VOCAL DETECTION IN MUSIC WITH SUPPORT VECTOR MACHINES

Mathieu Ramona

RTL (Ediradio)
22 rue Bayard, 75008 Paris, France

G. Richard, B. David

TELECOM ParisTech / LTCI-CNRS
37 rue Dareau, 75014 Paris, France

ABSTRACT

We propose a statistical learning approach for the automatic detection of vocal regions in a polyphonic musical signal. A support vector model, based on a large feature set, is employed to discriminate accompanied singing voice from pure instrumental regions. We propose a temporal smoothing of the posterior probabilities with a hidden Markov model that helps adapting the segmentation sequence to the precision of the manual annotation. Quantitative results on a copyright-free public musical corpus show a classification accuracy of 82%.

Index Terms— Vocal detection, Support Vector Machines, Hidden Markov Models

1. INTRODUCTION

Most of the western popular music consists in a leading singing voice accompanied by background music. In most cases, the singer is only active part of time leading to an alternation between purely instrumental sections and singing voice sections. The automatic detection of the singing voice regions is an essential step for a number of applications including singer automatic identification [1], singing voice separation [2] or query-by-humming. Another useful commercial application arises from the need of the broadcast radio stations to locate the point where singing starts and ends in a song, since the speaker usually fills on air the instrumental introduction and outing.

Even though this problem shares some characteristics with the speech/music discrimination task, it is more complex for several reasons and specific approaches need to be developed. In fact, the singing voice covers a much wider range of intrinsic variations than speech both in term of timbre and fundamental frequencies. It is also often highly correlated with the corresponding background music with therefore strong and durable overlaps on its frequency components.

Furthermore, the variety of artists and instruments complicates the exhaustive characterization of both singing voice and instrumental music.

The issue of locating singing voice regions in musical songs has already been addressed following traditional statistical approaches applied on widely-used speech features. For example, Gaussian Mixture Models ([1]), Neural networks and SVM ([2], [3]) or Hidden Markov Model [4] were used. However, these studies are based on a limited number of features, traditionally used in speech recognition systems (MFCC or/and PLP) and which may not be appropriate to capture the characteristics of the singing voice in background music as noted by [5] and [6]. The different methods are also difficult to compare since all published results are obtained on different corpora with different evalution protocols.

Therefore, the purpose of this paper is firstly to propose a novel two-step approach for locating singing voice segments where the output of a support vector machine (e.g. the posterior probabilities) are further processed to obtain smooth decision functions. Secondly, to cope with the high variabilities of the sources, our system integrates a much wider feature set than the previous studies. Finally, in order to permit future comparison with our results, we provide a direct link to the public music database used and provide all annotations and evaluation protocols.

The paper is organized as follows. An overview of the system is provided in section 2. Then, a brief presentation of the feature set used is given in section 2.1 while the classification scheme and the temporal smoothing are described in section 2.2. The experimental protocol and the results are then presented and discussed in section 3. Finally, some conclusions are proposed in section 4.

2. SYSTEM OVERVIEW

The architecture of the proposed system follows a traditionnal bag-of-frames approach where a machine learning technique (Support Vector Machine) is applied on a set of features computed on successives frames of the incoming music signal. The ouput of the classifier is then further processed in order to obtain smooth decision function to localise musical segments that contain singing voice. The different building blocks of our system are described in some details below.

2.1. Feature extraction

The audio signal is first segmented in overlapping frames of 32 ms with a 16 ms overlap. On each frame a FFT is computed with a Hamming window. Most of the features chosen were elected for their ability to discriminate speech from music in a previous study [7].

Most of the features were computed on the short-scale frames stated above. Those include spectral descriptors such as centroid, width, asymmetry, slope, decreasing, flux and similar temporal statistical moments, along with their first and second derivates. 13-order Mel-Frequency Ceptral Coefficients (with their derivates), Linear Predictive Coding Coefficients from a second order Auto-Regressive analysis of the signal and the Zero Crossing Rate are also included. The Sharpness and Spread are defined from the perceptual Loudness are also part of our feature set. Finally, we have used the monophonic F0 frequency and Aperiodicity measure extracted with the YIN library [8].
Some additional features that do not represent an instantaneous behaviour of the signal are computed on long-scale frames of 1s with overlap of 0.5s. They are repeated on short-scale frames in order to be used with the previous descriptors in a common feature vector. Those include amplitude modulation descriptors (4 features represent the maximal peak amplitude and frequency along with their product and the ratio of the peak amplitude on the mean amplitude, for both frequency bands of 4-8Hz and 10-40Hz, characterizing tremolo and granularity), temporal statistical moments computed on long-scale frames and on their estimated envelope, and the ZCR computed on long-scale frames as well.

