EURASIP Journal on Applied Signal Processing 2002:4, 372–378

Audio Classification in Speech and Music: A Comparison Between a Statistical and a Neural Approach

Alessandro Bugatti

Department of Electronics for Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy
Email: bugatti@ing.unibs.it

Alessandra Flammini

Department of Electronics for Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy
Email: flammini@ing.unibs.it

Pierangelo Migliorati

Department of Electronics for Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy
Email: pier@ing.unibs.it

Received 27 July 2001 and in revised form 8 January 2002

We focus the attention on the problem of audio classification in speech and music for multimedia applications. In particular, we present a comparison between two different techniques for speech/music discrimination. The first method is based on zero crossing rate and Bayesian classification. It is very simple from a computational point of view, and gives good results in case of pure music or speech. The simulation results show that some performance degradation arises when the music segment contains also some speech superimposed on music, or strong rhythmic components. To overcome these problems, we propose a second method, that uses more features, and is based on neural networks (specifically a multi-layer Perceptron). In this case we obtain better performance, at the expense of a limited growth in the computational complexity. In practice, the proposed neural network is simple to be implemented if a suitable polynomial is used as the activation function, and a real-time implementation is possible even if low-cost embedded systems are used.

Keywords and phrases: speech/music discrimination, indexing of audio-visual documents, neural networks, multimedia applications.

1. INTRODUCTION

Effective navigation through multimedia documents is necessary to enable widespread use and access to richer and novel information sources.

Design of efficient indexing techniques to retrieve relevant information is another important requirement. Allowing for possible automatic procedures to semantically index audio-video material represents therefore a very important challenge. Such methods should be designed to create indices of the audio-visual material, which characterize the temporal structure of a multimedia document from a semantic point of view.

The International Standard Organization (ISO) started in October 1996 a standardization process for the description of the content of multimedia documents, namely MPEG-7: the "Multimedia Content Description Interface" [1, 2]. However, the standard specifications do not indicate methods for the automatic selection of indices.

A possible mean is to identify series of consecutive segments, which exhibit a certain coherence, according to some property of the audio-visual material. By organizing the degree of coherence, according to more abstract criteria, it is possible to construct a hierarchical representation of information, so as to create a Table of Content description of the document. Such description appears quite adequate for the sake of navigation through the multimedia document, thanks to the multi-layered summary that it provides [3, 4].

Traditionally, the most common approach to create an index of an audio-visual document has been based on the automatic detection of the changes of camera records and the types of involved editing effects. This kind of approach has generally demonstrated satisfactory performance and lead to a good low-level temporal characterization of the visual content. However, the reached semantic level remains poor since the description is very fragmented considering the high number of shot transitions occurring in typical audio-visual programs.

Alternatively, there have been recent research efforts to base the analysis of audio-visual documents by a joint audio and video processing so as to provide for a higher-level organization of information [5, 6, 7, 8]. In [7, 8] these two sources of information have been jointly considered for the identification of simple scenes that compose an audio-visual program. The video analysis associated to cross-modal procedures can be very computationally intensive (by relying, e.g., on identifying correlation between nonconsecutive shots).

We believe that audio information carries out by itself a rich level of semantic significance, and this paper focuses on this issue.

In particular, we propose and compare two simple speech/music discrimination schemes for audio segments.

The first approach, based mainly on Zero Crossing Rate (ZCR) and Bayesian classification, is very simple from a computational complexity point of view, and gives good results in case of pure music or speech. Some problems arises when the music segment contains also some speech superimposed on music, or strong rhythmic components.

To overcome this problem, we propose an alternative method, that uses more features and is based on neural networks (specifically a Multi Layer Perceptron, MLP). In this case we obtain better performance, at the expense of an increased computational complexity. Anyway, the proposed neural network is simple to be implemented if a suitable polynomial is used as the activation function, and a real-time implementation is possible even if low-cost embedded systems are used.

The paper is organized as follows. Section 2 is devoted to a brief description of the solutions for speech/music discrimination presented in the literature. The proposed algorithms are described, respectively, in Sections 3 and 4, whereas in Section 5 we report and discuss the experimental results. Some concluding remarks are given in Section 6.

