A Speech/Music Discriminator Based on RMS and Zero-Crossings

Costas Panagiotakis and George Tziritas, *Senior Member, IEEE*

***Abstract*—Over the last several years, major efforts have been made to develop methods for extracting information from audiovisual media, in order that they may be stored and retrieved in databases automatically, based on their content. In this work we deal with the characterization of an audio signal, which may be part of a larger audiovisual system or may be autonomous, as for example in the case of an audio recording stored digitally on disk. Our goal was to first develop a system for segmentation of the audio signal, and then classification into one of two main categories: speech or music. Among the system's requirements are its processing speed and its ability to function in a real-time environment with a small responding delay. Because of the restriction to two classes, the characteristics that are extracted are considerably reduced and moreover the required computations are straightforward. Experimental results show that efficiency is exceptionally good, without sacrificing performance.**

Segmentation is based on mean signal amplitude distribution, whereas classification utilizes an additional characteristic related to the frequency. The classification algorithm may be used either in conjunction with the segmentation algorithm, in which case it verifies or refutes a music-speech or speech-music change, or autonomously, with given audio segments. The basic characteristics are computed in 20 ms intervals, resulting in the segments' limits being specified within an accuracy of 20 ms. The smallest segment length is one second. The segmentation and classification algorithms were benchmarked on a large data set, with correct segmentation about 97% of the time and correct classification about 95%.

***Index Terms*—Audio segmentation, speech/music classification, zero-crossing rate.**

IEEE Proof

I. Introduction

A. Problem Position

IN MANY applications, there is a strong interest in segmenting and classifying audio signals. A first content characterization could be the categorization of an audio signal as one of speech, music, or silence. Hierarchically, these main classes could be subdivided, for example, into various music genres, or by recognition of the speaker. In the present work, only the first level in the hierarchy is considered.

A variety of systems for audio segmentation and/or classification have been proposed and implemented in the past for the needs of various applications. We present some of them in the following paragraphs, permitting a methodological comparison with the techniques proposed in this paper. We also report their performance for related comparisons. However, the test data set is different and the conclusions are hindered by this fact.

Saunders [6] proposed a technique for discrimination of audio as speech or music using the energy contour and the zero-crossing (ZC) rate. This technique was applied to broadcast radio divided into segments of 2.4 s, which were classified using features extracted from intervals of 16 ms. Four measures of the skewness of the distribution of the ZC rate were used with a 90% correct classification rate. When a probability measure on signal energy was added a performance of 98% is reported.

Zhang and Kuo [14] proposed a method for audio segmentation and classification in music, speech, song, environmental sound and silence, etc. They used features like the energy function, average ZC rate, the fundamental frequency and the spectral peaks tracks. A heuristic rule-based method was proposed. In audio classification, they achieved an accuracy rate of more than 90%, and 95% in audio segmentation.

Scheirer and Slaney [7] used 13 features, of which eight are extracted from the power spectrum density, for classifying audio segments. A correct classification percentage of 94.2% is reported for 20 ms segments and 98.6% for 2.4 s segments. Tzanetakis and Cook [10] proposed a general framework for integrating, experimenting with and evaluating different techniques of audio segmentation and classification. In addition, they proposed a segmentation method based on feature change detection. They used energy-spectral based features, ZC, etc. For their experiments on a large data set, a classifier performance of about 90% is reported. In a more recent work, Tzanetakis and Cook [11] proposed a whole file and real-time frame based classification method using three feature sets (timbral texture, rhythmic content, and pitch content). They achieved 61% for ten music genres. This result is considered comparable to results reported for human musical genre classification. Also, their music/speech classifier has 86% accuracy and male/female/sports announcing classifier has 74% accuracy.

In [12], a system for content-based classification, search, and retrieval of audio signals is presented. The sound analysis uses the signal energy, pitch, central frequency, spectral bandwidth, and harmonicity. This system is applied mainly in audio data collections. More general framework related issues are reviewed in [1].

In [4] and [8], cepstral coefficients are used for classifying or segmenting speech and music. Moreno and Rifkin [4] model these data using Gaussian mixtures and train a support vector machine for the classification. On a set of 173 hours of audio signals collected from the WWW, a performance of 81.8% is

Manuscript received January 11, 2001; revised May 20, 2003. The associaste editor coordinating the review of this manuscript and approving it for publication was Prof. Wayne Wolf.

The authors are with the Department of Computer Science, University of Crete, Heraklion, Crete GR-714 09 Greece (e-mail: cpanag@csd.uoc.gr).

Digital Object Identifier 10.1109/TMM.2004.840604

1520-9210/05$20.00 © 2005 IEEE

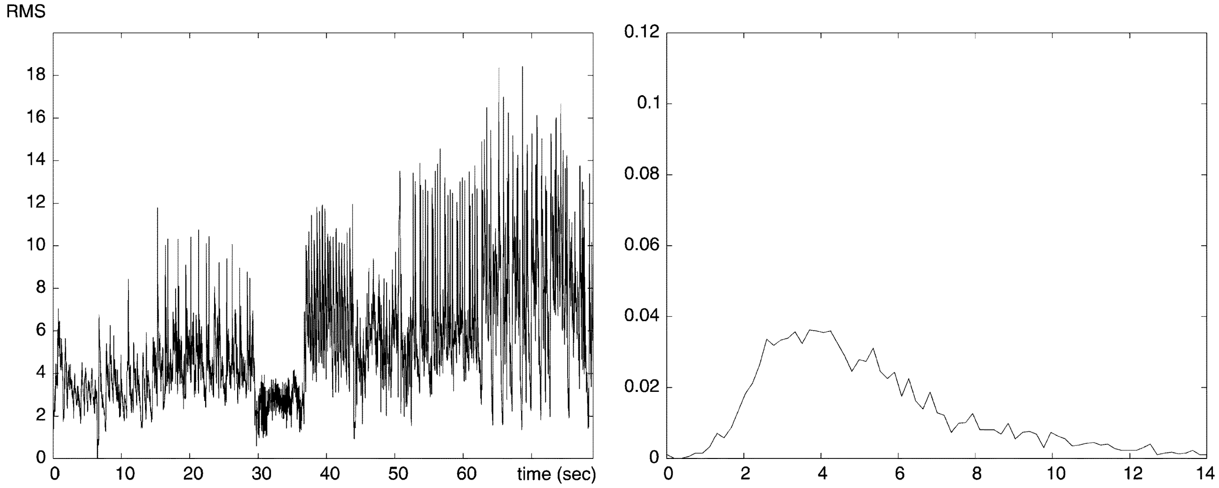

Fig. 1. RMS of a music signal and its histogram.

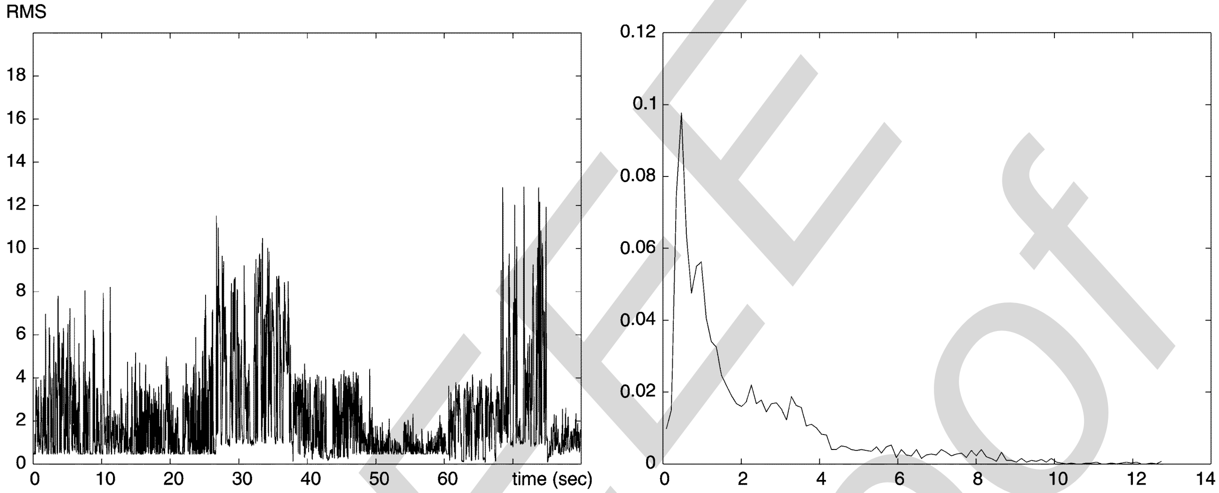

Fig. 2. RMS of a voice signal and its histogram.

