AUTOMATIC DETECTION OF VOCAL SEGMENTS IN POPULAR SONGS

Tin Lay Nwe[*]
School of Computing
National University of Singapore
3 Science Drive 2,Singapore 117543
tlnma@i2r.a-star.edu.sg

Ye Wang
School of Computing
National University of Singapore
3 Science Drive 2,Singapore 117543
wangye@comp.nus.edu.sg

ABSTRACT

This paper presents a technique for the automatic classification of vocal and non-vocal regions in an acoustic musical signal. The proposed technique uses acoustic features which are suitable to distinguish vocal and non-vocal signals. We employ the Hidden Markov Model (HMM) classifier for vocal and non-vocal classification. In contrast to conventional HMM training methods which employ one model for each class, we create an HMM model space (multi-model HMMs) for segmentation with improved accuracy. In addition, we employ an automatic bootstrapping process which adapts the test song's own models for better classification accuracy. Experimental evaluations conducted on a database of 20 popular music songs show the validity of the proposed approach.

1. INTRODUCTION

Rapid progress in computer and Internet technology has enabled the circulation of large amounts of music data on the Internet. With the immense and growing body of music data, automatic analysis of song content is important for music retrieval and many other applications. The singing voice (vocal) is one of the most important characteristics of music [1]. It is still a challenge to detect the vocal segments within a song automatically.

The problem of vocal detection can be stated as follows: given a song, classify each segment of the song in terms of whether it is pure instrumental (referred to as a non-vocal segment in this paper) or a mixture of vocals with/without background instrumental (referred to as the vocal segment).

The basic procedure for any vocals /non-vocals segmentation system includes the extraction of feature parameters from audio signals with a time resolution constrained by the analysis window length. Then, segments of the song are classified as vocal or non-vocal using a statistical classifier or a relatively simple threshold method.

A large number of features have been proposed to represent audio signals. Some of the methods originate from the area of speech recognition. These include Mel Frequency Cepstral Coefficients (MFCC) [1, 2], Linear Prediction Coefficients (LPC) [1, 2, 3], perceptual linear prediction coefficients [4, 5], energy function [6] and the average zero-crossing rate [6]. In addition, features that have been used in the area of music analysis are spectral flux [6], relative subband energy [1] and features that can differentiate the harmonic structure of music signals [3, 6, 7, 8]. Zhang [6] mentioned that voice signal tends to have a higher rate of change than instrumental music, and the start of vocals can be indicated by the appearance of high peaks in the spectral flux value. Kim [8] stated that the straightforward method to detect vocals is to note the energy within the frequencies bounded by the range of vocal energy. Collectively, the studies [3, 6, 7, 8] stated that features that can measure the harmonic content of the music signal are important for detecting vocals in a song. To measure harmonicity, the frequency range of 200Hz ~2 KHz is considered in [8]. The highest frequency usually considered for the analysis of normal speech is 3 kHz [9]. Due to the fact that the harmonic content of vocals is higher than normal speech [6], the frequencies which are higher than 3 kHz are important to take into consideration for vocals analysis. Based on these studies, features that can capture the harmonic content and spectral structure of the audio seem to be suitable features for vocal detection.

Early audio segmentation algorithms such as [16] are specifically designed for speech signals. These algorithms detect the several acoustic events such as speaker changing points in the audio. More recent works often use pattern classifiers such as HMM, Neural Network (NN), Support Vector Machine (SVM) for song segmentation [1, 2, 4, 5]. In those studies use pattern classifiers such as HMM, Neural Network (NN), Support Vector Machine (SVM) for song segmentation [1, 2, 4, 5]. In those studies, a statistical model for each of the vocal or non-vocal class was created using entire songs in the training data. However, popular songs usually have a structure comprising of intro, verse, chorus, bridge and outro,

[*] Tin Lay Nwe is currently with I2R, Singapore

and different sections are of different signal strengths [10]. The signal strength of the chorus section is usually much higher than the intro or verse section. Therefore, statistical models of vocal and non-vocal classes should be built based on the structure of the song. The method in [5] uses the classifier of a speech recognizer trained on normal speech to detect speech-like sounds of music. However, there are significant differences between singing and speech signals. Since singing voice is a relatively poor match to normal speech, it could be more effective if we use singing voice instead of speech for statistical modelling. Tzanetakis [1] used a bootstrapping process for the identification of vocal segments. A small random sampling of the song was annotated by the user and these samples were used to train the song-specific vocal and non-vocal models. This approach requires manual annotation for every song it processes and is therefore not fully automatic. Furthermore, since only a small number of samples can be annotated by user, it affects the quality of the training data.