2. STATE OF THE ART SOLUTIONS

In this section, we focus the attention on the solutions proposed in the literature to the problem of speech/music discrimination.

Saunders [9] proposed a method based on the statistical parameters of the ZCR, plus a measure of the short time energy contour. Then, using a multivariate Gaussian classifier, he obtained a good percentage of class discrimination. This approach is successful for discriminating speech from music on a broadcast FM radio program, and it allows achieving the goal for the low computational complexity and for the relative homogeneity of this type of audio signal.

Scheirer and Slaney [10] developed another approach to the same problem, which exploits different features still achieving similar results. Even in this case the algorithm achieves real-time performance and uses time domain features (short-term energy, ZCR) and frequency domain features (4 Hz modulation energy, spectral rolloff point, centroid and flux, . . .), extracting also their variance in one second segments. In this case, they use some methods for the classification (Gaussian mixture model, K-nearest neighbor), and they obtain similar results.

Foote [11] adopted a technique purely data-driven, and he did not extract subjectively "meaningful" acoustic parameters. In his work, the audio signal is first parameterized into Mel-scaled Cepstral coefficients plus an energy term, obtaining a 13-dimensional feature vector (12 MFCC plus energy) at a 100 Hz frame rate. Then using a tree-based quantization the audio is classified into speech, music, and novocal sounds.

Saraceno and Leonardi [7], and Zhang and Kuo [12] proposed more sophisticated approaches to achieve a finest decomposition of the audio stream. In both works the audio signal is decomposed at least in four classes: silence, music, speech, and environmental sounds.

In the first work, at the first stage, a silence detector is used, which divides the silence frames from the others with a measure of the short time energy. It considers also their temporal evolution by dynamic updating of the statistical parameters, and by means of a finite state machine, to avoid misclassification errors. Hence, the three remaining classes are divided using autocorrelation measures, local as well as contextual, and the ZCR, obtaining good results, where misclassifications occur mainly at the boundary between segments belonging to different classes.

In [12] the classification is performed at two levels: a coarse and a fine level. For the first level, it is used a morphological and statistical analysis of the energy function, average ZCR and the fundamental frequency. Then a rule-based heuristic procedure is proposed to classify audio signals based on these features. At the second level, a further classification is performed for each type of sounds. Because this finest classification is inherently semantic, for each class a different approach could be used. The results for the coarse level show a good accuracy, and misclassification usually occurs in hybrid sounds, which contains more than one basic type of audio.

Liu et al. [13] used another kind of approach, because their aim was to analyze the audio signal for a scene classification of TV programs. The features selected for this task are both in time and frequency domain, and they are meaningful for the scene separation and classification. These features are: no-silence ratio, volume standard deviation, volume dynamic range, frequency component at 4 Hz, pitch standard deviation, voice of music ratio, noise or unvoiced ratio, frequency centroid, bandwidth and energy in 4 sub-bands of the signal. Feedforward neural networks are used successfully as pattern classifiers in this work. The recognized classes are advertisement, basketball, football, news, weather forecasts, and the results show the usefulness of using audio features for the purpose of scene classifications.

An alternative approach in audio data partitioning consists in a supervised partitioning. The supervision concerns the ability to train the models of the various clusters considered in the partitioning. In literature, the Gaussian Mixture Models (GMM) [14] are frequently used to train the models of the chosen clusters. From a reference segmented and labeled database, the GMMs are trained on acoustic data for modeling characterized clusters (e.g., speech, music, and background). The great variability of noises (e.g., rumbling, explosion, creaking), and of music (e.g., classic, pop) observed on the audio-video databases (e.g., broadcast news, movie films) makes difficult to select a suitable training strategy of the models of the various clusters characterizing these sounds. The main problem to train the models is the segmentation/labeling of large audio databases allowing a statistical training. So long as the automatic partitioning is not perfect, the labeling of databases is time consuming of human experts. To avoid this cost and to cover the processing of any audio document, the characterization must be generic, and an adaptation of the techniques of data partitioning on the audio signals is required to minimize the training of the various clusters of sounds.

