, divided by the cube of the standard deviation) and the skewness of the difference magnitude between consecutive analysis frames are also computed.

For the energy we also measure the low short time energy ratio (LSTER, [9]). The LSTER is defined as the percentage of frames within the segment whose energy level is below one third of the average energy level across the segment.

2.5. Threshold Setting and Probability Density Function Estimation. In the learning phase, training data is collected for speech segments and for music segments separately. For each feature and each statistical parameter the corresponding probability density functions (PDFs) are estimated—one for speech segments and one for music segments. The PDFs are computed using a nonparametric technique with a Gaussian kernel function for smoothing.

Five thresholds are computed for each feature, based on the estimated PDFs (Figure 2).

(1) Extreme speech threshold—defined as the value beyond which there are only speech segments, that is, 0% error based on the learning data.

(2) Extreme music threshold—same as 1, for music.

(3) High probability speech threshold—defined as the point in the distribution where the difference between the height of the speech PDF and the height of the music PDF is maximal. This threshold is more permissive than the extreme speech threshold: values beyond this threshold are typically exhibited by speech, but a small error of music segments is usually present. If this error is small enough, and on the other hand a significant percentage of speech segments are beyond this threshold, the feature may be a good candidate for separation between speech and music.

(4) High probability music threshold—same as 3, for music.

(5) Separation threshold—defined as the value that minimizes the joint decision error, assuming that the prior probabilities for speech and for music are equal.

For each of the first four thresholds the following parameters are computed from the training data:

(i) inclusion fraction (I)—the percentage of correct segments that exceed the threshold (for the speech threshold this refers to speech segments, and for the music threshold this refers to music segments);

(ii) error fraction (Er)—the percentage of incorrect segments that exceed the threshold. For speech thresholds these are the music segments, and for music thresholds these are the speech segments. Note that by the definition of the extreme thresholds, their error fractions are 0.

2.6. Feature Selection. With a total of over 20 features computed on the frame level and 4–6 statistical parameters computed per feature on the segment level, the feature space is quite large. More importantly, not all features contribute equally, and some features may be very good in certain aspects, and bad in others. For example, a specific feature may have a very high value of I for the extreme speech threshold, but a very low value of I for the extreme music threshold, making it suitable for one feature group, but not the other.

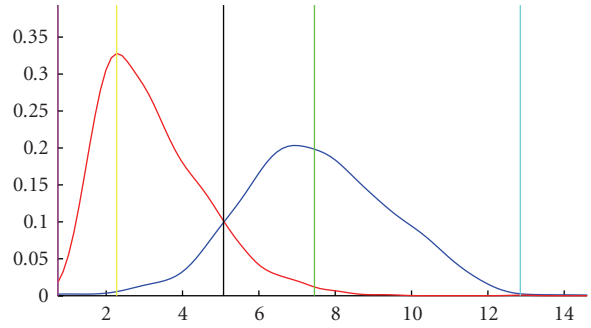

Figure 2: Probability density function for a selected feature (standard deviation of the short-time energy). Left curve: music data; right curve: speech data. Thresholds (left to right): music extreme, music high probability, separation, speech high probability, speech extreme.

The main task here is to select the best features for each classification stage to ensure high discrimination accuracy and to reduce the dimension of the feature space.

The "usefulness" score is computed separately for each feature and each of the thresholds. The feature ranking method is different for each of the three threshold types (extreme speech/music, high probability speech/music, and separation).

For the extreme thresholds, the features are ranked according to the value of the corresponding inclusion fraction I. When I is large, the feature is likely to be more useful for identifying typical speech (resp., music) frames.

