Exploiting Semantic Content for Singing Voice Detection

Leonidas Ioannidis, Jean-Luc Rouas

Exploiting Semantic Content for Singing Voice Detection

Ioannidis Leonidas
LaBRI
Univ. Bordeaux,UMR 5800
Talence, France
Email: ioannidis.leonidas@labri.fr

Jean-Luc Rouas
LaBRI
CNRS, UMR 5800, Talence, France
Email: jean-luc.rouas@labri.fr

***Abstract*—In this paper we propose a method for singing voice detection in popular music recordings. The method is based on statistical learning of spectral features extracted from the audio tracks. In our method we use Mel Frequency Cepstrum Coefficients (MFCC) to train two Gaussian Mixture Models (GMM). Special attention is brought to our novel approach for smoothing the errors produced by the automatic classification by exploiting semantic content from the songs, which will significantly boost the overall performance of the system.**

I. Introduction

Music characterization via high-level semantic content is a need dictated from the ongoing rise of digital information needed to be retrieved by its content. Accessing and labelling data automatically, according to its content, has been necessary since large databases of multimedia are easily transferred and stored. Singing voice is one of the most memorable features of music. As a high-level semantic feature, its accurate detection can be difficult, especially when mixed with accompanying music. It is a problem that has been long studied since it is related to many applications of music information retrieval tasks such as singer identification, automatic lyrics transcription and alignment. Knowing where the singing voice is, becomes a preliminary step to other retrieval systems of high-level semantic features such as music genre, singing style and emotion.

To the present day many machine-learning techniques have been used for this task. Among them Gaussian mixture models(GMM) [1], [2], Support vector machines(SVM) [3], [4], Hidden markov models(HMM) [5] and more. Concerning the spectral features we find a large variety of sound descriptors that have been used to model the vocal and non-vocal signals. In [1] a series of experiments are conducted in an effort to find the most appropriate feature set for this task. It is concluded that MFCC are the best performing descriptors along with their delta coefficients among other descriptors such as spectral roll-off, spectral flatness, Linear and Perceptual Predictive coefficients(LPC, PLP) and others.

A step that is strongly being discussed here is the temporal smoothing in similar systems. The windows, used for frame-cutting the audio signals, can vary in length but are usually selected to be from 15ms to 40ms long. During such period of time the singing voice can be often interrupted by respiratory sounds and small pauses that cannot be modeled accurately from the feature set. Frame-based classification is bound to fall into errors due to over-segmentation. This has been acknowledged since various existing approaches intent to filter this noisy classification. To overcome this problem in [4] it is proposed the selection of larger portions of sound with length of 190ms and finding the maximum likelihood from the mean of the log-likelihood over the frames in a portion. Similarly, filtering the results with a median filter for each consecutive 30 frames, was able to add a 10% in the overall accuracy in Ramona [3]. In the same article, another post-processing method is proposed. A HMM system is used to model the out-coming results from an SVM classification. The Viterbi decoding is able to significantly augment the accuracy. As disadvantages, the median filtering still shows erroneous results especially in the vocal regions, while the HMM-modeling misclassifies the borders between vocal and pure instrumental parts, while still achieving better accuracy levels.

The need for morphologically coherent segmentation of the songs is first investigated in Nwe [5], where he achieves high accuracy by integrating the songs structure, tempo and loudness in a Multi-model HMM classifier. We also refer to [7] where a segmentation based on spectral change relies in the assumption that voice is bound to bring a big change in the spectrum.

II. System Description

Our model follows the standard methodology, found in state-of-the-art approaches, which includes the training of a classifier for the vocal and pure instrumental parts (PI) of a dataset by a set of spectral features extracted on every consecutive frame and the classification of a test set, which is done by scoring each feature vector according to the classifier's ability to discriminate the two classes. As discussed earlier, the frame-based nature of such experiments results in over-segmentation of the excerpts, something that is not representative of the "human-like" segmentation that it is aimed and can permit pauses sometimes as long as 1 second [6]. By observing the hand-made labelling, it is sometimes trivial to decide where the real boundaries are, especially when the voice is dying away in a continuous instrumental accompaniment [8]. As shown in

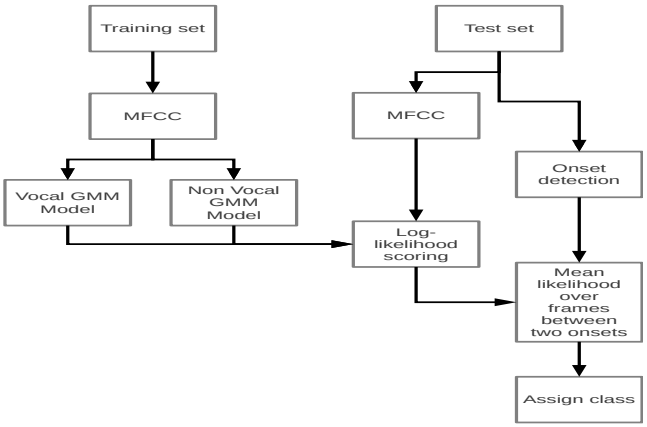

Fig. 1. Block diagram of the proposed model.

the introduction, smoothing the posterior-probability produced from the statistical training can bring great improvement in this type of systems, but it is usually performed without considering any content information. The block-diagram of the model proposed is shown in Fig. 1.