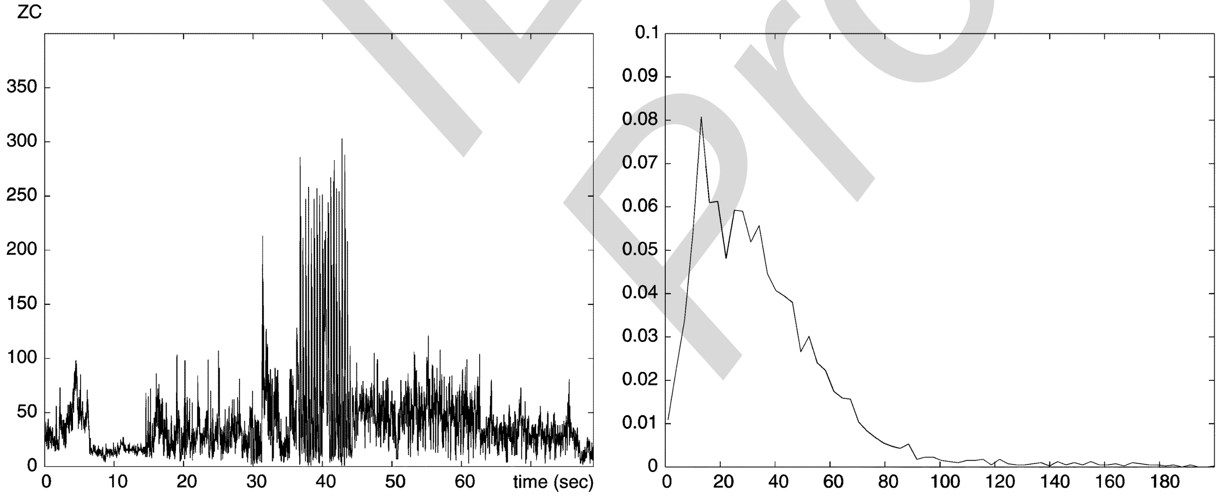

Fig. 3. Number of ZCs for a music signal and its histogram.

IEEE Proof

reported. In [8], Gaussian mixtures are used too, but the segmentation is obtained by the likelihood ratio. For very short (26 ms) segments, a correct classification rate of 80% is reported.

A general remark concerning the above techniques is that often a large number of features is used for discriminating a certain number of audio classes. Furthermore, the classification tests are frequently heuristic-based and not derived from an analysis of the data. In our work, we tried at first to limit the number of features, as we have limited our task to the music/speech discrimination. We concluded that a reliable

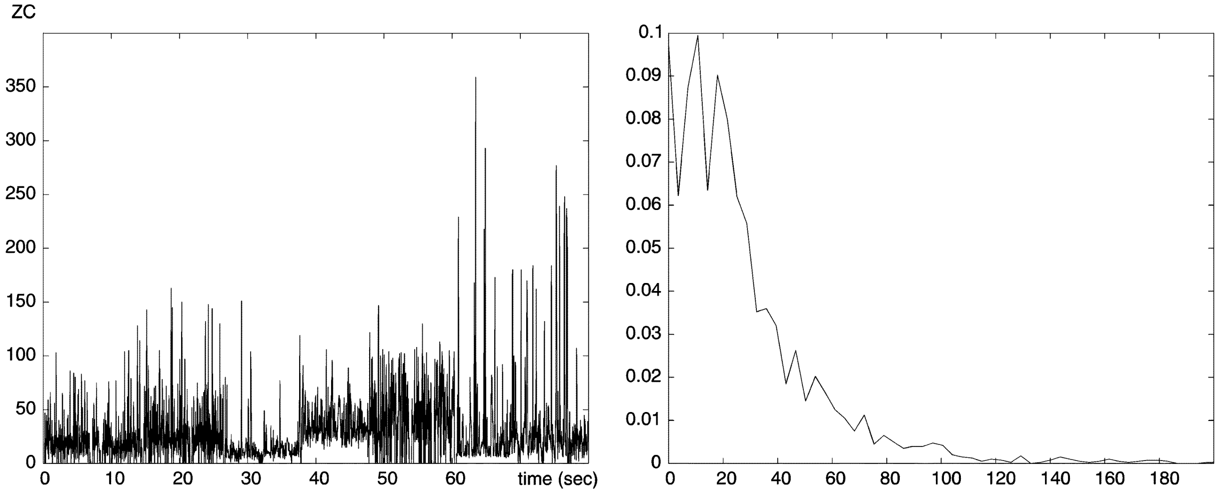

Fig. 4. Number of ZCs for a voice signal and its histogram.

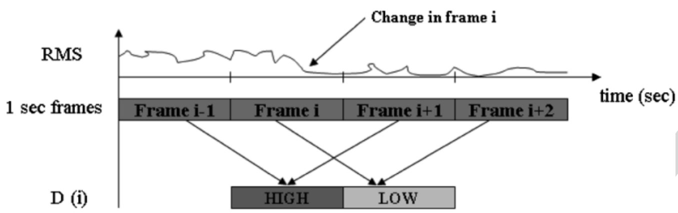

Fig. 5. First stage of the segmentation method.

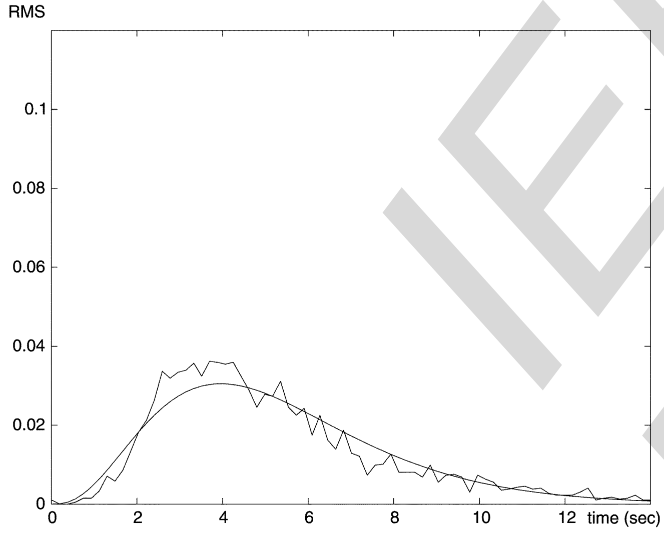

Fig. 6. RMS histogram for a collection of music data and its fitting by the generalized χ^2 distribution.

discriminator can be designed using only the signal amplitude, equivalent to the energy used in [6], and the central frequency, measured by the ZC rate, a feature already exploited in previous work. In addition, we analyzed the data in order to extract relevant parameters for making the statistical tests as effective as possible. However, some of the proposed tests are mainly heuristic, while other are well defined and based on appropriate models.

We conclude this introduction by describing the signal and its basic characteristics as utilized in our work. In Section II, we present the proposed segmentation method, which is a change detector based on a dissimilarity measure of the signal amplitude distribution. In Section III, the classification technique is presented, which could either complete the segmentation, or be used independently. Features extracted from the ZC rate are added and combined with the amplitude parameters.

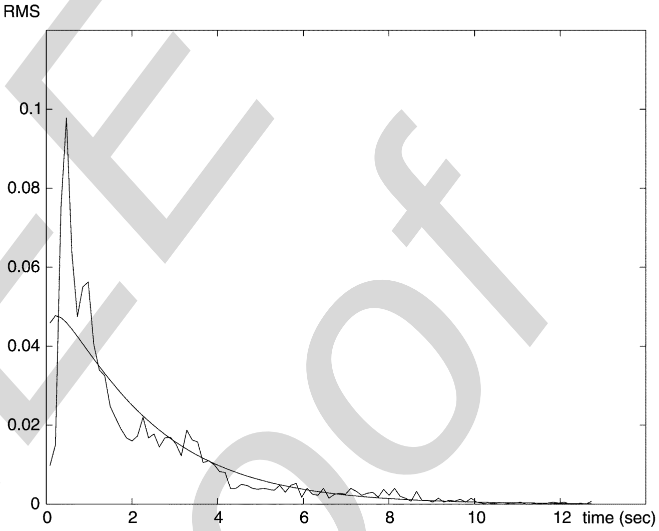

Fig. 7. RMS histogram for a collection of voice data and its fitting by the generalized χ^2 distribution.