The raw feature vector has 116 components. We have used the IRMFSP algorithm to sort the features according to their ability of discriminating the two classes (Pure instrumental and Singing voice with instrumental background). The IRMFSP algorithm was originally proposed for a musical instruments classification system based on a large feature set [9]. We then feed the Support Vector Machine with the most discriminating features found on the training set.

2.2. Classification

After a simple silence detection process, based on heuristic rules on the frames energy sequence, the remaining frames are discriminated using a one-vs-one Support Vector Machine with Radial Basis Function kernel. Only the pure instrumental (PI) and singing voice with instrumental background (VI) regions are kept for classification. There is no pure singing voice in our experimental set, since the challenge here lies in the presence of a largely overlapping musical background, as explained earlier. Spoken or rap voice regions are discarded for this experiment.

If $(\boldsymbol{x_i})$ is the collection of support vectors and $\boldsymbol{k}$ the kernel function, the decision function f for a vector $\boldsymbol{x}$ has the following expression :

$$f(\boldsymbol{x}) = \sum_i y_i \alpha_i \boldsymbol{k}(\boldsymbol{x_i}, \boldsymbol{x}) + b,$$

The output of the decision function is an unbounded value that is not a probability. A sigmoid bijection has been proposed in [10] and is now widely used to get probabilistic outputs from SVMs :

$$p(x) = \frac{1}{1 + \exp(Af(x) + B)} \quad (1)$$

The A and B are computed from the distribution of the SVM output on a specific training corpus (called validation set). The posterior probability is thus thresholded with 0.5 value for a maximum likelihood decision.

Nevertheless, as shown on figure 1(a), the posterior probability evolution for the singing voice (the other class of course has a symmetric profile), is very chaotic and results in a poor classification accuracy and the detection of many undesired segments. A first idea is to apply a simple classical smoothing with a median filter of 30 frames long (i.e. about 0.5s). The resulting measure illustrated in figure 1(b) shows better behaviour but this process does not adapt to the proper sequentially of the singing voice frames : class changes occur much more frequently in a "mainly" singing region than in music regions, if annotated precisely.

We thus propose the temporal smoothing of the resulting posterior probabilities with a Hidden Markov Model with two states (PI and SV). The observation distributions are modeled by a mixture of 5 Gaussians[1], fitted with the Expectation Maximization algorithm. The best path of states is then deduced from the SVM output sequence with the Viterbi algorithm, as shown in figure 1(c). The class sequence computed with HMM post-processing still presents misclassifications at the segment borders, but the segment sequence has a structure adapted to the ground truth annotation.

Finally, since the manual annotation necessarily encountered a precision limit, we have discarded the pure instrumental segments shorter than 0.5s, after empirical observation of the annotated segments duration.

3. EXPERIMENTAL RESULTS

3.1. Audio collection and evaluation protocol

We have collected a set of 93 songs with Creative Commons license from the Jamendo free music sharing website [11], which constitute a total of about 6 hours of music. The files are all from different artists and represent various genres from mainstream commercial music. Each file has been manually annotated by the same person with high precision, with the Transcriber free software, developed for the annotation

[1] Increasing the number of Gaussians up to 30 did not show a clear gain in performance.

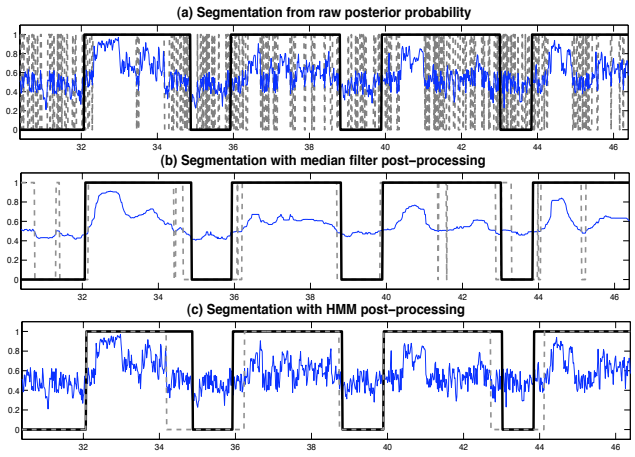

Fig. 1. The thin blue curve shows the posterior probability of class SI (Voice+Instrument) based on the SVM output. Black bold crenel shows the ground truth annotated segmentation (0 stands for PI and 1 for SI) and the dash gray crenel shows the estimated segmentation.

of speech transcription [12]. All the audio material is described and available at http://www.enst.fr/~ramona/icassp08/ along with the annotation files. The jamendo audio files are coded in stereo Vorbis OGG 44.1kHz with 112KB/s bitrate. The files are converted to mono and downsampled to 16kHz in this experiment.