Taking the existing research a step further, our focus here is to construct a statistical classifier with parametric models that learn the specific vocal characteristics of a song without the need for manual annotation. In addition, we employ the multi-model HMM (MM-HMM) training approach to tackle the intra-song and inter-song variations for improved classification performance. Our approach consists of two steps. First, MM-HMM is trained using the vocal and non-vocal segments of songs from a training database. Then, the test song is segmented and classified using MM-HMM. Following that, the first classification result of the test song is used to train its own vocal and non-vocal bootstrapped HMM models. Finally, the song is segmented again using its own models.

The rest of the paper is organized as follows. The process of feature extraction from an audio signal is presented in Section 2. Details of the MM-HMM classifier and the bootstrapping process are given in Section 3. The song database used in the experiments is described in Section 4. The experiment set-up and results are given in Section 5. Section 6 concludes the paper.

2. ACOUSTIC FEATURES

Our technique of feature extraction is based on sub-band processing that uses the Log Frequency Power Coefficients (LFPC) to provide an indication of the energy distribution among subbands.

A digital waveform is converted into an acoustic feature vector for classification. For high accuracy in vocals detection, the features suitable for the task should be selected. We assume that the spectral characteristics of different segments (pure vocals, vocals with instrumental and pure instrumental) are different. If vocals begin while instrumental is going on, a sudden increase in the energy level of the audio signal is observed [6]. Therefore, we extract feature parameters based on the distribution of energy in different frequency bands in the range from 130Hz to 16 kHz. We use these parameters to facilitate the classification of vocal and non-vocal segments.

A music signal is divided into frames of 20 ms in length with a 13ms overlap. Each frame is multiplied with a hamming window to minimize signal discontinuities at the end of each frame, and then, fast Fourier Transform (FFT) is computed. Each audio frame is passed through a bank of 12 bandpass filters spaced logarithmically from 130Hz to 16 kHz. Figure 1 is a diagrammatic representation of 12 subband filters.

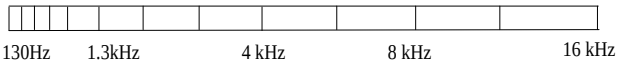

Figure 1. Subband frequency divisions

Subband-based Log Frequency Power Coefficients (LFPC) [11] are then computed using Equations (1) and (2).

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

where, $X_t(k)$ is the k^{th} spectral component of the hamming windowed 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 centre frequency and bandwidth of the m^{th} subband, respectively.

The LFPC parameters which provide an indication of energy distribution among subbands are calculated as follows:

$$LFPC_t(m) = 10\log_{10}\left[\frac{S_t(m)}{N_m}\right] \qquad (2)$$

where N_m is the number of spectral components in the m^{th} subband. For each frame, 12 LFPCs are obtained.

Figure 2 shows the energy distribution of non-vocal and vocals segments over above-defined 12 subband filters. The segments are extracted from six different songs. The total length of each vocal/non-vocal segment is 90 seconds. The figure shows that the vocal segments have relatively higher energy values in the higher frequency bands in comparison with the non-vocal segments. Therefore, as can be seen in Figure 2,

LFPC is a quite effective feature for the discrimination of vocal and non-vocal segments.

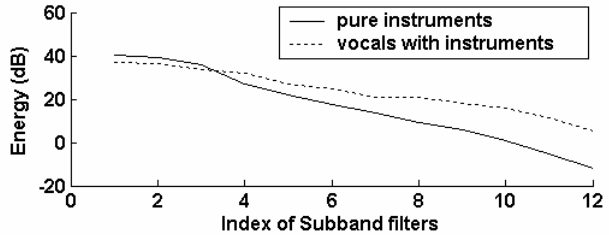

Figure 2. Energy distribution of pure instrumental segments and vocals with instrumental segments over 12 subband filters

3. CLASSIFIER FORMULATION

3.1. Multi-model HMM classifier

Most studies on vocals detection use statistical pattern classifiers [1, 4, 5]. However, to our knowledge, none of the studies takes into account song structure information in song modelling. An important observation is that vocal and non-vocal segments display intra-song signal characteristics variation. For example, signal strengths in different sections (verse, chorus, bridge and outro) are usually different. In our observation, for most songs, the signal strength of the verse is relatively low compared to that of the chorus. Chorus sections are usually of stronger signal strength in comparison with verse sections since they may have busier drums, some additional percussion, a fuller string arrangement and an additional melody line [10]. The verse section usually has lighter arrangement than the chorus section. Sample waveforms extracted from different verse and chorus sections of a popular song are depicted in Figure 3.

