Automatic Detection of Song Changes in Music Mixes Using Stochastic Models

Thomas Plötz and Gernot A. Fink, and
Peter Husemann, Sven Kanies, Kai Lienemann, Tobias Marschall,
Marcel Martin, Lars Schillingmann, Matthias Steinrücken, Henner Sudek
Faculty of Technology, Bielefeld University, 33 594, Germany
{tploetz,gernot}@techfak.uni-bielefeld.de

Abstract

The annotation of song changes in music mixes created by DJs or radio stations for direct access in digital recordings is, usually, a very tedious work. In order to support this process we developed an automatic song change detection method which can be used for arbitrary music mixes. Stochastic models are applied to music data aiming at their segmentation with respect to automatically obtained abstract generic acoustic units. The local analysis of these stochastic music models provides hypotheses for song changes.

Results of an experimental evaluation processing music mix data demonstrate the effectiveness of our method for supporting the annotation with respect to song changes.

1. Introduction

In the "pre-digital" era (legal) recording of music mixes was usually performed by means of magnetic tape recorders. The mixes consist of songs cross-faded either automatically or by a DJ at least avoiding silence between two songs. Once recorded the segmentation into separate songs, e.g. for direct access, was performed manually, i.e. listening to the mix and annotating the song changes.

Meanwhile digitally recording music mixes directly to e.g. CDs has replaced the old-fashioned tape recording. Furthermore, DJ culture has developed enormously and music mixing has been established as an art. Today, on many radio channels (including internet streams) a large amount of high-quality music mixes is provided.

When aiming at direct access to separate songs within digital recordings of music mixes the abovementioned segmentation problem remains the same as for tape recording. In the worst case the whole mix needs to be listened to for manual annotation. In order to support this segmentation process we developed a method for automatic song change detection based on stochastic models. Basically the focus of the system is on high sensitivity, i.e. the number of missed song changes should be as small as possible. Furthermore, the system should work on arbitrary music mixes, i.e. the songs included can be unknown and their genres are not fixed beforehand. The hypotheses of song changes delivered by the automatic segmentation system are the basis for accelerated manual post-processing. Given our song change detector the annotation of arbitrary music mixes can be performed very efficiently.

In this paper the automatic song change detection in music mixes is treated as a general pattern recognition problem. Based on a parametric representation using general audio features stochastic models, namely Hidden Markov Models, are estimated covering general acoustic events. Therefore the concept of generic acoustic generators (GAGs, first described in [1]) is further developed. By means of these GAGs an abstract description of music is obtained which is used for local modeling of small parts of music (approximately one minute). Using an extended sliding window approach, the so-called *Snip-Snap* technique, local analysis of differences in N-gram scores or competitive evaluation of original and adapted GAG sets is used for song change detection. The larger the difference of these scores, the higher the evidence for song change. Based on an experimental evaluation on real data, i.e. both music mixes created automatically by cross-fading and those created by professional DJs, the capabilities of our automatic song change detection method is demonstrated. Song change annotation in music mixes can be performed very efficiently when using the newly developed system.

In the following section we describe related work and our method for automatic song change

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

detection is introduced in 3. This includes the features used, the modeling of generic acoustic events using HMMs, and the *Snip-Snap* technique for local music chunk discrimination, i.e. song change detection. In section 4 the results of our experimental evaluation are presented.

2. Related Work

With respect to our general goal of automatic song change detection in *arbitrary* music mixes, i.e. including songs which are neither known in advance nor limited to specific music genres, there is hardly any literature available. Most of the music transcription or song identification systems which might be used for song change detection as addressed by this paper can only be applied to songs which are known to the system while training.

According to [7] Mel Frequency Cepstral Coefficients (MFCCs) originally developed for automatic speech recognition applications (cf. e.g. [5]) can also be used for music modeling. Furthermore, various statistical features calculated when analyzing the underlying music signal either directly or in its spectral representation can be defined (e.g. energy-based features, or auto correlation values). In [1, 2], for example, MFCCs are used for HMM-based song identification whereas in [9] statistical features describing pitch, rhythm etc. are applied for mixture-density based genre classification.

In [1, 2] a system for song identification in radio broadcasts has been presented. The authors use an MFCC-based feature representation of music for un-supervised training of HMMs for modeling *generic acoustic generators* (GAGs). Using these GAGs song signatures are established which can be used for identification of known songs in radio broadcasts even when they are distorted by radio edits, speaker over audio, or when parts of the songs are changed or removed. As the main purpose of this system is copyright enforcement it has been designed to give almost no false positives to achieve very high accuracy.

3. Automatic Song Change Detection

In order to detect song changes the method described in this paper is based on a high-level description of general music. Therefore, we generalize the concept of GAGs modeled by HMMs as briefly summarized in the previous section. Using GAGs which are trained on feature representation of music data, we segment the particular music mix into general acoustic units. This segmentation serves as the basis for actual song change detection. We developed two variants either utilizing local N-gram scores or analyzing GAG scores directly.