The files are divided into three non-overlapping sets, a training set, a validation set and a testing set, composed respectively of 61, 16 and 16 songs. The validation set is used to tune the parameters after the SVM training on an unknown audio material, in order to simulate the behaviour on the test set, on which the final result is measured. Empirical experiments have shown that the increase of the training data set volume does not have a noticeble impact on the performances over a certain limit. A random subset of 20000 features vectors from the training set have thus been used for computational reasons. In each experiment, the sigmoid (equation 1) is fitted on the results of the classification of the validation set with the trained model. The HMM probability density functions are also evaluated from the posterior probabilities computed on the validation set. However, the rest of the HMM model (initial state distribution π and state transition probability A) is estimated from the annotation informations of the training set.

The classification accuracy is then calculated on all the frames of the test set belonging to either of the two PI and SV classes. We have also used the F-measure as a segment-based criterion.

3.2. Results

Figure 2 shows the evolution of the classification error rate with the feature vector dimension varying from 10 to 110. The σ parameter of the RBF kernel has been tuned empirically by logarithmically griding the values between 2^{-4} and 2^4, and optimizes the performance for $\sigma = 1$. We also show here the classification error rate after HMM post-processing. The error rate decreases in both cases with the feature vector dimension d, but remains quite unaffected over d = 40. The classification accuracy increases from 82.2% with $d = 40$ to 83% with $d = 110$, but obviously the first case represents a good trade-off between computational load and performances. In the remainder of the paper, all result are then given with feature vectors of dimension d=40.

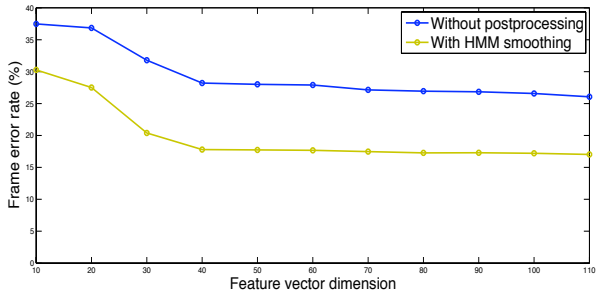

Fig. 2. Frame classification error rate.

Table 1 summarizes the classification accuracy and F-measure after various post-processings on the sequence of posterior probabilities. We show the raw performances and the results after HMM post-processing, along with the results after median-filtering, or following a segment-decision scheme, inspired from [1], on the segments bounded by the onset detection algorithm described in [13]. We also show, as a comparison, the results computed using our own implementation of the algorithm described in [2] (designated as "Alternate"), implementing a SVM applied on a 38 components classical speech processing feature vector (PLP, LFPC and MFCC), with $C = 2^8$ and $\sigma = 2^4$. The decisions are based on segments of about 190ms, after the averaging of the feature vectors on a segment.

Our system reaches 71.8% without post-processing, which is significantly above the 62.7% obtained with the alternate algorithm, with a higher temporal resolution. This demonstrates the relevance of our feature set, compared to more straightforward descriptors. The HMM post-processing proposed shows better performances than the other post-processings, with a frame accuracy of 82.2% and an F-measure of 83.2. We observe that the homogeneous segment-based decision scheme performs much better on the Singing voice detection. The alternate algorithm has the same bias on the SI class, reaching 87.7% of good classification with only 35.8% on the PI class.

Table 2 shows the detailed results on each audio file of the test set, with the HMM post-processing. It is clear first, that our system shows greater efficiency in locating the pure instrumental regions than singing voice. However some files (highlighted in the table) show very poor classification of the

Class	Sing+Instr		Pure Instr		ALL	
Criterion	fr%	F	fr%	F	fr%	F
Raw	74.8	72.6	68.5	70.4	71.8	71.6
Alternate	87.7	70.5	35.8	47.4	62.7	62.4
Median filter	84.6	81.6	76.4	80.5	80.7	81.1
Segment based	88.0	84.8	74.0	80.3	81.3	82.7
HMM	80.9	84.4	84.0	82.0	82.2	83.2

Table 1. Frame classification accuracy (fr%), and F-measure (F) on the audio test set without post-processing, and with the various post-processing schemes proposed. The *Alternate* line brings a comparison to the results of the alternate algorithm [2] on our test set.

instrumental regions, compared to the other files average. After audition of the misclassified regions on these files, we have noted that they all contain an instrument that has a somewhat similar timbre than singing voice. This shows the limitations of our modeling of the instrumental class, that does not sufficiently take into account the richness and variability of the music background. We have tried, with no success, to boost our algorithm by training another classifier on each song of the test set, fed with the frames with the most confident result (highest posterior probability), but the poor modeling of a class cannot be corrected with test-based boosting. We believe that the feature set proposed here, although better than other approaches, is still the main limit of our approach.

