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

C. Panagiotakis and G. Tziritas*

Department of Computer Science, University of Crete,

P.O. Box 2208, Heraklion, Greece

E-mails: {*cpanag,tziritas*}@*csd.uoc.gr*

Abstract

Over the last years major efforts have been made to develop methods for extracting information from audio-visual 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 audio-visual 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. 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 msec intervals, resulting in the segments' limits being specified within an accuracy of 20 msec. 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 – speech/music classification, audio segmentation, zero-crossing rate

EDICS: 4-KNOW

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

Scheirer and Slaney [5] used thirteen 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 msec segments and 98.6% for 2.4 sec segments. Tzanetakis and Cook [8] proposed a general framework for integrating, experimenting and evaluating different techniques of audio segmentation and classification. In addition they proposed a segmentation method based on feature change detection. For their experiments on a large data set a classifier performance of about 90% is reported.

In [9] a system for content-based classification, search and retreival 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. In a more general framework related issues are reviewed in [1].

In [3] and [6] cepstral coefficients are used for classifying or segmenting speech and music. Moreno and Rifkin [3] 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 reported. In [6] Gaussian mixtures are used too, but the segmentation is obtained by the likelihood ratio. For very short (26 msec) segments a correct classification rate of 80% is reported.

A general remark concerning the above techniques is that often a large number of features are used. 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. We concluded that a reliable discriminator can be designed using only the signal amplitude, equivalent to the energy reported previously, and the central frequency, measured by the zero-crossing rate, a feature already exploited in previous work. In addition we analysed the data in order to extract relevant parameters for making the statistical tests as effective as possible.

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 used independently. Features extracted from the zero-crossing rate are added and combined to the amplitude parameters.

B. Description of signal and its characteristics

The signal is assumed to be monophonic. In the case of multi-channel audio signals the average value is taken as input. There are no restrictions on the sampling frequency functioning equally well from 11025 Hz to 44100 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.

Two signal characteristics are used: the amplitude, measured by the Root-Mean-Square (RMS), and the mean frequency, measured by the average density of zero-crossings. One measure of each is acquired every 20 msec. We describe these in the following paragraphs.

The signal amplitude, A, is defined as follows:

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

Voice and music are distinguished by the distribution of amplitude values. Figures 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 may be exploited for both segmentation and classification. The mean frequency is approximated by the average number of zero-crossings

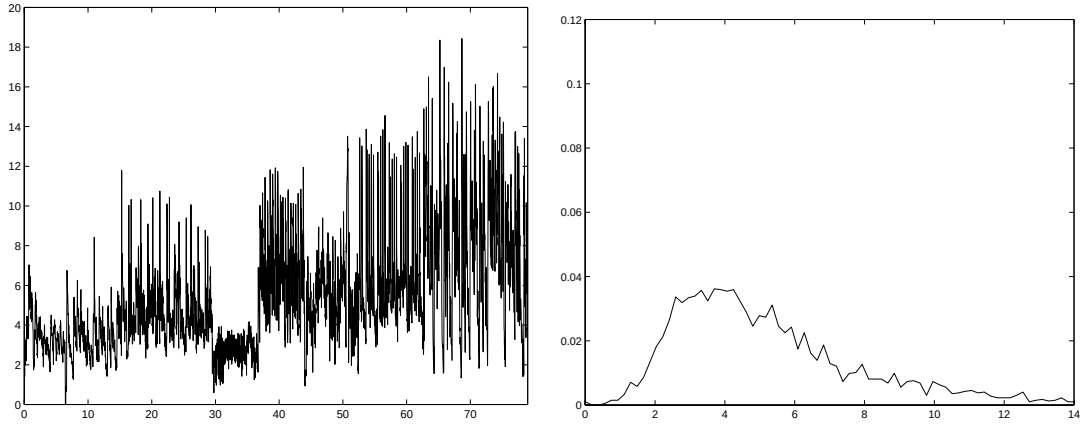

Fig. 1. The RMS of a music signal and its histogram.

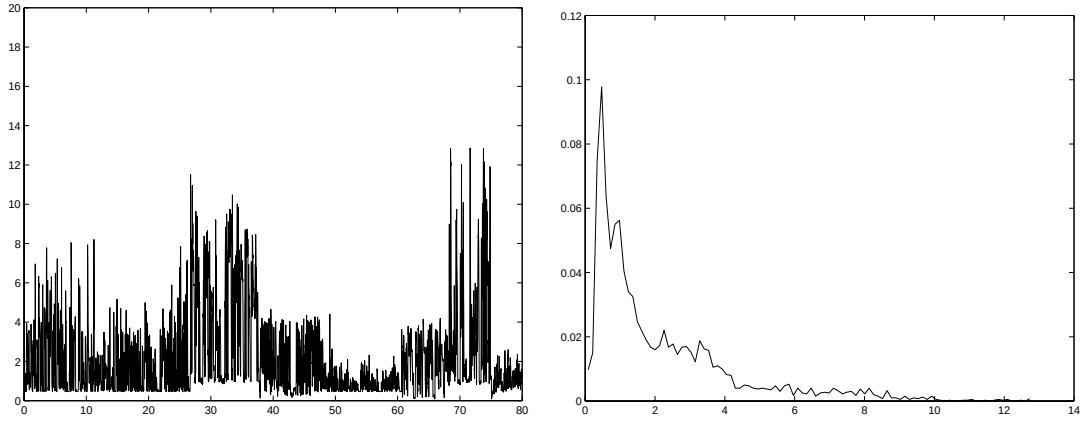

Fig. 2. The RMS of a voice signal and its histogram.

