A Speech/Music Discriminator using 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

An audio segmentation method and a speech/music classifier are proposed. The characteristics used are considerably reduced. Segmentation is based on mean signal amplitude distribution, whereas classification utilizes an additional characteristic related to the mean frequency. 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%.

1 Introduction

1.1 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 speech or music. A variety of systems for audio segmentation and/or classification have been proposed and implemented in the past for the needs of various applications. Some of them are presented below.

Saunders [2] proposed a technique for discrimination of audio as speech or music using the energy contour and the zero-crossing rate. Scheirer and Slaney [3] used thirteen features, of which eight are extracted from the power spectrum density, for classifying audio segments. Tzanetakis and Cook [4] 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. In [5] 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. In a more general framework related issues are reviewed in [1].

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 2 we present the proposed segmentation method which is a change detector based on a dissimilarity measure of the signal amplitude distribution. In Section 3 the classification technique is presented which could either complete the segmentation, or used independently.

1.2 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, while the sound volume may differ from one recording to another. Two signal characteristics are used: the amplitude, A, measured by the Root-Mean-Square (RMS), and the mean frequency, measured by the average density of zero-crossings (ZC). One measure of each is acquired every 20 msec.

Voice and music are distinguished by the distribution of amplitude values. Fig. 1 shows the histograms of the RMS for a music and for a speech signal. The distributions are different and may be exploited for both segmentation and classification. Fig. 2 shows the histograms of the average zero-crossings for a music and for a voice signal.

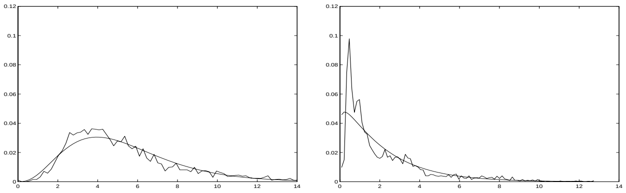

Figure 1: RMS histogram for a collection of music (resp. voice) data in the left (resp. right) and its fitting by the generalized χ^2 distribution.

The two characteristics used in our work are almost independent. We have verified this hypothesis using two different independence tests. 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

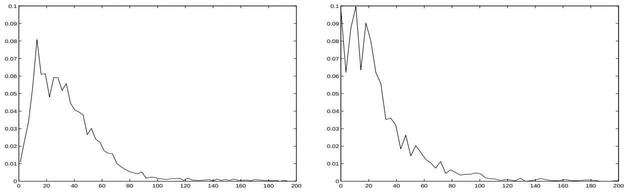

Figure 2: The histogram of the average number of zero-crossings for a music and a voice signal.

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.

2 Segmentation

Segmentation is based only on RMS. The mean and the variance of the RMS are calculated for frames of 1 sec. 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.

In the first 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. 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 search for an appropriate model for the distributions 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

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

fits well the histograms for both music and speech (Fig. 1). The parameters a, b are related to the the mean and the variance values of the RMS.

The segmentation will be based on a dissimilarity measure which is applied between frames. We propose to use a similarity measure defined on the probability density functions, $\rho(p_1, p_2) = \int \sqrt{p_1(x)p_2(x)}dx$. 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 [6], can be interpreted as the distance between the two signal segments.

At first the dissimilarity measure is used for localizing a candidate change frame. Therefore, we compute for each frame i a value, which gives the possibility of a change within that frame, $D(i) = 1 - \rho(p_{i-1}, p_{i+1})$. Basically, if there is a single change within frame i, then frames $i-1$ and $i+1$ must differ. Such a change locally maximizes 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. We introduce a locally normalized distance, named $D_n(i)$. The comparison of the distance $D(i)$ and the normalized distance is illustrated in Fig. 3. 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.

