Identification of Live or Studio Versions of a Song via Supervised Learning

Nicolas Auguin, Shilei Huang and Pascale Fung
Department of Electrical and Computer Engineering, HKUST, Clear Water Bay, Hong Kong
E-mail: njlpauguin@ust.hk, eehuangs@ust.hk, pascale@ece.ust.hk

***Abstract*—We aim to distinguish between the "live" and "studio" versions of songs by using supervised techniques. We show which segments of a song are the most relevant to this classification task, and we also discuss the relative importance of audio, music and acoustic features, given this challenge. This distinction is crucial in practice since the listening experience of the user of online streaming services is often affected, depending on whether the song played is the original studio version or a secondary live recording. However, manual labelling can be tedious and challenging. Therefore, we propose to classify automatically a music data set by using Machine Learning techniques under a supervised setting. To the best of our knowledge, this issue has never been addressed before. Our proposed system is proven to perform with high accuracy on a 1066-song data set with distinct genres and across different languages.**

I. Introduction

Music Information Retrieval (MIR) has gained more and more interest over the last few years. Progress in music, audio and acoustic feature extraction, as well as state-of-the-art Machine Learning techniques, have boosted the development of information retrieval. This assists Music Genre Classification or Music Emotion Classification for example.

Meanwhile, with the fast-spreading development of the Internet have come music streaming services, which have deeply modified the experience of everyday music listeners. Though the listening experience of the users of such services may be enhanced through recommendation tools, for example, it may also be affected by detrimental factors, such as bad quality (due to a poor recording, unexpected stops, etc.) or song-specific disappointment (a played song is different from or is not the one that the listener was expecting). In this paper, we are interested in targeting song-specific user disappointment. In particular, we focus on the following problem: how to guarantee that the song played is the original studio version, and not a live recording. To the best of our knowledge, this issue has never been addressed before. Therefore, we need to give further insight into the problem.

Live versions of a song can differ greatly from the original ("acoustic") version; outward signs of the crowd presence, inherent to any public performance, often interfere with the music content. Most of the time, these signs are applause or cheering.

Identifying those signs is a challenging issue, addressed in the more general framework of audio event detection. For example, in [1], the authors propose to measure the audience's appreciation through its applause or loud cheering so as to detect the highlights of a baseball game. In [2], applause identification is used to distinguish the music pieces' boundaries and highlights in Carnatic concerts (with a single, continuous recording). For instance, in [3], the authors propose to use Mel Frequency Cepstrum Coefficients (MFCCs) to detect applause. In [4], an SVM-based audio event detection system is implemented to differentiate speech, music, cheering and applause, using Linear Predictive Cepstrum Coefficients (LPCCs), pitch, and audio signal energy-related features. The same issue is addressed in [5], though they use MFCC and MPEG-7 audio features, and their system aims to classify audio events classification in sport. However, on our music classification task, such audio event detection is more challenging since applause, cheering, music and the artist's voice are mixed. Nonetheless, such strategies give us some insight into the features that may be useful to discriminate between live and studio versions of songs.

Meanwhile, previous works in music classification offer us some insightful perspectives. In music genre classification, for example, supervised systems have been successfully built using timbral features (including MFCCs, roloff, flux, zero-crossings and spectral centroid), beat histogram features, and rhythmic and pitch content features [6]. Psycho-acoustic features (describing the loudness, the pitch, the sharpness, etc.), aimed at modeling the parameters of human auditory sensation, have been used in music emotion classification [7], in addition to some of the aforementioned features.

These works, in both the audio and music processing fields, will guide our steps in building the most relevant feature set for our given task.

The rest of this paper is organized as follows: in section II, we present our methodology, our music data set and our feature set. The experiments presented in section III show us which parts of a song are particularly useful for information retrieval. In section IV, we propose an experimental study of the relative importance of the features introduced in section II. We then show that our system performance is independent of the language used by the singer(s), using some English and French music data sets. We conclude in section VI.

II. Live and studio versions identification system

A. Music data set

We constitute a music data set composed of 1066 unique songs from various genres (rock, pop, jazz...) and in different

languages (English, French, Spanish, Chinese, Portuguese). The class distribution can be seen in Table I.