For the high probability thresholds, we define the "separation power" of a feature as I^2/Er. This particular definition is chosen due to its tendency, for a given inclusion/error ratio, to prefer features with higher inclusion fraction I. Since independence of features cannot be assumed a priori, we adjusted the selection procedure to consider the mutual correlation between features as well as their separation power. In each stage, a feature is chosen from the pool of remaining features, based on a linear combination of its separation power and its mutual correlation with all previously selected features. This is formalized as follows:

(i) let C be the separation power (for the extreme thresholds we set $C = \mathrm{I}$, whereas for the high probability thresholds we use $C = \mathrm{I}^2/\mathrm{Er}$);

(ii) in the first stage select the first feature that maximizes C: $i_1 = \mathrm{argmax}_j\{C(j)\}$;

(iii) the second feature x_{i_2} is computed so that

$$i_2 = \mathrm{argmax}_j \left\{\alpha \cdot C(j) - \beta \cdot \left|\rho_{i_1,j}\right|\right\}, \quad j \neq i_1, \tag{17}$$

where α and β are weighting factors (typically $\alpha = \beta = 0.5$), determining the relative contributions of C and of the mutual correlation $\rho_{i,j}$:

$$\rho_{i,j} = \frac{\sum_{n=1}^{N} x_{ni} \cdot x_{nj}}{\sqrt{\sum_{n=1}^{N} (x_{ni})^2 \sum_{n=1}^{N} (x_{nj})^2}} \tag{18}$$

(iv) the kth feature x_{i_k} is computed using

$$i_k = \mathrm{argmax}_j \left\{\alpha \cdot C(j) - \frac{\beta}{k-1} \sum_{r=1}^{k-1} \left|\rho_{i_r,j}\right|\right\} \tag{19}$$

$$j \neq i_r, \quad r = 1, 2, \ldots, k-1.$$

For the separation threshold we originally tried the Fisher Discriminant Ratio (FDR, [20]) as a measure of the feature separation power:

$$F_d = \frac{(\mu_S - \mu_M)^2}{\sigma_S{}^2 + \sigma_M{}^2}, \tag{20}$$

where μ_S, μ_M are the mean values and σ_S, σ_M are the standard deviations, for speech and music, respectively. As in the first two stages, the features were selected according to a combination of the FDR and the mutual correlation.

A small improvement was achieved by using the sequential floating (forward-backward) selection procedure detailed in [20, 21]. The advantage of this procedure is that it considers separation power of entire feature vector as a whole, and not just as combination of individual features.

To measure the separation power of the feature vectors, we computed the *scatter matrices* S_w and S_m. S_w is the within-class scatter matrix, defined as the normalized sum of the class covariance matrices S_{SPEECH} and S_{MUSIC}:

$$S_w = \frac{S_{\mathrm{SPEECH}} + S_{\mathrm{MUSIC}}}{2}. \tag{21}$$

S_m is the mixture scatter matrix, defined as the covariance matrix of the feature vector (all samples, both speech and music) around the global mean.

Finally, the separation criterion is defined as

$$J_2 = \frac{|S_m|}{|S_w|}. \tag{22}$$

This criterion tends to take large values when the within-class scatter is small, that is, the samples are well-clustered around the class mean, but the overall scatter is large, implying that the clusters are well-separated. More details can be found in [20].

2.7. Best Features for Discrimination. Using the above selection procedure, the best features for each of the five thresholds were chosen. The optimal number of features in each group is typically selected by trying different combinations in a cross-validation setting over the training set, to achieve the best detection rates. Table 2 lists these features in descending order (best is first). Note that it is possible to take a smaller subset of the features.

As can be seen from the table, certain features, for example, the energy, the autocorrelation, and the 9th MFCC are useful in multiple stages, while some, like the spectral rolloff point, are used only in one of the stages. Also it can be noticed that some of the features considered in the learning phase were found inefficient in practice and were eliminated from the features set in the test phase. As the procedure is automatic, the user does not even have to know which features are selected, and in fact very different sets of features were sometimes selected for different thresholds.

Table 2: Best features for each of the five thresholds.

