SPEECH/MUSIC DISCRIMINATION IN A LARGE DATABASE OF RADIO BROADCASTS FROM THE WILD

Ewald Wieser, Matthias Husinsky, Markus Seidl

University of Applied Sciences St. Pölten
Institute for Creative Media Technologies
3100 St. Pölten, Austria
{ewald.wieser, matthias.husinsky, markus.seidl}@fhstp.ac.at

ABSTRACT

This paper describes the development, implementation and evaluation of a speech/music detector. We aim at audio from different sources with different qualities - i.e. audio from "the wild". We examine existing approaches for audio classification and select a recent feature. We modify the feature and evaluate the classification accuracy on a random test set of more than 60 hours of audio material against a standard speech/music detection approach. With our approach, we reach a classification accuracy of 96,6%. We provide a performant open source implementation of our detector.

Index Terms— Audio classification, speech/music discrimination, radio broadcast

1. INTRODUCTION

The Cultural Broadcasting Archive[1] is a collection of radio broadcasts of over 23 Austrian and international community radio stations with more than 30,000 hours of audio data that is open for public access on the Internet. Typical radio programs contain not only speech, but also music. Due to legal reasons, only the spoken parts can be offered online for on-demand usage permanently. The archive is already quite large and is growing daily. Hence, we aim at an automatic music detection algorithm in order to delete the music parts from stored broadcasts. The algorithm has to be computationally efficient and easily implementable into the existing infrastructure of the archive. The main challenge is a large quality variation in the broadcasts as the programs in the archive come from many different sources. Many of them are user generated content which are produced by untrained community members using their own equipment. Not being audio technicians, the quality of the broadcasts has a random element, depending on recording situation and user settings. Besides low audio quality due to standard mistakes, like the usage of wrong gain levels or bad placing of microphones, we find many other errors in the material, e.g. hums, DC offsets, etc.

Several approaches for speech/music detection exist. These usually focus on high quality material from commercial or public broadcasters. None of the existing approaches focuses on audio from the wild. With the growth of video material from the wild on platforms like Youtube, we assume that our work is of general and growing interest. Our main contribution is the adaption of an existing speech/music classifier and the embedding of the modified classifier in a processing chain to work robustly on material from the wild. We apply several pre- and post-processing steps and evaluate the performance of the classifier as well as the pre- and post-processing against a standard approach. Furthermore, we provide an open source implementation of our detector[2]. The paper is organized as follows: In Section 2 we survey the existing state of the art of speech/music detection. In Section 3, we describe our approach. In sections 4 and 5, we present our experiments and the results. Section 6 contains conclusions.

2. STATE OF THE ART

The topic of speech/music detection has found many applications since first attempts in the mid 90's. Saunders [1] was one of the first to implement a real-time speech-music discriminator for FM-radio broadcasts. He uses the zero crossing rate (ZCR) and some derived statistical features, combined with the energy contour of the waveform. He reports a classification accuracy of over 98% for the training data and 95-96% for a real-time test run but never mentions the length and quality of his training and test set. Scheirer and Slaney [2] evaluate thirteen audio features for their capability of creating a real-time audio classifier for radio broadcasts. They use single features and multivariate combinations of them with a Gaussian MAP estimator, a Gaussian mixture model (GMM), a k-nearest neighbor (kNN) estimator and a k-dimensional tree. With a data set of 40 minutes of audio, they report a classifica-

This work was supported by the Austrian Research Promotion Agency (FFG) under the program "Innovationsscheck Plus", number 835537.

[1] http://cba.fro.at/

[2] http://mcrg-fhstp.github.io/cba-yaafe-extension/

tion accuracy between 94.2% and 98.6%. Williams and Ellis [3] use the same test data set and reach about the same classification results with features derived from phone-posterior estimation but only for clean speech segments. El-Maleh et al. [4] develop a real-time classifier based on line spectral frequencies (LSF) and zero-crossing based features and obtain a classification accuracy of 95.9% with a test data set of about 20 minutes at a frame delay of only 20ms. Another real-time classifier is proposed by Panagiotakis and Tziritas [5], who report a classification performance of 97% using an algorithm that applies root mean square (RMS) and ZCR features with simple threshold classification. More recent work by Ravindran et al. [6] and Khan et al. [7] states that the use of the Mel frequency cepstrum coefficients (MFCCs) for speech/music discrimination shows good classification results applying different classifiers (hidden Markov models (HMM), GMMs and neuronal networks). Pikrakis et al. [8] use energy, ZCR, spectral entropy and MFCCs and a HMM combined with a Bayesian network to classify 340 minutes of radio recordings with an accuracy of 94.95%. Izumitani at al. [9] concentrate on detecting music in the background of speech and test various standard audio features against MFCCs, the spectral powers of linear-scaled frequencies and the spectral powers of Mel-frequencies with a GMM and a k-NN discriminator. Thus they reach a classification accuracy of 92% on an artificially generated ground truth of 900 seconds of music. Based on the work of Keum et al. [10], who use a spectral peak feature together with a multilayer perceptron, Seyerlehner et al. [11] develop a new feature for music detection in television productions called continuous frequency activation (CFA). The result is a single numeric value for each computed time frame, which quantifies the presence of steady frequency components. The value is higher for blocks with music that typically contains steady frequencies and lower for blocks without music. They tested their feature with quite a large ground truth of 545 minutes of audio data and achieved a classification accuracy of 89.93%. Work conducted after the end of our project include Schlüter and Sonnleitner [12], who use a mean-covariance restricted Boltzmann machine to evaluate the use of unsupervised feature learning methods from computer vision against engineered features such as Seyerlehner's [11] and reach a classification accuracy of up to 98.4% on a dataset of 42 hours of radio broadcasts. Han and Coover [13] propose a new feature called Transient Activation (TAC) to use along with the CFA and other standard audio features. Thus they reach a classification result of 84.01% on a ground truth of 65 minutes of recordings of radio broadcasts.