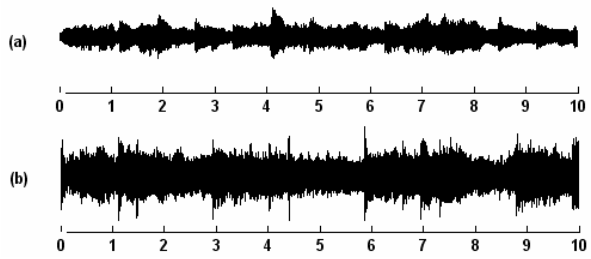

Figure 3. Waveforms of 10-second segments extracted from (a) the verse, (b) the chorus sections of the song '25 Minutes'. The horizontal axis represents time in seconds.

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 the section type (intro, verse, chorus, bridge and outro), tempo and loudness. We assume the tempo of the input song to be constrained between 40~185 beats per minutes (BPM). We divide music into high and low tempo classes according to a fixed threshold, which is 70BPM in our current implementation. Similarly, we divide music into loud and soft classes according to a threshold, which is determined by each individual song in the training dataset. Finally, a model is created for each class as shown in Figure 4. In our current implementation, we use 20 models for modelling vocal and non-vocal segments respectively.

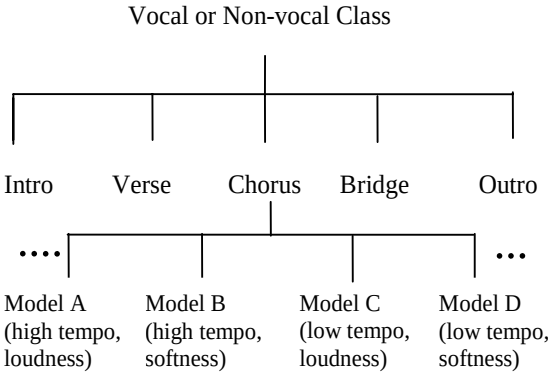

Figure 4. Creating several variants of the vocal or non-vocal HMM model

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 modelling in comparison to the single-model baseline.

3.2. Bootstrapping process

In the above section, we have created a classifier that learns the characteristics of the vocal and instrumental components in the training data. We may use this classifier to segment songs. However, the variations in tempo, timbre properties and loudness of different songs may affect the classification accuracy. By employing a classifier that can learn the specific characteristics of the test song, we can anticipate a better classification performance. With similar motivations, a method of bootstrapping was proposed in which song-specific vocal characteristics were learned by the classifier to segment the song [1]. This process requires human annotated vocal and non-vocal segments (bootstrapped samples) of every test song to train their model. In addition, this method depends on the number of bootstrapped samples available to learn the vocal characteristics of the song. In our approach, we first segment the song into vocal and non-vocal segments using the MM-HMM classifier. We then use the initial segmentation as bootstrapped samples to

build song-specific vocal and non-vocal models of the test song with a bootstrapping process. This process allows us to use a song's own model for classification as shown in Figure 5. This bootstrapping process makes the algorithm adaptive and capable of achieving higher vocals detection accuracy.

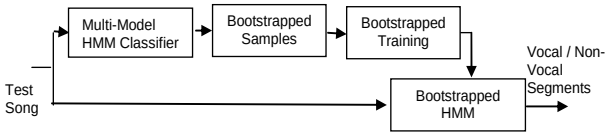

Figure 5: Bootstrapped training and segmentation process

4. SONG DATABASE

In order to conduct the experiments, we compile a small song database which includes 20 popular songs. The songs are selected to obtain a variety in time period and artists. On average, the vocal segments occupy 67% of the total duration of a song, and the rest 33% are non-vocal segments. Each song is annotated manually to obtain the vocal and non-vocal segments to provide ground truth data. This ground truth data is used to evaluate system performance. Typical segment durations range from 0.8 seconds to 12 seconds. The sampling frequency of the songs is 44.1 kHz, stereo channel and 16 bit per sample. In our experiments, the songs with original sampling frequency are used without re-sampling. Six songs of different artists are allocated to the training dataset and the remaining 14 songs to the test set. There is no overlap between the two datasets.