in the 20 msec interval. Figures 3 and 4 show the zero-crossing 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 [2], defined as

$$V = \frac{|n_1 - n_2|}{n} \tag{2}$$

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

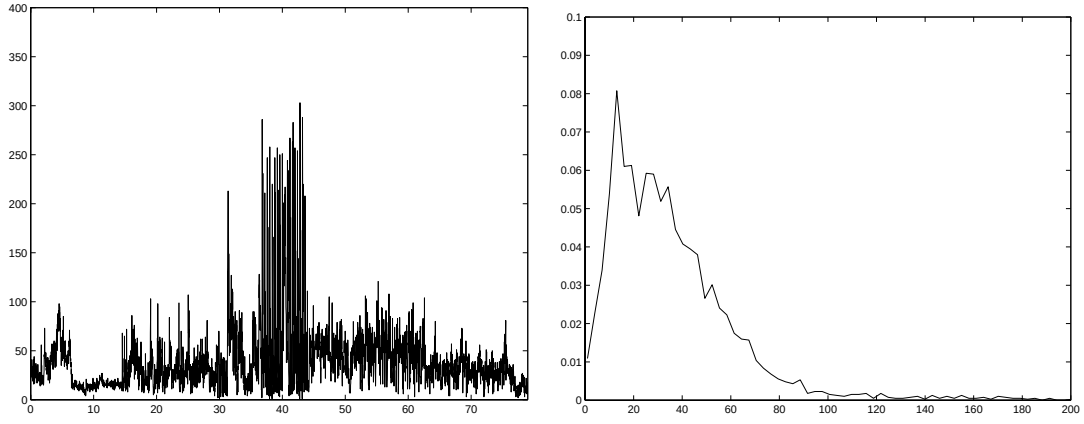

Fig. 3. The average number of zero-crossings for a music signal and its histogram.

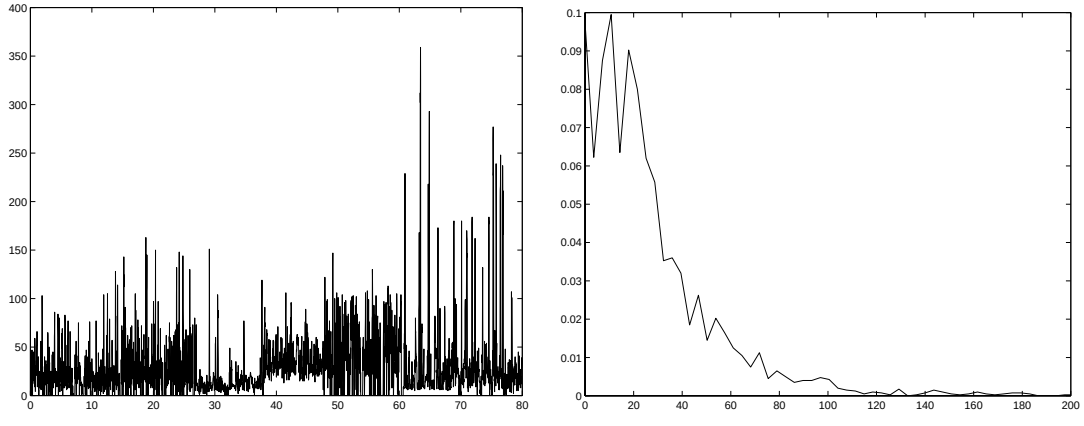

Fig. 4. The average number of zero-crossings for a voice signal and its histogram.

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

$$I = \sum P_i \log \frac{1}{P_i} + \sum Q_j \log \frac{1}{Q_j} \tag{3}$$

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. We exploit this possible discrimination in a feature defined for the classification.

In [5], [8] and [9] 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 zero-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 syllabical 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 [3] [6].

II. Segmentation

Segmentation is implemented in real-time and is based only on RMS. For each 1 sec frame 50 values of the RMS are computed from successive intervals of 20 msec. The mean and the variance of the RMS is 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 msec, is marked. The last stage is more time consuming, but is employed only in case of frame change detection.

The instantaneous accuracy is fixed at 20 msec because the human perceptual system is generally not more precise, and moreover because speech signals remain stationary for 5–20 msec [7]. The maximal interval for measuring speech characteristics should therefore be limited to intervals of 20 msec.

A. Change detection between frames

In this stage, frames containing probable transitions are sought. A change is detected 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 pevious 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 do not attempt to measure the distance between the distributions, which could be expensive, but rather search for an appropriate model for them in order to reduce the problem to the estimation of some parameters, and obtain the dissimilarity as a function of these parameters. We have observed that the generalized χ^2 distribution fits well the histograms for both music and speech (Figures 5 and 6). 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

$$p(x) = \frac{x^a e^{-bx}}{b^{a+1}\Gamma(a+1)}, \qquad x \geq 0. \tag{4}$$

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

$$a = \frac{\mu^2}{\sigma^2} - 1 \quad \text{and} \quad b = \frac{\sigma^2}{\mu}. \tag{5}$$

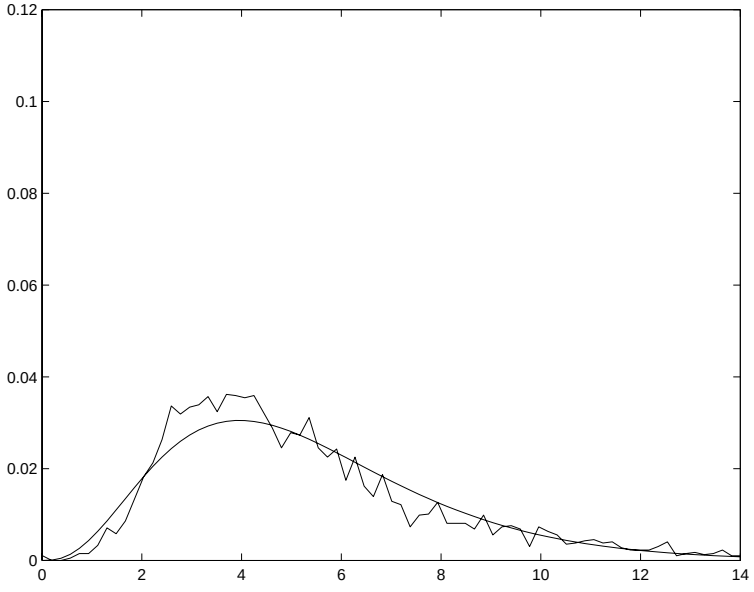

Fig. 5. RMS histogram for a collection of music data and its fitting by the generalized χ^2 distribution.