In general, these algorithms suffer some performance degradation when the music segment contains some speech superimposed on music, or strong rhythmic components. As previously mentioned, in our work, to overcome these problems, we propose a method that is based on neural networks, that gives good performance also in these specific cases, at the expense of a limited growth in the computational complexity. The performance of this method are then compared to that obtained using a statistical approach based on ZCR and Bayesian classification.

3. ZCR WITH BAYESIAN CLASSIFIER

As previously mentioned, several researchers assume an audio model composed of four classes: silence, music, speech, and noise.

In this work, we focus the attention on the specific problem of audio classification in music and speech, assuming that the silence segments have already been identified using, for example, the method proposed in [8].

For this purpose, we use a speech characteristic to discriminate it from the music; the speech shows a very regular structure where the music does not show it. Indeed, the speech is composed of a succession of vowels and consonants: while the vowels are high energy events with most of the spectral energy contained at low frequencies, the consonant are noise-like, with the spectral energy distributed more towards the higher frequencies.

Saunders [9] used the ZCR, which is a good indicator of this behavior, as shown in Figure 1.

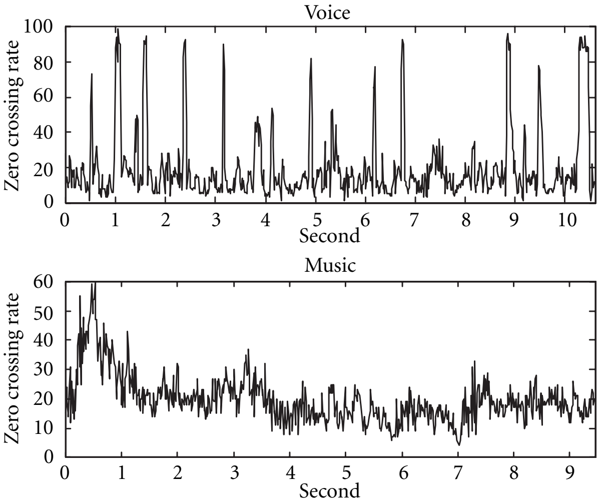

Figure 1: The ZCR behaviour for voice and music segments.

In our algorithm, depicted in Figure 2, the audio file is partitioned into segments of 2.04 second; each of them is composed of 150 consecutive nonoverlapping frames. These values allow a statistical significance of the frame number and, using a 22050 Hz sample frequency, each frame contains 300 samples, which is an adequate tradeoff between the quasi-stationary properties of the signal and a sufficient length to accurately evaluate the ZCR. For every frame, the value of the ZCR is then calculated using the definition given in [9].

These 150 values of the ZCR are then used to estimate the following statistical measures:

- *variance*: which indicates the dispersion with respect to the mean value;
- *third-order moment*: which indicates the degree of skewness with respect to the mean value;
- difference between the number of ZCR samples, which are above and below the mean value.

Each segment of 2.04 seconds is thus associated with a 3-dimensional vector.

To achieve the separation between speech and music using a computationally efficient implementation, a multivariate Gaussian classifier has been used. A set of about 400 4-second-long audio sample, equally distributed between speech and music, have been used to characterize the classifier. At the end of this step we obtain a set of consecutive segments labeled like speech or no-speech.

The next step is justified by an empirical observation: the probability to observe a single segment of speech surrounded of music segments is very low, and vice versa. Therefore, a simple regularization procedure is applied to properly set the labels of these spurious segments.