A. Features

The MFCCs are commonly used as feature set for speech recognition systems. They are a cepstral representation of the signal mapped onto the mel scale, which is a perceptual scale of pitches. Its computation is done by mapping the powers of the Fourier spectrum of a signal onto the mel scale, using triangular overlapping windows. The resulting amplitudes are first logged and its discrete cosine transform (DCT) is computed. The computation of the coefficients is made on a frame basis. They are computed each 10ms using overlapping 30ms hamming windows. Here, the MFCC consists of 13 coefficients that were calculated from 22 Mel scale bands excluding the zero coefficient, considering only the frequencies from 40Hz to 5000Hz, where the human voice is normally found. In the feature set the first and second delta derivatives of the MFCCs are added. The final feature vector is of 39 components.

B. Classifier

A Gaussian Mixture Model classifier is chosen for the discrimination of vocal and pure instrumental frames task. Diagonal covariance matrices of the 13 components of MFCC were used and the Expectation Maximization (EM) is used for training mixture models of 512 gaussians for the vocal and pure instrumental model respectively. The number of gaussians is dictated as best performing after preliminary experiments. The frame classification is done by scoring each feature vector in the test set for each of the two models "voice" and "pure instrumental".

C. Post-Processing

The temporal smoothing proposed here is based on the segmentation of each song according to peak locations detected from an onset detection algorithm, so that our segmentation consists with the musical events found in each piece. For the experiments held here, an onset detection procedure using the complex domain function is used [9]. The idea here is to introduce the temporal morphology of the song as an actual feature extracted directly from the audio rather than to smooth the results in an arbitrary way, just by assuming larger-than-a-frame portions of audio. This idea is promising for the singing voice detection since a voice more likely joins the accompaniment at beat times [7]. After preliminary experiments we tuned the onset detection with parameters -80dB as silence threshold and 0.46 peak picking, which sets the threshold for choosing the peaks in the complex domain function which will be considered as onsets [9]. We must state here that the *a priori* goal of this onset-detection procedure is not to detect all the onsets in the songs with maximum accuracy rather than to obtain a musically coherent representation of the structure of the pieces in order to segment them according to the events onsets detected. For the purpose of this experiment the accuracy of the onset-detection is not evaluated, although it is assumed to be fairly good as it achieves 95% accuracy in an onset detection task as reported in [9].

The final automated labelling is done as follows. For each song a segment of length L is accepted as vocal only if equation 1 is true, elsewhere it is assigned as pure instrumental (PI).

$$\sum_{i=1}^{L} \log p(c_i|\theta_{voice}) > \sum_{i=1}^{L} \log p(c_i|\theta_{PI}) \qquad (1)$$

, where L denotes the length in frames in-between two onsets and c is the feature vector of a single frame and $\theta_{voice}, \theta_{PI}$ are the two trained models.

III. Experiments

A. Dataset

The dataset used here is a freely distributed music library from [3]. This dataset is initially collected for the same task of singing detection. The ground truth annotation is made manually and the library is available for download. For the purpose of this experiment the library is converted into raw mono audio sampled at 16KHz by using the sum of channels. The training set and the test set consist of 60 and 16 songs respectively. Both contain popular tracks of various musical genres and have an overall equilibrated ratio of vocal, 49.61% and pure intrumental parts, 51.39%. A capella music is not included in this dataset.

B. Results

F-measure and accuracy levels are computed to evaluate the success of the classifier of the proposed method. The overall results for the test set are summarized in the Table I. It is seen that the raw frame accuracy from the GMM modelling is

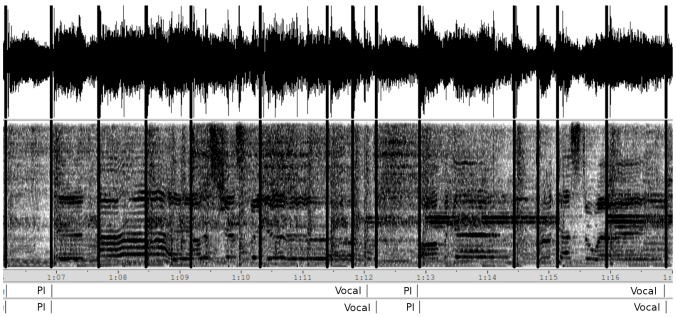

Fig. 2. Demonstration of the onset segmentation and the automated annotation from the song castaway. The onsets are marked as black lines in the waveform. Ground truth labels are found in the first row and the automated labelling is seen in the second one.

relatively low compared to the literature, 67.64%, something that is expected by our reduced feature set. Nevertheless, our onset-based segmentation can help in the discrimination between vocal and pure instrumental parts, as it brought a large amount of corrected instances, +16 percent points. We note that the classifier's ability to discriminate between the two classes is biased towards the vocal class. This drawback is observed before and after the post-process where a lower classification on the pure instrumental segments is obtained. This is also probably due to the feature set chosen which performs well for the modeling of vocal signals but not for the pure instrumental ones. On the other hand, a high accuracy level for the vocal class is obtained reaching 91% of correctly classified instances.