While former mentioned work concentrates on professionally produced audio material, consumer generated audio of low quality is seldomly taken into account. Related is work of Lee and Ellis [14], who classify consumer videos from Youtube by analyzing their audio data only. Classification is done into 25 semantic categories by creating clip level features of summarized MFCCs using single Gaussian modeling, GMM and probabilistic latent semantic analysis of a Gaussian component histogram. With support vector machine (SVM) classifiers on a ground truth of 1873 random video clips having an average duration of 145 seconds, adding up to 75 hours of material, they achieve average precisions way above chance.

However, all the reported classification results are hard to compare, as they all use different test sets for their approaches and there is no standard test set for audio classification.

3. OUR APPROACH

We choose the most promising approach from our literature review in Section 2, which is CFA [11]. The CFA feature performs excellently on high quality TV broadcasts. It does not need training and delivers a single value of music-likelihood for a given chunk of audio. We add several pre- and post-processing steps and change the parameters used for the calculation steps of the CFA feature massively. To evaluate the performance of our implementation and modification of the CFA and the surrounding processing chain, we employ a classic machine learning approach for speech/music detection, utilizing MFCCs with SVM classification.

4. EXPERIMENTS

4.1. Base material and ground truth

Public access radio stations from all over the world, who are members of the CBA network, upload their broadcasts to the platform. The archive grew to a large database with audio files from many different sources and differing quality. The base material is a collection of over 30,000 hours of radio broadcasts from which a large ground truth data set was annotated.

A random sample of 50 programs was taken from the existing database. In this material we annotated music and speech. Sections where speech and music are mixed were annotated as speech to meet the requirements of our project partner to delete only pure music parts from the programs but no parts containing speech. This resulted in a training set of about 30 hours of audio data with a proportion of 40.7% music. On this set we performed our initial processing and comparisons. In a second step we took another random sample of 50 programs to be able to test the obtained parameter set for the CFA feature against a larger dataset. This resulted in a test set of 33 hours of audio data with 43.7% music. In total we had 63 hours of annotated audio for our experiments.

4.2. Feature extraction and classification

Several experiments using different combinations of features and classifiers were conducted. First, simple classification using 13 MFCCs and an SVM was performed. Using repeated random sampling, we tested models which were trained on

Fig. 1. Consecutive processing steps for the analysis of an audio file.

each program alone and models on random selection of parts of all programs.

In a second step we implemented the CFA feature according to [11] and evaluated its performance with the parameter set suggested by its authors. It performed suboptimal on our material. Hence, we modified the parameter set and added several pre- and post-processing steps.

5. RESULTS

5.1. Classification accuracy

First classification tests using the MFCCs on our files resulted in an average accuracy of 82,7% with an SVM model trained on random chunks of each audio file, and evaluation against the remaining chunks of the same file. Subsequently, we trained the SVM model with random chunks of our training set and finally with random chunks of our complete grund truth (see Table 1).

In a second step we extended the MFCC feature vector with the raw CFA value. Classification results increased significantly, which suggests a high relevance of the CFA-feature.

Finally, we tested the CFA-feature on its own using a simple threshold classification without the need for a model to be trained. This way even better classification results were achieved on average over all files than in combination with the MFCC-feature (see Table 2). Applying an extensive grid search over the CFA's 14 parameters led us to values that differ greatly from the original but yield much better results. E.g for the binarization threshold we found the value 10 to work best, instead of 0.1 in the original paper. For the number of peaks to calculate the CFA value we found 40 peaks to yield better results than the original 5 peaks. The obtained parameter set can be provided on request.