TABLE I
CLASS DISTRIBUTION OF THE MUSIC DATA SET

Live songs	Studio songs	Total songs
378	688	1066

The class "live" refers to the songs recorded during live performances (concerts), while the class "studio" refers to the original songs recorded in a studio. Note that to guarantee a fair classification, all songs have been extracted from official albums. In other words, the "live" songs have not been recorded by amateurs and are of similar quality to the "studio" songs.

We consider a very general framework, where all songs are unique *a priori*, regardless of their class ("studio" or "live"). Therefore, we do not consider two (or more) different versions of the same song, but rather a set of songs unrelated to each other (except for the class label).

B. Feature set

Since this classification task has not been addressed before, we decided to extract a wide range of features that have been proven to be useful in both sound event detection and music classification. Thus, we extracted MFCC features, plus other timbral features (zero crossings, spectral centroid, flux, roloff, etc.) LPCC features, MPEG-7 features (Spectral Flatness Measure and Spectral Crest Factor), psycho-acoustic features (loudness, sharpness, spectral and tonal dissonance, etc.), beat histograms and signal energy-based features. These features were extracted using three toolkits: Marsyas [8], OpenSMILE [9] and PsySound [10].

We first extracted 30-second samples of each music piece prior to converting them to 22,050 Hz and 16 bits format and mono channel PCM WAV files. The features were then extracted from these files and concatenated in one single vector (following an early integration strategy).

C. Classifiers

In the following section, we detail how we ran different experiments under a supervised setting. We compare the performance of three classifiers, namely, Naive Bayes (NB), k-Nearest Neighbors (k-NNs) and Support Vector Machines (SVMs). For SVM training, we used the Sequential Minimal Optimization algorithm introduced in [11], and we used two different kernels for SVMs, linear and polynomial (with exponent 2).

III. EXPERIMENTS AND RESULTS

A. Comparison of different classifiers

As is often done in Music Classification (see [12] for genre classification and [13] for mood classification), we chose the second 30 seconds of each song as the sample for feature extraction. This choice assumes that the audio/music content does not differ too much during this period.

We performed a feature selection using the method described in [14]. This correlation-based attribute subset selection method was implemented using weka [15]. We used a greedy forward search for the features subset selection. Feature selection was conducted using stratified 10-fold cross validation. It implies that, though the folds were chosen randomly, the proportion of "live" and "studio" songs within each fold is roughly the same as in the original data set.

Classification was then performed using the same folds as used for the feature selection. Results can be seen in Table II.

TABLE II
GLOBAL ACCURACY COMPARISON USING VARIOUS CLASSIFIERS

1-NN	78.80
10-NN	79.27
15-NN	80.68
Naive Bayes	75.05
SVM (linear kernel)	80.88
SVM (polynomial kernel, exponent 2)	**82.55**

Experimental results show that we achieve very good accuracy. SVMs with polynomial kernel turn out to outperform both k-NN ($k = 1$, $k = 10$ and $k = 15$) and Naive Bayes.

However, we may wonder whether other segments of the song may be more meaningful than these second 30 seconds. This is the purpose of the approach in the following subsection.

B. Alternative segmentation approach

Identifying whether a given song constitutes a live performance recording or not can sometimes be done by listening only to the first seconds of the song. This motivates an alternative segmentation approach. We thus propose to perform our classification task on other segments of each song. Therefore, we extracted the features from samples lasting from 0 until 30 seconds (referred to period A), from 30 until 60 seconds (referred to B) and from 60 until 90 seconds (referred to C). AB refers to samples lasting from 0 to 60 seconds, where features from periods A and B have been concatenated in one single vector. BC refers to samples from 30 to 90 seconds and ABC to samples from 0 to 90 seconds.

From here on, we used only SVMs. These classifiers performed better than the other classifiers in further experiments, but comparison results are omitted for the sake of space and readability.

Results are presented in Table III.

