Comparison of four classifiers for speech-music discrimination: a first case study for costa rican radio broadcasting

Comparación de cuatro clasificadores para la discriminación de voz y música: un primer estudio de caso para la radiodifusión costarricense

Joseline Sánchez-Solís[1], Marvin Coto-Jiménez[2]

Sánchez-Solís, J.; Coto-Jiménez, M. Comparison of four classifiers for speech-music discrimination: a first case study for costa rican radio broadcasting. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 119-127.

https://doi.org/10.18845/tm.v35i8.6463

[1] University of Costa Rica. Costa Rica. Costa Rica. E-mail: joseline.sanchez@ucr.ac.cr

[2] University of Costa Rica. Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr
https://orcid.org/0000-0002-6833-9938

Keywords

Classification; music; radio broadcasting; speech.

Abstract

During the past decades, a vast amount of audio data has be- come available in most languages and regions of the world. The efficient organization and manipulation of this data are important for tasks such as data classification, searching for information, diarization among many others, but also can be relevant for building corpora for training models for automatic speech recognition or building speech synthesis systems. Several of those tasks require extensive testing and data for specific languages and accents, especially when the development of communication systems with machines is a goal. In this work, we explore the application of several classifiers for the task of discriminating speech and music in Costa Rican radio broadcast. This discrimination is a first task in the exploration of a large corpus, to determine whether or not the available information is useful for particular research areas. The main contribution of this exploratory work is the general procedure and selection of algorithms for the Costa Rican radio corpus, which can lead to the extensive use of this source of data in many own applications and systems.

Palabras clave

Clasificación; música; radiodifusión; habla.

Resumen

Durante las últimas décadas, una gran cantidad de datos de audio ha estado disponible en la mayoría de los idiomas y regiones del mundo. La organización y manipulación eficiente de estos datos son importantes para tareas como clasificación de datos, búsqueda de información, diarización entre muchas otras, pero también pueden ser relevantes para construir corpus para modelos de entrenamiento para reconocimiento automático de voz o construir sistemas de síntesis de voz. Varias de esas tareas requieren pruebas y datos exhaustivos para idiomas y acentos específicos, especialmente cuando el objetivo es el desarrollo de sistemas de comunicación con máquinas. En este trabajo, exploramos la aplicación de varios clasificadores para la tarea de discriminar el habla y la música en la radiodifusión costarricense. Esta discriminación es una primera tarea en la exploración de un gran corpus, para determinar si la información disponible es útil o no para áreas de investigación particulares. El principal aporte de este trabajo exploratorio es el procedimiento general y la selección de algoritmos para el corpus de radio costarricense, lo que puede llevar al uso extensivo de esta fuente de datos en muchas aplicaciones y sistemas propios.

Introduction

In our days, there is a vast amount of multimedia data, such as images, audio, and video, which are available on the Internet, radio and television broadcasts. The amount of information of this kind during the last years has seen an exponential growth [1]. The manipulation and organization of this data are required in many tasks, for example, in classification for storage, summarizing and describing the content. A large portion of the data is audio, from resources such as broadcasting radio, audio-books, internet streams, and commercial music recordings.

Due to massive amounts of this data it is impossible to generate classes, la- bels, descriptions, transcriptions manually, or many of the main tasks that are required to take advantage of the information [2]. To answer the demands for handling the data, a field of research, known as audio content analysis (ACA), or machine listening, has recently emerged [1]. The purpose of ACA can be established as extracting information directly from the acoustic signal and automatically create descriptions, detect types of content, semantic annotation, speaker diarization or other tasks according to specific requirements.

One of the first tasks that need to be solved in ACA applications is the automatic distinction between music and speech. This is, due to the very different content and applications of both classes, such as genre classification (in music) or automatic speech recognition. A broadcast audio content can be annotated as music or speech once input audio is classified as speech/ music segment [3]. This problem, with only two classes has been of interest in academia, and also has industrial applications, for example general purpose audio codecs [4].