3.1. Features

According to the literature (cf. section 2) there is no standard method for music feature extraction. Thus, in our approach MFCCs, energy-based as well as spectral features including first and second derivatives are integrated into a 105-dimensional vector. In table 1 an overview of the features used is given whereas a detailed description of the features is omitted due to space limitations. The feature vectors are calculated on 25ms frames which are extracted using a sliding window technique with an overlap of 10ms. For redundancy reduction a PCA-based automatic decorrelation is performed resulting in 76-D feature vectors.

Dimension	Feature
1 – 48	16 MFCCs incl. derivatives
49 – 78	energy-based incl. derivatives
79 – 105	spectral incl. derivatives

Table 1. Summary of features

3.2. GAG Training

GAGs are trained on *general* music data originating from various sources (not necessarily mixes). Since GAGs are defined as HMMs modeling abstract acoustic events their actual meaning cannot be specified in advance. Thus, standard HMM training using annotated sample sets of music is not suitable. Contrary to [1] we estimate a set of GAGs as some kind of fixed inventory used for the symbolic description of music in general by means of a model-based clustering technique (cf. [8]). Therefore, free model alignment and training based on preceding annotation is alternated iteratively until convergence. Here, convergence means that annotations obtained in two subsequent iteration steps do not differ substantially. As the result of the clustering a set of well-trained HMMs covering generic acoustic events is obtained.

Aiming at robust clustering results and reasonable convergence speed the abovementioned iterative training process is initialized as follows. Based on general music data a mixture density is estimated where every component corresponds to an initial GAG prototype. This model is used for computing a mixture-component based annotation of the training data which corresponds to an initial GAG labeling. In figure 1 the training process is illustrated by means of a visualization of the GAG-based annotations of one part of an exemplary song. The first row of the diagram represents the mixture-classifier based initial annotation. In the subsequent iterations the annotations change until convergence at the 15th iteration where a stable alignment is reached.

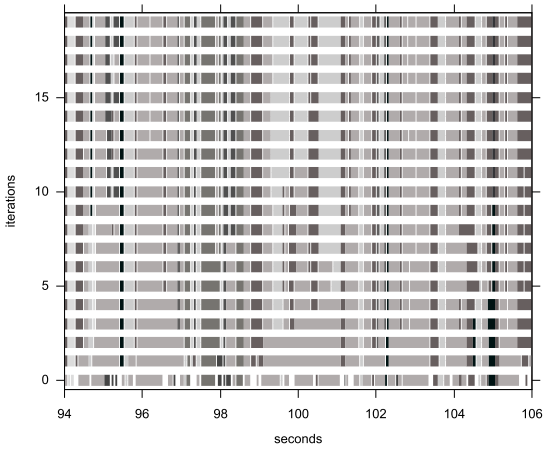

Figure 1. Progress of GAG-based sample alignments until convergence

3.3. The Snip-Snap Approach

When analyzing music mixes, basically, local discontinuities within the signal are the major evidence for a song change. Thus, instead of modeling and tracking the changes of global properties of specific songs in the approach presented here an explicit local analysis is addressed.

Therefore, two consecutive windows each covering approximately one minute of music, the *Snip* followed by the *Snap*, are moved frame-wise along the particular music mix. Given the symbolic GAG-related description of the underlying music mix, some stochastic music model is estimated using the Snip data. Following this the probability for a song change can be computed for every position within the music mix by evaluating the Snip model for the Snap window. If the Snap score, depending on its actual meaning, drops below or rises above some threshold, the current Snip model cannot predict the Snap data properly which gives evidence for a song change. In figure 2 the general Snip-Snap approach is summarized graphically.

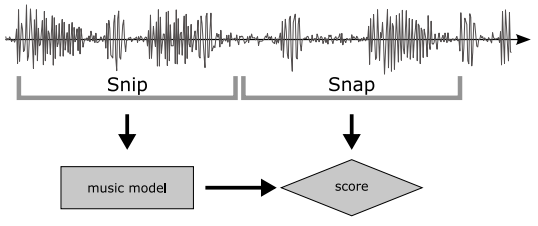

Figure 2. The Snip-Snap Approach

For one variant of our song change detection system the GAG-related music representation of the Snip window is used for estimating a statistical N-gram model, namely a Bi-gram (cf. e.g. [3]). Using the Snip-based Bi-gram the perplexity is calculated using the GAG representation of the Snap data. When moving both windows along the whole music mix a numerical representation consisting of Bi-gram perplexity values for every time step is created in which, finally, a peak detection technique provides the song change hypotheses.

3.4. Model Specialization

In addition to the N-gram related Snip-Snap approach a Snip-Snap variant based on competitive evaluation of different GAG model sets has been developed. The Snip feature data is utilized for specialization of the general set of GAG HMMs estimated during training (cf. section 3.2), i.e. adaptation towards the music data covered by a particular Snip. Following this, both the original set λ_{org} of GAG HMMs and the Snip adapted set λ_{special} are evaluated for the Snap data providing two HMM scores $P(\text{Snap}|\lambda_{\text{org}})$ and $P(\text{Snap}|\lambda_{\text{special}})$, respectively. If $P(\text{Snap}|\lambda_{\text{org}}) \geq P(\text{Snap}|\lambda_{\text{special}})$, the Snip adapted model predicts the Snap worse than the original GAG sets corresponding to a putative song change. Similar to the N-gram variant of the Snip-Snap approach both windows are moved along the whole mix and a peak detection technique is applied to the score differences.

