Singing Voice Detection in Popular Music

Tin Lay Nwe* Arun Shenoy Ye Wang

Department of Computer Science, School of Computing
National University of Singapore, Singapore 117543
{nwetl, arunshen, wangye}@comp.nus.edu.sg

ABSTRACT

We propose a novel technique for the automatic classification of vocal and non-vocal regions in an acoustic musical signal. Our technique uses a combination of harmonic content attenuation using higher level musical knowledge of key followed by sub-band energy processing to obtain features from the musical audio signal. We employ a Multi-Model Hidden Markov Model (MM-HMM) classifier for vocal and non-vocal classification that utilizes song structure information to create multiple models as opposed to conventional HMM training methods that employ only one model for each class. A statistical hypothesis testing approach followed by an automatic bootstrapping process is employed to further improve the accuracy of classification. An experimental evaluation on a database of 20 popular songs shows the validity of the proposed approach with an average classification accuracy of 86.7%

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing - Methodologies and Techniques; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Design, Experimentation

1. INTRODUCTION

The singing voice is one of the most important characteristics of music [13]. With the immense and growing body of music data, information on the singing voice could be used as a valuable tool for the automatic analysis of song content in the field of music information retrieval and many other applications. The problem of singing voice detection can be stated as follows: for a given song, classify each segment as being of either the pure instrumental type (referred to as a non-vocal segment in the rest of this paper) or as a mixture of vocals with/without background instrumental (referred to as the vocal segment in the rest of this paper).

In [2], Berenzweig and Ellis used Posterior Probability Features (PPF) obtained from the acoustic classifier of a general-purpose speech recognizer to derive a variety of statistics and models which allowed them to train a vocal detection system. In [3], Berenzweig et al. used a multi-layer perceptron (MLP) neural network to segment songs into vocal and non-vocal regions using perceptual linear prediction (PLP) coefficients based features. Tzanetakis [13] developed a technique to detect the presence of singing voice using a bootstrapping process that trains a different classifier for each song. This technique is semi-automated requiring a small random sampling of every song to be annotated by the user for training. Zhang [15] used the features of energy, average zero-crossing rate (ZCR), harmonic coefficient and spectral flux computed at regular intervals and compared with a set of predetermined thresholds to detect the start of the singing voice. Chou and Gu [4] have proposed a technique using a combination of harmonic coefficient based features, conventional Mel-frequency Cepstral Coefficients (MFCC) and log energy features in a GMM-based Speech/Music Discriminator (SMD) system to detect the singing voice. Kim and Whitman [6] have proposed a technique to detect the singing voice based on an analysis of the energy within the frequencies bounded by the range of vocal energy. This has been achieved using a combination of an IIR filter and an inverse comb filter bank. In [8], Maddage et al. have developed an SVM based classification approach to detect singing voice using the musical audio features of Linear Prediction Coefficients (LPC), LPC derived cepstrum (LPCC), MFCC, spectral power (SP), short time energy (STE) and ZCR. In [7], Maddage et al. have proposed a Twice-Iterated Composite Fourier Transform (TICFT) technique to detect the singing voice boundaries by showing that the cumulative TICFT energy in the lower coefficients is capable of differentiating the harmonic structure of vocal and instrumental music in higher octaves.

The technique presented in this paper is based on the observation that popular songs usually have a structure comprising of intro, verse, chorus, bridge and outro, and different sections display differences in characteristics [14]. Therefore, statistical models of vocal and non-vocal classes should be built based on the structure of the song. Towards this

*Tin Lay Nwe is now with the *Institute for Infocomm Research* (I^2R), Singapore.

MM'04, October 10-16, 2004, New York, New York, USA.

end, we employ a multi-model-HMM (MM-HMM) training approach to tackle the intra-song and inter-song variations for improved vocal and non-vocal classification performance. This is followed by a statistical hypothesis testing method and bootstrapping technique to further increase accuracy. This proposed technique uses acoustic features which are a combination of harmonic content attenuation using higher level musical knowledge of key followed by sub-band energy processing that we have found suitable to distinguish vocal and non-vocal signals.