B. Description of Signal and Its Characteristics

The signal is assumed to be monophonic. In the case of multichannel audio signals, the average value per-sample across multiple channels is taken as input. This may fail in cases where special effects could affect the difference between two stereo channels. There are no restrictions on the sampling frequency functioning equally well from 11 025 Hz to 44 100 Hz, while the sound volume may differ from one recording to another. The system is designed to fulfill the requirement of independence on the sampling frequency and on the sound volume, and to depend only on the audio content. The changes in volume are recognized (Section II), but, if the segment before and the segment

IEEE Proof

IEEE Proof

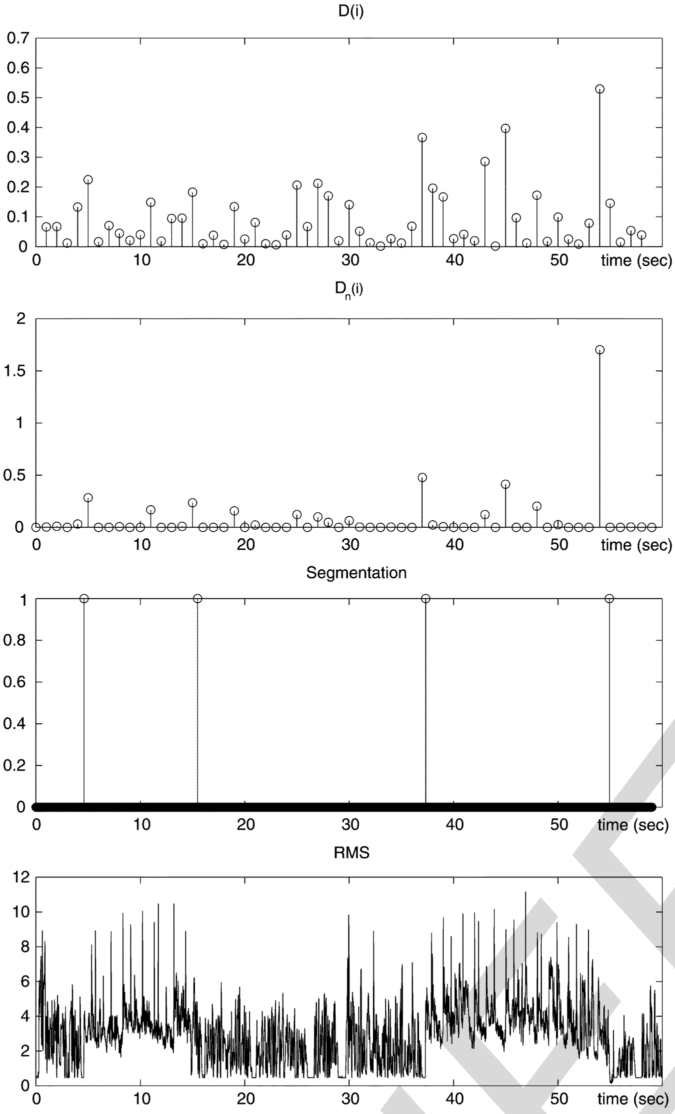

Fig. 8. Example of segmentation with four transitions. Shown are the distance $D(i)$, the normalized distance $D_n(i)$, the change detection result, and the RMS data.

after the change belong to the same class, the change will be ignored (Section III-B).

Two signal characteristics are used: the amplitude, measured by the root mean square (RMS), and the mean frequency, measured by the average density of ZCs. One measure of each is acquired every 20 ms. For simplifying the calculation, the average across all the samples of the considered interval is omitted without any data reduction. The signal amplitude, RMS, and the ZCs, are therefore defined as follows:

$$\mathrm{RMS} \triangleq \sqrt{\sum_{n=1}^{N} x^2(n)} \tag{1}$$

$$\mathrm{ZC} \triangleq \frac{1}{2} \cdot \sum_{n=2}^{N} |\mathrm{sign}(x(n)) - \mathrm{sign}(x(n-1))| \tag{2}$$

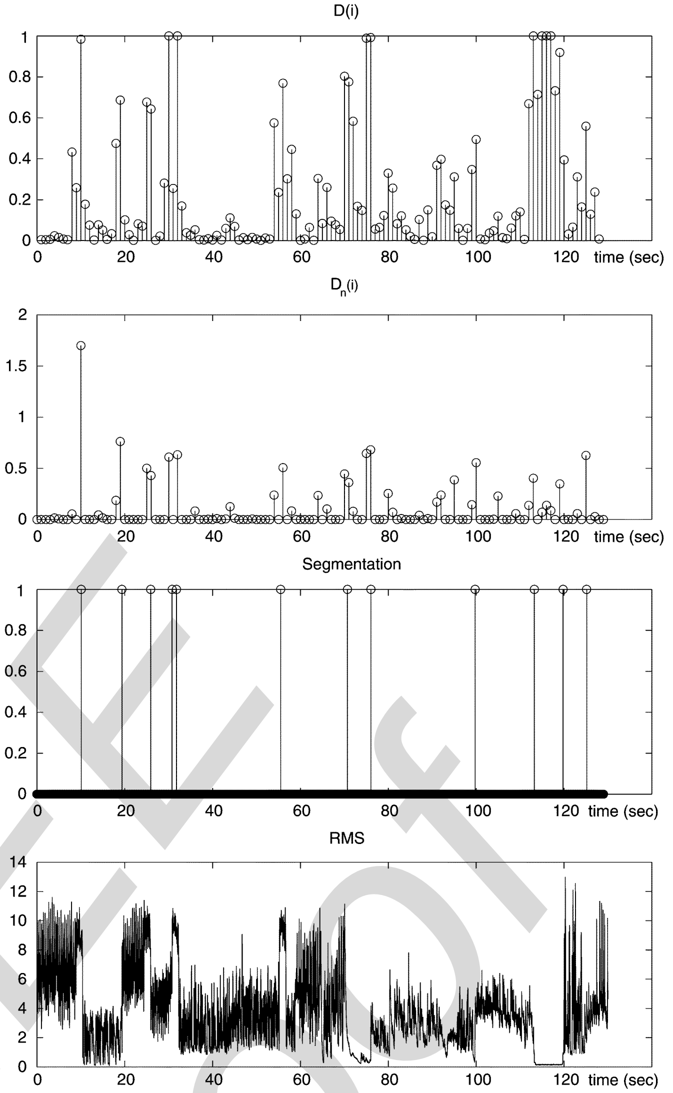

Fig. 9. Example of segmentation with many transitions. Shown are the distance $D(i)$, the normalized distance $D_n(i)$, the change detection result, and the RMS data.

where

$$\mathrm{sign}(a) = \begin{cases} 1, & a > 0 \\ 0, & a = 0 \\ -1, & a < 0. \end{cases}$$

Voice and music are distinguished by the distribution of amplitude values. Figs. 1 and 2 show the RMS measured as described above and the corresponding histogram for a music and for a speech signal. The distributions are different and this fact may be exploited for both segmentation and classification. The mean frequency is approximated by the number of ZCs in the 20 ms interval. Figs. 3 and 4 show the ZC rate and the corresponding histograms for a music and for a voice signal.

The two characteristics used in our work are almost independent. We have tested two measures of independence for the verification of this hypothesis. The first is the Blomquist measure [3], defined as

$$V = \frac{|n_1 - n_2|}{n} \tag{3}$$

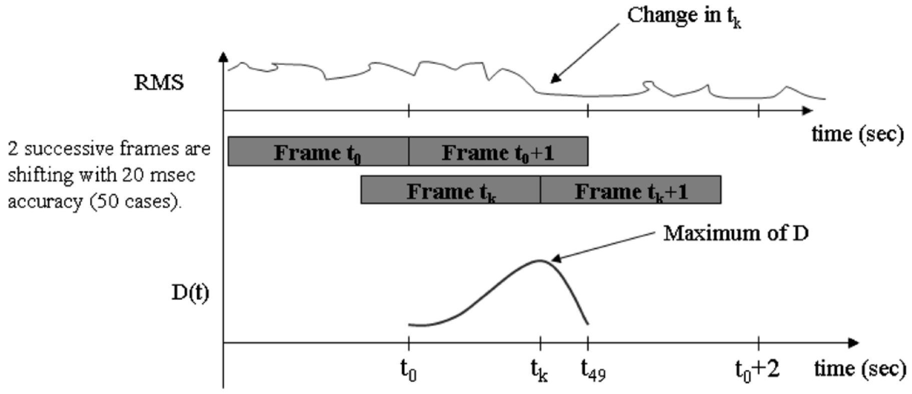

Fig. 10. Second stage of the segmentation method.

