An Evaluation of Using Chroma- and MFCC-based Features for Classifying Radio Transmissions

Frank Kurth, Dirk von Zeddelmann

FGAN-FKIE, 53343 Wachtberg-Werthhoven, Germany, Email: {kurth,zeddelmann}@fgan.de

Introduction

In this contribution, we evaluate the usability of chroma- and MFCC-based features for the task of classifying radio transmissions containing music, speech and mixed audio content. In our system, an incoming audio stream is analyzed by a cascade of binary classifiers, each based on a different type of audio feature. Here, the type of feature used by an individual classifier is choosen according to the particular classification task, based on the hypothesis that the well-known MFCC features better characterize speech contents whereas the comparatively new chroma-based features [KM08] better characterize music audio. The latter hypothesis is validated in a series of experiments which we describe subsequently. To compensate possible strong fluctuations which are inherent in the classical MFCC features, we describe the novel class of MFCC-ENS features which are constructed by calculating suitable short-time MFCC-statistics. The novel features are shown to significantly improve the MFCCs' classification results. Our classifiers are evaluated in an audio segmentation scenario, where an incoming broadcast radio recording is to be segmented into a sequence of regions each assigned one of the class labels *music*, *speech*, or *mixed forms*. The latter class includes simultaneous speech and music as it typically occurs in commercials, jingles, or when background music is present during speech activity. For evaluation we use a larger-scale corpus of manually annotated audio material recorded from live radio transmissions.

Feature Extraction

Chroma-based audio features have turned out to be a powerful feature representation in the music retrieval context, where the chroma correspond to the twelve traditional pitch classes $\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}$ of the equal-tempered scale. To construct chroma features, the audio signal is converted into a sequence of twelve-dimensional chroma vectors by first analysing the signal using a semi-tone filter-bank (i.e., one subband for each musical note) and then summing up the energies of all subbands for a common chroma. To obtain features that robustly represent the harmonic progression of a piece of music, chroma features have been extended in [KM08] by computing local statistics of the chroma energies using a sliding time-window and quantization, followed by an energy normalization and subsequent downsampling. By choosing the size of the statistics window and the downsampling factor, the resulting CENS-features (Chroma Energy Normalized Statistics) may be calculated in different time resolutions and may hence be used for very robustly matching music audio containing strong temporal variations [KM08].

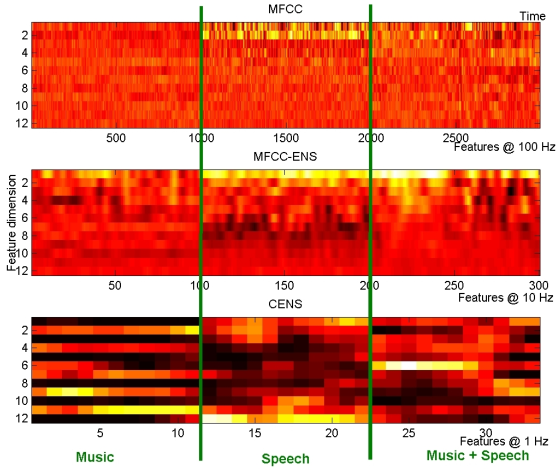

Figure 1: Three feature sets, MFCCs (top), MFCC-ENS (center), CENS (bottom), extracted from an audio fragment of mixed content.

The approach of using short-time statistics to construct robust audio features may by applied to different feature types. For the classification problem considered in this paper, we propose to use short-time feature statistics with the MFCC features which have been sucessfully used in speech processing applications [RJ93]. The MFCC construction consists of a short time Fourier transform where, for each Fourier vector, the logarithmic amplitude spectrum is binned into 40 mel-spaced bands which are subsequently decorrelated using a DCT. To obtain local statistics, the above CENS-construction is performed using the 40 mel-subbands instead of the chroma subbands. This step is followed by the usual DCT. The corresponding features will be called MFCC-ENS (MFCC Energy Normalized Statistics). Fig. 1 illustrates three different feature sets, MFCC, MFCC-ENS, and CENS (each 12 dimensional), extracted from a short piece of audio consisting of 11 seconds of each music, speech and a radio jingle (mix of two speakers and background music). In constructing the MFCC-ENS, the local statistics window has a duration of 800 ms while the resulting feature rate is 10 Hz.