The boundaries between segments of different classes are placed in fixed positions, inherently to the nature of the ZCR algorithm. Obviously these boundaries are not placed in a sharp manner, thus a fine-level analysis of the segments across the boundaries is needed to determine a sharp placement of them. In particular, the ZCR values of the neighboring segments are processed to identify the exact position of the transition between speech and music signal. A new signal is obtained from these ZCR values, applying this function

$$y[n] = \frac{1}{P}\sum_{m=n-P/2}^{n+P/2}\left(x[m]-\bar{x}_n\right)^2 \quad \text{with } \frac{P}{2} < n < 300 - \frac{P}{2}, \tag{1}$$

where $x[n]$ is the nth ZCR value of the current segment, and $\bar{x}_n$ is defined as

$$\bar{x}_n = \frac{1}{P}\sum_{m=n-p/2}^{n+P/2} x[m]. \tag{2}$$

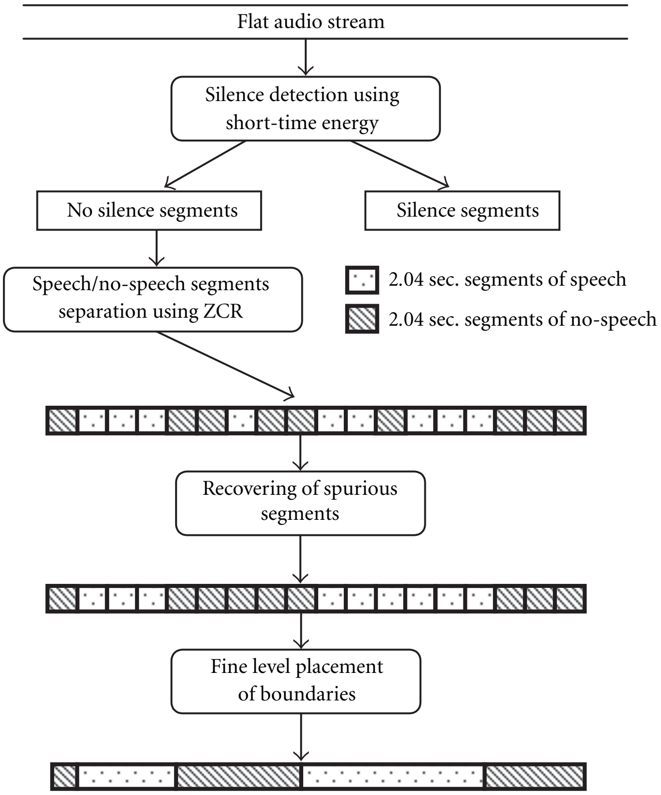
Flat audio stream
Silence detection using short-time energy
No silence segments
Silence segments
Speech/no-speech segments separation using ZCR
2.04 sec. segments of speech
2.04 sec. segments of no-speech
Recovering of spurious segments
Fine level placement of boundaries

Figure 2: The proposed ZB algorithm.

Therefore, $y[n]$ is an estimation of the ZCR variance in a short window. A low-pass filter is then applied to this signal to obtain a smoother version of it, and finally a peak extractor is used to identify the transition between speech and music.

4. NEURAL NETWORK CLASSIFIER

The second approach we propose is based on a Multi-Layer Perceptron (MLP) network [15].

The MLP has been trained using five classes of audio traces, supposing other audio sources, as silence or noise, to be previously removed. The classes of audio traces considered have been, namely: instrumental music without voice, as Beethoven symphony no. 6 (class labeled as "Am"), melodic songs, as "My heart will go on" from Titanic (class labeled as "Bm"), rhythmic songs, as rap music or Dire Straits song "Sultans of swing" (class labeled as "Cm"), pure speech (class labeled as "Av"), and speech superimposed on music (class labeled as "Bv"), as commercials.

In the literature main features have been suggested for speech/music discrimination, for example, see [16]. In this work, we have analysed more than 30 features, and eight of them have been selected as the neural network inputs. These parameters have been computed considering 86 frames by 1024 points each (sampling frequency $f_s = 22050$ Hz), with a total observing time of about 4 seconds.