Threshold type	Features
Extreme speech	(1) 9th MFCC (mean val. of diff. mag.)
	(2) Energy (std)
	(3) 9th MFCC (std of diff. mag.)
	(4) LSTER
Extreme music	(1) High Band Energy Ratio (mean value)
	(2) Spectral rolloff point (mean value)
	(3) Spectral centroid (mean value)
	(4) LSTER
High probability speech	(1) Energy (std)
	(2) 9th MFCC (mean val. of diff. mag.)
	(3) Energy (mean val. of diff. mag.)
	(4) Autocorrelation (std)
	(5) LSTER
High probability music	(1) Energy (mean val. of diff. mag.)
	(2) Energy (std)
	(3) 9th MFCC (std of diff. mag.)
	(4) Autocorrelation (std of diff. mag.)
	(5) ZCR (skewness)
	(6) ZCR (skewness of diff. mag.)
	(7) LSTER
Separation	(1) Energy (std)
	(2) Energy (mean val. of diff. mag.)
	(3) Autocorrelation (std)
	(4) 9th MFCC (std of diff. mag.)
	(5) Energy (std of diff. mag.)
	(6) 9th MFCC (mean val. of diff. mag.)
	(7) 7th MFCC (mean val. of diff. mag.)
	(8) 4th MFCC (std)
	(9) 7th MFCC (std of diff. mag.)
	(10) Autocorrelation (std of diff. mag.)
	(11) LSTER

3. Test Phase and Speech/Music Segmentation

The aim of the test phase is to perform segmentation of a given audio signal into "speech" and "music". There are no prior assumptions on the signal content or the probabilities of each of the two classes. Each segment is classified separately and almost independently of other segments. A block diagram describing the classification algorithm is shown on Figure 3.

3.1. Streaming and Feature Computation. The input signal is divided into consecutive segments, with segment size S. Each segment is further divided into consecutive and overlapping frames with frame size N (as in the learning phase) and hop size h_f, where typically $h_f = \mathrm{N}/2$. For each such frame, the features that were chosen by the feature selection procedure (Section 2.6) are computed. Consequently, the feature statistics are computed over the segment (of length S) and compared to the predefined thresholds, which were also set during the learning phase. This comparison is used as a basis for classification of the segment either as speech or as music, as described below.

In order to provide better tracking of the changes in the signal, the segment hop size h_s, which represents the resolution of the decision, is set to a small fraction of the segment size (typically 100–400 ms).

For the evaluation tests (Section 4) the following values were used: N = 40 milliseconds, $h_f = 20$ milliseconds, $S = 4$ seconds.

3.2. Initial Classification. The initial decision is carried out for each segment independently of other segments. In this decision each segment receives a grade between -1 and 1, where positive grades indicate music and negative grades indicate speech, and the actual value represents the degree of certainty in the decision (e.g., ± 1 means speech/music with high certainty).

For each of the five threshold types computed in the learning phase (see Section 2.5) a set of features is selected, that are compared to their corresponding thresholds. The features for each set are chosen according to the feature selection procedure (see Section 2.6). After comparing all features to the thresholds, the values are computed as shown in Table 3 .

A segment receives a grade of $D_i = -1$ if one of the following takes place:

(i) $S_X > 0$ and $M_X = M_H = 0$ (i.e., at least one of the features is above its corresponding extreme speech threshold; whereas no feature surpasses the extreme or the high probability music thresholds);

(ii) $S_X > 1$ and $M_X = 0$ (we allow $M_H > 0$ if S_X is at least 2);

(iii) $S_H > \alpha |A_S|$ and $M_H = 0$, where $\alpha \in (0.5, 1)$, A_S is the set of all features used with the high probability speech threshold (i.e., if a decision cannot be made using the extreme thresholds, we demand a large majority of the high probability thresholds to classify the segment as speech with high certainty).

The above combination of rules allows classifying a segment as speech in cases where its feature vector is located far inside the speech half-space along some of the feature axes, and at the same time, is not far inside the music half-space along any of the axes. In such cases we can be fairly certain of the classification. It is expected that if the analyzed segment is indeed speech, it will rarely exhibit any features above the extreme or high probability music thresholds.