Among the main challenges in performing speech/music classification is to obtain high accuracy with characteristics of short-time delay and low complexity [5], due to the vast amount of information that need to be processed and its efficient utilization in the applications. For discriminating the audio content in categories such as music and speech, data two successive stages have to be performed: i) the extraction of features from the input audio data, and ii) the classification of the data into established categories [2].

And beyond classification and semantic annotation of information, speech/music discrimination is also an essential part of speech coding, to efficiently utilize the bandwidth resources. For example, different bit rate allocations for different in- put formats can be applied according to the content [6]. If an automatic speech recognition system is applied to broadcast news, it is also desirable to consider disabling the input to the speech recognizer during the non-speech portion of the audio stream [7].

In this paper, we compare the performance of several algorithms for speech/music classification for a Costa Rican radio broadcasting, using feature vectors ex- tracted from the audio [8]. To our knowledge, this is the first report for this task with Costa Rican data, which is relevant due to the dependency on many kinds

of music and speech-related tasks in the language and particular accents.

The rest of this paper is organized as follows: Section 2 presents a review of the literature related to music and speech classification. Section 3 presents the Experimental Setup used in the research. The results and discussion are presented in Section 4. Finally, the conclusions are found in Section 5.

Related Work

For discrimination of speech and music in audio signals, numerous techniques exploring several features and classifiers have been proposed. It is also important to consider that there could be a dependency on the type of recordings (radio or television broadcasts, audio quality, audio books, among others), as well as the type of music and language.

Some of the most common features include speech-specific parameters, such as zero-crossing rate (ZCR), spectral centroid, spectral roll-off, and Mel fre- quency cepstral coefficients (MFCC) [3]. One of the first works, presented in [9] applied a calculation of ZCR and its statistics in short segments of FM audio broadcasting, obtaining error rates below 10%.

The fixed-size set of features for audio segments is also known as i-vector, as reported in [10]. There, an analysis of algorithms such as cosine distance score (CDS), support vector machine (SVM), and linear discriminant analysis (LDA) is performed for speech and music classification in the English language, with a database derived from TIMIT.

Among the most recent features applied to this tasks, [11] includes some more complex features derived from speech analysis, such as the normalized auto- correlation peak strength (NAPS), the zero frequency filtered signal (ZFFS), the peak-to-sidelobe ratio (PSR), and Hilbert envelope (HE) of linear prediction(LP) coefficients. It is important to remark that these more complex features are applied for languages and databases where previous experimentation has been done.

A measure based on energy, called Minimum Energy Density (MED), was applied to the binary speech/music classification in the radio broadcasts in [12], for the Polish language. The primary motivation of the study was to provide information about the contents of the radio stations.

The possibility of discriminate segments of speech and music can also be approached from short segments. For example, in [4], segments in the order of tens of milliseconds are explored. The authors adopted a statistical approach, and the features were found using an unsupervised procedure. Also, in [13], a two- step segmentation approach is employed to identify transition points between homogeneous regions, and then successfully apply the algorithms for classifying the segments.

Some of the best results in the task of binary speech/music classification have been obtained using machine learning techniques, such as the support vector machine (SVM), Gaussian mixture model (GMM), and deep belief networks (DBN) [6]. In most of the references, a comparative study is performed using several features or algorithms for classification, in particular when a new set of features is proposed. Also, it is likely to experiment on databases made from clean speech and pure music, leaving aside some interesting cases, such as music plus background music.

More recently [3], experimented with chromagram-based music-specific features, claiming that such features outperform other existing approaches in terms of classification accuracy. In [14] some new algorithms, based on deep learning, also increase the accuracy of the classification.

Experimental Setup

Database used

The tests were made using a music/speech database created for this work, using audio recordings from ”Comunidad 870” podcasts, a program that belongs to ”Radio 870” of the University of Costa Rica. The podcasts are publicly available at https://radios.ucr.ac.cr/.