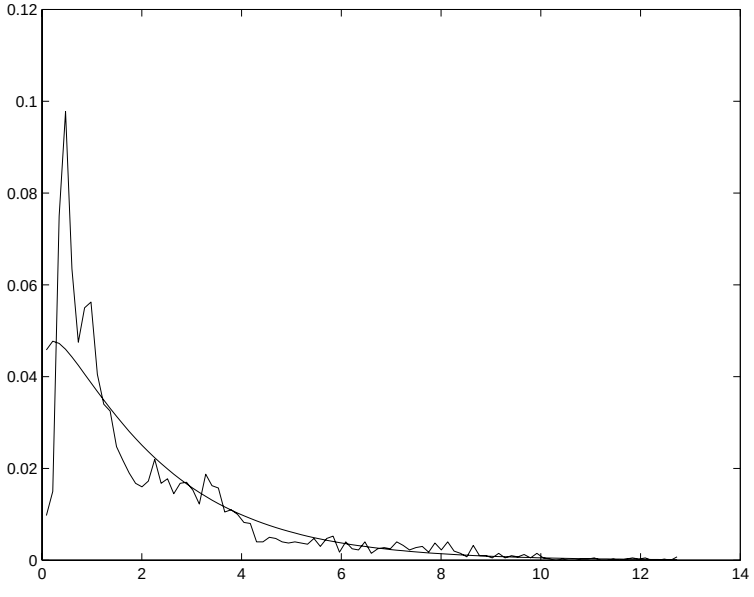

Fig. 6. RMS histogram for a collection of voice data and its fitting by the generalized χ^2 distribution.

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

$$\rho(p_1, p_2) = \int \sqrt{p_1(x)p_2(x)}dx \tag{6}$$

The similarity takes values in the interval $[0, 1]$, where the value 1 means identical distributions, and zero means completely non-intersecting distributions. For this reason, the value $1-\rho$, known as the Matusita distance [10], 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 [10]. For the case of two equiprobable hypotheses the classification error is bounded by

$$P_e \leq \frac{\rho(p_1, p_2)}{2}. \tag{7}$$

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

$$\rho(p_1, p_2) = \frac{\Gamma(\frac{a_1+a_2}{2}+1)}{\sqrt{\Gamma(a_1+1)\Gamma(a_2+1)}} \frac{2^{\frac{a_1+a_2}{2}+1} b_1^{\frac{a_2+1}{2}} b_2^{\frac{a_1+1}{2}}}{(b_1+b_2)^{\frac{a_1+a_2}{2}+1}}. \tag{8}$$

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,

$$D(i) = 1 - \rho(p_{i-1}, p_{i+1}). \tag{9}$$

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 neighbouring 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:

$$D_n(i) = \frac{D(i)V(i)}{D_M(i)}, \tag{10}$$

where $V(i)$ measures the (positive) difference of $D(i)$ from the mean value of the neighbouring 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 Figures 7 and 8. 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. 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 msec, the maximal accuracy of our method. 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 sec and the distance measure is that of Equation (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. The final segmentation is completed by a classification stage, which could also be used independently for the characterization of audio signals. In Figures 9 and 10 we show the instant change detections 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. We have tested our technique extensively, 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 sec. Some examples of segmenation results are shown in Figures 7, 8 and 11.

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 zero-crossings. 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.

A.1 Normalized RMS variance

The normalized RMS variance is defined as the ratio of the RMS variance to the square of RMS mean. It is therefore equal to the inverse of parameter $a + 1$. This feature is volume invariant. In Figure 12 we show two typical histograms of the normalized variance for speech and music signals. We observe that the two distributions are almost non-overlapping, 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 variance of RMS is used as the final test in our algorithm.

A.2 The probability of null zero-crossings

The zero-crossing rate is related to the mean frequency for a given segment. In the case of a silent interval the number of zero-crossings is null. In speech there are always some silent intervals, thus the occurence of null zero-crossings 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 is set to 0.1. Our experiments showed that about 40% of speech verifies this criterion, while we have not found any music segment exceeding the threshold. Comparing the histograms in Figures 3 and 4, we see the discriminating capability of the null zero-crossings feature.

A.3 Joint RMS/ZC measure

Together with the RMS and null zero-crossings features we exploit the fact that RMS and ZC are somewhat correlated for speech signals, while essentially independent for music signals. Thus we define a feature related to the product of RMS and ZC,

$$C_Z = e^{-\psi C}, \quad 1 \geq \psi \geq 5, \tag{11}$$

where

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

with $A_x = \max\{A(i) : 1 \leq i \leq N\}$, $A_n = \min\{A(i) : 1 \leq i \leq N\}$ and $A_m = \text{median}\{A(i) : 1 \leq i \leq N\}$. The normalization by $2A_x - A_n - A_m$ is used because in speech signals the denominator usually takes on large values, as the median and the minimum values are small for such a signal. The test consists of comparing this feature to some threshold. If C_Z is close to 1, then the segment is classified as speech.