To test the effectiveness of the various features, and to train the MLP, a set of about 400 4-second-long audio samples have been considered belonging to the five classes labeled as Am, Bm, Cm, Av, Bv, and equally distributed between speech (Av, Bv) and music (Am, Bm, Cm). The discrimination power of the selected features has been firstly evaluated by computing the index α, defined by (3), for each feature P_j, with $j = 1$ to 8, where μ_m and σ_m are, respectively, the mean value and standard deviation of parameter P_j for music samples, and μ_v and σ_v are the same for speech. If parameter P_j follows a Gaussian distribution, an α-value equal to 1 yields to a statistical classification error of about 15%. α-values between 0.7 and 1 result for the selected features

$$\alpha = \left|\frac{\mu_m - \mu_v}{\sigma_m + \sigma_v}\right|. \tag{3}$$

A short description of the eight selected features follows. Parameter P_1 is the spectral flux, as suggested in [10]. It indicates how rapidly changes the frequency spectrum, with particular attention to the low frequencies (up to 2.5 kHz), and it generally assumes higher values for speech.

Parameters P_2 and P_3 are related to the short-time energy [17]. Function $E(n)$, with $n = 1$ to 86, is computed as the sum of the square value of the previous 1024 signal samples. A fourth-order high-pass Chebyshev filter is applied with about 100 Hz as the cutting frequency. Parameter P_2 is computed as the standard deviation of the absolute value of the resulting signal, and it is generally higher in speech. Parameter P_3 is the minimum of the short-time energy and it is generally lower in speech, due to the pauses that occur among words or syllables.

Parameters P_4 and P_5 are related to the cepstrum coefficients, evaluated using

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X\left(e^{jw}\right)\right| e^{jwn} dw. \tag{4}$$

Cepstrum coefficients $c_j(n)$, suggested in [18] as good speech detectors, have been computed for each frame, then the mean value $c_\mu(n)$ and the standard deviation $c_\sigma(n)$ have been calculated, and parameters P_4 and P_5 result as indicated in

$$\begin{aligned} P_4 &= c_\mu(9)\cdot c_\mu(11)\cdot c_\mu(13),\\ P_5 &= c_\sigma(2)\cdot c_\sigma(5)\cdot c_\sigma(9)\cdot c_\sigma(12). \end{aligned} \tag{5}$$

Parameter P_6 is related to the centroid that is computed starting from the spectrum module of each frame.

Parameter P_6 is the product of the mean value by the standard deviation computed by the 86 values of barycentre. In fact, due to the speech discontinuity, standard deviation makes this parameter more distinctive.

Parameter P_7 is related to the ratio of the high-frequency power spectrum (7.5 kHz $< f <$ 11 kHz) to the whole power spectrum. The speech spectrum is usually considered up to 4 kHz, but the lowest limit has been increased to consider signals with speech over music. To consider the speech discontinuity and increase the discrimination between speech and music, P_7 is the ratio of the mean value to the standard deviation obtained by the 86 values of the relative high-frequency power spectrum.

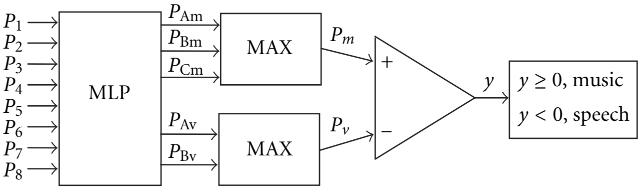

Figure 3: The decision algorithm.

Parameter P_8 is the syllabic frequency [10] computed starting from the short-time energy calculated on 256 samples ($\approx$ 12 ms) instead of 1024. A 5-taps median filter has filtered this signal, and the computed syllabic frequency (P_8) is the number of peaks detected in 4 seconds. As it is known, music should present a greater number of peaks [10].

The proposed MLP has eight input, corresponding to the normalized features $P_1 \div P_8$, fifteen hidden neurons, five output neurons, corresponding to the five considered classes, and uses normalized sigmoid activation function.

The 400 audio samples, that have been used also for the ZCR with Bayesian classifier, have been divided into three sets: training (200 samples), validation (100 samples), and test (100 samples). Each sample is formatted as $\{P_1 \div P_8, P_{\text{Av}}, P_{\text{Bv}}, P_{\text{Am}}, P_{\text{Bm}}, P_{\text{Cm}}\}$, where P_{Av} is the probability that sample belongs to class Av.