From 10 hours of audio recordings, an automatic segmentation based on detection of silence, with the SoX program was performed to produce 178 music files and 773 clean speech files. The resulting files do not have a standard du- ration. All files have different duration, with some lasting several minutes while others only seconds. 10% of the total music or speech files were used for the test files. All the files were manually classified as clean speech or music to establish the ground-truth for the classifiers.

Due that the main purpose of this exploration is the assessment of the audio resources of this podcast for further research in speech technologies, the speech files with background music were considered as music segments.

Feature extraction

The sound features used to train the classifiers were selected upon bibliographic criteria. Each set of features was tested alone and in combination with the others in every classifier, with the aim of finding the best set of features and in case of similitude, those with the simplest or least amount of features. The features were extracted using pyAudioAnalysis [15], and are described within each of the following groups:

1. ***Spectral and Energy Features.*** The mean value of the following features was obtained for each audio file containing music of speech:
    - Zero Crossing Rate: The rate of sign-changes of the signal during the dura- tion of a particular frame.
    - Energy: The sum of squares of the signal values.
    - Entropy of Energy: The entropy of the sub-frames, which can be interpreted as a measure of abrupt changes.
    - Spectral Centroid: The center of gravity of the spectrum.
    - Spectral Spread: The second central moment of the spectrum.
    - Spectral Entropy: Entropy of the normalized spectral energies.
    - Spectral Flux: The squared difference between the normalized magnitudes of the spectra of the two successive frames.
    - Spectral Rolloff: The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
2. ***MFCC.*** The Mel Frequency Cepstrum Coefficients (MFCC) are a set of the Discrete Cosine Transform parameters, computed using a perceptually spaced triangular filter bank that processes the Discrete Fourier Transform of the speech signal [16]. Every frame of speech of music signal can be expressed as a set of MFCC coefficients. In our work, we used the mean value of each of 13 coefficients in each audio file.
3. ***Chroma Vectors.*** Chroma features are twelve-element vectors, where each di- mension represents the intensity associated with a particular frequency of tem- pered musical scale; thus, each element can be associated with a semitone of the musical scale, regardless of octave [17].

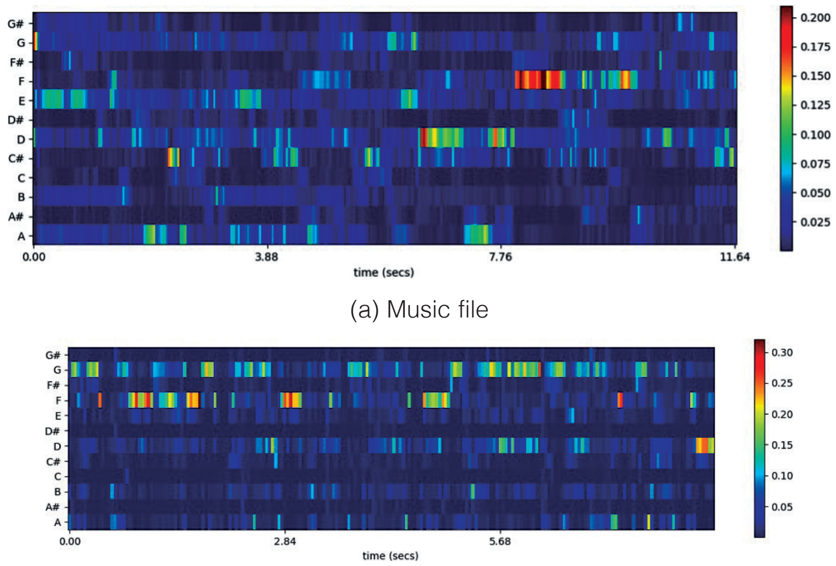
    Chroma vectors have been applied successfully in music gender classification and several tasks related to music identification. In our work, we used the mean values of the twelve chroma dimensions for each audio file.

Classifiers

For this case study, we chose four classifiers commonly applied in exploratory studies: 1) Support Vector Machines (SVM) with a polikernel implementation. 2) K-nearest neighbors (KNN). 3) Random Forest with a number of trees of 100 and seed = 1. 4) Naive Bayes.