Similarly, a segment gets a grade of $D_i = 1$ (that is considered as music with high certainty) if one of the following takes place:

(i) $M_X > 0$ and $S_X = S_H = 0$,

(ii) $M_X > 1$ and $S_X = 0$,

(iii) $M_H > \alpha |A_M|$, and $S_H = 0$, where $\alpha \in (0.5, 1)$, A_M is the set of all features used with the high probability speech threshold.

Table 3

$S_X(M_X)$	No. of features in the extreme speech (music) set that surpass their thresholds
$S_H(M_H)$	No. of features in the high probability speech (music) set that surpass their thresholds
$S_P(M_P)$	No. of features in the separation set that are classified as speech (music)

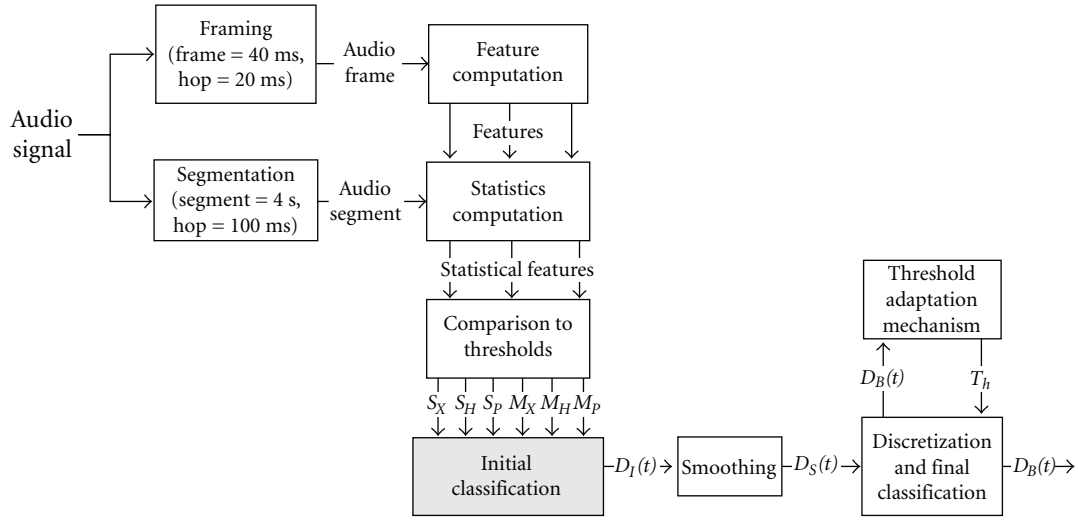

Figure 3: General block diagram of the classficication algorithm.

If none of the above applies, the decision is based on the separation threshold as follows:

$$D_i = \frac{M_P - S_P}{|A_P|}, \tag{23}$$

where A_P is the set of features used with the separation threshold. Note that $0 \le M_P,\ S_P \le |A_P|$, and $M_P + S_P = |A_P|$, so the received grade is always between -1 and 1, and in some way reflects the certainty with which the segment can be classified as speech or music.

This procedure of assigning a grade to each segment is summarized in Figure 4.

3.3. Smoothing and Final Classification. In most audio signals, speech-music and music-speech transitions are not very common (for instance, musical segments are usually at least one minute long, and typically several minutes or longer). When the classification of an individual segment is based solely on data collected from that segment (as described above), erroneous decisions may lead to classification results that alternate more rapidly than normally expected. To avoid this, the initial decision is smoothed by a weighted average with past decisions, using an exponentially decaying "forgetting factor," which gives more weight to recent segments:

$$D_s(t) = \frac{1}{F}\sum_{k=0}^{K} D_i(t-k)\, e^{-k/\tau}, \tag{24}$$

where K is the length of the averaging period, τ is the time constant, and $F = \sum_{k=0}^{K} e^{-k/\tau}$ is the normalizing constant. Alternatively, we tried a median filter for the smoothing. Both approaches achieved comparable results.