We assume the meter to be 4/4, this being the most frequent meter of popular songs and the tempo of the input song is assumed to be constrained between 40-185 beats per minute (BPM) and almost constant. The audio signal is framed into beat-length segments to extract metadata in the form of quarter note detection of the music [11]. The basis for this technique of audio framing is that within the quarter note, the harmonic structure of the music can be considered as quasi-stationary. This is based on the premise that musical changes are more likely to occur on beat times in accordance with the rhythm structure than on other positions. For a comprehensive description of each segment, we extract acoustic features at 20 ms frame length intervals (13 ms frame overlap) within the inter-beat interval and group all of these together. Each frame is multiplied with a hamming window to minimize signal discontinuities at its ends.

The rest of this paper is organized as follows. The process of feature extraction from the audio signal is presented in Section 2. Details of the classifier formulation which includes the MM-HMM classifier, the classification decision verification using statistical hypothesis testing and the bootstrapping process is discussed in Section 3. We present the empirical evaluation of our approach in Section 4. Section 5 concludes the paper.

2. ACOUSTIC FEATURES

Both, musical instrument sounds and the human singing voice are rich in harmonic content. However, we observe that the sound signals produced by instruments have more regular harmonic patterns compared to the singing voice. Figures 1(a) and 2(a) show the frequency content of non-vocal and vocal signals respectively for the first 1 kHz of audible range. A more regular harmonic pattern is observed for non-vocal signals as compared to the vocal signals. Therefore, we use the technique of harmonic content attenuation in our approach to be able to distinguish better between vocal and non-vocal regions. Further, the spectral characteristics of vocal and non-vocal segments are different. If vocals begin while instrumental is going on, a sudden increase in the energy level of the audio signal is observed [15]. Thus, we follow up our harmonic attenuation with an analysis of energy distribution in different frequency bands. We shall now discuss the implementation of harmonic attenuation followed by energy analysis.

2.1 Harmonic Attenuation

In [11], we have demonstrated a system to determine the key of acoustical musical signals. The key defines the diatonic scale that a piece of music uses. The diatonic scale is a seven note scale and is more familiar as the Major/Minor scale in music. Since every song is in a certain key, we use this information to attenuate only those harmonic patterns originating from the pitch notes in the key. To reduce the complexity in implementation, we have used triangular bandpass filters. This filter has the highest attenuation of the signals at regular harmonic frequencies and the least attenuation of the signals at irregular harmonic frequencies (those that are highly deviated from regular harmonic frequency locations). The more the deviation occurs, the less the signal is attenuated. After attenuation, non-vocal signals have lower energy content than vocal signals as shown in Figures 1(c) and 2(c) respectively.

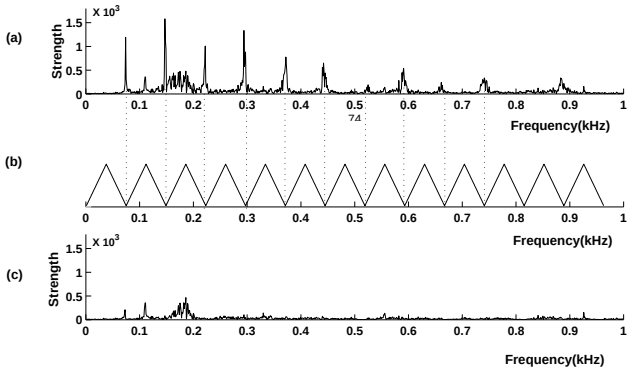

Figure 1: (a)Non-vocal signal in frequency domain (b) Frequency response of triangular bandpass filter (c) Non-Vocal signal after attenuating harmonic content