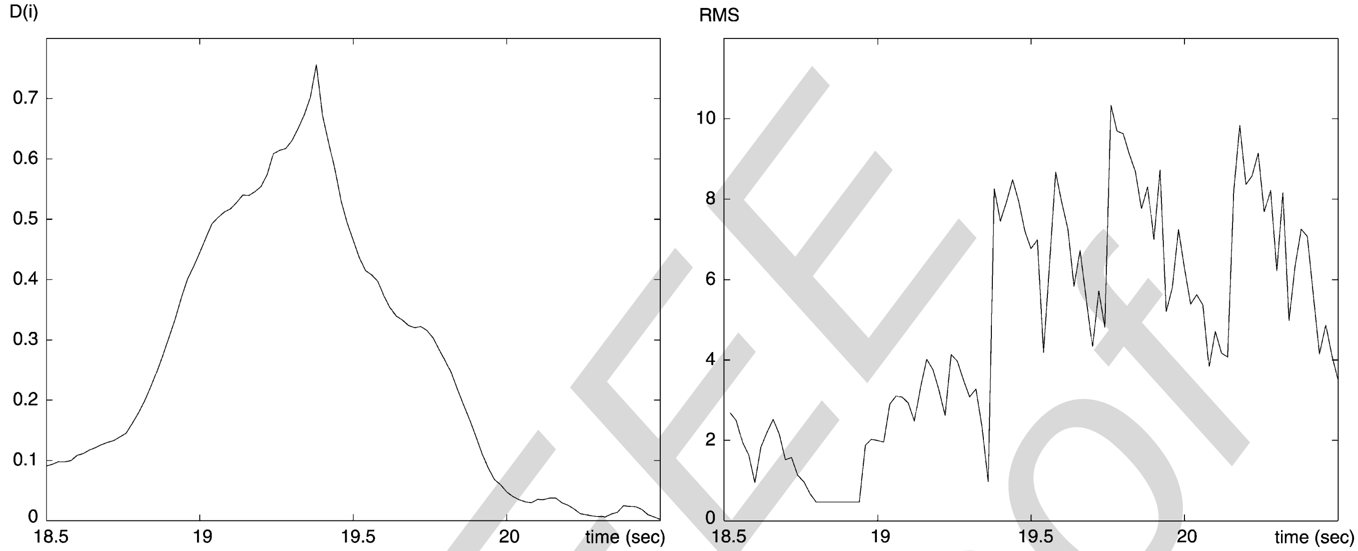

Fig. 11. Shown on the left is the distance $D(i)$ for the RMS shown in the right plot. The accuracy is excellent for this transition from speech to music.

IEEE Proof

where n is the number of data pairs, n_1 is the number of pairs with the same sign related to the median values of the two variables, and n_2 is the number of pairs with opposite sign. The empirical value obtained for V was about 0.1, showing an almost sure independence. We have also used the ratio of the mutual information to the sum of entropies of the two variables

$$I = 2\frac{\sum\sum P_{ij}\log\frac{P_{ij}}{P_iQ_j}}{\sum P_i\log\frac{1}{P_i}+\sum Q_j\log\frac{1}{Q_j}} \tag{4}$$

and have obtained a value of about 0.05, again near the independence condition. The independence between the RMS and ZC of the signal is more clear in music than in speech. This is due to the fact that speech contains frequent short pauses, where both the RMS and ZC are close to zero, and therefore correlated in this case. Also the above values were 10% lower in music data than in speech data. We exploit this possible discrimination in a feature defined for the classification.

In [7], [10], and [12] the classification uses features extracted from the power spectrum density computed by the FFT as the spectral centroid, which however is strongly correlated with the ZC rate [2], [6]. In the Appendix we have examined the relation between ZC rate and spectral centroid for a class of zero-mean random signals. In cases where there are noisy impulsive sounds, such as drum hits, the ZC rate is much more affected than the spectral centroid, and they might not be strongly correlated. The mean value of sound signals, that we used, was close to zero, so it was not needed to subtract the mean value in order to compute the ZCs. In the general case, it is needed to subtract the mean value [11], therefore the feature should be the mean-crossing rate. The maximal frequency and the pitch have been also used, as well as the power spectrum density at 4 Hz, which is roughly the syllabic speech frequency. On the other hand, the LPC coefficients and the cepstrum analysis, as they are used for speech analysis, can discriminate speech from music [4], [8].

II. Segmentation

Segmentation is implemented in real-time and is based only on RMS. For each 1 s frame, 50 values of the RMS are computed from successive intervals of 20 ms. The mean and the variance of the RMS are calculated for each frame. The segmentation algorithm is separated in two stages. In the first stage, the transition frame is detected. In the second stage, the instant of transition, with an accuracy of 20 ms, is marked. The last stage is more time consuming, but it is employed only in case of frame change detection.

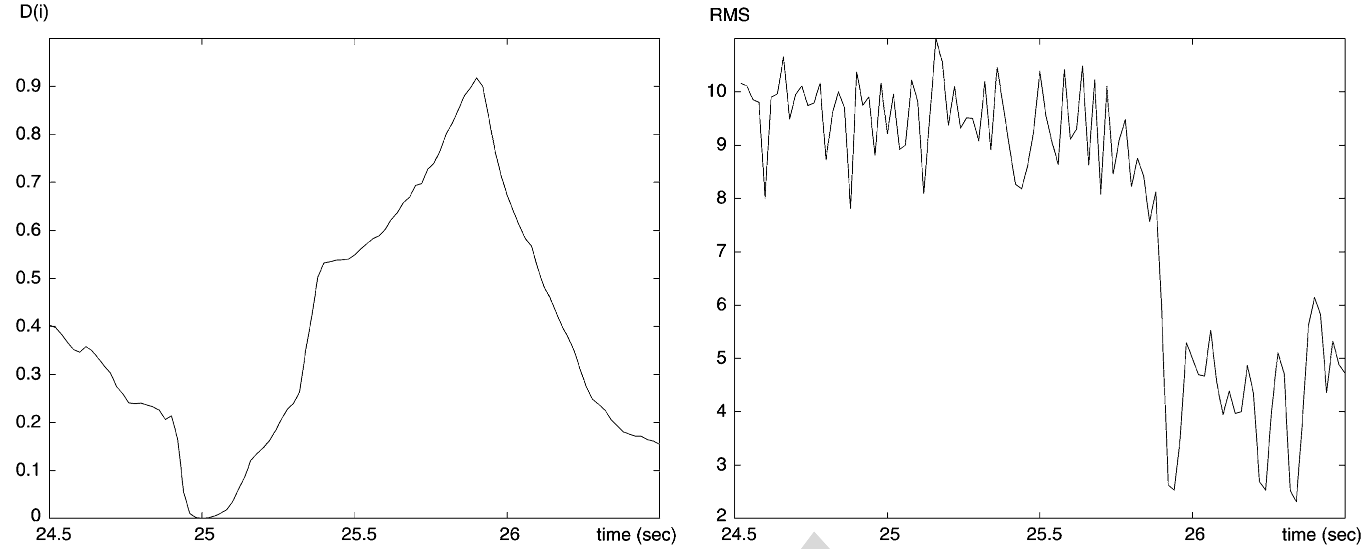

Fig. 12. Shown on the left is the distance $D(i)$ for the RMS shown in the right plot. The accuracy is very good for this transition from music to speech.

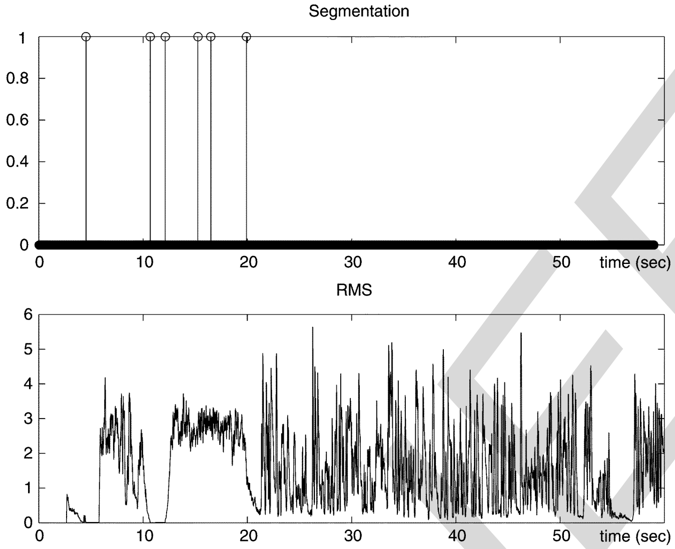

Fig. 13. Change detection is illustrated and the signal amplitude shown. No transition loss occurs, but some segments are over-segmented.

The instantaneous accuracy is fixed at 20 ms because the human perceptual system is generally not more precise, and moreover because speech signals remain stationary for 5–20 ms [9]. The maximal interval for measuring speech characteristics should therefore be limited to intervals of 20 ms.

A. Change Detection Between Frames

The technique used for detecting potential frames containing a change is represented in Fig. 5. A change is detected in frame i if the previous and the next frames are sufficiently different. The detection is based on the distribution of the RMS values, which differ between speech and music, as seen in the previous section. In speech the variance is large in comparison with the mean value, because there are pauses between syllables and words, while in music the variation of the amplitude remains in general moderated.