A.4 Void intervals frequency

The void 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, void intervals are detected. A test is defined on RMS and ZC, as follows:

$$(RMS < T_1) \text{ or } (RMS < 0.1 \max(RMS) \text{ and } RMS < T_2) \text{ or } (ZC = 0) \tag{12}$$

This test is applied over intervals of 20 msec. The $\max(RMS)$ is determined for the whole segment. After detecting the void intervals, neighbouring silent intervals are grouped, as well as successive audible intervals. The number of void intervals reported over the whole segment defines the so-called *void 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, $F_v < 0.6$. Figure 13 shows a transition from music to speech, very well discriminated by the described feature.

A.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 can be used for taking advantage of this property. This feature can be estimated using the

zero-crossing 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.

B.1 Silent segments recognition

A measure of signal amplitude for a given segment is used for testing the signal presence

$$E = 0.7A_m + \frac{0.3}{N}\sum_{i=1}^{N} A(i) \tag{13}$$

This is a robust estimate of signal amplitude as a weighted sum of mean and median of the RMS. A threshold is set for detecting the effective signal presence.

B.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:

Void 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 exceeds an empirically preset threshold, the segment is classified as speech.

Probability of null zero-crossings If this probability is greater than 0.1, the segment is classified as speech.

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 thresholds are selected in order to obtain a decision with near certainty. 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

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%

TABLE I

THE PERFORMANCE OF THE VARIOUS FEATURES INDIVIDUALLY AND IN CONJUCTION.

segments, and here classification errors may occur. These 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 CDs (70%). The sampling frequency ranged from 11025 Hz to 44100 Hz. The total speech duration was 11328 sec (3 h, 9 min) which was subdivided by the segmentation algorithm into about 800 segments. 97% of these segments were correctly classified as speech. The total music duration was 3131 sec (52 min) which was subdivided by the segmentation algorithm into about 400 segments. 92% of these segments were correctly classified as music.

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 zero-crossing rate, the combined

use of the five features does not significantly increase the computation time.

Further results are given in Figures 14, 15 and 16. Each of these Figures 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 Figures 17 and 18. 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 assessement 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. Conclusions

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 zero-crossing rate. The fact that the signal amplitude and the zero-crossing rate are almost independent is appropriately exploited in the design of the implemented sequential tests.

One possible application of the developed methods, which can be implemented in real-time, is in content-based indexing and retreival of audio signals. The algorithms could also be used for broadcast radio monitoring, or as a pre-processing stage for speech recognition.

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 audio-visual analysis.

References

[1] J. Foote. An overview of audio information retrieval. *Multimedia Systems*, pages 2–10, 1999.

[2] P.R. Krishnaiah and P.K. Sen (eds). *Handbook of statistics: Nonparametric methods.* North-Holland, 1984.

[3] P. Moreno and R. Rifkin. Using the fisher kernel method for web audio classification. In *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing*, pages 1921–1924, 2000.

[4] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, 1996.

[5] E. Scheier and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, 1997.

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

[7] A. Spanias. Speech coding: a tutorial review. *Proc. of the IEEE*, 82:1541–1582, Oct. 1994.

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

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

[10] T. Young and K.-S. Fu (eds). *Handbook of pattern recognition and image processing.* Academic Press, 1986.

