A computationally efficient speech/music discriminator for radio recordings

Aggelos Pikrakis, Theodoros Giannakopoulos and Sergios Theodoridis
University of Athens
Department of Informatics and Telecommunications
Panepistimioupolis, 15784, Athens, Greece
{pikrakis, tyiannak, stheodor}@di.uoa.gr

Abstract

This paper presents a speech/music discriminator for radio recordings, based on a new and computationally efficient region growing technique, that bears its origins in the field of image segmentation. The proposed scheme operates on a single feature, a variant of the spectral entropy, which is extracted from the audio recording by means of a short-term processing technique. The proposed method has been tested on recordings from radio stations broadcasting over the Internet and, despite its simplicity, has proved to yield performance results comparable to more sophisticated approaches.

Keywords: Speech/music discrimination, spectral-entropy, region growing techniques

1. Introduction

The problem of speech/music discrimination is important in a number of audio content characterization applications. Since the first attempts in the mid 90's, a number of algorithms have been implemented in various application fields. The majority of the proposed methods deal with the problem in two separate steps: firstly, the audio signal is split into segments by detecting abrupt changes in the signal statistics and at a second step the extracted segments are classified as speech or music by using standard classification schemes.

One of the first methods focused on the real-time, automatic monitoring of radio channels, using energy and zero-crossing rate (ZCR) as features [1]. In [2], thirteen audio features were used to train different types of multidimensional classifiers, including a Gaussian MAP estimator and a nearest neighbor classifier. In [3], energy, ZCR and fundamental frequency were used as features and segmentation/classification was achieved by means of a procedure based on heuristic rules. A similar approach was proposed in [4]. Frameworks based on combinations of standard Hidden Markov Models, Multilayer Perceptrons and Bayesian Networks were used in [5] and [6]. An Adaboost-based algorithm, applied on the spectrogram of the audio samples, was used in [7]. The authors in [8] have used Gaussian Mixture Modeling on a single feature, called Warped LPC-based spectral centroid, for the classification of pre-segmented audio data to speech and music.

In this paper, a different philosophy is adopted, that bears its origins in the field of image segmentation. The main idea is that, if speech/music discrimination is treated as a segmentation problem (where each segment is labeled as either speech or music), then each of the segments can be the result of a segment (region) growing technique, where one starts from small regions (segments) and keeps expanding them as long as certain criteria are fulfilled. This approach has been used in the past in the context of image segmentation, where a number of pixels are usually selected as candidates (seeds) for region growing. In image segmentation, regions grow by attaching neighboring pixels, provided that certain criteria are fulfilled. These criteria usually examine the relationship between statistics drawn from the region and the pixel values to be attached.

Following this philosophy, a feature sequence is first extracted from the audio recording, by means of a short-term processing technique. To this end, a variant of the spectral entropy is extracted per short-term frame. Once the feature sequence is generated, a number of frames are selected as candidates for region expansion. Starting from these seeds, segments grow and keep expanding as long as the standard deviation of the feature values in each region remains below a pre-defined threshold. In the end, adjacent segments are merged and short segments are eliminated. All segments that have survived are labeled as music, whereas the rest of the feature sequence is tagged as speech. The novelty of our approach lies in the fact that ideas from the field of image segmentation are applied in the context of off-line speech/music discrimination, yielding a computationally efficient algorithm that achieves high discrimination accuracy.

The paper is organized as follows: the next Section focuses on feature extraction, Section 3 presents the region growing technique, the performance of the proposed algorithm is discussed in Section 4 and finally conclusions are drawn in Section 5.

2. Feature extraction

At a first step, the audio recording is broken into a sequence of non-overlapping short-term frames (46.4ms long). From each frame, a variant of the spectral entropy [9] is extracted as follows, by taking into account the frequency range up to approximately 2KHz (by definition, entropy is a measure of the uncertainty or disorder in a given distribution [10]):

- All computations are carried on a mel-scale, i.e., the frequency axis is warped according to the equation

  $$f = 1127.01048 * \log(f_l/700 + 1)$$

  where f_l is the frequency value on a linear scale.

- The mel-scaled spectrum of the short-term frame is divided into L sub-bands (bins). The center frequencies of the sub-bands are chosen to coincide with the frequencies of semitones of the chromatic scale, i.e.,

  $$f_k = 1127.01048 * \log(\frac{f_0 * 2^{\frac{k}{12}}}{700} + 1), k = 0, \ldots, L-1$$

  where f_0 is the center frequency of the lowest sub-band of interest (on a linear scale).