The goal is to distinguish between speech and music and not to identify the class; for this purpose a different and more complex set of parameters should be designed. To perform the proposed binary classification, target has been assigned with "1" to the selected class, "0" to the farest class, a value between 0.8 and 0.9 to the similar classes, and a value between 0.1 and 0.2 to the other classes. For instance, if a sample of Bm (melodic songs) is considered, $P_{\text{Bm}} = 1$, $P_{\text{Am}} = P_{\text{Cm}} = 0.8$ because music is dominant, $P_{\text{Bv}} = 0.2$ because it is anyway a mix of music and voice, and $P_{\text{Av}} = 0.1$, because the selected sample contains voice.

If a pure music sample is considered (class Am), $P_{\text{Am}} = 1$, $P_{\text{Bm}} = P_{\text{Cm}} = 0.8$ because it is a mix of music and voice where music is dominant, $P_{\text{Bv}} = 0.1$ because it contains music, and $P_{\text{Av}} = 0$, because pure speech is the farest class. In fact, classifying the speech over music as speech inclines the MLP to classify as speech some rhythmic songs: by adjusting the sample target it is possible to incline to one side or another the MLP response.

The MLP has been trained using the Levenberg-Marquardt method [19] with a starting value of μ equal to 1000 (slow and accurate behavior). The decision algorithm is depicted in Figure 3.

The mean square error related to the 400 samples was about 4%. It should be noticed that most of the music samples wrongly classified as speech belonged to the class Cm, that is, rhythmic songs as, for example, rap music.

The selected features are rather simple to be computed even by a low-cost device (DSP, microcontroller), except for parameters P_4 and P_5, related to the cepstrum coefficients. If P_4 and P_5 are neglected, and a 6-inputs MLP is used, the mean square error related to the 400 samples increases to about 5%.

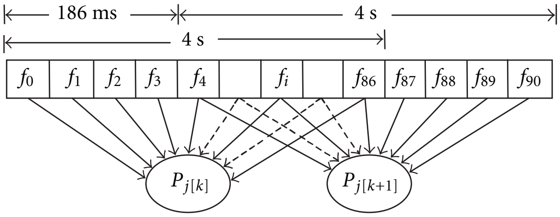

Figure 4: Features P_j updating frequency.

The neural network is simple to be implemented if a suitable polynomial is used as the activation function [20], and a real-time implementation is possible even if low-cost embedded systems are used.

Output y is updated every 4 seconds, and this could be a limit to finely detect the exact position of class changes. To increase the output updating frequency, a circular frame buffer has been provided, and features p_j, in terms of mean value and standard deviation, are updated every 186 ms, corresponding to 4 frames f_i, as shown in Figure 4.

The new updating frequency has been chosen as the fastest to be implemented on a low-cost DSP (TMS320C31). In addition, this operation allow low-pass filters to be applied to the MLP output before the maximum value has been computed.

5. SIMULATION RESULTS

The proposed algorithms have been tested by computer simulations to estimate the classification performance. The tests carried out can be divided into two categories: the first one is about the misclassification errors, while the second one is about the precision in music-speech and speech-music change detection.

Considering the misclassification errors, we defined three parameters as follows:

• MER (Music Error Rate): it represents the ratio between the total duration of the music segments misclassified, and the total duration of the music test file.

• SER (Speech Error Rate): it represents the ratio between the total duration of the speech segments misclassified, and the total duration of the speech test file.

• TER (Total Error Rate): it represents the ratio between the total duration of the segments misclassified in the wrong category (both music and speech), and the total duration of the test file.

The selection of the test files was carried out "manually," that is, each file is composed of many pieces of different types of audio (different speakers over different environmental noise, different kinds of music such as classical, pop, rap, funky, etc.) concatenated in order to have a five minutes segment of speech followed by a five minutes segment of music, and so on, for a total duration of 30 minutes.

All the content of this file has been recorded from an FM radio receiver, and it has been sampled at a frequency of 22050 Hz, with a 16-bit uniform quantization.

The classification results for both the proposed methods are shown in Table 1.

