FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Speech/Music classification by using the C4.5 decision tree algorithm

Diego Castán, Alfonso Ortega, Eduardo Lleida

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
dcastan@unizar.es, ortega@unizar.es, lleida@unizar.es

Abstract

In this work a study about different features for classification of audio frames into speech or music is presented. This paper focuses on the following set of features: High Zero-Crossing Rate Ratio (HZZCR), Variation of Spectral Flux (VSF), Low Short Time Energy Ratio (LSTER), Amplitude Modulation Ratio (AMR), Mel-Frequency Cepstrum Coefficients Variation (Var.MFCC) and Minimum-Energy Tracking (MET). In addition, we propose the use of a system based on a decision tree in order to combine the proposed set of features getting an improvement in the number of correct classifications. Experimental results on a broadcast radio database are presented showing that the selected features along with the use of the decision tree classifier allows the segregation of speech from music with a high degree of accuracy.

Index Terms: Speech/Music classification, time-domain features, frequency-domain features, cepstral-domain features and C4.5 decision tree.

1. Introduction

The growth and development of the Internet and Information and Communication Technologies (ICTs) in recent years has led to a dramatic increase of the number of multimedia documents. Therefore it is necessary to develop efficient systems that allow the organization, search and manipulation for a proper classification and information storage. Many of the current indexing systems seek for the desired information using tags that were defined by the users so the semantic description may be limited to a few words or phrases. As an alternative to this type of systems there exists a new field of research, Automatic Audio Content Analysis (ACA) that tries to develop information retrieval systems that analyze the audio data and extract the information directly from the audio signals. An example of automatic indexing system is given in Figure 1.

In our case, we are involved in developing an automatic system able to classify broadcast radio data which often includes sections with different types of signals like music, voice or music and voice. To develop our system we focus on broadcast ratio data transferred by "Aragón Radio", the public radio station in Aragón. As a first step, a fundamental task is to segment this type of signals and classify those segments into different classes according to acoustical criteria. After this preprocessing different subsystems can be applied like speaker recognition and automatic transcription of sections classified as voice.

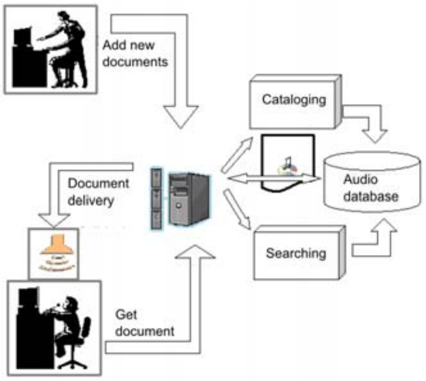

Figure 1: *Indexing system*

One of the difficulties in speech/music classification is to create a robust model for the identification of music signals. Speech is composed of a selection of typical sounds, therefore, it can be represented quite well by using simple models. However, music signals are composed of a big amount of different sounds produced by many different instruments with very different nature and often by many simultaneous sources. So music is defined by different styles and it can become difficult to create a single robust model. Moreover, the difficulty of our goal increases considering that we want to identify sections where speech and music overlap.

Former studies offer a rich set of papers about speech/music classification with a variety of applications. They describe which acoustic features can be used to emphasize the differences between music and speech. [1] presents a review of different solutions and the acoustic features used in each one of them. It can be found that many of the proposed solutions use the same features by varying the length of the feature vectors and the classification algorithm.

The solution proposed in this paper uses a set of six features. The combination of them is made by means of a C4.5 binary decision tree to provide the minimum classification error. The reduced dimensionality of the feature vector allows this algorithm to be used in real time. In addition, the model has been trained to identify the three most common classes that can be found on broadcast radio data: Music (M), speech (V) and the overlapping of both (A).

The authors would like to thank Aragón Radio for providing the data. This work has been partially supported by the national project TIN2008-06856-C05-04

The paper is arranged as follows: section 2 discusses the set of features used in the system, section 3 describes the classification algorithm, section 4 provides an evaluation of the system and is followed by the conclusion in section 5.

2. Features

Music and speech are described differently in both time and frequency domains. Speech signals are stationary for short periods of time (between 5 and 100 ms) while music signals should have larger stationary periods (around 200 ms). This means that speech has remarkable energy changes due to the alternation between voiced and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure.

In the spectral domain, speech signals present a rapidly and constantly changing energy distribution while music has a tonal structure which results in a more structured spectral distribution. We propose here a system that combines both temporal and spectral characteristics in order to be able to get the benefits from both descriptions.