5. EXPERIMENTS

5.1. Experimental configuration

Several experiments are conducted to evaluate the effectiveness of the proposed approach. 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 [9]. In our experiments, 30% of the database which includes six songs, selected randomly, is used as training data, and the system is tested on the remaining 14 songs. Using this 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 4.

First, the test song is blocked into 200ms analysis frames, and then, LFPC features are calculated from 20ms with 13ms overlapping subframes. Then, the song is segmented into vocal and non-vocal frames of 200ms in duration using the MM-HMM classifier. In the MM-HMM classifier, every analysis frame of the song is matched with models of the classifier, and the frame is assigned to the model having the best match. As shown in Figure 4, each model of the MM-HMM classifier is associated with a vocal/non-vocal class, section type (verse, chorus, etc.), tempo and loudness level. Therefore, in the classification process, the MM-HMM classifier assigns a vocal/non-vocal label as well as section type, tempo and loudness level labels to a classified analysis frame. As a result, the frames of the song classified by the MM-HMM classifier are associated with song structure information.

This initial segmentation produces a bootstrapped database which includes vocal and non-vocal segments (bootstrapped samples) of the test song. Then, the bootstrapped samples are used to train the HMM. This process provides song-specific vocal and non-vocal models of the test song. Since bootstrapped samples are associated with song structure information, the bootstrapping training process takes care of song structure information in song modelling. Finally, the test song is segmented into vocal and non-vocal regions using the song-specific vocal and non-vocal models. The same analysis frame length used in the MM-HMM classifier is also used here.

To find the best matched model for each analysis frame, frame log-likelihoods are calculated for all models and the likelihoods are compared in the HMM classifiers. Accumulating the frame log-likelihoods over a longer period is more statistically reliable for decision making [12]. In addition, the feature parameters of a relatively short frame length do not capture information about melody, rhythm or long-term song structure [13]. To observe the classification accuracy of using longer analysis frame lengths, additional experiments are carried out using analysis frames of 400 ms, 600 ms, 800 ms, 1000 ms, 1200 ms and 1400 ms. The experimental results are presented in Table 1.

5.2. Results and discussion

Table 1 shows the average detection accuracy of the vocal and non-vocal segments of 14 pop songs by the MM-HMM classifier and the bootstrapping method using different analysis frame lengths.

The results show that long-term acoustic features are more capable of differentiating vocal and non-vocal segments. The optimal frame length seems to be around 1 second. The reason of the decreased performance for longer frame length is that the assumption of stationarity in a frame is not longer valid. With long frame length it is more likely that the vocal and non-vocal segments are present in a frame.

Table 1: Vocal/ non-vocal segment detection average accuracies of different analysis frames over 14 songs
N= Non-vocal segments, V=Vocal segments, Avg=Average

Frame size (ms)	MM-HMM			Bootstrapped HMM		
	N	V	Avg	N	V	Avg
200	78.8	73.5	76.2	79.2	75.9	77.6
400	80.9	78.9	79.9	80.4	82.1	81.3
600	80.4	82	81.2	80.8	84.2	82.5
800	80.6	84.3	82.5	79.2	87	83.1
1000	81.9	84.3	83.1	82	86.6	84.3
1200	80.1	85.2	82.7	79.3	87.4	83.4
1400	78.1	86.4	82.3	78.3	88.2	83.3

Relatively high accuracy is obtained using the MM-HMM classifier. After employing the bootstrapping process, the accuracy of the system is improved. Repeating the bootstrapping process several times improves performance but with the penalty of increased computational cost. As our preliminary experiments show that the performance improvement is marginal. Therefore, we apply bootstrapping only once in the experiment.