TABLE I
FRAME ACCURACY AND F-MEASURE. RAW SHOWS THE FRAME CLASSIFICATION BEFORE THE POST-PROCESSING STEP. F-MEASURE IS CALCULATED FOR BOTH VOCAL AND PI CLASSES. RESULTS FROM RAMONA AND REGNIER [3], [6], WHEREVER WERE AVAILABLE, ARE PRESENTED HERE FOR COMPARISON.

Class/Method	Vocal	PI	Overall	F-measure
Raw	70.48	64.8	67.6	67.2
Ramona (HMM)	80.9	84.0	82.2	83.3
Regnier	-	-	-	77.4
model proposed	91.6	75.1	83.3	84.5

C. Evaluation

The dataset has already been used for the same task which permits a direct comparison to our results. We see that our system achieves 83.8% accuracy, which is comparable to the 82.2% obtained in [3]. In Fig. 2 we observe the success of our system to overcome misclassification in the borders while also smoothing out errors in-between. Regnier, although proposes a different approach to the problem, reports results from a similar machine-learning method for the same dataset. We must also note the simplicity of our method, which uses less sophisticated techniques for training and annotation parts, as it does not include a large feature set or any feature selection procedure and avoids computationally expensive calculations such as HMM-modelling while still achieving a better overall performance. Table II shows a detailed comparison with the results obtained by Ramona[3] and the ones from the method proposed. In bold are noted the songs where the HMM post-processing method did out-perform the method proposed here. After listening to these tracks and also the ones that showed the lowest performance, it is observed that some do not have a strong tempo sensation and in some cases there is absence of percussive sounds, something that have affected the performance of the onset detection.

TABLE II
DETAILED F-MEASURE RESULTS. COMPARISON IS BROUGHT WITH RESULTS FROM [3]

method/song	Ramona	Onset-filtering
Say me Good Bye	85.8	91.5
School	87.3	91.8
Si Dieu	80.7	89.6
Une charogne	91.7	91.4
castaway	87.3	89.1
Believe	88.5	90.4
Healing Luna	81.6	83.6
Inside	68.2	72.8
You are	91.9	93.0
LIrlandaise	64.2	**60.0**
16 ans	84.8	**78.9**
Circons[...]	88.2	**83.5**
A Poings Fermes	92.2	**73.1**
Crepuscule	88.8	**76.9**
Dance	83.2	83.7
Elles disent	78.7	87.8

IV. DISCUSSION

Conventional singing detection methods do not take into account any content information from the song other than spectral, even though singing, especially in popular music, is bound to start and/or stop at beat times. We presented here a method that takes advantage of the temporal structure to detect the vocal regions within a song. The results show that such information can lead to a better performance, especially when a song has strong rhythmic patterns. We intent to further evaluate these assumptions with ground truth data for the tempo and morphological structure of the songs to verify the amount of significant information they can bring in order to tackle this task. Future work can also be dedicated to further improve the above system by the use of more efficient feature sets able to better model both vocal and instrumental parts.

REFERENCES

[1] M. Rocamora and P. Herrera, Comparing audio descriptors for singing voice detection in music audio files, in Brazilian Symposium on Computer Music 11th, 2007, pp. 27.

[2] L. Regnier and G. Peeters, Partial clustering using a time-varying frequency model for singing voice detection, in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 441444.

[3] M. Ramona, G. Richard, and B. David, Vocal detection in music with support vector machines, in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 2008, pp. 18851888.

[4] S. Vembu, S. Baumann, Separation of vocals from polyphonic audio recordings, in Proc. ISMIR, 2005.

[5] T. L. Nwe, A. Shenoy, and Y. Wang, Singing voice detection in popular music, in Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004, pp. 324327.

[6] L. Regnier and G. Peeters, Singing voice detection in music tracks using direct voice vibrato detection, in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009, pp. 16851688.

[7] Yipeng Li and DeLiang Wang, Separation of Singing Voice From Music Accompaniment for Monaural Recordings, Audio, Speech, and Language Processing, IEEE Transactions on DOI, vol. 15, no. 4, pp. 14751487, 2007.

[8] A. Holzapfel and Y. Stylianou, Singer identification in rembetiko music, Proc. SMC, vol. 7, pp. 2326, 2007.

[9] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, Complex domain onset detection for musical signals, in Proc. 6th Conf. Digital Audio Effect (DAFx-03), London, U.K., 2003.