We need an appropriate model for the RMS distribution, since we have only 50 values per frame, in order to measure frames' dissimilarity. Then the dissimilarity is obtained as a function of the models' parameters. We have observed that the generalized χ^2 distribution fits well the histograms for both music and speech (Figs. 6 and 7). We can see that the approximation is acceptable. The good fit is due to the Laplacian (symmetric exponential) distribution of the audio signals. The generalized χ^2 distribution is defined by the probability density function

$$p(x) = \frac{x^a e^{-bx}}{b^{a+1}\Gamma(a+1)}, \quad x \geq 0. \tag{5}$$

The parameters a, b are related to the the mean and the variance values of the RMS,

$$a = \frac{\mu^2}{\sigma^2} - 1 \quad \text{and} \quad b = \frac{\sigma^2}{\mu}. \tag{6}$$

The segmentation will be based on a dissimilarity measure, which is applied between frames. We propose to use a known similarity measure defined on the probability density functions

$$\rho(p_1, p_2) = \int \sqrt{p_1(x)p_2(x)}dx \tag{7}$$

The similarity takes values in the interval $[0, 1]$, where the value 1 means identical distributions, and zero means completely nonintersecting distributions. For this reason, the value $1 - \rho(p_1, p_2)$, known as the Matusita distance [13], can be interpreted as the distance between the content of the two frames. It is well-known that the above similarity measure is related to the classification error [13]. For the case of two equiprobable hypotheses the classification error is bounded by

$$P_e \leq \frac{\rho(p_1, p_2)}{2}. \tag{8}$$

IEEE Proof

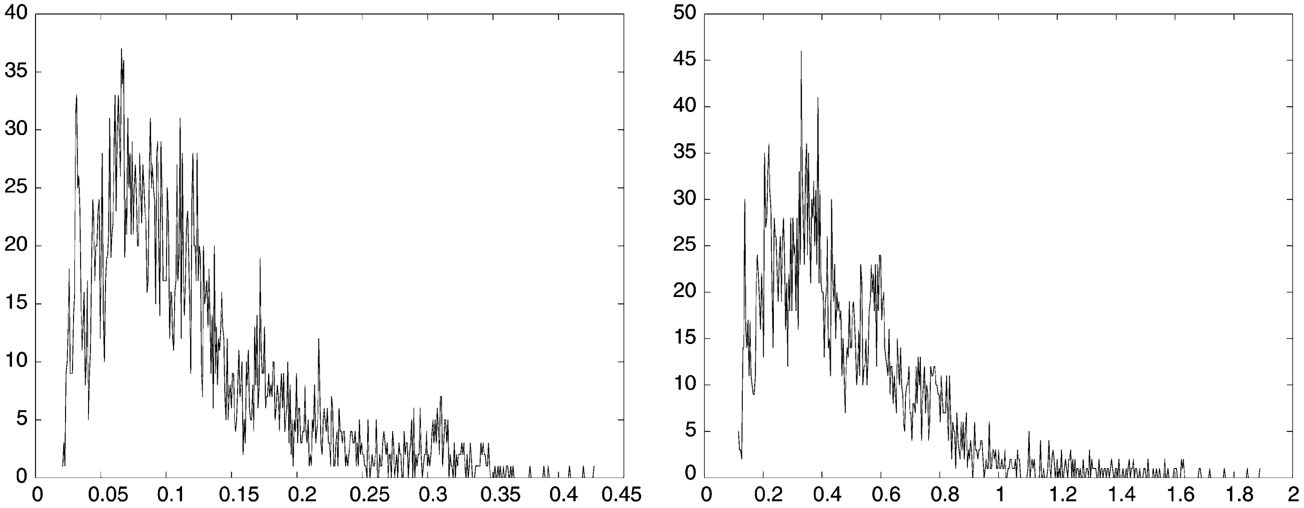

Fig. 14. Histograms of the normalized RMS variance for (left) music and (right) voice.

For the generalized χ^2 distribution the similarity measure depends on the parameters a and b,

$$\rho(p_1, p_2) = \frac{\Gamma(\frac{a_1+a_2}{2} + 1)}{\sqrt{\Gamma(a_1+1)\Gamma(a_2+1)}} \frac{2^{\frac{a_1+a_2}{2}+1} b_1^{\frac{a_2+1}{2}} b_2^{\frac{a_1+1}{2}}}{(b_1+b_2)^{\frac{a_1+a_2}{2}+1}}. \tag{9}$$

At first the similarity measure, or the corresponding distance, is used for localizing a candidate change frame. Therefore, we compute for each frame i a value $D(i)$, which gives the possibility of a change within that frame

$$D(i) = 1 - \rho(p_{i-1}, p_{i+1}). \tag{10}$$

Basically, if there is a single change within frame i, then frames $i-1$ and $i+1$ must differ. On the other hand, if the change is instantaneous, e.g., a very brief interval within the frame, then frames $i-1$ and $i+1$ will be similar and the factor $\rho(p_{i-1}, p_{i+1})$ will be close to 1 and the $D(i)$ will be small. The system is designed to extract any important change from music to voice, and vice versa, or very large changes in volume, as for example from silence to an audible sound. These changes locally maximize the $D(i)$ and can be detected with a suitable threshold.

However, some filtering or normalization is needed. One reason is that relatively large distances are also expected in the neighboring frames of a change frame. Furthermore an adaptation of the threshold should be introduced, since the audio signal activity is time-variant. The latter is more relevant for voice signals. In any case the nonstationarity of the audio signals should be taken into consideration. We introduce the locally normalized distance, as follows:

$$D_n(i) = \frac{D(i)V(i)}{D_M(i)} \tag{11}$$

where $V(i)$ measures the (positive) difference of $D(i)$ from the mean value of the neighboring frames. If the difference is negative, it is set to zero. $D_M(i)$ is the maximal value of distances in the same neighborhood of the examined frame. In the current implementation we use a neighborhood of two frames before and two frames after the current one. The comparison of the distance $D(i)$ and the normalized distance is illustrated for two examples in Figs. 8 and 9. The local maxima of $D_n(i)$ are determined provided that they exceed some threshold. The threshold on $D_n(i)$ is set according to the local variation of the similarity measure. If the similarity variation is small, the detector is more sensitive, while in the case of large similarity variation, the threshold is larger. This procedure introduces a delay of 3 s, which is necessary for the change detection. It is needed to examine the next frames of frame i, in order to determine if there is a change in frame i. The method is remaining a real-time process with 3 s delay. At the end of this procedure we have the change candidate frames.

B. Change Instant Detection

The next step is detecting the change within an accuracy of 20 ms, the maximal accuracy of our method (Fig. 10). For each of the frames, we find the time instant where two successive frames, located before and after this instant, have the maximum distance. The duration of the two frames is always 1 s and the distance measure is based on the similarity measure defined in (9). At the end of the segmentation stage, homogeneous segments of RMS have been obtained. Our aim was to find all possible audible changes, even those based only on volume or other features. An oversegmentation is very probable, if we are interested only on the main discrimination between speech and music. If just the volume changes, the segmentation method will find a change. The final segmentation is completed by a classification stage, which could also be used independently for the characterization of audio signals. In Figs. 11 and 12, we show the instant change detection for two frames.

C. Segmentation Results

In our experiments we obtained reliable detection results. Because in our scheme segmentation is completed by the classification, false detections can be corrected by the classification module. Thus the detection probability is the appropriate quality evaluation measure.

Our data set is described in Section IV. The segmentation algorithm was tested by test files that were created by our data

IEEE Proof

set. These files contained speech, music, and silent transitions. There were about 100 speech/music transitions and about 20 silence/(speech-music) transitions. The results for this last case were always correct. The duration of each segment varied from 2 to 30 s. In most cases the volume in speech/music transitions was similar in order to drive the segmentor to detect changes in form of RMS distribution.

We have tested our technique on the above test files, and obtained a 97% detection probability, i.e., only 3% of real changes have been missed. Accuracy in the determination of the change instant was very good, almost always within an interval of 0.2 s. Some examples of segmentation results are shown in Figs. 8, 9, and 13.

III. Classification

A. Features

For each segment extracted by the segmentation stage some features are computed and used for classifying the segment. We call these features the *actual* features, which are obtained from the basic characteristics, i.e., the signal amplitude and the ZCs. We will define some tests, which will be implemented in sequential order, taking into consideration that the basic characteristics are nearly independent. The discrimination is based mainly on the pauses, which occur in speech signals due to syllables and word separation.