Class	Sing+Instr		Pure Instr		ALL	
Criterion	fr%	F%	fr%	F%	fr%	F%
03 - Say me Good Bye.wav	77.0	85.8	85.6	76.7	80.1	82.3
03 - School.wav	76.2	87.3	94.9	84.1	84.3	85.7
03 - Si Dieu.wav	66.1	80.7	97.4	72.6	76.4	77.3
03 - Une charogne.wav	87.1	91.7	79.4	72.3	85.3	86.8
03 - castaway.wav	94.4	87.3	**55.1**	73.4	79.0	82.8
04 - Believe.wav	94.4	88.5	**52.3**	65.2	80.0	82.3
04 - Healing Luna.wav	70.7	81.6	96.8	86.1	85.5	84.4
04 - Inside.wav	76.5	68.2	85.3	87.6	83.3	82.9
04 - You are.wav	86.6	91.9	87.7	79.0	87.0	87.0
05 - 05 LIrlandaise.wav	93.8	64.2	**33.4**	47.5	57.7	57.1
05 - 16 ans.wav	69.3	84.8	99.6	94.7	91.5	92.7
05 - 2003-Circons[...].wav	85.5	88.2	89.2	91.2	87.6	89.9
05 - A Poings Fermes.wav	84.5	92.2	98.7	95.7	93.7	94.6
05 - Crepuscule.wav	81.6	88.8	89.3	85.6	85.2	87.2
05 - Dance.wav	75.3	83.2	81.4	61.5	77.0	75.8
05 - Elles disent.wav	76.1	78.7	**63.6**	59.9	71.8	72.1
ALL	80.9	84.3	83.6	81.8	82.2	83.1

Table 2. Details results on each file of the test set

4. CONCLUSION

In this paper we have proposed a statistical learning approach for detecting singing voice regions in monaural polyphonic music. The use of a two state HMM on the posterior probabilities calculated from the SVM output allows us to take into account the temporal structure of the annotated segments and thus adapt properly the estimated class sequence. Quantitative experiments show a clear increase of the overall performances and our system reaches over 82% of frame classification accuracy. Comparison to a similar algorithm based on a classical speech processing feature set shows the relevance of our feature set. However, examination of the test audio files automatic annotation reveals that on a few audio files, one of the musical instruments is almost always mistaken for singing voice. This problem reveals the need for a more adapted set of features that better discriminate musical singing voice. Other future research directions include the coupling of source separation algorithm to enhance the singing voice signal and to tackle the more general case of stereo signals. Finally, specific efforts will be dedicated to the modeling of the background music and to its efficient adaptation to the test data.

References

[1] W. Tsai and H. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *IEEE Trans. on ASLP*, vol. 14 (1), 2006.

[2] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. ISMIR*, 2005.

[3] T. Leung, C. Ngo, and R.W.H. Lau, "Ica-fx features for classification of singing voice and instrumental sound," in *Proc. ICPR*, 2004, vol. 2.

[4] A. Berenzweig and D.P.W. Ellis, "Locating singing voice segments within music signals," in *Proc. WASPAA*, 2001.

[5] T.L.Nwe, A. Shenoy, and Y. Wang, "Singing voice detection in popular music," in *Proc. ACM Multimedia*, 2004.

[6] N.C. Maddage, K. Wan, C. Xu, and Y. Wang, "Singing voice detection using twice-iterated composite fourier transform," in *Proc. ICME*, 2004, vol. 2.

[7] G. Richard, M. Ramona, and S. Essid, "Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams," in *Proc. ICASSP*, 2007, vol. 2, pp. 461–464.

[8] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Proc. JASA*, vol. 111(4), pp. 1917–1930, 2002.

[9] G. Peeters and X. Rodet, "Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instrument database," in *Proc. DAFX*, 2003.

[10] J.C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999.

[11] http://www.jamendo.com, "Jamendo, open your ears," .

[12] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Proc. Speech Com*, vol. 33 (1-2), 2000.

[13] C. Duxbury, J.P. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *Proc. DAFX*, 2003.