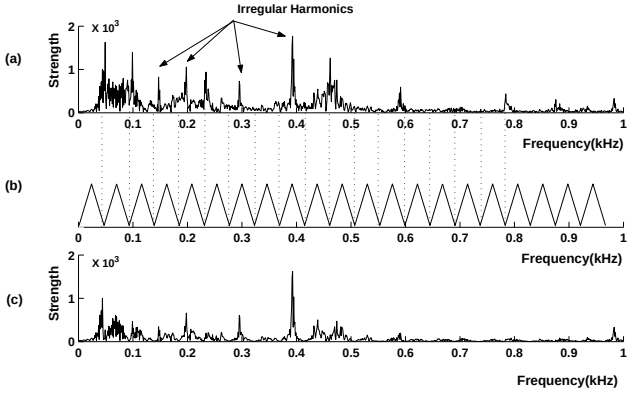

Figure 2: (a)Vocal signal in frequency domain (b) Frequency response of triangular bandpass filter (c) Vocal signal after attenuating harmonic content

2.2 Energy Distribution Analysis

After attenuating the harmonic content, each audio frame is passed through bandpass filters spaced logarithmically from 130 Hz to 16 kHz. Sub-band based Harmonic Attenuated Log Frequency Power Coefficients (HA-LFPC) are then computed using Equations (1) and (2) which we have defined previously for LFPC calculation in [9].

$$S_t(m) = \sum_{k=f_m-\frac{bm}{2}}^{f_m+\frac{bm}{2}} X_t(k)^2, m = 1...12 \qquad (1)$$

where, $X_t(k)$ is the k^{th} spectral component of the signal, t is

the frame number, $S_t(m)$ is the output of the m^{th} subband, and f_m and b_m are the center frequency and bandwidth of the m^{th} subband respectively. The HA-LFPC parameters which provide an indication of energy distribution among subbands are calculated as follows:

$$HA - LFPC_t(m) = 10log_{10}[\frac{S_t(m)}{N_m}] \quad (2)$$

where N_m is the number of spectral components in the m^{th} subband. For each frame, 12 HA-LFPCs are obtained. Figure 3 shows the energy distribution obtained using HA-LFPC for 90 second vocal and non-vocal segments extracted from six different songs.

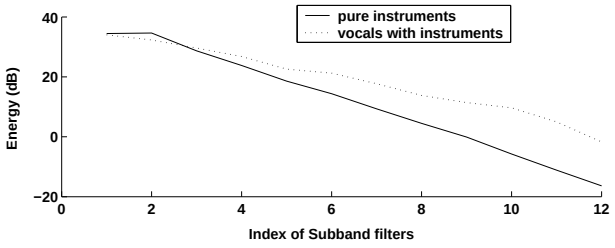

Figure 3: Energy distribution for vocal/non-vocal segments using HA-LFPC over 12 subband filters

The figure shows that the vocal segments have relatively higher energy values in the higher frequency bands as compared to the non-vocal segments. Therefore, HA-LFPC is an effective feature for the discrimination of vocal and non-vocal segments in subsequent steps.

3. CLASSIFER FORMULATION

3.1 Multi-Model HMM Classifier

Most studies on vocals detection use statistical pattern classifiers [2, 3, 13]. However, to our knowledge, none of the studies takes into account song structure information in song modeling. An important observation is that vocal and non-vocal segments display variation in intra-song signal characteristics. For example, signal strengths in different sections (intro, verse, chorus, bridge and outro) are usually different. In our observation, for most songs, the signal strength of the intro is relatively low compared to that of the verse or the chorus. Chorus sections are usually of stronger signal strength in comparison with the verse and bridge sections since they have a "fuller" musical arrangement with busier drums, some additional percussion, a fuller string arrangement and additional melody lines [14]. The outro section might repeat a vocal phrase from the chorus and it usually ends with a fade-out. Sample waveforms extracted from different sections of a popular song are depicted in Figure 4.

Tempo and loudness are important attributes accounting for inter-song variation. Therefore, we integrate the song structure, inter-song and intra-song variation into our models. The training data (vocal or non-vocal segments) are manually classified based on three parameters - the section type (intro, verse, chorus, bridge and outro), tempo and loudness, and a model is created for each class as shown in Figure 5. In our current implementation, we use 40 models - 20 each for modelling vocal and non-vocal segments. This process results in multiple HMM models for each vocal and non-vocal class. Several models for each class form an HMM model space, to allow more accurate modeling in comparison to the single-model baseline.