2.1. Amplitude Modulation Ratio (AMR)

This feature calculates the relationship between local minima and local maxima in the envelope signal as detailed in (1). The envelope is obtained by filtering the signal with a lowpass filter with 25 Hz cutoff frequency. The alternation of high energy and low energy segments (vowels and consonants) in a speech signal causes the amplitude modulation ratio to be higher for speech and lower for music signals.

$$ID = \frac{V_{max} - V_{min}}{V_{max} + V_{min}} \tag{1}$$

2.2. High Zero-Crossing Rate Ratio (HZZCR)

One of the most widely used acoustic feature in music/speech classification is the Zero-Crossing Rate (ZCR) [2]. However, there are several variants of this feature like the High Ratio Zero Crossing Rate (HZCRR) [3]. HZCRR is defined as the ratio of the number of frames whose ZCR are above 1.5 times the average zero-crossing rate as described in (2). The advantage of HZCRR against ZCR is that HZCRR provides a more robust description of music and speech by removing the small variations that may occur with ZCR. Hence, for speech signal, HZCRR variation will be greater than that of music.

$$HZCRR = \frac{1}{2N} \sum_{n=0}^{N-1} \left[sgn(ZCR(n) - 1.5\overline{ZCR}) + 1 \right] \tag{2}$$

To obtain the HZCRR, first we calculate the ZCR in 20 ms windows and 10 ms frame shift. HZCRR is calculated then in 1 second windows. A median filter is applied to smooth the resulting values.

2.3. Variation of Spectral Flux (VSF)

The spectral flux provides an idea about the changes that occur in the shape of the spectrum frame by frame [4]. Speech exhibits an alternating sequence of noise-like segment with some others that present a more stationary behaviour. On the other hand, music presents a tonal structure due to a succession of periods of relative stability (notes or chords). In other words, the speech signal is distributed along the spectrum in a more random way than music does. The Spectrum Flux can be defined as the ordinary Euclidean norm of the delta spectrum magnitude which is calculated as:

$$SF = \|\mathbf{S}_i - \mathbf{S}_{i-1}\| = \frac{1}{N} \left(\sum_{k=0}^{N-1} (S_i(k) - S_{i-1}(k))^2 \right)^{\frac{1}{2}}, \tag{3}$$

where S_i is the spectrum magnitude vector of frame *i*, which is defined as the DFT with a frame size of 20 ms and frame shift of 10 ms. Finally, the variation is calculated every 0.2 seconds.

2.4. Low Short Time Energy Ratio (LSTER)

The features based on the energy of the input signal are very popular for speech/music classification. A reasonable generalization is that speech follows a pattern of high-energy periods for voiced sounds followed by low-energy segments for unvoiced sounds. On the other hand, the envelope of music is less likely to exhibit this behaviour. In this solution we use the Low Short Time Energy Ratio [5] that is defined as the ratio of the number of frames whose short time energy is less than 0.5 times the average short-time energy,

$$LSTER = \frac{1}{2N} \sum_{n=0}^{N-1} \left[sgn(0.5\overline{STE} - STE(n)) + 1 \right] \tag{4}$$

In the same way we did for the HZCRR, we calculate the energy in 20 ms windows with a frame shift of 10 ms. LSTER is calculated in 1 second windows. A median filter is applied to smooth the resulting values.

2.5. Mel-Frequency Cepstrum Coefficients Variation (Var.MFCC)

It is well known that the mel frequency cepstral coefficients are a compact and efficient representation of speech [6].

In our approach, MFCCs are extracted every 10 ms using 20 ms windows with a 40 channel filter-bank. Then, we used 13 coefficients that are summed and we calculate the variation every 0.1 seconds.

$$\sigma^2_{sumMFCC_i} = \frac{1}{N} \sum_{k=0}^{N-1} \left(sumMFCC_i(k) - \overline{sumMFCC_i} \right) \tag{5}$$

2.6. Minimum-Energy Tracking (MET)

This feature tries to describe the natural pauses that can be found in speech due to the nature of the human speech production mechanisms. These segments are characterized by the reduction of energy (represented by MFCC coefficient C_0) below a certain threshold. If C_0 is above the threshold for longer than 1.5 seconds, the frame will be classified as music, otherwise it will be classified as speech. This feature has certain limitations in sections of speech with background music and if the music level is high, the section can be classified as music. However, on those sections where only speech is present, this feature provides good results.