- The energy X_i of the i-th sub-band, $i = 0, \ldots, L-1$, is then normalized by the total energy of all the sub-bands, yielding $n_i = \frac{X_i}{\sum_{i=0}^{L-1} X_i}$, $i = 0, \ldots, L-1$. The entropy of the normalized spectral energy is then computed by the equation:

  $$H = -\sum_{i=0}^{L-1} n_i \cdot log_2(n_i) \qquad (1)$$

In the sequel we will also refer to this feature by the term “chromatic entropy”. At the end of the feature extraction stage, the audio recording is thus represented by the feature sequence $\mathbf{F}$, i.e., $\mathbf{F} = \{\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_T\}$, where T is the number of short-term frames. Figure 1 presents the feature sequence that has been extracted from a BBC radio recording, the first half of which corresponds to speech and the second half corresponds to music. It can be observed that the standard deviation of chromatic entropy is significantly lower for the case of music.

3. Segmentation Algorithm

Once the feature sequence has been extracted, speech/music discrimination is achieved by means of a region growing technique. The main idea behind this approach is that, at an initialization stage, a number of frames are selected as “seeds”, i.e., as candidates, that will serve as the basis to form regions (segments). Subsequently, by means of an iterative procedure, these regions will grow (expand) while a criterion related to the standard deviation of the chromatic entropy is fulfilled. The procedure is repeated until no more region growing takes place. At a final step, neighboring regions are merged and after merging, regions that do not exceed a pre-specified length are eliminated. At the end of this procedure, all segments that have survived correspond to music, whereas all frames that do not belong to such segments are considered to be speech segments. The procedure is described in detail as follows:

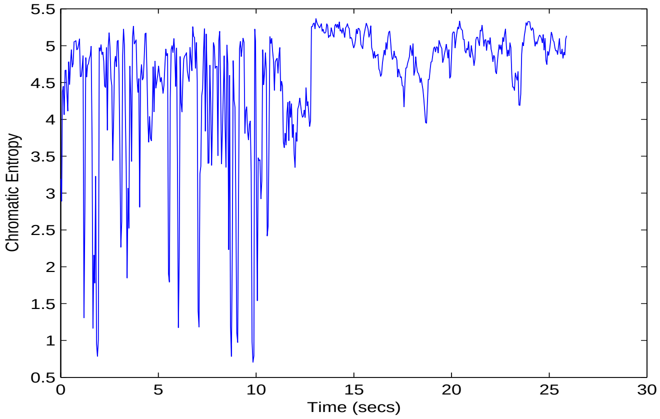

Figure 1. Chromatic entropy for 26 seconds of a BBC radio recording.

Initialization step - Seed generation: If T is the length of the feature sequence, a “seed” is chosen every M frames, M being a pre-defined constant. If K is the total number of seeds and i_k is the frame index of the k-th seed, then the frame indexes of the seeds form the set

$$\{i_1, i_2, \ldots, i_K\}$$

The k-th seed is considered to form a region, R_k consisting of a single frame, i.e., $R_k = \{O_{i_k}\}$ where O_{i_k} is the feature value of the respective frame.

Iteration: In this step, every region, R_k, is expanded by examining the feature values of the two frames that are adjacent to the boundaries of R_k. To this end, let l_k and r_k be the indexes that correspond to the leftmost and rightmost frames of R_k respectively. Clearly, if R_k consists of a single frame, then $l_k = r_k = i_k$. Following this notation, $l_k - 1$ and $r_k + 1$ are the indexes of the two frames which are adjacent to the left and right boundary of R_k, respectively. Our algorithm decides to expand R_k to include O_{l_k-1}, if O_{l_k-1} is not already part of any other region and if the standard deviation of the feature values of this expanded region is below a pre-defined threshold T_h, common to all regions. In other words, if the standard deviation of feature values for $O_{l_k-1} \cup R_k$ is less than T_h, then, at the end of this step R_k will have grown one frame to left. Similarly, if O_{r_k+1} is not already part of any other region and if the standard deviation of the feature values in $R_k \cup O_{r_k+1}$ is less than T_h, then R_k will also grow by one frame to its right. At the end of this step, each R_k will have grown by at most two frames. It has to be noted that, certain regions may not grow at all, because both frames that are adjacent to their boundaries, already belong to other regions. At the end of the step, it is examined whether at least one region has grown by at least one frame. If this is the case, this step is repeated until no more region growing takes place.