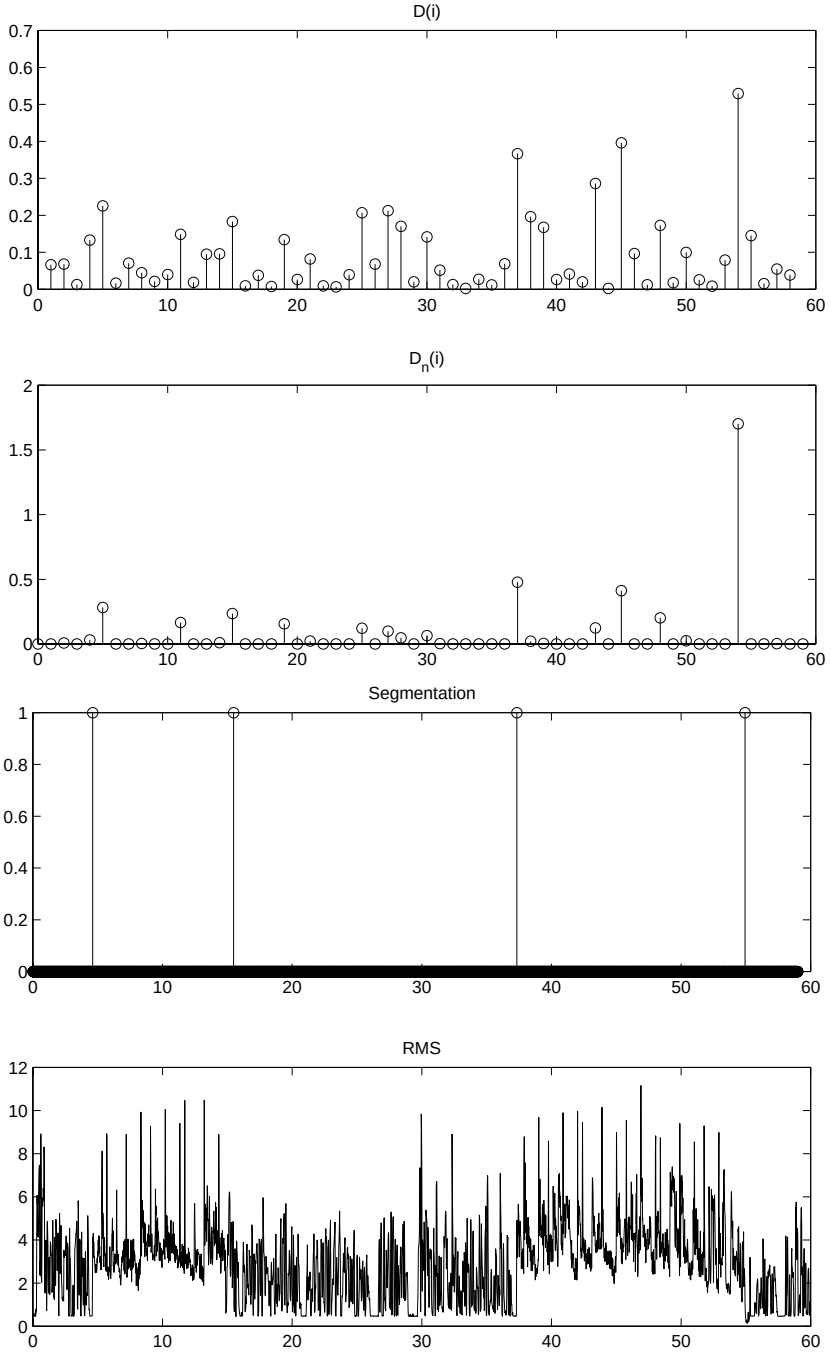

Fig. 7. An example of segmentation with four transitions. Are shown: the distance $D(i)$, the normalized distance $D_n(i)$, the change detection result, and the RMS data.

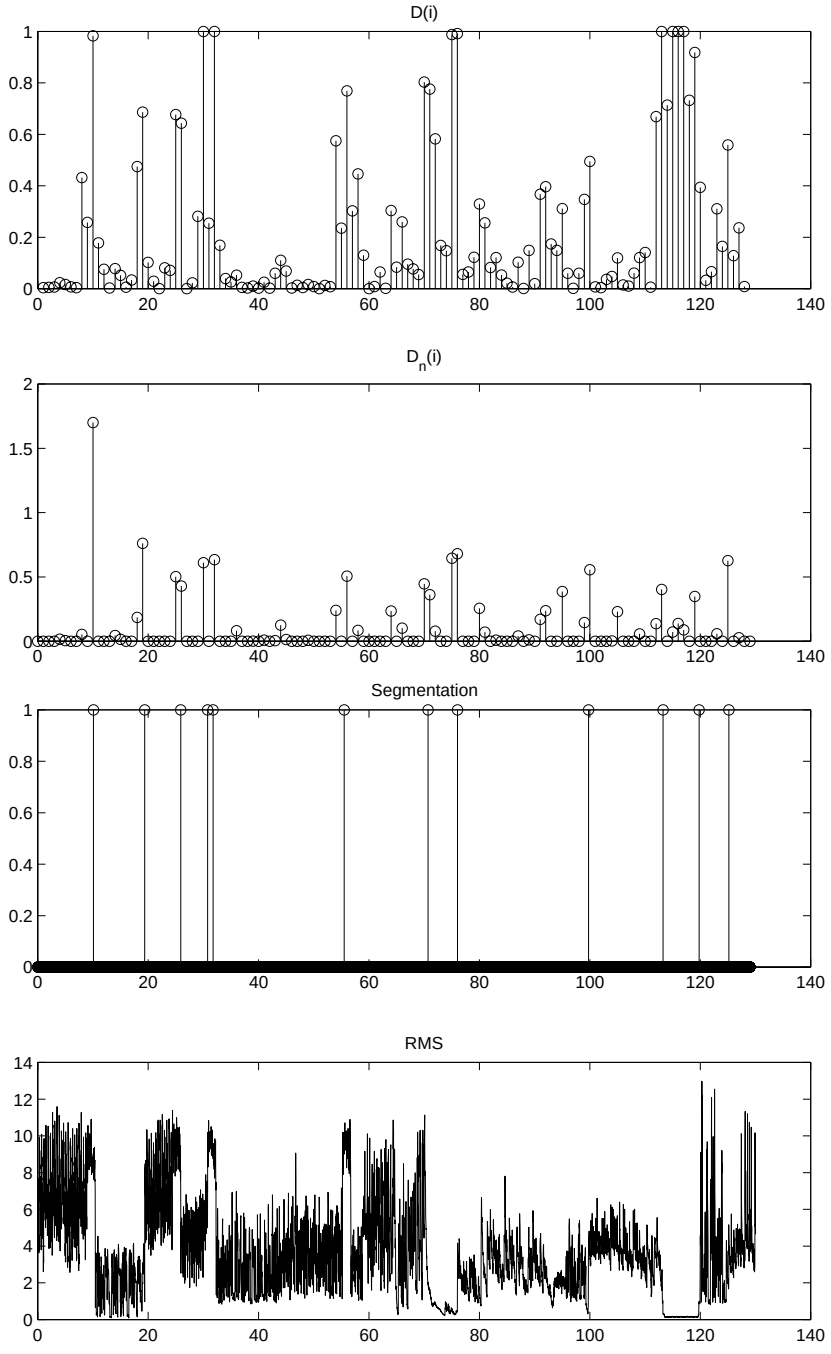

Fig. 8. Another example of segmentation with many transitions. Shown are: the distance $D(i)$, the normalized distance $D_n(i)$, the change detection result, and the RMS data

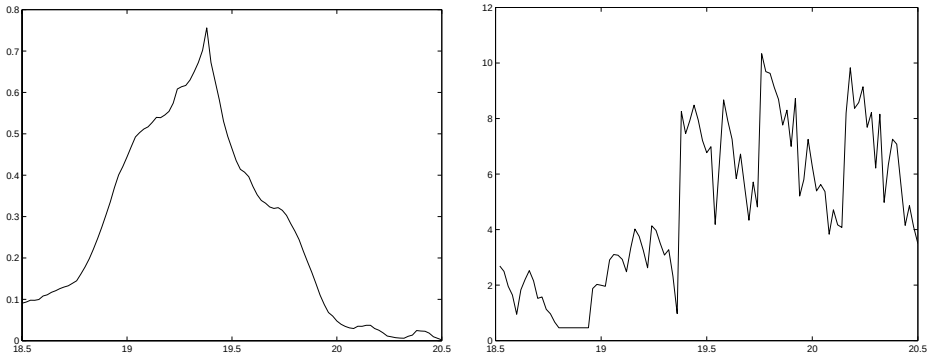

Fig. 9. Shown on the left is the distance $D(i)$ for the RMS shown in the right plot. The accuracy is excellent for this transition from speech to music.