By taking into account temporal properties (i.e. music and speech usually last longer than a second) we could successfully apply post-processing steps that improved classification results even further. By smoothing the CFA results of consecutive frames using a window convolution and applying an dilation/erosion filter after the threshold classification, we could eliminate short parts of one group surrounded by longer parts of the opposite group (also see Table 2). While overall improvement is marginal, these measures simplify human inspection mentioned in section 5.3. The final sequence of the calculation and classification steps can be seen in Figure 1.

SVM with MFCC	single model for each file	combined model over all files
training set	82,7%	78,3%
whole ground truth	-	74,5%
SVM with MFCC + CFA		
training set	94,2%	91,1%
whole ground truth	-	90,5%

Table 1. Classification accuracy of MFCC-feature and MFCC + CFA-feature using SVM-classification.

CFA	average over all files with indiv. threshold	average over all files with same threshold
w/o. post-processing	95,8%	90,8%
w. post-processing	96,6%	92,2%

Table 2. Classification accuracy of the CFA-feature with a simple threshold classifier.

5.2. Encountered problems with the CFA-feature

In our continuous tests during the implementation of the CFA-feature we encountered problems, some of which had tremendous influence on the classification results. In the following sections we describe the problems as well as the approaches to solve them.

5.2.1. Mains hum

In several audio files we observed noise during parts of silence (see Figure 2). In some cases it turned out to be mains hum, probably caused by a ground loop in the recording equipment. Having a steady frequency around 50Hz it is amplified by the CFA computation algorithm and leads to misclassification of silent parts as music. To overcome that problem we inserted an adjustable noise gate as pre-processing step (see Figure 1), eliminating silent noise.

5.2.2. DC-Offset

Another problem we encountered that seemed to have an influence on our classification results were changing DC-offsets within programs (see Figure 3). We do not know its cause for sure but we believe it to be a result of mixing different audio sources. It cannot be heard and is usually eliminated by a

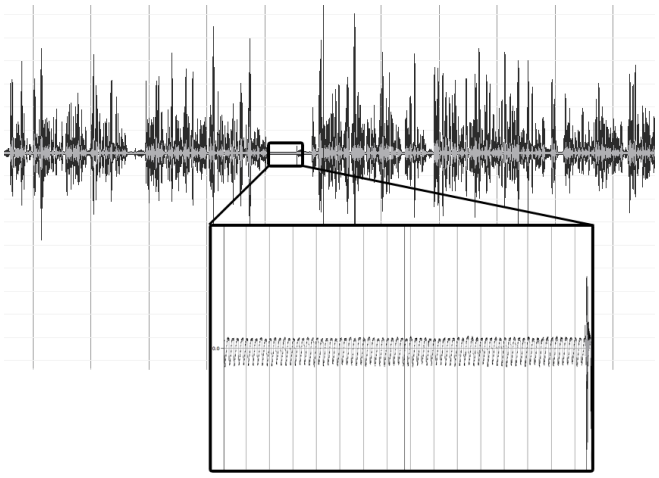

Fig. 2. Mains hum during parts of silence leads to misclassification of the parts as music.

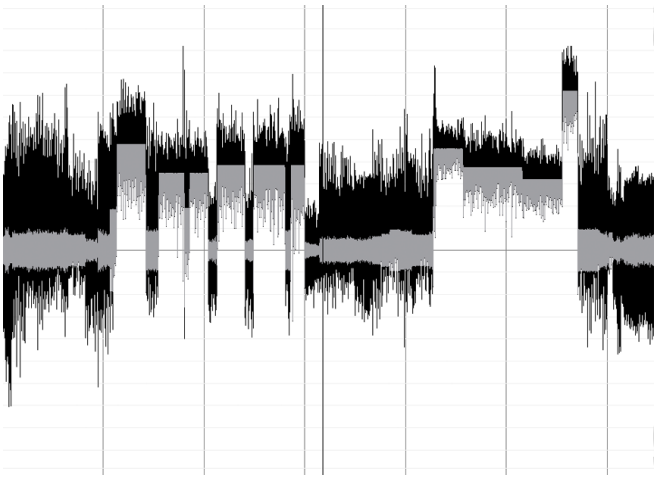

Fig. 3. A changing DC-offset in an audio file thought of being a reason for major misclassification.

filter in the digital-analog conversion circuit on a sound card, but does have an influence on the CFA calculation.

We assumed this to be the main reason for a major misclassification of one single audio file with large DC-offsets. We employed a high pass-filter on all of our files, but instead of increasing our classification accuracy dramatically, it was only raised by 2.0% for that single file and the average value over all files even decreased by 0.5%. After closer analysis of this audio file and the computed CFA-values we found that most of the misclassification occurred in a part of the audio file where music was played with highly rhythmical structure but only little tonal portions. As the CFA-feature is designed to find steady frequencies to identify musical parts in an audio file, it is literally doomed to fail for music that does not contain any or only little steady frequencies. This circumstance has already been indicated by Seyerlehner et al. in [11], who suggest to combine it with other features that focus on rhythmical properties of music. One possible solution for this problem was proposed after the end of our project by Han and Coover in [13], who apply transient activation (TAC) for detection of rhythmical content in speech/music detection.