Experimental results show that period A (corresponding to the very beginning of the song) is more meaningful than periods B and C ($0 - 30$s and $30 - 60$s). This is induced by the fact that, in most cases, audio events such as cheering

TABLE III
GLOBAL ACCURACY COMPARISON USING CONSECUTIVE SEGMENTS OF A SONG

SVM classifiers	Segments of the song					
	A	B	C	AB	BC	ABC
Linear SVM	90.05	80.88	80.30	**91.28**	84.24	91.37
Polynomial SVM	**91.56**	82.55	81.05	91.18	85.46	**91.84**

and applause, are present at the beginning of a live version, whereas they are absent from a studio version of a song. This validates the observation that humans usually only need the few first seconds of a song to know whether it is a live performance or not. However, the accuracy when using the B or the C samples also shows that we can achieve very good results, while human labelling is much more challenging if we only consider such parts of the song.

Samples B and C also prove to bring some discriminative information with respect to the A sample, since the results achieved with the AB and ABC samples ($0-60$s and $0-90$s) are better than A's accuracy result.

IV. FEATURES RELEVANCE STUDY

In this section, we are interested in evaluating the relative performance of some subsets of the features. Indeed, previously, we relied on the feature selection method to guarantee a high-correlation between features and classes.

For the following experiments, we used only SVMs on the first 30 seconds of the songs (sample A, as denoted in section III).

A. Features subsets significance

We first wanted to measure the actual relevance of the MFCC, the MPEG-7, the LPCC and the psycho-acoustic features. We thus classified our data set using these subsets of features separately. Results are shown in Table IV.

While MFCC features lead to relatively good classification, it turns out that the LPCC features are of limited interest (with an accuracy of only 64.54% using a linear kernel). Psycho-acoustic features give some fairly good results, which may confirm the assumption that we actually "perceive" differently a live version and a studio version of a song, especially regarding factors such as loudness or pitch. We observe similar results with only MPEG-7 features, timbral features (excluding MFCCs), or beat histograms features (with a global accuracy ranging from 79.36% to 83.49%). However, the signal energy-based features lead to a relatively lower accuracy (73.86% with a linear kernel).

B. Influence of the MFCC and LPCC features on the global accuracy

In this part, we study empirically the influence of the removal of some subsets of features on the global accuracy.

We first removed all the MFCC-based features from the original feature set, built using the first 30 seconds of a song. We then removed the LPCC-based features and the psycho-acoustic features. We also present the global accuracy results when we remove both MFCCs and LPCCs, and finally all LPCCs, MFCCs and psycho-acoustic features (noted PA features). We also recall the global accuracy of the complete original feature set (without feature selection) for comparative purpose. Results are shown in Table V.

TABLE IV
INDIVIDUAL SUBSET RELEVANCE (ACCURACY)

Subsets of features	SVM classifiers	
	Linear kernel	polynomial kernel
MFCCs	**86.96**	80.30
LPCCs	64.54	71.01
MPEG-7	78.48	**79.93**
PA[a]	79.46	**80.21**
Timbral[b]	81.23	**83.49**
BH[c]	**79.36**	79.17
SE[d]	73.86	73.54

[a] Psycho-acoustic features
[b] Excluding MFCCs
[c] Beat histograms features
[d] Signal energy-based features

TABLE V
INFLUENCE OF THE REMOVAL OF SUBSETS OF FEATURES ON THE GLOBAL CLASSIFICATION ACCURACY

Original feature set	SVM classifiers	
	Linear kernel	polynomial kernel
Complete	90.90	89.87
Without MFCCs	90.81	90.71
Without LPCCs	90.415	**91.56**
w/o PA features	90.72	**92.50**
w/o MFCCs/LPCCs	90.71	90.52
w/o PA/MFCCs/LPCCs	**91.75**	89.96

The experimental results show that the removal of the MFCC features does not imply a decrease in performance accuracy. Interestingly, MFFCs alone outperform other subsets of features (as we showed in section IV.A.). It suggests the idea that MFCCs may be redundant to other subsets of features, like other timbral features for example. The same conclusion holds when it comes to the LPCC features and to the psycho-acoustic features.