Following the smoothing procedure, discretization to a binary decision is performed as follows: a threshold value $0 < T < 1$ is determined. Values above T or below $-T$ are set to 1 or -1, respectively, whereas values between $-T$ and T are treated according to the current trend of $D_s(t)$, that is, if $D_s(t)$ is on the rise, $D_b(t) = 1$ and $D_b(t) = -1$ otherwise, where $D_b(t)$ is the binary desicion.

Additionally, a four-level decision is possible, where values in $(-T, T)$ are treated as "weakly speech" or "weakly music." The four-level decision mode is useful for mixed content signals, which are difficult to firmly classify as speech or music.

To avoid erroneous transitions in long periods of either music or speech, we adapt the threshold over time as follows: let $T_h(t)$ be the threshold at time t, and $D_b(t)$, $D_b(t-1)$ be the binary decision values of the current and the previous time instants, respectively. We have the following:

$$\begin{aligned} &\text{if} \quad D_b(t) = D_b(t-1), \\ &\text{then} \quad T_h(t) \Leftarrow \max\left(M \cdot T_h(t), T_{\min}\right), \\ &\text{else} \quad T_h(t) \Leftarrow T_{\text{init}}, \end{aligned} \tag{25}$$

where $0 < M < 1$ is a predefined multiplier, T_{init} is the initial value of the threshold, and $T_{\min}$ is a minimal value, which is set so that the threshold will not reach a value of zero. This mechanism ensures that whenever a prolonged music (or speech) period is processed, the absolute value of the threshold is slowly decreased towards the minimal value. When the decision is changed, the threshold value is reset to T_{init}.

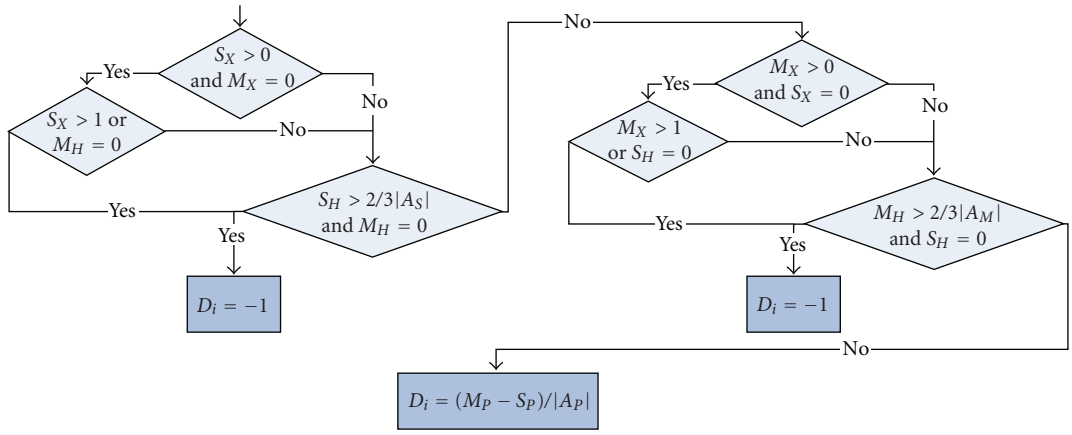

Figure 4: A schematic flowchart of the initial decision made for each segment in the test phase.

Table 4: Total correct identification rate for music and for speech.

	Duration (seconds)	correct %
Music	78028	97.8%
Speech	42430	99.4%

4. Evaluation and Results

4.1. Speech and Music Material. The speech material for the evaluation was collected from various free internet speech databases, such as the International Dialects of English Archive (http://www.ku.edu/~idea/), the World Voices Collection (http://www.world-voices.com/), the Indian institute of scientific heritage (http://www.iish.org/), and others. The test set contained both clean speech and noisy speech. Most of the data were MP3 files at bitrates of 128Kbps and higher or uncompressed PCM WAV files.

