Singer Identification Based on Vocal and Instrumental Models

Namunu Chinthaka Maddage[1, 2], Changsheng Xu[1], Ye Wang[2]
[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{maddage, xucs}@i2r.a-star.edu.sg
[2]School of Computing, National University of Singapore, Singapore 117543
wangye@comp.nus.edu.sg

Abstract

In this paper, we propose a novel method to identify the singer of a query song from the audio database. The database contains over 100 popular songs of solo singers. The rhythm structure of the song is analyzed using our proposed rhythm tracking method and the song is segmented into beat space time frames, where within the beat space time length the harmonic structure is quasi stationary. This inter-beat time resolution of the song is used for both feature extraction and training of the classifiers (i.e. Support Vector Machine (SVM) for vocal/instrumental boundary detection and Gaussian Mixture Models (GMMs) for modeling the singer). Combining the instrumental music similarities in the songs of the same singer with the vocal model can improve the identification of the singer with an accuracy of over 87%.

1. Introduction

Automatic singer identification is important for music indexing and retrieval. The singing voice is the oldest musical instrument. Therefore, human auditory physiology and perceptual apparatus have evolved to a high level of sensitivity to the human voice. After over three decades of extensive research on speech recognition, the technology has matured to the level of practical applications. However speech recognition techniques have limitations when applied to singing voice identification, because speech and singing voice differ significantly in terms of their production and perception [10].

Several approaches have been proposed to identify the singer of a query song from databases. Zhang [12] trained GMMs using Linear Prediction derived Cepstral Coefficients (LPCC) calculated from manually labeled vocal sections of each singer. In this method, the beginnings of the vocal sections were detected using simple threshold settings which were calculated from extracted features i.e. energy, zero crossing rate, spectral flux and harmonic coefficients. It was assumed that vocal sections lasted for up to 10~30 seconds and these vocal sections were fed into GMMs for further singer identification. Berenzweig *et. al.,* [1] trained multilayer perceptron neural network with LPCC to detect the vocal passages in the song and same neural network was trained with Mel-Frequency Cepstral Coefficients (MFCC) for singer identification. Kim *et. al.* [6] used inverse combo filter bands to analyze the harmonicity and the vocal regions were detected by setting a threshold to the harmonicity against a fixed value. Then GMM and SVM classifiers were trained with warped Linear Prediction Coefficient to identify the singer.

Although above mentioned methods have achieved up to 80% of frame level accuracy, their performances are inefficient due to reasons given below.

- Experiments are performed on the studio recorded pure vocal music, not on the normal instrumental mixed vocal music.
- Inaccurate vocal boundary detection in the music.
- Music knowledge has not been effectively exploited for modeling the singers in existing (mostly bottom-up) methods.

We believe that a combination of bottom-up and top-down approach, which combines the strength of low-level features and high-level music knowledge, can provide us a powerful tool to improve the system performance.

In this paper we propose a novel method to identify the singer using both low-level features and music structure knowledge. Usually, popular singers in their albums, follow similar instrumental setup and music patterns such as chord combinations and music scale changes. Therefore the melody contours in the song is closely correlated with the formant structures of the singer. In our proposed singer identification technique, in addition to vocal tract characteristics such as formant and harmonic structures of the singing voice, we use the structure similarity of the instrumental music sections of the same singer to identify the singer with high confidence level.

The process of singer information modeling is shown in Figure 3. We assume the meter of the song to be 4/4, being most frequent of popular songs and the tempo of the input song to be constrained between 30-240 M.M (Mälzel's Metronome: the number of quarter notes per minute) and almost constant. This reveals that the temporal properties (pitch contours / harmonic structure variation) more likely vary on inter-beat intervals. Therefore, first we analyze the beat / note onset structures of the music and segment the music according to the beat space time resolution (similar method in [7]). The SVM

IEEE COMPUTER SOCIETY

based learning method is implemented for the detection of vocal and instrumental sections. Finally two GMMs (one for vocal sections and the other for instrumental sections) are trained to represent the singer characteristics.

The rest of the paper is organized as follows. The vocal and instrumental boundary detection, singer characteristics modeling and singer identification are described in Section 2, 3 and 4 respectively. Experimental results are reported in Section 5. We conclude with future work in Section 6.