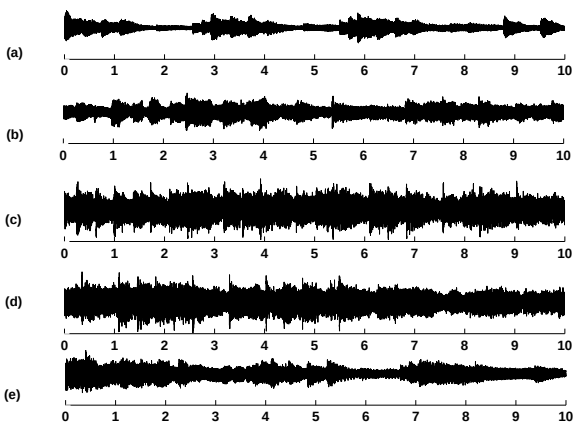

Figure 4: Waveforms of 10 second segments extracted from (a)Intro (b)Verse (c)Chorus (d)Bridge (e)Outro/ending of the song *25 Minutes* by *MLTR*

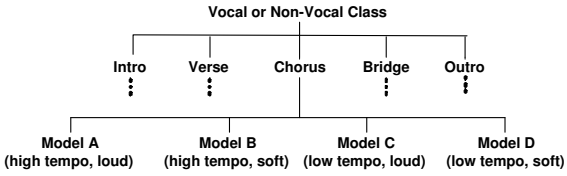

Figure 5: Multiple HMM models for each vocal and non-vocal class

3.2 Classification Decision Verification

The test song is segmented using the above mentioned MM-HMM classifier. However, some of the vocal and non-vocal segments might be wrongly classified. Therefore, in this step, we evaluate the reliability of the classification decision. The most effective way to measure the confidence of the classification decision is based on how much the classification decision significantly overtakes the other possible competitors. We use the neighborhood information in HMM model space discussed in [5] to determine the properties of the possible competing source distribution of the target model. Hypothesis test [12] is then performed for each segment of audio to obtain a confidence score for its current classification as obtained from the MM-HMM. This confidence score is compared with a predetermined threshold to retain only the frames that have a high confidence of being classified as either vocal or non-vocal segments.

3.3 Bootstrapping Process

The frames with high confidence score that are retained in the previous step, are used to build song-specific vocal and non-vocal models of the test song with a bootstrapping process [13] to further improve accuracy. This process allows us to use a song's own model for classification as shown in Figure 6. The bootstrapping process makes the algorithm adaptive and capable of achieving high vocal detection accuracy.

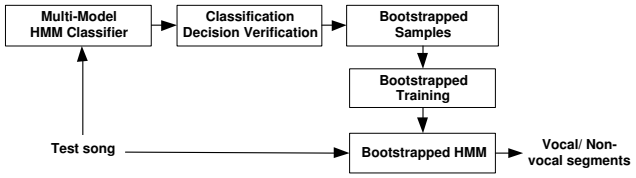

Figure 6: Bootstrapped training and segmentation

4. EXPERIMENTS

Our experimental database includes 20 popular songs, carefully selected for their variety in artist and time spans. Of these, 6 songs are allocated to the training dataset and the remaining 14 songs to the test dataset. There is no overlap between the two datasets. We use the continuous density HMM with four states and two Gaussian mixtures per state for all HMM models in all our experiments. The standard procedures for HMM training and classification are well documented in [10]. Using the training database, the MM-HMM classifier is trained to obtain several variants of the vocal and non-vocal HMM models which are shown in Figure 5. The results of our analysis are tabulated in Table 1 below. To compare the vocal detection performance of HA-LFPC features with traditional features, experiments are conducted using simple LFPC (without harmonic attenuation) [9], MFCC [1] and LPCC [10].