Results and Discussion

In this section, we present the results of the four classifiers and the combinations of features. Table 1 shows the results of the test music/speech file classification, for each of the four algorithms applied and all the subsets of features available.



(a) Music file

(b) Speech file

Figure 1. Chroma vector for two files of the database.

The purpose of this comparison is to find the best algorithm and ideally the simpler set of features that perform properly for this problem.

It can be seen that the best results are obtained with KNN, since it presents error rates as low as 1.9%, with the combinations of features Energy + MFCC, MFCC + Chroma and Energy + MFCC + Chroma. The Random Forest classifier was the classifier with the second-best results, with the Energy + MFCC + Chroma features.

Table 1. Evaluation of results for each classification algorithm and set of features.* Is the best result.

KNN				
Features	Error rate	Precision	Recall	F1 score
Energy	11.8812	0.876	0.881	0.874
MFCC	2.9701	0.927	0.981*	0.969
Chroma	8.9109	0.909	0.911	0.907
Energy+MFCC	1.9802*	0.981*	0.980	0.980*
Energy+Chroma	4.9505	0.950	0.950	0.950
MFCC+Chroma	1.9802*	0.981*	0.980	0.980*
Energy+MFCC+Chroma	1.9802*	0.981*	0.980	0.980*
SVM				
Features	Error rate	Precision	Recall	F1 score
Energy	11.8812	0.897	0.881	0.863
MFCC	14.8515	0.857	0.851	0.826
Chroma	21.7822	0.612	0.782	0.687
Energy+MFCC	10.8911*	0.904*	0.891*	0.876*
Energy+Chroma	10.8911*	0.904*	0.891*	0.876*
MFCC+Chroma	13.8614	0.882	0.861	0.831
Energy+MFCC+Chroma	10.8911*	0.904*	0.891*	0.876*
Random Forest				
Features	Error rate	Precision	Recall	F1 score
Energy	9.901	0.899	0.901	0.895
MFCC	4.9505	0.950	0.950	0.949
Chroma	8.9109	0.920	0.911	0.902
Energy+MFCC	3.9604	0.962	0.960	0.956
Energy+Chroma	4.9505	0.953	0.950	0.948
MFCC+Chroma	3.9604	0.962	0.960	0.959
Energy+MFCC+Chroma	2.9703*	0.971*	0.970*	0.969*
Naive Bayes				
Features	Error rate	Preision	Recall	F1 score
Energy	18.8119	0.792	0.812	0.793
MFCC	20.7921	0.803	0.792	0.797
Chroma	22.7723	0.726	0.772	0.733
Energy+MFCC	16.8317*	0.829*	0.832*	0.830*
Energy+Chroma	21.7822	0.769	0.782	0.774

In Figure 1, the speech chromogram shows a region where the highest in- tensity frequencies are more concentrated, unlike music, where a wider diversity of frequencies are highlighted. Even though this visually evident differentiation, the Chroma-vector isolated parameters are not performing properly, compared to the combination of features.

According to the results, the best combination of features for this problem in the dataset developed is Energy + MFCC + Chroma, which presents a lower error rate in three of the four classifiers used. No specific combination of features presents the lower error rates for all classifiers. But on the other hand, it is clear from the results that the Energy and Chroma features are not efficient for this problem when used individually.

Other features such as individual MFCC presents error percentages of less than 10% for the KNN classifier, but for the rest classifiers, the error rates are greater than 10%.

In the case of the MFCC + Chroma features, the results are the best for the KNN classifier. However, for the Naive Bayes classifier, the percentage of error and other indexes are the most inefficient. SVM present error greater than 10%, but this is not the worst result for this classifier. For Random Forest the error percentage is 3.9604%, but it is not the most functional feature for this classifier.