1) Normalized RMS Variance: The normalized RMS variance (σ_A^2) is defined as the ratio of the RMS variance to the square of RMS mean. It is therefore equal to the inverse of parameter $a + 1$ defined in (6). This feature is volume invariant. In Fig. 14, we show two typical histograms of the normalized RMS variance for speech and music signals. We observe that the two distributions are almost nonoverlapping, and thus the normalized variance discriminates very well the two classes. In our experiments 88% of speech segments have a value of normalized RMS variance greater than a separation threshold of 0.24, while 84% of music segments have a value less than the same threshold. In addition the two distributions can be approximated by the generalized χ^2 distribution, and using the maximum likelihood principle we obtain the aforementioned separating threshold. The normalized RMS variance is used as the last test in the proposed algorithm.

2) The Probability of Null Zero-Crossings (ZC0): The ZC rate is related to the mean frequency for a given segment. In the case of a silent interval the number of ZCs is null. In speech there are always some silent intervals, thus the occurrence of null zero-crossings (ZC0) is a relevant feature for identifying speech. Thus, if this feature exceeds a certain threshold, the tested segment almost certainly contains a voice signal. In our work the threshold on the probability of ZC0 is set to 0.1 (see the histogram shown in Fig. 4). Our experiments showed that about 40% of speech segments verify this criterion, while we have not found any music segment exceeding the threshold. Some speech segments don't satisfy the above criterion because of noise or fast speaking rate. Comparing the histograms in Figs. 3 and 4, we see the discriminating capability of the null ZCs feature.

3) Joint RMS/ZC Measure: Together with the RMS and null ZCs features we exploit the fact that RMS and ZC are somewhat correlated for speech signals, while essentially independent for

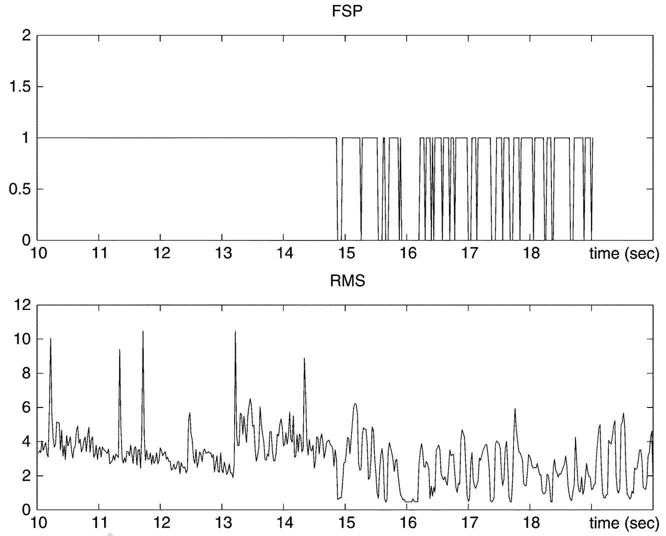

Fig. 15. Transition from speech to music. In the bottom the RMS is shown, and in the top the detected silent intervals. Silent intervals are more frequent in speech than in music.

TABLE I
Performance of the Various Features Individually and in Conjuction

Features	Performance in music	Performance in speech
$ZC0$	90%	60%
σ_A^2	84%	88%
C_Z	90%	60%
σ_A^2, $ZC0$	80%	97%
σ_A^2, C_Z	82%	97%
C_Z, σ_A^2	80%	97%
$ZC0$, σ_A^2	70%	97%
F_v, σ_A^2	88%	92%
F_v, C_Z, $\max(ZC)$, $ZC0$, σ_A^2	92%	97%

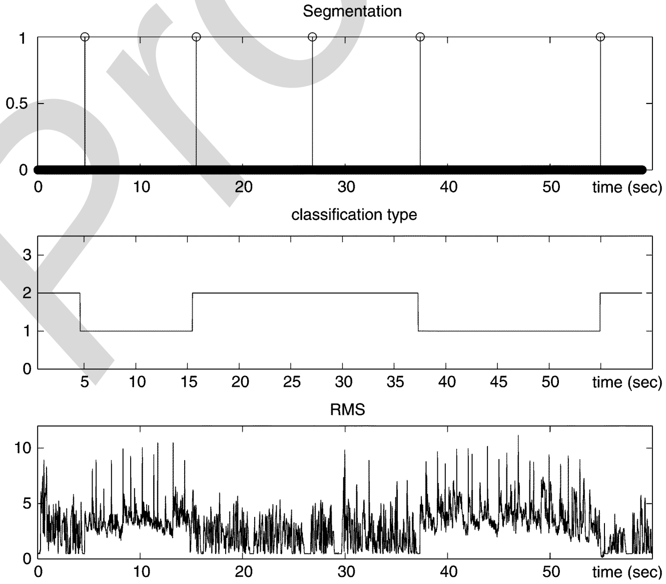

Fig. 16. Result of classification after the change detection. The second and the fourth segment are music, while the others are speech.

IEEE Proof

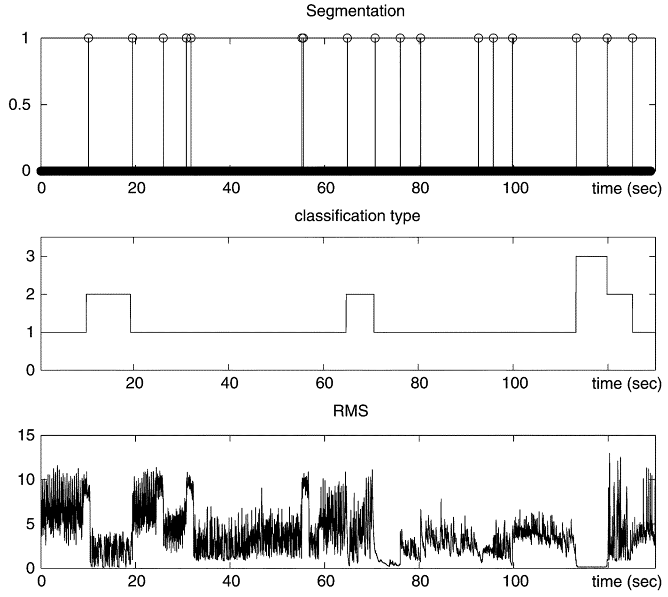

Fig. 17. Over-segmented signal for which all segments were correctly classified. 1: music, 2: speech, 3: silence.

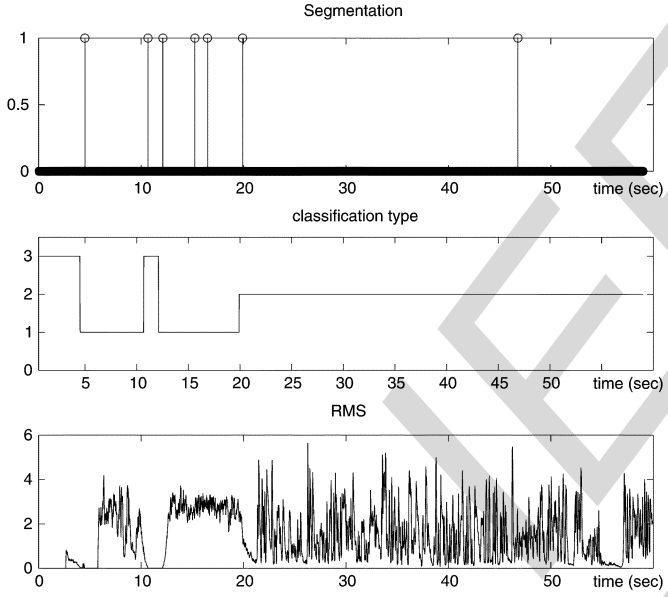

Fig. 18. Example of correct classification.

music signals. Thus we define a feature related to the product of RMS and ZC

$$C_Z = \frac{\sum_{i=1}^{N} A(i)z(i)}{2A_x - A_n - A_m} \tag{12}$$

with $A_x = \max\{A(i) : 1 \leq i \leq N\}, A_n = \min\{A(i) : 1 \leq i \leq N\}$, and $A_m = \text{median}\{A(i) : 1 \leq i \leq N\}$. This is a normalized correlation measure. The normalization by $2A_x - A_n - A_m$ is used because in speech signals the maximal RMS value is much larger than the median and the minimum values in comparison with the case of music signals. The test consists of comparing this feature to some empirically set threshold. If C_Z is close to 0, then the segment is classified as speech. Thus even if the correlation between RMS and ZC may be not negligible, the two classes are discriminated by the large deviations in speech signals.