Table 1: Experimental Results (% accuracy)

Feature	MM-HMM			Bootstrapped HMM		
	Non Vocal	Vocal	Avg	Non Vocal	Vocal	Avg
HA-LFPC	80.5	85.8	83.1	79.2	94.1	86.7
LFPC	81.2	83.2	82.2	78.2	91.9	85.1
MFCC	79.3	77.5	78.4	77.2	85.4	81.3
LPCC	75.6	78.9	77.3	73.4	86.6	80.0

It can be seen that the HA-LFPC feature, with an average accuracy of 86.7%, outperforms all the traditional features. The usage of the bootstrapping technique gives us a 3.6% increase in performance (83.1% to 86.7%) over the MM-HMM. It is observed that performing the harmonic attenuation gives 1.6% improvement in performance over simple LFPC (85.1% to 86.7%). Next, we investigate the effectiveness of employing song structure information in song modeling. We compare the performance of the MM-HMM against the baseline HMM training method in which only one model is created for each vocal and non-vocal class with no regard for song structure information. The results presented in Table 2 show that the MM-HMM training method outperforms the baseline HMM training approach by 3% (80.1% to 83.1%).

Table 2: Performance comparison between Baseline HMM and MM-HMM (% accuracy)

Feature	Baseline HMM			MM-HMM		
	Non Vocal	Vocal	Avg	Non Vocal	Vocal	Avg
HA-LFPC	77.6	82.6	80.1	80.5	85.8	83.1

5. CONCLUSION

We have presented an automatic approach for detecting vocal segments in a song. Using a combination of harmonic attenuation based on musical knowledge of key, MM-HMM, hypothesis testing and bootstrapping, our system is able to garner a net performance of 86.7% accuracy in vocal/non-vocal classification. One drawback of this framework is that it is computationally expensive since it entails two training steps: training the MM-HMM classifier and the bootstrapped classifier. To reduce the computational complexity, we could discard bootstrapping process in favor of using a more powerful hypothesis testing approach proposed in [5] instead. The current hypothesis testing method [12] removes a few correctly labeled segments in addition to the wrongly labeled segments. We believe that the alternate hypothesis testing method would overcome this problem. We could also consider an implementation using mixture modeling or classifiers such as neural networks or support vector machines.

6. REFERENCES

[1] C. Becchetti and L. P. Ricotti. *Speech Recognition : Theory and C++ Implementation*. John Wiley & Sons, New York, May 1999.

[2] A. Berenzweig and D. P. W. Ellis. Locating singing voice segments within music signals. In *WASPAA*, 2001.

[3] A. Berenzweig, D. P. W. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. In *AES*, 2002.

[4] W. Chou and L. Gu. Robust singing detection in speech/music discriminator design. In *ICASSP*, 2001.

[5] H. Jiang and C. H. Lee. A new approach to utterance verification based on neighborhood information in model space. *IEEE Transactions on Speech and Audio Processing*, 11(5):425–434, Sept 2003.

[6] Y. Kim and B. Whitman. Singer identification in popular music recordings using voice coding features. In *ISMIR*, 2002.

[7] N. C. Maddage, K. Wan, C. Xu, and Y. Wang. Singing voice detection using twice-iterated composite fourier transform. In *ICME*, 2004.

[8] N. C. Maddage, C. Xu, and Y. Wang. An svm-based classification approach to musical audio. In *ISMIR*, 2003.

[9] T. L. Nwe, F. S. Wei, and L. C. De-Silva. Stress classification using subband features. *IEICE Transactions on Information and Systems*, E86-D(3):565–573, March 2003.

[10] L. R. Rabiner and B. H. Juang. *Fundamentals of speech recognition*. Prentice Hall, Englewood Cliffs, N.J., 1993.

[11] A. Shenoy, R. Mohapatra, and Y. Wang. Key determination of acoustic musical signals. In *ICME*, 2004.

[12] R. A. Sukkar and C. H. Lee. Vocabulary independent discriminative utterance verification for non keyword rejection in subword based speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(6):420–429, Nov 1996.

[13] G. Tzanetakis. Song-specific bootstrapping of singing voice structure. In *ICME*, 2004.

[14] I. Waugh. Song structure. *Music tech magazine*, October 2003.

[15] T. Zhang. System and method for automatic singer identification. In *ICME*, 2003.