Table 1: Classification results of the proposed algorithms (MLP: Multi Layer Perceptron; ZB: ZCR with Bayesian classifier).

	MER	SER	TER
MLP	11.62%	0.17%	6.0%
ZB	29.3%	6.23%	17.7%

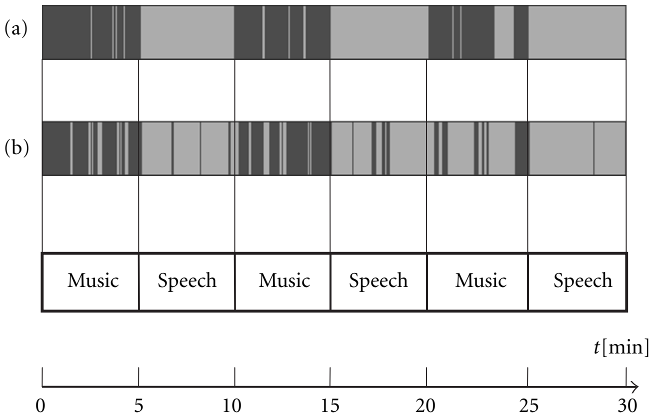

Figure 5: Graphical display of the classification results ((a) MLP, (b) ZB).

From the analysis of the simulation results, we can see that the MLP method gives better results compared to the ZB one, having a lower error rate both in music and speech.

Moreover, both the methods show the worst performance in the classification of the music segments, that is, many segments of music are classified as speech than viceversa. For a better understanding of these results, Figure 5.

In the figure, the white intervals represent the segments classified as speech, whereas the black ones show the segments classified as music. From this figure, it appears clearly that the worst classification results are obtained in the third music segment, between the minutes 20 and 25. The explanation is that these pieces of music contain strong voiced components, under a weak music component (e.g., rap and funky). The neural network makes some mistakes only with the rap song (minutes 23–24 referred in Figure 5), while the ZB approach misclassifies the funky song (minutes 20–23) too. Commercials, that includes speech with music in background, are present in the test file at minutes 17–18: in this case the ZB approach shows only some uncertainties.

The problem related to music identification is due mainly to the following reasons:

- The MLP has been trained to recognize also music with a voiced component, and it gets wrong only if the voiced component is too rhythmic (e.g., rap song in our case). On the other hand, the Bayesian classifier used in the ZB approach does not take into account cases with mixed component (music and voice), and therefore in this case the classification results are significantly affected by the relative strongness of the spurious components.
- Furthermore, the ZB approach, that uses very few parameters, is inherently unable to discriminate between pure speech and speech with music background, while the MLP network, which uses more features, is able to make it.

Table 2: MLP (a), and ZB (b) change detection results expressed in seconds.

	PM2S	PS2M
Min	0.56	0.19
Mean	1.30	1.53
Max	1.49	2.98

(a)

	PM2S	PS2M
Min	0.56	12.28
Mean	1.30	14.51
Max	2.79	16.74

(b)

Considering the precision of music-speech and speech-music change detection, we measured the distance between the correct point in the time scale when a change occurred, and the nearest change point automatically extracted from the proposed algorithms. In particular, we have measured the maximum, minimum, and the mean interval between the real change and the extracted one. The results are shown in Table 3(b), where PS2M (Precision Speech to Music) is the error in speech to music change detection, and PM2S (Precision Music to Speech) is the error in music to speech change detection.

Also in this case, the MLP obtains better performance than the ZB.

6. CONCLUSION

In this paper, we have proposed and compared two different algorithms for audio classification into speech and music. The first method is based mainly on ZCR and Bayesian classification (ZB). It is very simple from a computational point of view and gives good results in case of pure music or speech. Anyway some performance degradation arises when the music segment contains also some speech superimposed on music, or strong rhythmic components. We have proposed therefore a second method that is based on a Multi-Layer Perceptron. In this case we obtain better performance, at the expense of a limited growth in the computational complexity. In practice, a real-time implementation is possible even if low-cost embedded systems are used.