The music material for the test set contained music of different genres, such as classical, folk, pop, rock, metal, new age, and others. The source material was either in cd audio format (uncompressed) or mp3 (64 kbps and higher).

4.2. Evaluation of Long Tracks. The evaluation was performed on approximately 12 hours of speech and 22 hours of music from the above material. The results are summarized in Table 4. For music, the correct identification rate is 97.8% (98.9% excluding the electronic music). A breakdown of music detection rates into different genres is summarized in Table 5. It can be noticed that in electronic music and pop, the scores are lower than those of the other genres. For speech an identification rate of 99.4% has been achieved.

4.3. Evaluation of Alternating Sections. In this test, a signal was constructed from 80 alternating sections of speech (40 sections) and music (40 sections), each 30 seconds long. The proportion of correct detection is shown in Table 6. Since the first six seconds of each section are defined as transition time, only the steady-state portion of the signal is taken into account, that is, 960 seconds of each type, speech and music.

Table 5: Breakdown of identification rates for different musical genres.

Music genres	Duration (sec)	% correct
Classical	16732	99.77%
Folk	2475	98.55%
Pop	5999	93.37%
Rock	17932	99.17%
Metal	6827	99.53%
New age	5143	99.86%
Electronic	17526	93.94%
Misc.	5394	99.96%

Table 6: Correct identification rates for alternating sections.

Signal	Speech	Music
Total duration(sec)	960 seconds	960 seconds
Total error(sec)	7.01	13.73
Correct identification	**98.8**%	**98.6**%
Average latency	4.01 (std = 0.51)	5.13 (std = 0.71)

Most of the sections were identified perfectly. Among the speech sections only 2 out of 40 sections were problematic, where in one case the speech was excerpted from a TV soundtrack with background music, and the other contained strong background noise. Among the music, only 3 out of 40 sections had errors, 2 of them containing a strong vocal element. The high success rates are even more satisfying considering the fact that most of the speech sections were with a background noise.

4.4. Gradual Transitions. In radio and TV broadcasts, gradual speech-music and music-speech transitions are often present, for example between songs and the voice of the announcer. Because our algorithm supports soft speech/music classifications, it can recognize such gradual transitions. Figure 5 shows a transition between music and speech at the end of a song. After the first 8 seconds the music

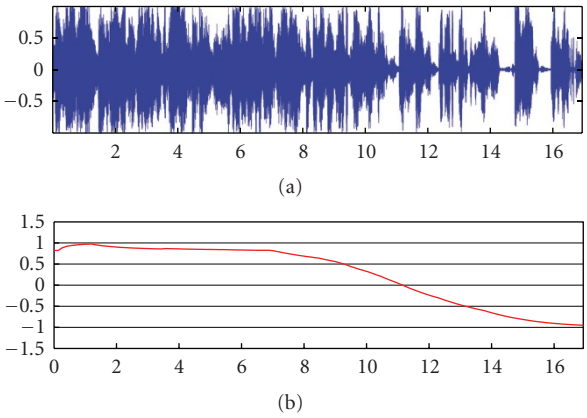

Figure 5: Gradual transition between music and speech. (a) The audio section (17 seconds). (b) The continuous speech/music classification curve.

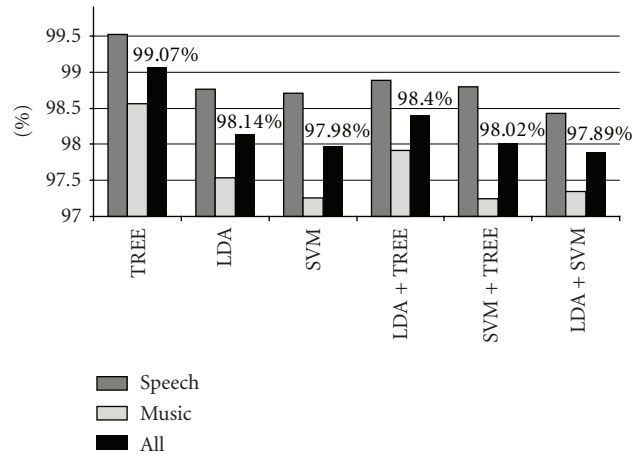

Figure 6: Comparison of five approaches: the proposed decision tree algorithm (TREE), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and a combination of the SVM/LDA classifiers with the decision tree, in which the classifiers take the place of the separation stage. The best detection rates achieved with each classifier are shown, for speech, music and overall.