Two-Stage Classifier

The above features are now used to construct several binary audio classifiers. Those are then combined into a two-stage classifier which is used for segmenting broadcast radio programmes. In our segmentation scenario we consider the classes of *Music (C1)*, *Speech (C2)* and

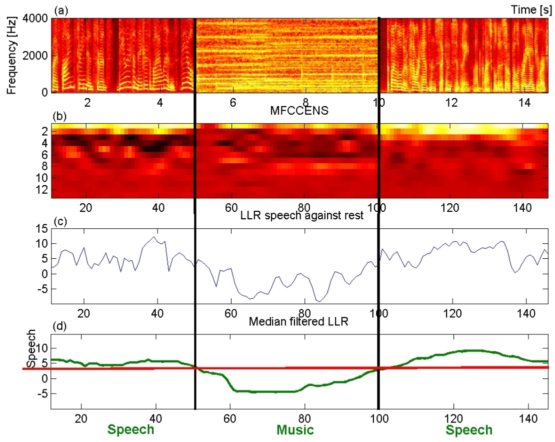

Figure 2: (a) Signal containing speech and music, (b) extracted MFCC-ENS features, (c) log likelihood ratio (LLR) of speech against mixed forms GMMs, (d) median-smoothed LLR and classification result.

Classification task ↓	Accuracy in % for feature type		
	CENS	MFCC	MFCC-ENS
Music-Speech	99.23	94.27	97.10
Speech-Mixed Forms	72.06	91.98	94.86

Table 1: Correct classifications in % of three considered feature types.

Mixed-Forms (C3). Each of our binary classifiers is constructed to discriminate between two of those classes.

For classification, we train Gaussian Mixture Models (GMMs) for each of the three audio classes and for each of the considered feature classes, resulting in a total of 9 GMMs. Each consists of a mixture of 16 densities with parameters and weights adapted by EM-based learning using features extracted from 20, 40, and 100 minutes (classes music, speech, mixed forms) of training data. The different durations are due to the different sampling rates of the resulting feature sequences. For binary classification of a sequence of input features, two GMMs are selected and log-likelihood ratios between the GMMs' class probabilities are calculated. Fig. 2 shows an example of a classification between classes (C2) and (C3) based on MFCC-ENS features. The MFCC-ENS features (b) extracted from the input signal (a) yield a log likelihood ratio (LLR) as depicted in (c). The LLR is smoothed by median filtering (d), and classification is subsequently performed by thresholding (red line).

Following the above considerations that the particular features are adapted to specific audio classes, we now concentrate on classifiers for discrimination between both *Music-Speech*, i.e., (C1) against (C2), and *Speech-Mixed Forms*, i.e., (C2) against (C3). Table 1 summarizes the resulting rates of correct identifications obtained from a per second evaluation of the classifier outputs. It is observed that the CENS-based classifier is most suitable for music while the MFCC-ENS yield the best results for discriminating between speech and mixed forms. As a consequence, our proposed two-stage classifier illustrated in Fig. 3 uses a cascade of, first, a CENS-based GMM-classifier for detecting music segments and, second, a MFCC-ENS-based GMM-classifier for partitioning the remaining signal parts into speech and mixed-form segments. In our application scenario, we found that the first classifier can be suitably realized as a binary classifier between the music and speech classes.

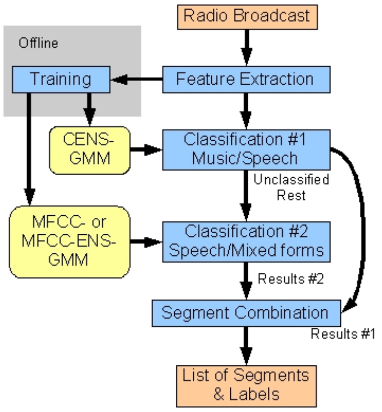

Figure 3: Two-stage segmentation system using CENS- and MFCC-ENS-based GMM classifiers.

Seg. result [%] ↓	True class		
	C1	*C2*	*C3*
C1	98.3	0	0
C2	1.7	97.55	6.19
C3	0	2.45	93.81

Table 2: Confusion matrix for results of proposed CENS- and MFCC-ENS-based segmenter (right). Used classes: *Music (C1)*, *Speech (C2)* and *Mixed forms (C3)*.

Evaluation

For evaluation, audio segmentation was performed using the above procedure. Our test data consisted of 4:09 hours of a contiguous audio programme recorded from a classic radio station and labeled manually. The material comprises 206.42 minutes of music *(C1)*, 13.5 minutes of speech *(C2)* and 30.15 minutes of *(C3)*-segments (mainly jingles and commercials consisting of mixed speech and music). For this data, the overall rate of correct classifications using the two-stage approach was 97.72%, where we again evaluated one classification result per second. Table 2 shows the confusion matrix for the three involved classes. As might be expected, the class *C3* containing superpositions of music and spoken language causes the largest classification errors.

As a conclusion, the proposed two-stage approach to audio classification using content-adapted feature types yields good results. Furthermore, the described novel MFCC-ENS features significantly improve the MFCC-based part of the classification.

References

[KM08] Kurth, F., Müller, M.: Efficient Index-based Audio Matching. IEEE Trans. on Audio, Speech, and Language Processing **16**(2), 2008, 382–395.

[RJ93] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.