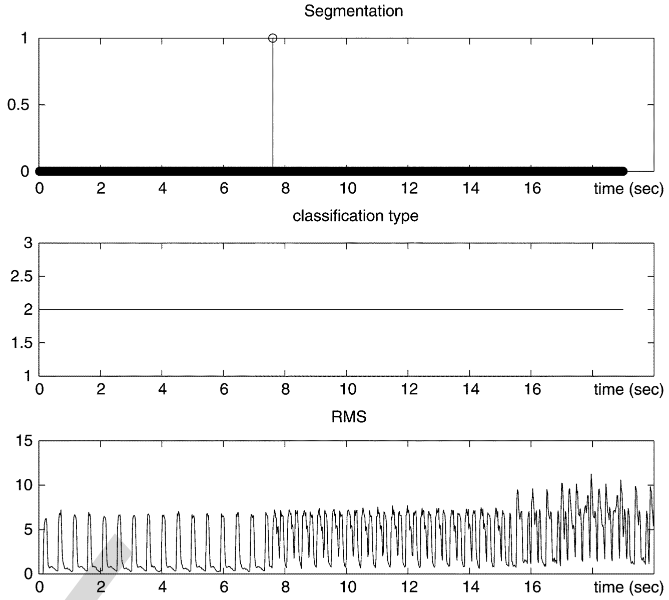

Fig. 19. Example of correct segmentation and erroneous classification.

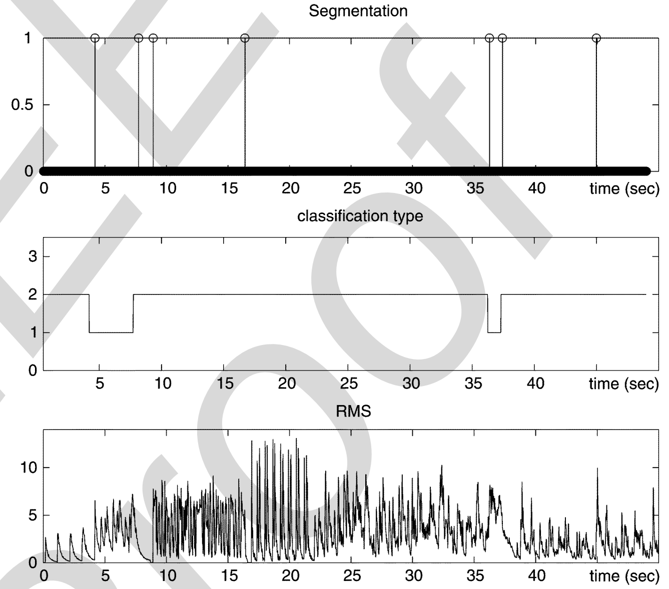

Fig. 20. False classifications due to a highly variant amplitude and to the presence of pauses in a music signal.

4) Silent Intervals Frequency: The silent intervals frequency, F_v, can discriminate music from speech, as it is in general greater for speech than for music. It is intended to measure the frequency of syllables. For music this feature almost always takes on a small value. Firstly, silent intervals are detected. A test is defined on the RMS value A and the ZC rate, as follows:

$$(\mathrm{A} < \mathrm{T}_1) \text{ or } (\mathrm{A} < 0.1\mathrm{A}_\mathrm{x} \text{ and } \mathrm{A} < \mathrm{T}_2) \text{ or } (\mathrm{ZC} = 0) \tag{13}$$

where A_x is the maximum RMS value on the whole segment. This test is applied over intervals of 20 ms. Using the above test a silent interval can be detected if its energy is very low or the number of zero crossings is null $(\mathrm{ZC} = 0)$. Because of noise,

IEEE Proof

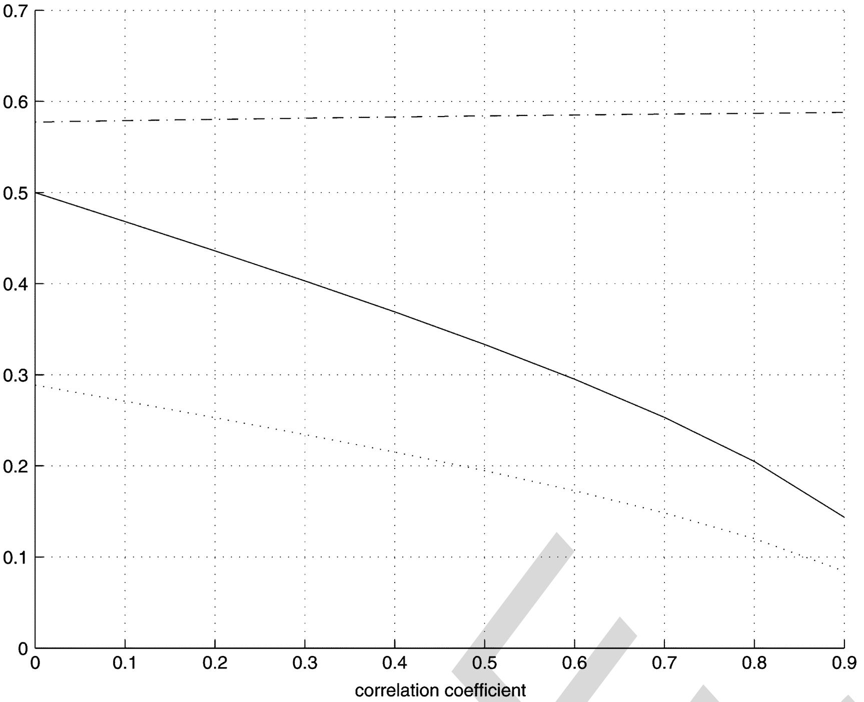

Fig. 21. Probability of ZC (solid line), the central frequency (dashed line) and their ratio (dashdot line) as a function of the correlation coefficient of a Gauss–Markov first order process.

IEEE Proof

there are cases where $A > T_1$, but these segments are silent intervals. This is solved by using the statement ($A < 0.1A_x$ and $A < T_2$). After detecting the silent intervals, neighboring silent intervals are grouped, as well as successive audible intervals. The number of silent intervals reported over the whole segment defines the so-called *silent intervals frequency*. In our experiments we found that almost always for speech signals $F_v > 0.6$, while for at least 65% of music segments, $F_v < 0.6$. This feature is highly correlated to the above defined ZC0. F_v is the rate of the silent intervals meanwhile ZC0 measures the duration of the silent intervals. Fig. 15 shows a transition from music to speech very well discriminated by the described feature.

5) Maximal Mean Frequency: One of the basic characteristics of speech waveforms is that they are bandlimited to about 3.2 kHz. The mean frequency is therefore smaller than this limit, and the maximal mean frequency $(\max(\mathrm{ZC}))$ can be used for taking advantage of this property. This feature can be estimated using the ZC rate. In order to reduce noise effects, only intervals with a large RMS value are considered. For speech signals the maximal mean frequency is almost always less than 2.4 kHz, while for music segments it can be much greater.

B. Classification Algorithm

Each segment is classified into one of three classes: silence, speech, or music. First, it is decided whether a signal is present and if so, the speech/music discrimination takes place.

1) Silent Segments Recognition: A measure of signal amplitude for a given segment is used for testing the signal presence

$$E = 0.7A_m + \frac{0.3}{N}\sum_{i=1}^{N} A(i). \tag{14}$$

This is a robust estimate of signal amplitude as a weighted sum of mean and median of the RMS. If the volume of silent segment is low and the segmentation method gives accurate boundaries for the silent segment its classification will be easy (using just the mean of RMS). In the opposite case (there is an error in boundary computation or noise in silent segment), we need a more robust criterion (a combination of mean value and median value of RMS). The weights were set according to the experimental results. A threshold is set for detecting the effective signal presence.

2) Speech/Music Discrimination: When the presence of a signal is verified, the discrimination in speech or music follows. The speech/music discriminator consists of a sequence of tests based on the above features. The tests performed are the following:

Silent intervals frequency: If $F_v < 0.6$, the segment is classified as music. This test classifies about 50% of music segments.

*RMS*ZC product:* If the feature C_Z is less than an empirically preset threshold, the segment is classified as speech.

Probability of ZC0: If this probability is greater than 0.1, the segment is classified as music.

Maximal mean frequency: If this frequency exceeds 2.4 kHz, the segment is classified as music.

Normalized RMS variance If the normalized RMS variance is greater than 0.24, the segment is classified as speech, otherwise it is classified as music.

The first four tests are positive classification criteria, i.e., if satisfied they indicate a particular class, otherwise we proceed to the next test for classification. Their order was determined by their performance (the first test has 100% performance, meanwhile the last one has 86%). The first four tests, which classify only in case of positive response, have almost 100% performance, i.e. a positive response is almost sure. This means that the thresholds are selected in order to obtain an almost sure result. The last test on the normalized RMS variance may lead to misclassifications. For this reason we choose the above simple and sequential algorithm instead of a more sophisticated procedure using machine learning techniques or neural networks. In our experiments the first four tests classified roughly 60% of the music segments and 40% of speech. The final test must decide the remaining segments, and here classification errors may occur. The different results are presented in the following section.