Termination: After region growing has been completed, adjacent regions (if any) are merged to form larger segments. Finally, after merging is complete, short regions are eliminated by comparing their length with a pre-defined threshold, say T_{min}.

Ideally, all segments (regions) that survive at the end of the algorithm should correspond to music and any frame outside these regions should correspond to speech. This is because the proposed scheme relies on the assumption that music segments exhibit low standard deviation in terms of the adopted feature (see Figure 1). As will be explained in the next section, our approach, despite its simplicity, exhibits a high discrimination accuracy.

Finally it has to be noted that the proposed algorithm is dependent on three parameters, namely M, the distance (measured in frames) between successive seeds, T_h, the threshold for standard deviation (based on which region growing takes place) and T_{min}, the minimum segment length (used in the final stage of the algorithm). The choice of values for these parameters is explained in the next section.

4. Experiments

We carried out two sets of experiments, each of which on a separate dataset, in order to:
a) Determine the values of the parameters of the method, subject to two different maximization criteria, namely overall discrimination accuracy and music precision. The latter refers to the proportion of audio data in the recording that corresponds to music and was also classified as music (see also subsections 4.2 and 4.3). It has to be noted that by maximizing music precision, overall discrimination accuracy is likely to decrease. Although this is undesirable if the proposed method is used as a standalone discriminator, it may not be a restriction if it is used as a low-complexity preprocessing step for music detection. In this case, high music precision (close to 100%) ensures that all detected segments are correctly classified as music, and all remaining parts of the audio recording can be subsequently fed to other, more sophisticated discrimination schemes, for further processing.
b) Assess the algorithm's performance, using the parameters extracted from maximizing the desired criteria.

4.1. Datasets

The audio data used for the above purposes was collected from seven different BBC Internet radio stations, covering a wide range of music genres and speakers. Obviously, the dataset used for testing system performance was different from the dataset used for parameter tuning. More specifically, 30 minutes of audio recordings (dataset D_1), were used for estimating parameter values. To this end, an exhaustive approach was adopted, i.e., each parameter was allowed to vary within a predefined range of values. For system testing, a different dataset, D_2, was created, consisting of audio recordings of a total duration of 160 minutes. A $16KHz$ sampling rate was used in all cases. All recordings (of both datasets) were manually segmented and labeled as speech or music. It has to be noted that "silent" segments, i.e. segments with almost zero energy) were treated as speech, under the observation that such segments usually occur between speech segments. This manual segmentation procedure revealed that 69.77% of the data was music and 30.23% was speech.

4.2. Segmentation results for maximizing the overall accuracy

The parameter estimation process, subject to maximizing discrimination accuracy for dataset D_1, led to the values of Table 1.

Threshold	Min. Duration	Seed Dist
0.50	3.0 sec	2.0 sec

Table 1. Parameter values subject to maximizing discrimination accuracy over D_1

Using the above parameter values, our method was then tested on D_2. The proposed scheme classified 75.10% of the data in D_2 as music and 24.90% as speech. Table 2 presents the average confusion matrix C, of the discrimination results. Each element $C_{i,j}$ of the matrix corresponds to the percentage of data whose true class label was i and was classified to class j. From C one can directly extract the following measures for each class:

	Music	Speech
Music	69.13%	0.65%
Speech	5.97%	24.25%

Table 2. Average confusion matrix for D_2 using the parameter values in Table 1

1. **Recall** (R_i). R_i is the proportion of data with true class label i, that were correctly classified in that class. For example, the recall of music is calculated as $R_1 = \frac{C_{1,1}}{C_{1,1}+C_{1,2}}$.
2. **Precision** (P_i). P_i is the proportion of data classified as class i, whose true class label is indeed i. Therefore, music precision is $P_1 = \frac{C_{1,1}}{C_{1,1}+C_{2,1}}$.

According to the confusion matrix, the overall discrimination accuracy of our system is equal to 93.38% ($C_{1,1} + C_{2,2}$). Table 3 presents recall and precision values for both speech and music. From this table, it can be seen that, more than 99% of the "true" music data was detected, while the "false alarm" for the music class was below 8%.