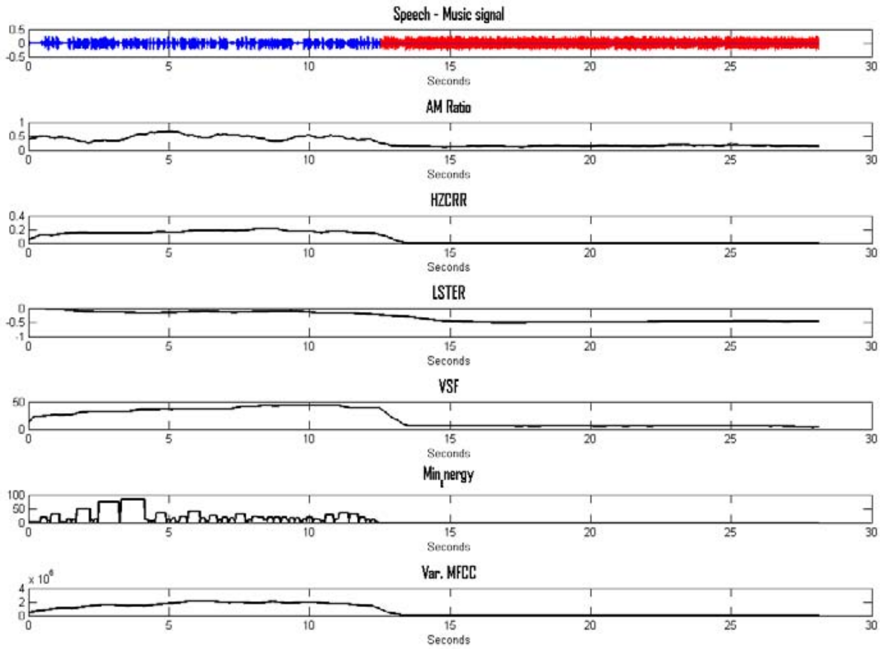

Figure 2: Behavior of the features with speech signal (blue) and music signal (red)

Finally, in Figure 2 the evolution of different features in a signal, where half of the samples are speech and the other half belong to music, is presented. Due to the asynchronous nature of some of the selected features like amplitude modulation ratio and minimum-energy tracking we have proceeded to the synchronization feature vector resampling all signals to 100 Hz.

3. C4.5 classification tree

When comparing different speech/music classification systems previously reported in the literature, the main difference that can be found among them is the classification algorithm they use. Artifitial Neural Networks (ANN), K Nearest-Neighbors (KNN) and Gaussian Mixture Models (GMM) are widely used [1].

In addition, decision tree methods have also been used for speech/music classification. In axis-parallel decision tree methods, a binary tree is constructed in which at each node a single parameter is compared to some constant. If the feature value is greater than the threshold, the right branch of the tree is taken; if the value is smaller, the left branch is followed. After a series of these tests, one reaches a leaf node of the tree where all the objects are labeled as belonging to a particular class. These are called axis-parallel trees because they correspond to partitioning the parameter space with a set of hyperplanes that are parallel to all of the feature axes except for the one being tested.

The C4.5 algorithm builds binary decision trees from a set of training data using the concept of *information entropy* that was developed by Quinlan in 1993 [7]. The training data is a set $S = s_1, s_2, ...$ of already classified samples. Each sample $s_i = x_1, x_2, ...$ is a vector where $x_1, x_2, ...$ represent our five features of the sample. The training data is augmented with a vector $C = c_1, c_2, ...$ where $c_1, c_2, ...$ represent the class to which each sample belongs. In our approach, the classes are *Speech*, *Music* and *Both*.

The algorithm considers all possible tests that can divide the set of data and select the test that contains greater information gain. For each continuous attribute binary testing is performed. At each node, the system must decide which test should be chosen to split the data. In our case, the features are continuous numerical values so the algorithm searches a threshold to get the best result as shown in Figure 3.

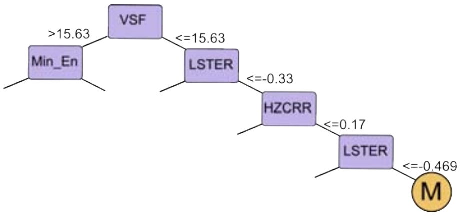

Figure 3: *C4.5 classification tree*

An additional constraint has been applied successfully: for any division, at least two of the subsets Ss_i must contain a reasonable number of cases. It uses a technique known as *Gain Ratio*. It consist on an information based measure that considers different numbers (and different probabilities) of the test results.