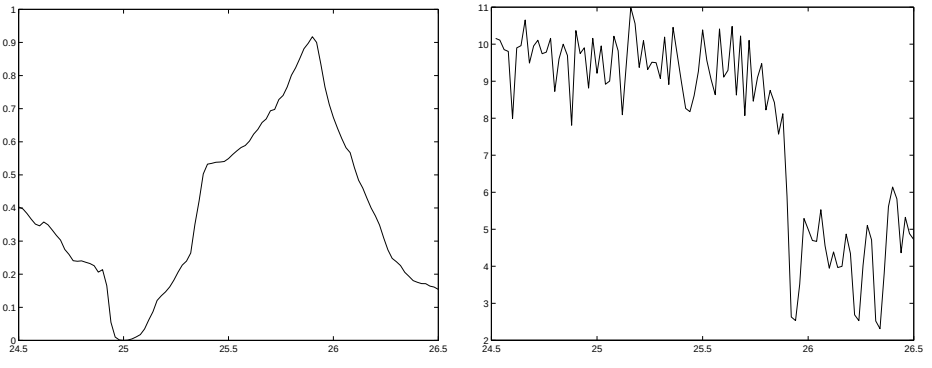

Fig. 10. Shown on the left is the distance $D(i)$ for the RMS shown in the right plot. The accuracy is very good for this transition from music to speech.

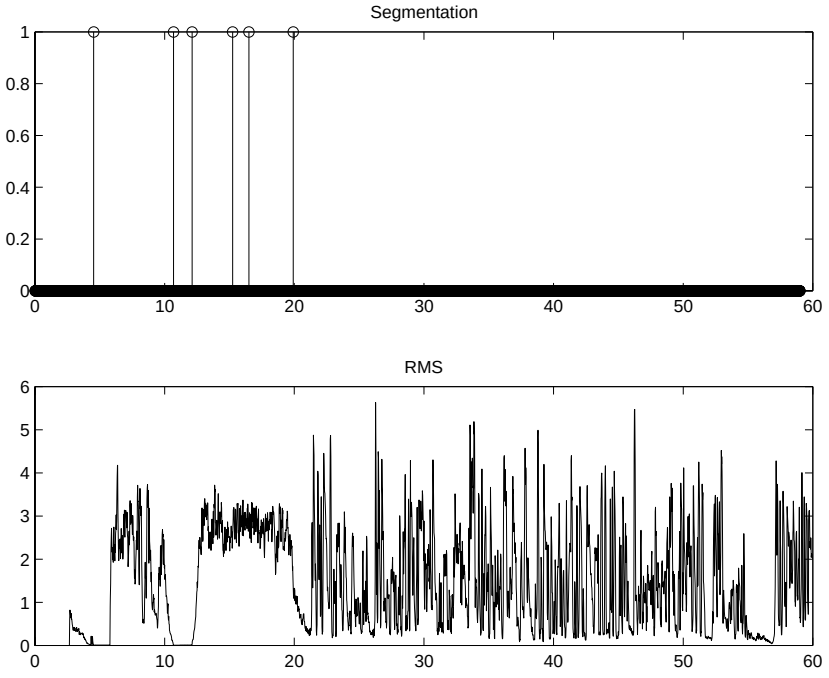

Fig. 11. The change detection is illustrated and the signal amplitude shown. No transition loss occurs but some segments are over-segmented.

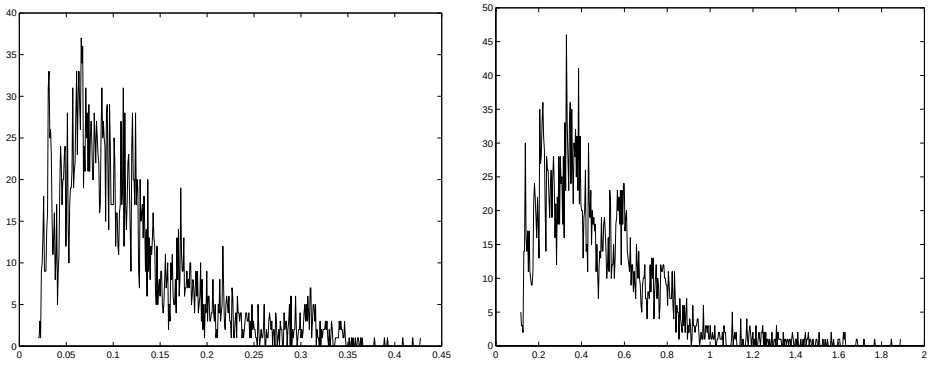

Fig. 12. Histograms of the normalized RMS variance for music (left) and voice (right).