5.3. Integration into existing infrastructure

We implemented our approach into the program upload procedure of the Cultural Broadcasting Archive. The initial computation run for all 30,000 files that already existed in the archive took about 5 days. In the current state of the archive the algorithm is executed in the same moment a user uploads a new audio file. As we evaluated in our experiments that most of the 3.8% misclassified audio parts can be classified correctly using a different threshold, the system provides ten different speech/music segmentations with ten thresholds for classification. The standard threshold is set to the value that we obtained with our best average evaluation results. The output is visualized in a waveform editor with a colored overlay for segments that our detector estimates as music. The uploader can then change the threshold with a slider and see the impact on the segmentation of the audio material directly. This manual post-processing step allows an accuracy close to 100%.

6. CONCLUSION

In this paper we examined the use of the continuous frequency activation feature developed by Seyerlehner et al. [11] for speech/music detection in radio programs stored in the Cultural Broadcasting Archive. The CFA feature was implemented in the YAAFE[3] feature extraction framework and a parameter set was determined and tested against a ground truth of over 60 hours of audio material. The achieved classification accuracy of **96,6%** was tested against a speech/music-detector using MFCCs and a SVM classifier.

During the tests some problems with the quality of different audio files from the ground truth became apparent. Some of them had a tremendous influence on the calculation of the CFA value, which resulted in misclassification of parts of the audio material. By introducing some further processing steps and parameter tweaking some of the problems could be overcome. Others, like the non-recognition of music with mainly rhythmical content, need further improvement. Still, the implemented classification algorithm yields very good classification results, even for non-professional radio programs.

For other, comparative work the ground truth and the obtained parameter set can be provided on request. The implementation of the computing algorithm was done as YAAFE extension and can be downloaded as open source package[4].

[3] http://yaafe.sourceforge.net/

[4] http://mcrg-fhstp.github.io/cba-yaafe-extension/

7. REFERENCES

[1] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP '96*, Atlanta, Georgia, USA, 1996, vol. 2, pp. 993–996, IEEE.

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature Speech/Music discriminator," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP '97*, Munich, Germany, 1997, pp. 1331–1334.

[3] Gethin Williams and Daniel PW Ellis, "Speech/music discrimination based on posterior probability features," in *Eurospeech 99: 6th European Conference on Speech Communication and Technology: Budapest, Hungary, September 5-9, 1999*, 1999.

[4] Khaled El-Maleh, Mark Klein, Grace Petrucci, and Peter Kabal, "Speech/music discrimination for multimedia applications," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 6, p. 24452448.

[5] C. Panagiotakis and G. Tziritas, "A Speech/Music discriminator based on RMS and zero-crossings," *IEEE Transactions on Multimedia*, vol. 7, pp. 155166, 2004.

[6] Sourabh Ravindran, Kristopher Schlemmer, and David V. Anderson, "A physiologically inspired method for audio classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1374–1381, June 2005.

[7] M. Kashif Saeed Khan and Wasfi G. Al-Khatib, "Machine-learning based classification of speech and music," *Multimedia Systems*, vol. 12, no. 1, pp. 55–67, Apr. 2006.

[8] Aggelos Pikrakis, Theodoros Giannakopoulos, and Sergios Theodoridis, "Speech/Music discrimination for radio broadcasts using a hybrid HMM-Bayesian network architecture," in *Proc. EUSIPCO*, 2006, vol. 6.

[9] T. Izumitani, Ryo Mukai, and K. Kashino, "A background music detection method based on robust feature extraction," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, Mar. 2008, pp. 13–16.

[10] Ji-Soo Keum and Hyon-Soo Lee, "Speech/Music discrimination based on spectral peak analysis and multilayer perceptron," in *International Conference on Hybrid Information Technology, 2006. ICHIT '06*, Nov. 2006, vol. 2, pp. 56 –61.

[11] Klaus Seyerlehner, Tim Pohle, Markus Schedl, and Gerhard Widmer, "Automatic music detection in television productions," in *Proceedings of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, France, 2007, pp. 221–228.

[12] Jan Schlüter and Reinhard Sonnleitner, "Unsupervised feature learning for speech and music detection in radio broadcasts.," in *Proceedings of the 15th Int. Conference on Digital Audio Effects (DAFx-12),*, 2012.

[13] Jinyu Han and B. Coover, "Leveraging structural information in music-speech dectection," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, July 2013, pp. 1–6.

[14] Keansub Lee and Daniel P. W. Ellis, "Audio-based semantic concept classification for consumer video," *Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 6, pp. 14061416, Aug. 2010.