4. Evaluation and Results

The proposed approach has been evaluated by using an audio database supplied by Aragón Radio. From this database 20 tracks that alternate segments of music, speech or both, have been selected and manually annotated. Each segment has a duration between 10 and 30 seconds an the total amount of data is around one hour. This corpus contains clips with one voice, two voices, spots, speech-music mixed, instrumental music, songs, capella music and many other musical styles.

Feature	Accuracy
Var.MFCC	88.37%
HZCRR	74.34%
MET.	83.35%
LSTER	78.51%
VSF	79.11%
AMR	83.06%

Table 1: Features results

To check the reliability of the features in the classification of music and speech, we proceeded to the evaluation of those on the database described above. In this first experiment the overlapping of music and speech was not taken into account, the purpose was only to asses wether music or speech was present by making a simple threshold comparison for each one of the features. The results are presented in Table 1. As it can be seen, the variation of the MFCC coefficients is the feature that best segregate speech from music. It is also interesting to note that energy-based features and spectral features obtained around 80% of correct classifications.

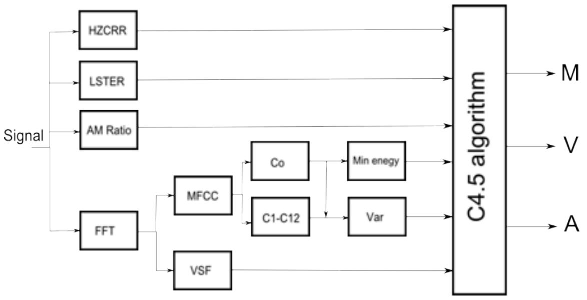

Figure 4: *Block diagram of feature extraction*

After evaluating the performance of each one of these features separately, it seems reasonable to implement a classification method that combines all of these features to improve the results and also to classify the overlapping between speech and music. The C4.5 algorithm consists of two phases: a training or definition of the tree and a test. This has been carried out with cross validation by dividing the database into four subsets randomly. The system used for this purpose can be seen in Figure 4. Thus, the results with the classification tree are presented in Tables 2 and 3.

Correctly Classified Frames	297113 → 99.83%
Incorrectly Classified Frames	487 → 0.16%
Total Number of Frames	297600

Table 2: Result of C4.5 tree

	Classification		
Original	Speech	Music	Both
Speech	99.77%	0.06%	0.17%
Music	0.03%	99.94%	0.02%
Both	0.35%	0.09%	99.56%

Table 3: Confusion Matrix

As can be seen in the results, the proposed classification method greatly improves the results of the individual features showing that the selected set of features complement each other. It is also interesting to note that the biggest classification error belongs to the class representing the segments in which music and speech overlap. This result could be expected beforehand since this class is less homogeneous than the other two and fewer samples are available for training.

5. Conclusion

In this work a set of acoustic features has been selected and evaluated for unsupervised speech/music classification on a broadcast radio database. Time-domain, frequency-domain and cepstral-domain features have been considered along with a decision tree based classification method. Experimental results show that the proposed approach can be successfully applied to the considered task. The use of the decision tree method outperforms the results obtained with each one of the features individually what highlights the complementarity among them. The proposed combination by using a decision tree classifier obtains an improvement of more than 10% over the most discriminative feature allowing also the introduction of a new class which is the overlapping of the two previously defined classes, music and speech, which occurs very often on broadcast radio.

6. References

[1] Lavner, Y and Ruinskiy, D, “A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation”, EURASIP Journal on Audio, Speech, and Music Processing, 2009.

[2] Saunders, J. , “Real-time discrimination of broadcast speech/music”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP, 1996.

[3] Lu ,L. and Zhang, H.J., “Content analysis for audio classification and segmentation”, IEEE Transactions on Audio, Speech, and Language Processing, 2002.

[4] Scheirer, E. and Slaney, M., “Construction and evaluation of a robust multifeature speech/music discriminator”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP, 1997.

[5] Alexandre, E. and Rosa, M., “Application of Fisher linear discriminant analysis to speech/music classification”, Proc. of the 120th Audio Engeneering Society Convention AES, 2006.

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

[7] Quinlan, R., “C4.5: Programs for Machine Learning”, in Morgan Kaufmann Publishers Springer Netherlands, 1993.