ACKNOWLEDGMENTS

The authors wish to thank Profs. D. Marioli and R. Leonardi for numerous valuable comments and suggestions, and Ing. C. Pasin for providing some of the simulation results.

REFERENCES

[1] MPEG Requirement Group. MPEG-7: Overview of the MPEG-7 Standard. ISO/IEC JTC1/SC29/WG11 N3752, France, October 1998.

[2] MPEG-7: ISO/IEC 15938-5 Final Commitee Draft-Information Technology-Multimedia Content Description Interface-Part 5 Multimedia Description Schemes, ISO/IEC JTC1/SC29/WG11 MPEG00/N3966, Singapore, May 2001.

[3] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, and L. A. Rossi, "Describing multimedia documents in natural and semantic-driven ordered hierarchies," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 2023–2026, Istanbul, Turkey, June 2000.

[4] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, and L. A. Rossi, "The ToCAI description scheme for indexing and retrieval of multimedia document," *Multimedia Tools and Applications*, vol. 14, no. 2, pp. 153–173, 2001.

[5] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using audio and visual information," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, 2000.

[6] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, "Video handling with music and speech detection," *IEEE Multimedia*, vol. 5, no. 3, pp. 17–25, 1998.

[7] C. Saraceno and R. Leonardi, "Indexing audio-visual databases through a joint audio and video processing," *Int. J. Image Syst. Technol.*, vol. 9, no. 5, pp. 320–331, 1998.

[8] A. Bugatti, R. Leonardi, and L. A. Rossi, "A video indexing approach based on audio classification," in *Proc. International Workshop on Very Low Bit Rate Video*, pp. 75–78, Kyoto, Japan, October 1999.

[9] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. 993–996, Atlanta, Ga, USA, May 1996.

[10] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1331–1334, Munich, Germany, April 1997.

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

[12] T. Zhang and C. C. J. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 6, pp. 3001–3004, Phoenix, Ariz, USA, March 1999.

[13] Z. Liu, J. Huang, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene classification," in *Proc. IEEE* 1997 *Workshop on Multimedia Signal Processing*, Princeton, NJ, USA, June 1997.

[14] J. L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. International Conference on Speech and Language Processing*, vol. 4, pp. 1335–1338, Sydney, Australia, December 1998.

[15] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition, 1999.

[16] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1432–1436, Phoenix, Ariz, USA, March 1999.

[17] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[18] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[19] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison Wesley, Ontario, Canada, 2nd edition, 1989.

[20] A. Flammini, D. Marioli, D. Pinelli, and A. Taroni, "A simple neural technique for sensor data processing," in *Proc. IEEE International Workshop on Emerging Technologies, Intelligent Measurement and Virtual Systems for Instrumentation and Measurement*, pp. 1–10, Minnesota Club, St. Paul, Minn, USA, May 1998.

Alessandro Bugatti was born in Brescia, Italy, in 1971. He received the Dr. Eng. degree in electronic engineering in 1998, from the University of Brescia. Currently he is working at University of Brescia as extern consultant. He was involved in the activities of the European project AVIR from 1998 to 2000 and in the MPEG-7 standardization process. His research interest is in audio segmentation and classification, user interfaces and audio-video indexing. His background includes also studies in AI and expert system fields. His efforts are currently devoted to both automatic audio sequences content analysis and cross-modal analysis.

Alessandra Flammini was born in Brescia, Italy, in 1960. She graduated with honors in Physics at the University of Rome, Italy, in 1985. From 1985 to 1995 she worked on industrial research and development on digital drive control. Since 1995, she has been a Researcher at the Department of Electronics for Automation of the University of Brescia. Her main field activity is the design of digital electronic circuits (FPGA, DSP, processors) for measurement instrumentation.

Pierangelo Migliorati got the Laurea (cum laude) in electronic engineering from the Politecnico di Milano in 1988, and the Master in information technology from the CEFRIEL Research Centre, Milan, 1989, respectively. He joined CEFRIEL research center in 1990. From 1995 he is Assistant Professor at University of Brescia, where he is involved in activities related to channel equalization and indexing of multimedia documents. He is a member of IEEE.