Conclusions

In this work, we developed a dataset and performed the first case study of Speech/Music classification in a Costa Rican radio broadcasting. Our purpose is to evaluate the applicability of such source of audio information for research in speech technologies development, where a source of clean speech in particular languages and accent is mandatory.

From the results of four classifiers, the best results were obtained with the simpler one: KNN. Additionally, in regards to the comparison of classifiers, we performed a comparison of features, and the simplest combination with the best results corresponded to Energy+MFCC and MFCC+Chroma vector features.

When implementing an individual set of features (such as only MFCC or only Chroma-vector features), the classifiers dropped its capacity to discriminate clean speech from music or speech with background music.

For future work, we intend to implement a live broadcasting discriminator for clean speech, and build databases of Costa Rican Spanish that help the development of speech technologies for this particular Spanish accent.

References

[1] Lavner, Yizhar, and Dima Ruinskiy. ”A decision-tree-based algorithm for speech/music classification and segmentation.” EURASIP Journal on Audio, Speech, and Music Processing 2009 (2009): 1-14.

[2] Ghosal, Arijit, and Suchibrota Dutta. ”Speech/music discrimination using per- ceptual feature.” Computational Science and Engineering: Proceedings of the In- ternational Conference on Computational Science and Engineering (Beliaghata, Kolkata, India, 4-6 October 2016). CRC Press, 2016.

[3] Birajdar, Gajanan K., and Mukesh D. Patil. ”Speech/music classification using visual and spectral chromagram features.” Journal of Ambient Intelligence and Humanized Computing 11.1 (2020): 329-347.

[4] Hirvonen, Toni. ”Speech/music classification of short audio segments.” 2014 IEEE International Symposium on Multimedia. IEEE, 2014.

[5] Wu, Qiong, et al. ”A combination of data mining method with decision trees build- ing for Speech/Music discrimination.” Computer Speech & Language 24.2 (2010): 257-272.

[6] Kang, Sang-Ick, and Sangmin Lee. ”Improvement of Speech/Music Classification for 3GPP EVS Based on LSTM.” Symmetry 10.11 (2018): 605.

[7] Ruiz-Reyes, Nicolas, et al. ”New speech/music discrimination approach based on fundamental frequency estimation.” Multimedia Tools and Applications 41.2 (2009): 253-286.

[8] Kim, S. B., and S. M. Lee. ”A Comparative Evaluation of Speech-Music Classi- fication Algorithms in the Noise Environment.” International Journal of Design, Analysis and Tools for Integrated Circuits and Systems 8.1 (2019): 36-37.

[9] Saunders, John. ”Real-time discrimination of broadcast speech/music.” 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Vol. 2. IEEE, 1996.

[10] Zhang, Hao, et al. ”Application of i-vector in speech and music classification.” 2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, 2016.

[11] Khonglah, Banriskhem K., and SR Mahadeva Prasanna. ”Speech/music classifica- tion using speech-specific features.” Digital Signal Processing 48 (2016): 71-83.

[12] Kacprzak, Stanis-law, B-laz˙ej Chwie´cko, and Bartosz Zi´o-lko. ”Speech/music discrim- ination for analysis of radio stations.” 2017 International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, 2017.

[13] Tsipas, Nikolaos, et al. ”Efficient audio-driven multimedia indexing through similarity-based speech/music discrimination.” Multimedia Tools and Applications 76.24 (2017): 25603-25621.

[14] Li, Zhitong, et al. ”Optimization of EVS speech/music classifier based on deep learning.” 2018 14th IEEE International Conference on Signal Processing (ICSP). IEEE, 2018.

[15] Giannakopoulos, Theodoros. ”pyaudioanalysis: An open-source python library for audio signal analysis.” PloS one 10.12 (2015).

[16] Hossan, Md Afzal, Sheeraz Memon, and Mark A. Gregory. ”A novel approach for MFCC feature extraction.” 2010 4th International Conference on Signal Processing and Communication Systems. IEEE, 2010.

[17] Ellis, Daniel PW. ”Classifying music audio with timbral and chroma features.” (2007): 339-340.