fades out and the announcer starts speaking over it, and for the last 4 seconds there is only the announcer's voice.

4.5. Comparison with Other Approaches. In order to evaluate the algorithm in the context of existing work, we compared it to traditional approaches, such as the Support Vector Machine (SVM, [22, 23]) and the Linear Discriminant Analysis (LDA, [20]). For the training, our original training databases were used, and for the testing we chose the Scheirer-Slaney database, provided to us by Dan Ellis. The database consists of radio recordings, with 20 minutes of speech and 20 minutes of music. This allows us to compare our results directly with past studies that used the same database such as [6, 12, 16].

In our comparative tests, the optimal feature sets for each classifier (the proposed algorithm, SVM and LDA) were chosen using the automatic feature selection procedure detailed in Section 2.6. For the SVM the Gaussian (radial basis function) kernel was used, as it has some advantages compared to the linear SVM and others. The data for the SVM training was automatically scaled to zero mean and unit variance, and the free parameters were selected using grid search, and evaluated through cross-validation [24, 25].

The SVM and LDA algorithms were tested individually, as stand-alone classifiers, as well as in the general framework of our decision tree, replacing the separation stage of the initial classification (Section 3.2), that is, D_i was assigned according to the output of the classifier, and the postprocessing techniques described in Section 3.3 were applied to it. Additionally, we tested a combination of the two techniques—with LDA being used for feature selection, and SVM for the classification.

The best results achieved for each classifier are shown in Figure 6. It can be seen that in all cases, the results are very good, and that the success rates attained by our algorithm are comparable and slightly better, with 99.07% correct detection. We further see that the usage of the full three-level decision tree, and the decision postprocessing techniques slightly improve the success rates of both SVM and LDA.

Table 7: Effect of number of features on performance of SVM classifier.

Number of features	Correct identifaction rate
9	96.94%
28	95.90%

The overall error rate of our decision tree algorithm (0.93%) is comparable, and somewhat surpasses the 1.3–1.5% rates reported in [6, 12, 16], although it should be noticed that, unlike in those studies, here the database was solely for testing, whereas the training was done on different data.

4.6. The Importance of Feature Selection. To verify the contribution of the feature selection procedure to the performance of the classifier, we conducted two experiments with the SVM classifier, using a base feature set of 28 different features. In the one experiment, the best 9 features were selected, and in the other one, all features were used. The results are summarized in Table 7. It can be seen that feature selection allows higher performance to be achieved.

5. Conclusion

In this paper we presented an algorithm for speech/music classification and segmentation, based on a multistage statistical approach. The algorithm consists of a training stage with an automatic selection of the most efficient features and the optimal thresholds, and a processing stage, during which an unknown audio signal is segmented into speech or music through a sieve-like rule-based procedure.

Tested on a large test set of about 34 hours of audio from different resources, and of varying quality, the algorithm demonstrated very high accuracy in the classification (98%, and over 99% under certain conditions), as well as fast

response to transitions from music to speech and vice versa. The test database is among the largest used in similar studies and suggests that the algorithm is robust and applicable in various testing conditions.

The algorithm was tested against standard approaches such as SVM and LDA, and showed comparable, often slightly better, results. On the other hand, the decision tree uses simple heuristics, which are easy to implement in a variety of hardware and software frameworks. This together with its fast and highly localized operation makes it suitable for real-time systems and for consumer-grade audio processing applications. A commercial product based on the presented algorithm is currently being developed by Waves Audio, and a provisional patent has been filed.