2. Beat space segmentation and vocal/ instrumental boundary detection

The beat can be represented as the sequence of equally spaced phenomenal impulses which define the tempo for the music [9]. In our rhythm tracking method, note onsets and beat positions are detected using a sub-band decomposition approach [1]. Then, based on statistical autocorrelation of detected strong and weak onset times, we obtain an inter-beat-interval corresponding to the temporal length of a quarter note described in Figure 1. The music signal is assumed to be quasi stationary between the inter-beat times, because the chords, music scales and harmonic structure change on beat times. Thus, we apply the following chord knowledge [4] to the timing of vocal passages:

1. Chords are more likely to change on beat times than on other positions.
2. Chords are more likely to change on half note times than on other positions of beat times.
3. Chords are more likely to change at the beginning of the measures than at other positions of half note times.

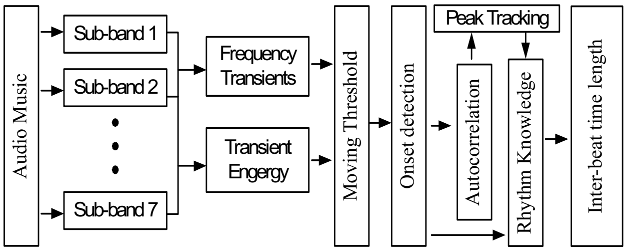

Figure 1: Rhythm tracking

Our rhythm tracking method can detect up to the 16th note time duration. We then segment the music into frames according to this inter-beat-interval instead of the conventional 20~30ms time window. The SVM is trained with the 10th order Octave Scale Cepstral Coefficients (OSCCs) calculated from each frame. In order to compute OSCCs, the magnitude spectrum of the signal is filtered by triangular filter bank where the filters are positioned in octaves [8] of the linear frequency scale. Then we calculate the cepstal coefficients [5] from the filter outputs. In the experiments, we found that OSCCs are more efficient than MFCCs for characterizing instrumental music from vocal music. The vocal/instrumental boundary detection steps are presented in Figure 2.

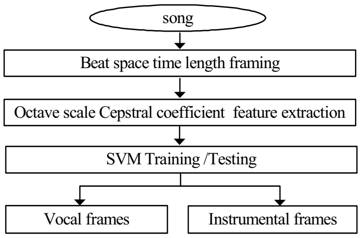
song
Beat space time length framing
Octave scale Cepstral coefficient feature extraction
SVM Training /Testing
Vocal frames
Instrumental frames

Figure 2: Vocal and instrumental boundary detection

3. Singer characteristic modeling

The GMMs are efficient in modeling the vocal tract than other statistical models [12] [6]. The vocal separation from the instrumental music is still a challenging task which is yet to overcome in the music research. Thus we model instrumental mixed vocal sections of the singer as vocal model and pure instrumental sections as instrumental model shown in Figure 3.

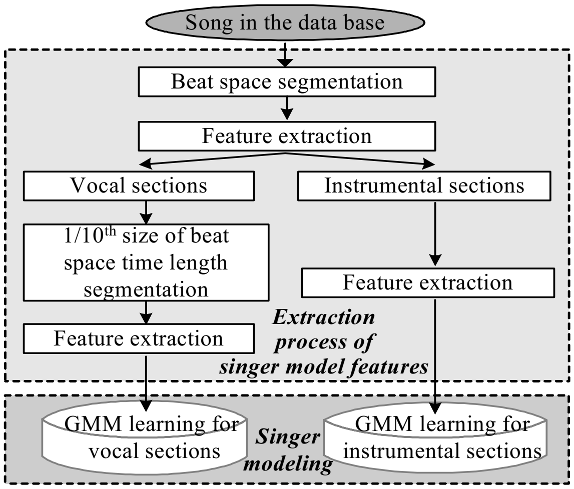
Song in the data base
Beat space segmentation
Feature extraction
Vocal sections
Instrumental sections
1/10th size of beat space time length segmentation
Feature extraction
Feature extraction
Extraction process of singer model features
GMM learning for vocal sections
Singer modeling
GMM learning for instrumental sections

Figure 3: Vocal and the relative instrumental section modeling of songs of same singer.

A singer has a unique formant structure. It is required to have shorter time resolution in order to extract formant structure [10]. Therefore we segment the vocal into sub-frames with 1/10th of beat-space vocal frames (quarter note time length 400-900ms) and extract LPCC (vocal tract model coefficients) from each vocal sub-frame. The harmonic structure and the pitch of the vocal are spaced according to the music scale (octave scale) in the spectrum. Therefore OSCCs are extracted from each sub-frame to characterize the singer's harmonic structure.

Since the instrumental harmonic structure is widely spread over the spectrum (i.e. 100 ~15000Hz) than it is in vocals and melody transition occurs in inter-beat time intervals, we use longer time length frames (i.e. initial beat space time frames) to calculate OSCCs. The features and their orders for training the GMMs for both vocal and instrumental models are described in Table 1. The mean, covariance and weights of the GMMs are estimated using expectation maximization (EM) algorithm [2]. The number of Gaussian mixtures for vocal and instrumental models is empirically found to be 15 and 25 respectively.