Table 2: Indices and titles of 20 songs in the database

Training Data	
Song Index	Title
1	[1978] - Village People - YMCA
2	[2002] - Blue - One Love
3	[1986] - Chris DeBurgh - Lady in Red
4	[1986] - Roxette - It must have been love
5	[1984] - Stevie Wonder - I just called to say I love you
6	[2000] - Ronan Keating - When you say nothing at all
Testing Data	
Song Index	Title
1	[1993] - MLTR - 25 Minutes
2	[1993] - MLTR - Wild Women
3	[1983] - The Police - Every breath you take
4	[1993] - MLTR - The Actor
5	[2000] N'Sync - This I promise you
6	[1993] - MLTR - Sleeping Child
7	[1980] - ABBA - Super Trouper
8	[1999] - Backstreet Boys - As Long As You Love Me
9	[2001] - Westlife - World Of Our Own
10	[1999] - Backstreet Boys - Back To Your Heart
11	[1989] - Phil Collins - Another day in Paradise
12	[1995] Take That - Back for good
13	[1998] Eagle Eye Cherry - Save Tonight
14	[2003] - Dido - White Flag

Figure 6 shows the results of the vocal/non-vocal segment classification for all the test songs in our database. Indices of the test songs as well as training songs and their titles are listed in Table 2. The classification performance is not consistent among the songs. This is because the songs in the database are of different characteristics. For example, some songs are associated with high tempo and loudness while some are associated with low tempo and softness. In addition, vocals in some songs are dominant in most part of the song while others having strong instrumental accompaniment throughout the song. Based on the characteristics of the song, the system achieves accuracies ranging from the highest of 91.1% (25 Minutes) to the lowest of 77.2% (White Flag). In general, songs with light background instrumental obtain higher accuracy than songs with strong background instrumental.

The bootstrapping process depends on the bootstrapped samples. For the last two songs (Indices 13 and 14) of Figure 6, the bootstrapped HMM is lower in accuracy than the MM-HMM classifier. The reason is that the accuracies of MM-HMM are relatively low for these songs and larger numbers of bootstrapped samples are incorrectly labelled compared to the other songs.

We give an example of vocal segments detected by the bootstrapped HMM together with manually annotated vocal segments of the chorus section of a song are shown in Figure 7.

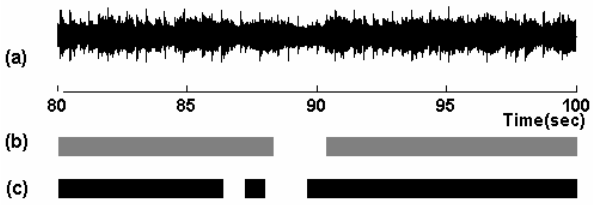

Figure 7: (a) The segment (Chorus, 20 sec) of the song '25 Minutes' (b) Manually annotated vocal segments (c) Automatically detected vocal segments

To compare the performance of LFPC feature with traditional MFCC feature, experiments are conducted using MFCC feature for the frame size of 1000ms. The results are summarized in Table 3. LFPC feature outperform the MFCC feature. MFCC feature is mainly used for speech recognition and its capability to capture

the spectral characteristics of the audio signal seems to be lower than LFPC feature.

Table 3: Performance comparison between LFPC feature and traditional MFCC feature (Frame size=1000ms)

Feature	MM-HMM			Bootstrapped HMM		
	N	V	Avg	N	V	Avg
LFPC	81.9	84.3	LFPC	81.9	84.3	LFPC
MFCC	59.6	83.3	MFCC	59.6	83.3	MFCC

FEA=Feature, N= Non-vocal segments, V=Vocal segments, Avg=Average

Next, we investigate the effectiveness of employing song structure information in song modelling. The experiments are conducted using the baseline HMM training method in which only one model is created for each vocal and non-vocal class. This approach disregards the structure of the song in song modelling. First, the experiments are conducted using the same number of mixtures per state (2 mixtures/state) for both MM-HMM and base line HMM. The results presented in Table 4 show that the MM-HMM training method outperforms the baseline HMM training approach for the same number of mixtures per state. The base line HMM has 20 times less free parameters in comparison with MM-HMM as MM-HMM has 40 models in total. To give the base line HMM a fair chance for comparison, we perform further experiment using base line HMM with 10 mixtures per state. The results presented in Table 4 show that using more free parameters for base line HMM can not improve performance. The reason is that automatic data clustering is not accurate in comparison with manual clustering.