Another advantage of the presented algorithm is the generic training procedure, which allows testing any number of additional features and selecting the optimal feature set using an automatic and flexible procedure.

One possible direction for improvement of the algorithm is increasing the segmentation accuracy in speech-music and music-speech transitions. In an offline application this could be achieved, without harming the overall identification rate, by postprocessing the boundary regions using smaller segments.

Acknowledgments

The authors would like to thank I. Neoran, Director of R&D of Waves Audio, for his valuable discussions and ideas, and Y. Yakir for technical assistance. They also thank D. Ellis for providing them the database they used in part of the experiments. Finally, they would like to thank the reviewers for many good remarks which helped them improve the paper.

References

[1] J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification," *Journal of the Audio Engineering Society*, vol. 52, no. 7-8, pp. 724–739, 2004.

[2] A. Bugatti, A. Flammini, and P. Migliorati, "Audio classification in speech and music: a comparison between a statistical and a neural approach," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 4, pp. 372–378, 2002.

[3] J. E. Muñoz-Expósito, S. G. Galán, N. R. Reyes, P. V. Candeas, and F. R. Peña, "A fuzzy rules-based speech/music discrimination approach for intelligent audio coding over the Internet," in *Proceedings of the 120th Audio Engineering Society Convention (AES '06)*, Paris, France, May 2006, paper number 6676.

[4] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 2, pp. 993–996, Atlanta, Ga, USA, May 1996.

[5] Z. Liu, J. Huang, Y. Wang, and I. T. Chen, "Audio feature extraction and analysis for scene classification," in *Proceedings of the 1st IEEE Workshop on Multimedia Signal Processing (MMSP '97)*, pp. 343–348, Princeton, NJ, USA, June 1997.

[6] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 2, pp. 1331–1334, Munich, Germany, April 1997.

[7] J. Ajmera, I. A. McCowan, and H. Bourlard, "Robust HMM-based speech/music segmentation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 1, pp. 297–300, Orlando, Fla, USA, May 2002.

[8] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 1, pp. 149–152, Phoenix, Ariz, USA, March 1999.

[9] 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.

[10] J. T. Foote, "A similarity measure for automatic audio classification," in *Proceedings of the AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, Stanford, Calif, USA, March 1997.

[11] T. Zhang and C.-C. J. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 6, pp. 3001–3004, Phoenix, Ariz, USA, March 1999.

[12] G. Williams and D. P. W. Ellis, "Speech/music discrimination based on posterior probability features," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, pp. 687–690, Budapest, Hungary, September 1999.

[13] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 6, pp. 2445–2448, Istanbul, Turkey, June 2000.

[14] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol. 40, no. 3, pp. 351–363, 2003.

[15] J. G. A. Barbedo and A. Lopes, "A robust and computationally efficient speech/music discriminator," *Journal of the Audio Engineering Society*, vol. 54, no. 7-8, pp. 571–588, 2006.

[16] E. Alexandre, M. Rosa, L. Caudra, and R. Gil-Pita, "Application of Fisher linear discriminant analysis to speech/music classification," in *Proceedings of the 120th Audio Engineering Society Convention (AES '06)*, Paris, France, May 2006, paper number 6678.

[17] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[18] T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2001.

[19] S. S. Stevens and J. Volkman, "The relation of pitch to frequency: a revised scale," *American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.

[20] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, New York, NY, USA, 4th edition, 2003.

[21] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[22] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK, 2000.

[23] C. J. C. Burges, "Simplified support vector decision rules," in *Proceedings of the 13th International Conference on Machine Learning (ICML '96)*, pp. 71–77, Bari, Italy, July 1996.
[24] C. Staelin, "Parameter selection for support vector machines," Tech. Rep. HPL-2002-354, HP Laboratories Israel, Haifa, Israel, 2003.
[25] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Tech. Rep., National Taiwan University, Taipei, Taiwan, 2003.