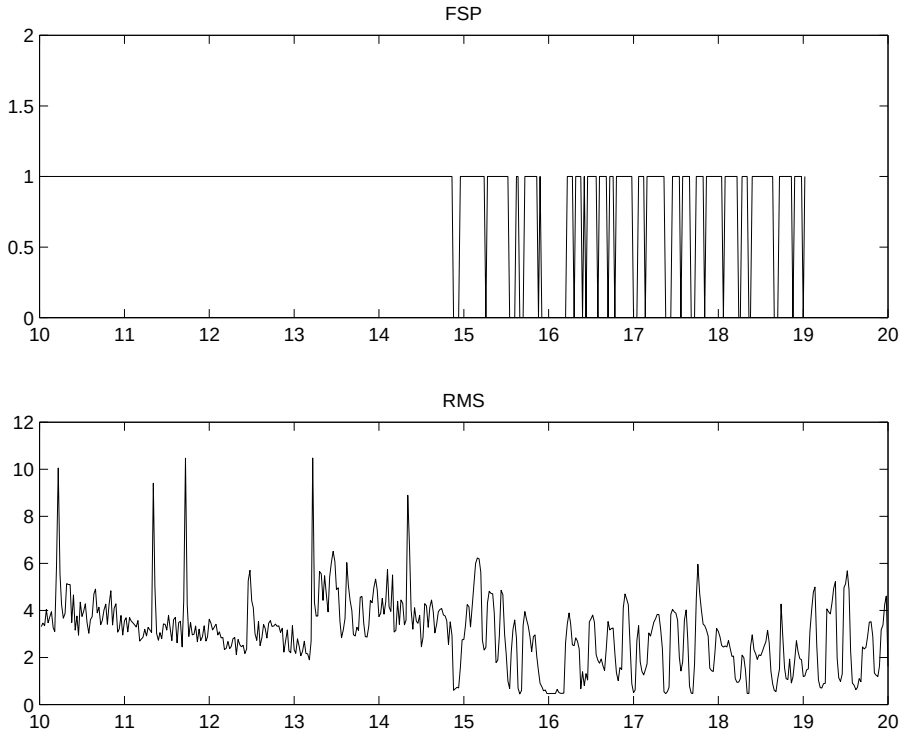

Fig. 13. Transition from speech to music. In the bottom the RMS is shown, and in the top the detected void intervals. Void intervals are more frequent in speech than in music.

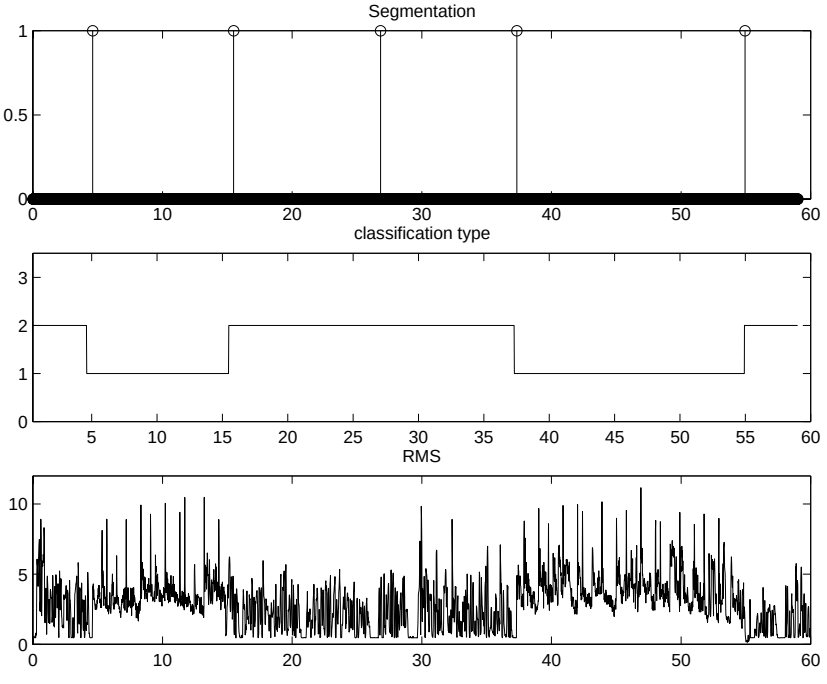

Fig. 14. A result of classification after the change detection. The second and the fourth segment are music, while the others are speech.

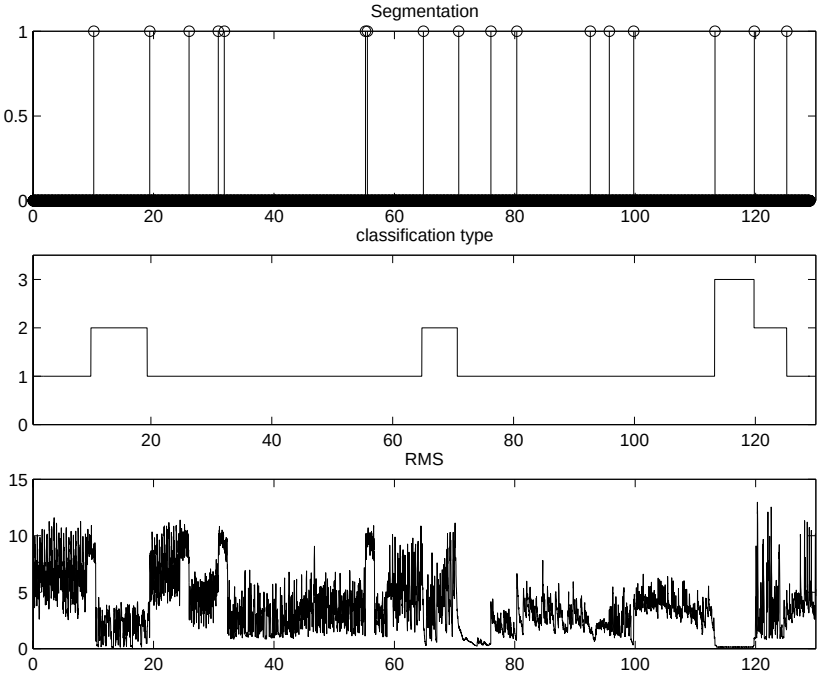

Fig. 15. An over-segmented signal for which all segments were correctly classified. 1: speech, 2: music, 3: silence.