Music Recall	99.07%
Speech Recall	80.25%
Music Precision	92.05%
Speech Precision	97.39%

Table 3. Recall and Precision of both classes for the parameter values in Table 1

4.3. Segmentation results for maximizing music precision

As explained above, we have also estimated the parameter values subject to maximizing music precision. The resulting values are presented in Table 4.

Threshold	Min. Duration	Seed Dist
0.30	5.0 sec	2.0 sec

Table 4. Parameter values subject to maximizing music precision

The average confusion matrix in this case is presented in Table 5. It can be seen that 58.09% of the data was classified as music and 41.91% as speech and the overall accuracy of the system was 87.58%, almost 6% lower than the accuracy presented in section 4.2. However, as can be seen in Table 6, *music precision* is now equal to 99.36%. This leads us to the conclusion, that, when the proposed algorithm is fed with this second set of parameter values, it can be used as a low-complexity preprocessing step in a more sophisticated audio characterization system (e.g. [6]), for the initial detection of a smaller proportion of the music segments (smaller music recall), but with an almost zero “false alarm” (0.4%).

	Music	**Speech**
Music	57.72%	12.05%
Speech	0.37%	29.86%

Table 5. Average confusion matrix for the results obtained with the parameter values in Table 4

Music Recall	82.73%
Speech Recall	98.78%
Music Precision	99.36%
Speech Precision	71.25%

Table 6. Recall and Precision of both classes for the parameter values in Table 4

5. Conclusions

This paper presented a computationally efficient, off-line speech/music discriminator, based on a region growing technique operating on a single feature that we call chromatic entropy. The system was tested on recorded Internet radio broadcasts (of almost 3 hours duration) and achieved an average discrimination accuracy of 93.38%. This is comparable to the performance obtained with other computationally more complex methods. It is worth noticing that, if the method’s parameters are tuned to maximize music precision, although the system’s overall accuracy drops to 87.58%, music precision almost reaches 100% (99.3%), i.e., all surviving segments correspond to music. Taking into account these results, it can be concluded that the proposed algorithm is capable of working both:

1. As a standalone speech/music discriminator of high performance.

2. As a computationally efficient preprocessing stage for music detection in audio streams (when the parameters are tuned to maximize music precision). In this latter case, non-music segments can be further processed by more complex discrimination schemes.

References

[1] J. Saunders, “Real-time discrimination of broadcast speech/music”, in *Proc. of ICASSP 1996*, Vol. 2, pp. 993-996, Atlanta, USA, May 1996.

[2] E. Scheirer and M. Slaney, “Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator”, in *Proc. ICASSP 1997*, pp. 1331-1334, Munich, Germany.

[3] Tong Zhang and C.-C. Jay Kuo, “Audio Content Analysis for Online Audiovisual Data Segmentation and Classification”, in *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 4, pp. 441-457, May 2001.

[4] C. Panagiotakis and G. Tziritas, “A Speech/Music Discriminator Based on RMS and Zero-Crossings”, *IEEE Transactions on Multimedia*, vol. 7(1), pp. 155-166, Feb. 2005.

[5] Jitendra Ajmera, Iain McCowan and Herve Bourlard, “Speech/music segmentation using entropy and dynamism features in a HMM classification framework”, *Speech Communication*, vol. 40, pp. 351-363, 2003.

[6] Aggelos Pikrakis, Theodoros Giannakopoulos and Sergios Theodoridis, “Speech/Music Discrimination for radio broadcasts using a hybrid HMM-Bayesian Network architecture”, in *Proc. of the 14th European Signal Processing Conference (EUSIPCO-06)*, September 4-8, 2006, Florence, Italy.

[7] N. Casagrande, D. Eck, and B. Kigl., “Frame-level audio feature extraction using AdaBoost”, in *Proc. of ISMIR 2005*, pp. 345-350, London, UK, 2005.

[8] J.E. Munoz-Exposito et al, “Speech/Music discrimination using a single Warped LPC-based feature”, *Proc. of ISMIR 2005*, pp. 614-617, London, UK, 2005.

[9] Hemant Misra, Shajith Ikbal, Herve Bourlard, and Hynek Hermansky, “Spectral entropy based feature for robust ASR”, in *Proc. of ICASSP 2004*, Vol. 1, pp. 193-196, Montreal, Canada, 2004.

[10] A. Papoulis and S. Unnikrishna Pillai, “Probability, Random Variables and Stohastic Processes, 4th edition”, McGraw-Hill, NY, 2001.