Currently two variants of model specialization are integrated into our song change detector:

Mixture adaptation: Given the Snip alignment using the general GAG set either maximum a-posteriori (MAP, cf. e.g. [4]) or maximum likelihood linear regression (MLLR [6]) adaptation is applied to the original HMMs.

GAG selection: The specialized set of GAGs is obtained by reducing the original set to the HMMs actually used for Snip alignment.

4. Experimental Evaluation

In order to set up our system and to evaluate its capabilities for automatic song change detection we performed experiments using music data obtained from internet radio streams.

For GAG training we used 240 separate songs (rock, techno, classic, folk, rap, 80s) corresponding to 1,055 minutes of general music. For parameter optimization, and evaluation of the overall system two different kinds of actual music *mixes* were processed. The first kind consists of $20\times$ 11 separate songs, not included in the abovementioned training set, automatically cross-faded which corresponds to common radio broadcasts. Based on this data we created two disjoint sets (cross-validation and evaluation) each consisting of approximately 470 minutes of music and 100 song changes per set.

The second kind of mixes was created by professional DJs. Again we created disjoint cross-validation (222 minutes of music with 65 song changes) and evaluation (117 minutes of music including 37 song changes) sets which were annotated manually (in a certainly subjective manner).

Due to space limitations not all but the most relevant parameters of the system adjusted during cross-validation are summarized in table 2.

Parameter	Value
# Mixtures for GAG HMMs	1,000
# GAG HMMs	200
Min. GAG length	30 - 120ms
Lengths of Snip/Snap	60s
Increment for Snip-Snap move	3s

Table 2. Optimal system configuration

Mix	Match [%]	Add [%]	Miss [%]
Bi-gram based Snip-Snap			
Artificial	93	25	7
DJ	70	110	30
Adaptation-based Snip-Snap			
Artificial	87	51	13
DJ	65	146	35
Combination			
Artificial	97	13	3
DJ	81	302	19

Table 3. Results of evaluation

Using the optimally configured system its final evaluation was performed processing either artificial or DJ mixes (evaluation sets). In table 3 the results are summarized for both Snip-Snap variants and, additionally, for the combination of both techniques. It can be seen that almost all song changes within the artificial mixes are detected correctly within a range of $\pm$10s around the actual change (corresponding to match rates of more than 90%). Although the detection rates drop significantly when processing professional DJ mixes the percentages of correctly predicted actual song changes is reasonably high (between 65 and 70%). The specificity of the system is good for the artificial mixes (20 – 50% false alarms). Unfortunately, the percentage of (erroneously predicted) additional song changes increases substantially when processing real DJ mixes. However, even for human listeners it is often really hard to detect the song changes correctly since the quality of the mixes is extraordinary. Furthermore, our system was designed to support the song change detection with special focus on minimizing the number of false negatives.

5. Conclusion

In this paper we presented an automatic song change detection system based on stochastic models. Therefore, local models (HMMs, and N-gram models) estimated using feature-based generic acoustic generators (GAGs) are trained in an unsupervised manner and evaluated by applying the newly developed *Snip-Snap* technique.

By means of an extensive experimental evaluation we demonstrated that the song change hypotheses provided by our system can be used very effectively for the original task of supporting song change annotation in arbitrary music mixes.

The focus of our system is on user *support* for fast song change annotation. Thus, the number of missed actual song changes needs to be minimized. Demonstrated by the experimental evaluation it becomes clear that our song change detection system fulfills this requirement very sufficiently. On the other hand, the number of false alarms is of minor importance only since a human listener can identify those false positives very quickly while listening the small music chunks which in combination greatly accelerates music mix annotation. However, reducing the false alarm rate is work in progress.

References

[1] E. Batlle, E. Guaus, and J. Masip. Automatic song identification in noisy broadcast audio. In*Proc. Int. Conf. Signal and Image Proc.*, 2002.

[2] E. Batlle, J. Masip, and E. Guaus. Amadeus: A scalable HMM-based audio information retrieval system. In *Proc. First Intern. Symp. Control, Communications and Signal Proc.*, 2004.

[3] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13, 1999.

[4] J.-L. Gauvain and C.-H. Lee. MAP estimation of continuous density HMM: Theory and applications. In *Proc. DARPA Speech and Natural Language Workshop*, 1992.

[5] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing – A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[6] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. *Computer Speech & Language*, 1995.

[7] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Int. Symp. Music Information Retrieval (ISMIR)*, 2000.

[8] M. P. Perrone and S. D. Connell. K-means clustering for Hidden Markov Models. In *Proc. Int. Workshop Frontiers in Handwriting Recog.*, 2000.

[9] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech and Audio Processing*, 10(5), 2002.