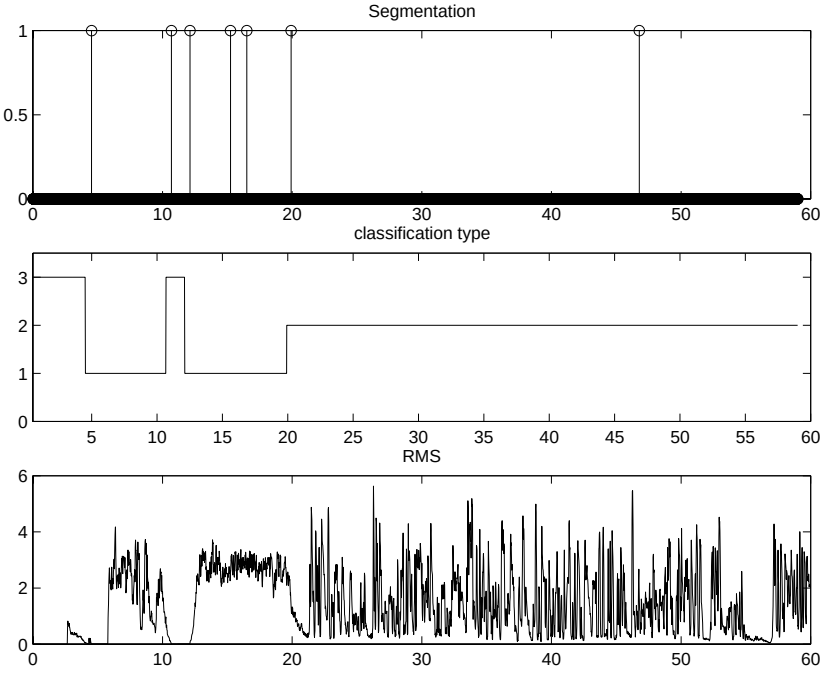

Fig. 16. An example of correct classification.

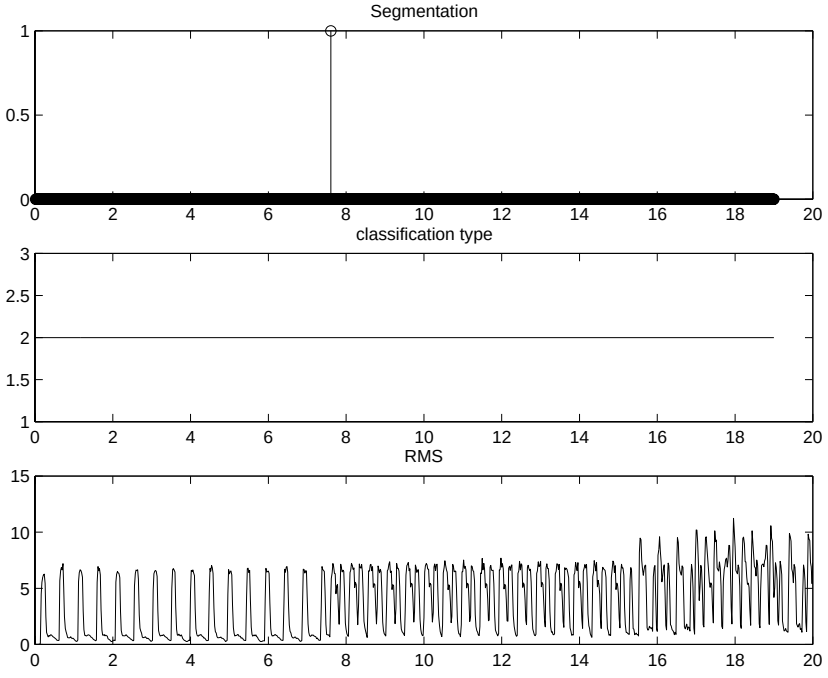

Fig. 17. An example of correct segmentation and erroneous classification.

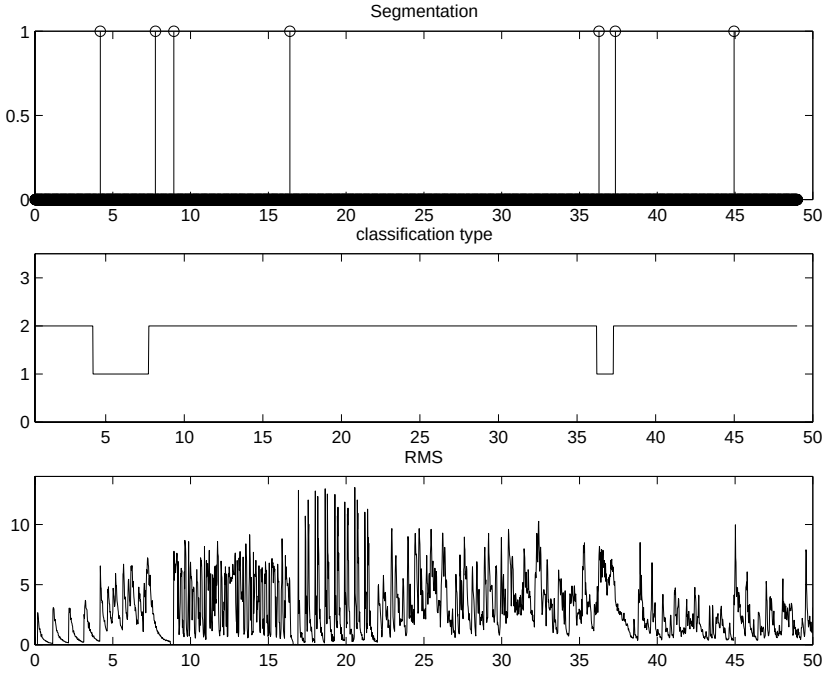

Fig. 18. False classifications due to a highly variant amplitude and to the presence of pauses in a music signal.