Table 4: Performance comparison between MM-HMM training and base line HMM training method (Frame size =1000ms)

	N	V	Avg
MM-HMM (2 mixtures/state)	81.9	84.3	83.1
Baseline HMM (2 mixtures/state)	79.2	83.4	81.3
Baseline HMM (10mixtures/state)	57.4	91.8	74.6

V=Vocal segments, N= Non-vocal segments

Figure 8 shows how well the test signal matches our MM-HMM. It displays the probability distribution of correctly matched models as well as of wrongly matched models for the test segments in five different section types. The darkness of the rectangles indicates the matching probability. As expected, segments from the verse section are more likely to match the verse models than others. In the same way, segments from the other sections tend to match their respective models rather than other models. However, chorus segments tend to match both the chorus and outro models. This is due to the fact that the chorus is repeated in the outro section before the song fades out [14].

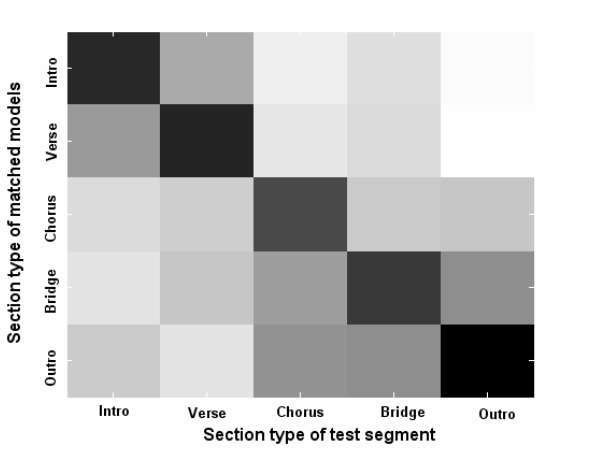

Figure 8: Probability distributions of the number of matched models for the test segments in five different sections

Based on our experimental results, we could consider the following options to improve system performance.

It is well-known from the literature that singing vocals are highly harmonic (energy exists at integer multiples of the fundamental frequency), and other high energy sounds – drums in particular – are not as harmonic and distribute their energy more widely in frequency [8]. Over 90% of the vocal signal is harmonic (much more than the case of a speech signal) [6]. It is believed that a vocal signal in general has higher values of the harmonic component, compared to the instrumental component [3]. Unfortunately, the relatively simple features used in our current system cannot capitalize the harmonic feature of singing vocal. Nevertheless, the promising results from our current system lead us to believe that we can further improve our system performance by incorporating the harmonic features.

The other possibility is to enable a semi-automatic system similar to that in [1]. However, instead of choosing bootstrapping samples randomly, we could allow the user to check and to select the bootstrapped samples (vocal and non-vocal segments) manually from the initial segmentation performed by the MM-HMM. The accuracy of the initial MM-HMM classifier is around 80%, with 20% of the bootstrapped samples wrongly labelled. The manual selection of automatically detected results is a less demanding job for the user in comparison with the manual annotation of singing vocals from the original song. By incorporating human intervention in the loop, it would improve the quality of the bootstrapped samples significantly as shown in Figure 5. As a consequence, we could expect a better system performance.

If we stick to an automatic system, we can use the neighbourhood information in HMM model space [15] and Bayes factors as tools to calculate the confidence measure on the output of the initial MM-HMM classifier. With these additional techniques, segments that have a high probability of being labelled wrongly by the classifier are rejected, and we can train the bootstrapped vocal detector using generally correctly labelled samples.

6. CONCLUSION

We have presented an automatic approach for detecting vocal segments in a song. The proposed approach combines the multi-model HMM classifier and the bootstrapping method. The key points are integration of song structure information and song-specific vocal characteristics in song modelling. The bootstrapping process is used to improve vocals detection accuracy.

In a test dataset comprising 14 popular songs, our approach has achieved an accuracy of 84.3% in identifying vocal segments from non-vocal ones.

One drawback of the proposed approach is that it is computationally expensive since it entails two training steps: training the MM-HMM classifier and training the bootstrapped classifier. To reduce computational complexity, the number of bootstrapped samples can be reduced. Instead of using all the segments of a song (of an average duration of 3 minutes) for bootstrapped training, only a certain number of samples with a very high confidence measure of correct classification could be used. This would reduce computation time and further improve system performance.

7. REFERENCES

[1] Tzanetakis, G. "Song-specific bootstrapping of singing voice structure", *IEEE International Conference On Multimedia And Expo*, 2004.