Table 1: Training parameters for GMMs

	Vocal	Instrumental
Features and their order	12-OSCC, 16-LPCC	20 – OSCC
No of Gaussian mixtures	15	25

4. Singer identification

In our perceptual analysis [11], we found that there are structure similarities (chords mixtures, music scales, rhythm and instrumental setup) in the instrumental music compositions at the songs of same singer. Thus the final response of a singer model to a test song is the combined response of vocal and instrumental models to the vocal and instrumental sections of the test song as shown in Figure 4. The block "Extraction process of singer model features" in Figure 4 is same as the "Feature extraction" in Figure 3. The extracted feature coefficients for each frame (see Table 1) of the test song are arranged as feature vector where the dimensions of both vocal and instrumental feature vectors are 28 and 20 respectively.

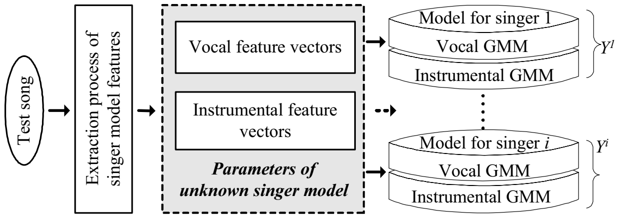

Figure 4: Singer identification

Y^i in Figure 4 is the final response of the i^{th} singer model to the test song and is calculated using Eq.(1) where the scalar weights α and β are the degrees of the model response of both vocal and instrumental respectively. The calculation of total vocal or instrumental model (i.e $GMM^i_{Vocal\ or\ inst}$) corresponding to the vocal or instrumental frames of the test song, is described in Eq.(2) where j is the frame number of either vocal or instrumental.

$$\alpha.GMM^i_{Vocal} + \beta.GMM^i_{instrumental} = Y^i \tag{1}$$

$$GMM^i_{Vocal\ or\ i\ nst} = \sum_j GMM^i_{Vocal\ or\ i\ nst}(j) \tag{2}$$

For a test song, if Y^i is greater than the response of the rest of the models, the i^{th} singer model will be assigned as the singer of the test song. In Figure 5, we have shown the two plots of normalized values of singer model response (Y^i) when (α=1 & β=0) and (α=1 & β=0.6) to a test song of *Eric Clapton* (singer database is shown in Table 2).

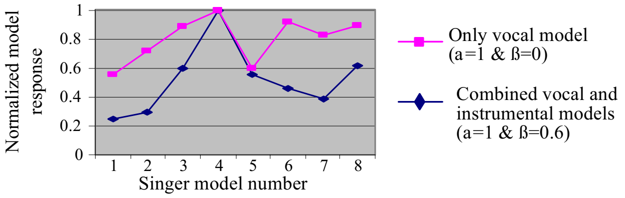

Figure 5: Normalized response of singer models

When vocal and instrumental models are combined, the singer of the test song is identified with high confidence. This can be seen from Figure 5. For combined vocal and instrumental model, except the highest response of the 4th singer model which is the correct singer of the test song, the responses of the rest of the models are 25% below the highest response. But for only vocal model, it is only 5% below the highest response of the 4th singer model.

5. Experiments

All the music data used in our experiments are collected from commercial audio CDs at a 44.1 kHz sample rate, and 16 bits per sample in stereo and the information of these songs is shown in Table 2. The songs are manually annotated with vocal/instrumental boundaries and singer identity.

Table 2: The training and testing data set

Singer Name	Language	Gender	Tracks
Bryan Adams	English	Male	14
Michael Bolton	English	Male	16
Eric Clapton	English	Male	16
Shania Twain	English	Female	18
Clarance Wijewardana	Sinhala	Male	18
Huang Pingyuan	Chinese	Male	10
Li Qi	Chinese	Female	10
Liu Ruoying	Chinese	Female	8

We perform two experiments and they are explained in EXP 1 and EXP 2.

- EXP 1: Vocal/instrumental boundary detection

The training set includes half of the total number of the tracks of each singer, which is actually half of the database (see Table 2). In the training and classification process, songs are first segmented into beat space time length frames and then the 10th order OSCCs are calculated to represent each frame as a feature vector. 81.34 % and 85.82% accuracies are achieved in correct classification of vocal frames and instrumental frames respectively.