It implies that the MFCC, the LPCC and the psycho-acoustic features (considered separately) are redundant to the

remaining features, namely: the MPEG-7 features, the timbral features (excluding the MFCCs), the beat histograms features and the signal energy-based features. These results show that extracting the MFCC, the LPCC and the psycho-acoustic features is not necessary to achieve a very good performance (in terms of accuracy) for live/studio versions identification.

V. Language correlation discussion

We mentioned earlier the fact that our data set was composed of distinct genres and languages. The language distribution of our data set is presented in Table VI. By language, we mean the actual language used by the singer(s) within the song.

TABLE VI
Language and class distribution of the music data set

Language	Live songs	Studio songs	Total songs
English	330	422	752
French	48	235	283
Miscellaneous	0	31	31
Total songs	378	688	1066

We applied our previous method to the English music data set and the French music data set separately. We ignored the other songs composing our data set since they did not represent a significant part of the music set.

Following the conclusions in section III, we used only the first 30 seconds of each song (sample A) for feature extraction, and ran Support Vector Machines as classifiers, using 10-fold cross validation. No feature selection was performed. Results are shown in Table VII. The line reading "English and French" refers to the set with English data and French data combined together.

TABLE VII
Global accuracy on the English and French data sets

Language(s)	SVM classifiers	
	Linear kernel	polynomial kernel
English	88.99	**90.85**
French	**91.93**	91.58
English and French	**91.05**	91.05

One can see that our system performs well on both English and French data sets, with a similar accuracy of close to 91% (with a polynomial kernel). This implies that our system may be language-independent, which is supported by the fact that no linguistic-related features have been considered for this classification task.

VI. Conclusion

We proposed to identify studio and live versions of a song using Support Vector Machines. We have shown that not all segments of a song are of equal relevance regarding the classification accuracy. We also have given some insight into the relative importance of some features to our given task. In particular, we showed that that some subsets of features (such as MFCCs, widely used in music classification, or LPCCs, used in speech-related audio events detection) are not necessary to efficiently distinguish between studio and live versions of a song.

The system described here also paves the way for further experiments. We plan to conduct experiments with our system under a semi-supervised setting, which may be more adapted to real-life situations, where only few labelled data are available. This system is very promising since it may help to enhance the listening experience of online streaming services users.

References

[1] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 5, pp. V–632.

[2] P. Sarala, V. Ishwar, A. Bellur, and H. A. Murthy, "Applause identification and its relevance to archival of Carnatic music," in *Serra X, Rao P, Murthy H, Bozkurt B, editors. Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12-13; Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra; 2012.* Universitat Pompeu Fabra, 2012.

[3] R. Jarina and J. Olajec, "Discriminative feature selection for applause sounds detection," in *Image Analysis for Multimedia Interactive Services, 2007. WIAMIS'07. Eighth International Workshop on*. IEEE, 2007, pp. 13–13.

[4] L. Lu, F. Ge, Q. Zhao, and Y. Yan, "A SVM-based audio event detection system," in *Electrical and Control Engineering (ICECE), 2010 International Conference on*. IEEE, 2010, pp. 292–295.

[5] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 5, pp. V–628.

[6] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.

[7] Y. Yang, C. Liu, and H. H. Chen, "Music emotion classification: a fuzzy approach," in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006, pp. 81–84.

[8] G. Tzanetakis, "Marsyas submissions to mirex 2009," *Proceedings of the Music Information Retrieval Evaluation EXchange*, 2009.

[9] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[10] D. Cabrera et al., "Psysound: A computer program for psychoacoustical analysis," in *Proceedings of the Australian Acoustical Society Conference*, 1999, vol. 24, pp. 47–54.

[11] J. Platt et al., "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.

[12] P. Ahrendt, J. Larsen, and C. Goutte, "Co-occurrence models in music genre classification," in *Machine Learning for Signal Processing, 2005 IEEE Workshop on*. IEEE, 2005, pp. 247–252.

[13] K. T. G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music into emotions," in *ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval*. Lulu. com, 2008, p. 325.

[14] M.A. Hall, *CORRELATION-BASED FEATURE SELECTION FOR MACHINE LEARNING*, Ph.D. thesis, The University of Waikato, 1999.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.