[2] Maddage, N., Xu, C., and Wang, Y. "An SVM-based classification approach to musical audio", *4th International Conference on Music Information Retrieval*, Maryland, USA, 2003.

[3] Chou, W., and Gu, L., "Robust singing detection in speech/music discriminator design", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp.865 - 868, 2001.

[4] Berenzweig, A., Ellis, D. P. W., and Lawrence, S., "Using voice segments to improve artist classification of music," *AES 22nd International Conference, Espoo*, Finland, 2002.

[5] Berenzweig, A.L. and Ellis, D.P.W. "Locating singing voice segments within music signals," *Proceedings of IEEE WASPAA'01*, pp.119-122, New York, Oct. 2001.

[6] Zhang, T, "System and method for automatic singer identification", *IEEE International Conference on Multimedia and Expo*, Baltimore, MD, 2003.

[7] Maddage, N., Wan, K.W., Xu, C., and Wang, Y., "Singing Voice Detection Using Twice-Iterated Composite Fourier Transform", *IEEE International Conference On Multimedia And Expo*, 2004.

[8] Kim, Y., and Whitman, B., "Singer identification in popular music recordings using voice coding features", *Proceedings of Int. Symposium on Music Information Retrieval*, 2002.

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

[10] Waugh, I. Song Structure. Music tech magazine, October 2003.

[11] Nwe, T.L, Foo, S.W, and De Silva, L.C. **"**Stress classification using subband based features**",** *IEICE Transactions on Information and Systems, Special Issue on Speech Information Processing*, Vol. E86-D, no.3, pp. 565-573, March 2003.

[12] Tsai, W.H., Wang, H.M., Rodgers, D., Cheng, S.S., and Yu, H.M. "Blind clustering of popular music recordings based on singer voice characteristics", *4th International Conference on Music Information Retrieval*, Maryland, USA, 2003.

[13] Berenzweig, A., Logan, B., Ellis, D.P.W., and Whitman, B., "A large-scale evaluation of acoustic and subjective music similarity measures," *Proc Intl Conf on Music Information Retrieval*, Washington DC, 2003.

[14] Watson, C.J., "The everything song writing book", Adams Media Corporation, Avon, Massachusettes, 2003.

[15] Jiang, H., and Lee, C.H., "A new approach to utterance verification based on neighborhood information in model space", *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp.425 – 434, 2003.

[16] Andre-Obrecht, R., "A New Statistical Approach for the Automatic Segmentation of

Continuous Speech Signals", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, No. 1, pp. 29-40, January 1988.

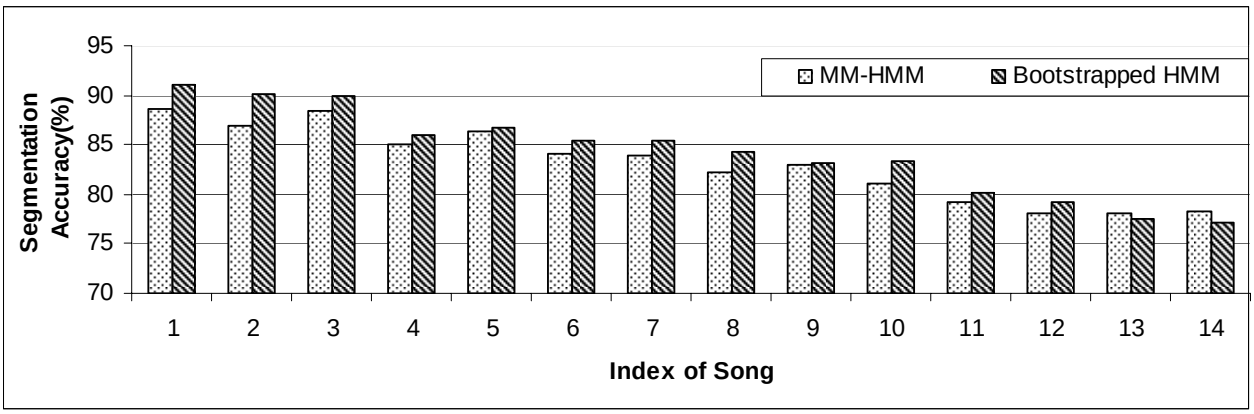

Figure 6: Vocal/ non-vocal segment detection accuracies for individual songs with analysis frame length of 1000ms