IV. Results

We have tested the proposed algorithms on a data set containing audio input through a computer's soundcard (15%), audio files from the WWW (15%), and recordings obtained from various archival audio CD's (70%). The sampling frequency ranged from 11 025 Hz to 44 100 Hz. The total speech duration was 11 328 s (3 h, 9 min) which was subdivided by the segmentation algorithm into about 800 segments (oversegmentation); 97% of these segments were correctly classified as speech. The total music duration was 3131 s (52 min), which was subdivided by the segmentation algorithm into about 400 segments (oversegmentation); 92% of these segments were correctly classified as music. The total number of speakers was 92 and the total number of music parts was 80. It has been used many different types of music like classical, jazz, rock, metal, blues, disco, techno, electronic.

In Table I, we present the experimental results. The various features are considered alone and in conjunction with others. The results with the complete above described algorithm are summarized in the last row of the table. The features are given in sequential order as processed. The normalized RMS variance alone has a success rate of about 86%. When it is combined with frequency measures, the correct classification rate reaches about 95%. Since all features are derived from the basic characteristics of signal amplitude and ZC rate, the combined use of the five features does not significantly increase the computation time.

Further results are given in Figs. 16–18. Each contains three plots: (a) the segmentation result, (b) the classification result, where 1 corresponds to music, 2 corresponds to speech and 3 corresponds to silence, and (c) the signal amplitude which alone determines the changes. The classification is always correct in these three files. Sometimes the signal is over-segmented, but the classifier retains only speech-to-music or music-to-speech transitions. We also present two results with erroneous classifications in Figs. 19 and 20. In both cases music with frequent instantaneous pauses and significant amplitude variations is falsely classified as speech.

The comparison with other methods could be unfair due to the variety of the data sets used. In the review of other methods presented in the Introduction, it appears that the correct classification percentage reported may vary from 80% to 99%, depending on the duration of the segments and of course on the data set. It should also depend on the features selected and the method applied, but no benchmark is available in order to have a definitive and reliable assessment of the different features and methods. Taking that into consideration, we can claim that we have proposed a new method which is simultaneously efficient, i.e., computable in real-time, and very effective.

V. Conclusion

In this paper, we have proposed a fast and effective algorithm for audio segmentation and classification as speech, music or silence. The energy distribution seems to suffice for segmenting the signal, with only about 3% transition loss. The segmentation is completed by the classification of the resulting segments. Some changes are verified by the classifier, and other segments are fused for retaining only the speech/music transitions. The classification needs the use of the central frequency, which is estimated efficiently by the ZC rate. The fact that the signal amplitude and the ZC rate are almost independent is appropriately exploited in the design of the implemented sequential tests. However, we have to note that for some musical genres the ZC rate could be low, while for impulsive musical sounds the ZC rate may be not so correlated to the spectral centroid as expected by our method. While the main advantage of the ZC rate is its simplicity, redundancy should be added in order to increase the robustness of the algorithm. A possible extension could be obtained by using the FFT with a few number of coefficients.

One possible application of the developed methods, which can be implemented in real-time, is in content-based indexing and retrieval of audio signals. The algorithms could also be used for monitoring broadcast radio, or as a preprocessing stage for speech recognition. Another possible application might be in portable devices with limited computing power such as cell phones, voice recorders, etc.

In the future, the methods introduced here could be extended to a more detailed characterization and description of audio. They may be used at the first hierarchical level of a classifier, and then continue by classifying into more specific categories, for example, classifying the music genre or identifying the speaker. The segmentation stage could be combined with video shot detection in audiovisual analysis.

Appendix
Correlation Between Zero-Crossing Rate and Central Frequency

The statistical characterization of the ZCs is a difficult problem. The ZC rate depends on the properties of the random process. In [5, pp. 485–487], it is proven that the density of ZCs for a continuous Gaussian process is

$$P(\mathrm{ZC}) = 2\sqrt{\frac{\int_{-\infty}^{\infty} f^2 S(f)\,df}{\int_{-\infty}^{\infty} S(f)\,df}} = 2f_c \qquad (15)$$

where $S(f)$ is the power spectrum density of the random process and f_c its spectrum centroid.

We examine in addition the correlation between the ZC rate and central frequency for a class of discrete-time random signals $(X(n))$. Let X_1 and X_2 be two random variables corresponding to two successive values of a first-order zero-mean Gauss Markov process.

IEEE Proof

Then the probability of a ZC $P(\mathrm{ZC})$ is given by

$$\begin{aligned} P(\mathrm{ZC}) &= \Pr\{X_1 X_2 < 0\} \\ &= \frac{1}{\pi\sqrt{1-\alpha^2}} \int_0^{\infty} \int_{-\infty}^{0} e^{-\frac{x^2-2\alpha xy+y^2}{2(1-\alpha^2)}}\, dx\, dy \quad (16) \end{aligned}$$

where α is the correlation coefficient between X_1 and X_2. The autocorrelation function $\gamma(m)$ of these signals is given by

$$\gamma(m) = E\{X(n) \cdot X(n-m)\} = \sigma^2 \alpha^{|m|} \qquad (17)$$

The central frequency of the power spectrum is given by

$$f_c = \sqrt{\frac{\int_0^{0.5} \frac{f^2}{1+\alpha^2-2\alpha\cos(2\pi f)}\, df}{\int_0^{0.5} \frac{1}{1+\alpha^2-2\alpha\cos(2\pi f)}\, df}}. \qquad (18)$$

The above integrals do not have a closed form, so we have computed them numerically for many values of α. In Fig. 21 we plot the $P(\mathrm{ZC})$ and f_c for many values of α. We observed that the $P(\mathrm{ZC})$ and f_c are strongly correlated $(P(\mathrm{ZC}) \approx 1.71 f_c)$.

REFERENCES

[1] J. Foote, "An overview of audio information retrieval," *Multimedia Syst.*, pp. 2–10, 1999.

[2] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proc. IEEE*, vol. 74, pp. 1477–1493, 1986.

[3] *Handbook of Statistics: Nonparametric Methods*, P. R. Krishnaiah and P. K. Sen, Eds., North-Holland, Amsterdam, The Netherlands, 1984.

[4] P. Moreno and R. Rifkin, "Using the fisher kernel method for web audio classification," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, pp. 1921–1924.

[5] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1965.

[6] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. IEEE ICASSP*, 1996.

[7] E. Scheier and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE ICASSP*, 1997.

[8] M. Seck, F. Bimbot, D. Zugah, and B. Delyon, "Two-class signal segmentation for speech/music detection in audio tracks," in *Proc. Eurospeech*, Sep. 1999, pp. 2801–2804.

[9] A. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.

[10] G. Tzanetakis and P. Cook, "A framework for audio analysis based on classification and temporal segmentation," in *Proc. 25th Euromicro Conf. Workshop on Music Technology and Audio Processing*, 1999.

[11] ——, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 4, pp. 293–302, Jul. 2002.

[12] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia Mag.*, pp. 27–36, 1996.

[13] *Handbook of Pattern Recognition and Image Processing*, T. Young and K.-S. Fu, Eds., Academic, New York, 1986.

[14] T. Zhang and J. Kuo, "Audio content analysis for on-line audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 441–457, May 2001.

Costas Panagiotakis was born in Heraklion, Crete, Greece, on April 27, 1979. He received the B.Sc. and M.Sc. degrees in computer science from the University of Crete in 2001 and 2003, respectively

Since 1999, he is involved in Research and Development European projects in the field of multimedia and image analysis. His research interests include signal processing, image processing and analysis, computer vision, algorithms, motion analysis, and neural networks.

Georgios Tziritas (M'89–SM'00) was born in Heraklion, Crete, Greece, on January 7, 1954. He received the Diploma of Electrical Engineering degree in 1977 from the Technical University of Athens, and the Diplome d'Etudes Approfondies (DBA) in 1978, the Diplome de Docteur Ingenieur in 1981, and the Diplome de Docteur d'Etat in 1985, all from the Institut Polytechnique de Grenoble, France.

From 1982 until August 1985, he was a Researcher of the Centre National de la Recherche Scientifique, with the Centre d'Etudes des Phenomenes Aleatoires (CEPHAG), with the Institut National de Recherche en Informatique et Automatique (INRA), until January 1987, and with the Laboratoire des Signaux et Systemes (LSS). From September 1992, he was Associate Professor and, from April 2003, he is Full Professor at the Department of Computer Science, University of Crete, teaching digital signal processing, digital image processing, digital video processing, and information and coding theory. He is coauthor (with C. Labit) of the book *Motion Analysis for Image Sequence Coding* (Amsterdam, The Netherlands: Elsevier, 1994), and of more than 70 journal and conference papers on signal and image processing, and image and video analysis. His research interests are in the areas of multimedia signal processing, image processing and analysis, computer vision, motion analysis, image, and video indexing, and image and video communication.

IEEE Proof