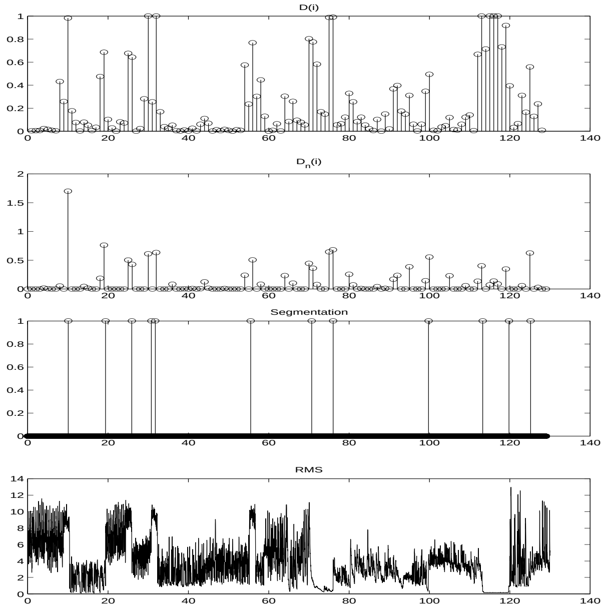

Figure 3: An example of segmentation where are shown: the distance $D(i)$, the normalized distance $D_n(i)$, the change detection result, and the RMS data.

The next step is detecting the change within an accuracy of 20 msec. For each change frame 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 dissimilarity is measured by the Matusita distance. 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. For this reason the segmentation is completed by a classification stage.

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. An example of segmentation results is shown in Fig. 3.

3 Classification

3.1 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.

Normalized RMS variance The normalized RMS variance is defined as the ratio of the RMS variance to the square of RMS mean. This feature is volume invariant. 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.

The probability of null ZC The ZC rate is related to the mean frequency for a given segment. In the case of a silent interval the number of ZC is null. In speech there are always some silent intervals, thus the occurence of null ZC is a relevant feature, noted $ZC0$, for identifying speech. Thus if this feature exceeds a certain threshold, the tested segment almost certainly contains a voice signal. 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 Fig. 2 we see the discriminating capability of the $ZC0$ feature.

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, noted C_Z.

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. 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, $F_v < 0.6$.

Maximal mean frequency Speech waveforms 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 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.

3.2 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.

A measure of signal amplitude for a given segment is used for testing the signal presence. We use a robust estimate of signal amplitude which is a weighted sum of mean and median of the RMS. A threshold is set for detecting the effective signal presence.

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: (1) Void intervals frequency: if $F_v < 0.6$, the segment is classified as music. (2) RMS*ZC product: if the feature C_Z exceeds an empirically preset threshold, the segment is classified as speech. (3) Probability of null zero-crossings: if this probability is greater than 0.1, the segment is classified as speech. (4) Maximal mean frequency: if this frequency exceeds 2.4 kHz, the segment is classified as music. (5) 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 segments.

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%
All	92%	97%

Table 1: The performance of the various features individually and in conjuction.

3.3 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 1 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.

Classification results are shown in Fig. 4. Three plots are shown: (a) the segmentation result, (b) the classification result, and (c) the signal amplitude which alone determines the changes. Sometimes the signal is over-segmented, but the classifier retains only speech-to-music or music-to-speech transitions.

4 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.

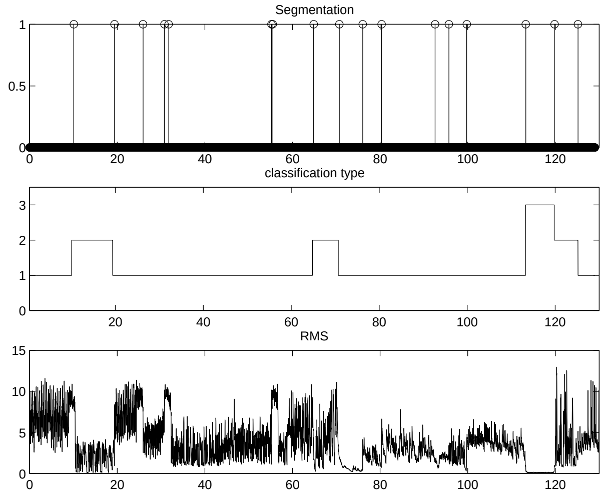

Figure 4: An over-segmented signal for which all segments were correctly classified. 1: speech, 2: music, 3: silence.

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.

References

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

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

[3] E. Scheirer 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.

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

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

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