- EXP 2: Singer identification

Two GMMs (one for vocal and one for instrumental) are trained with the half number of the songs of each singer and the other half of the songs are used for evaluating the accuracy of correct identification of the singers. Selected features and their orders for training GMMs are shown in Table 1. Previous automatic singer identification methods only model vocal sections and neglect the instrumental music structure (chords mixtures, music scales, rhythm and instrumental setup) which is usually identical to the singer. In our method, we increase the confidence of identifying the singer by combining both vocal GMM and instrumental GMM. This combination of two GMMs for the i^{th} singer is shown in Eq. (1), where $\alpha=1$ and $\beta=0.6$ are empirically found to be fit for our data set. Unknown singer is identified as a singer in the database, whose model response is the highest as explained in Section 4. The accuracy of the correct identification of the singer of the songs is shown in Figure 6. Based on our experimental results, it can be noticed that model-C (i.e. combined vocal and instrumental model) can increase the accuracy of the singer identification with high confidence level.

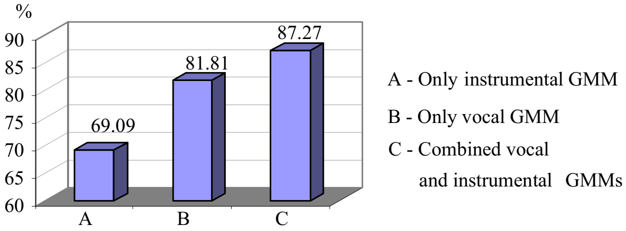

Figure 6: Correct identification of the singer

6. Conclusions and future work

We have presented a novel and efficient approach for singer identification. Our earlier proposed robust rhythm tracking and extraction method [7] is used to segment music signals at the beat level. This enables us to use musically meaningful inter-beat-interval, instead of conventional 20~30 ms frame length as the time resolution for music segmentation. Using both time resolution of the music signal and the SVM trained with OSCCs, we can detect the vocal sections and instrumental sections with over 80% accuracy.

In order to train GMM for vocal sections of the singer, the LPCCs and OSCCs are extracted from much shorter time frame, i.e. 1/10th of initial beat space time length. The GMM for instrumental sections of the same singer is trained with OSCCs extracted from initial beat space time length frames. By correlating both vocal and instrumental GMMs we can identify the singer with over 87% accuracy. This accuracy is obtained with 110 test songs.

When the test music sample is short (for example, less than 50% of the full song) and has more instrumental sections, our proposed approach may fail to identify the correct singer of the test sample due to the errors in calculating beat space time length and parameters in the combination of vocal and instrumental models.

Compared with previous singer identification methods, our proposed method can achieve a higher accuracy, but it is also more computationally complex. The overall accuracy in our method depends on the accuracy of beat space segmentation, vocal/instrumental boundary detection, and vocal and instrumental feature extraction and modeling. In the future we will focus on how to improve and validate the robustness of those sections in our system.

Since vocal extraction from the vocal/instrumental mixed music still remains a big challenge. Our future work will also focus on vocal separation from the mixed music to improve the accuracy of the vocal modeling and singer identification.

7. References

[1] A. Berenzweig, D.P.W. Ellis, and S. Lawrence, “Using Voice Segmentations to Improve Artist Classification of Music”, AES22nd ICVSEA 2002.

[2] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal statistical Society, 39, I-38, (1977).

[3] C. Duxburg, M. Sandler, and M. Davies, “A Hybrid Approach to Musical Note Onset Detection”, DAFx-02, Hamburg, Germany, September 26-28, 2002.

[4] M. Goto, “An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds”, JNMR Vol.30, No.2, pp.159-171, June 2001.

[5] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals.* IEEE Press, 2000.

[6] Y.K. Kim, & W. Brian, “Singer Identification in Popular Music Recordings Using Voice Coding Features”. ISMIR 2002.

[7] N.C. Maddage, K. Wan, C.S. Xu, and Ye. Wang, “Singing Voice Detection Using Twice-Iterated Composite Fourier Transform”, ICME 2004.

[8] J.L. Monzo, *JustMusic: A New Harmony - Representing Pitch as Prime Series*. Joseph L. Monzo. 1998.

[9] E.D. Scheirer, “Tempo and Beat Analysis of Acoustic Musical Signals”,JASA,103(1),1998

[10] J. Sundberg, *The Science of the Singing Voice*, Northern Illinois University Press, Dekalb, Illinois 1987.

[11] C.S. Xu, N.C. Maddage, and Xi. Shao, “Automatic Music Classification and Summarization”, *IEEE Transaction on Speech and Audio Processing* (accepted).

[12] T. Zhang, “Automatic singer identification”, ICME 2003.