Identifying Singers of Popular Songs

Tin Lay Nwe and Haizhou Li

Institute for Infocomm Research, Singapore 119613
{tlnma,hli}@i2r.a-star.edu.sg

Abstract

In this paper, we propose to identify the singers of popular songs using vibrato characteristics and high level musical knowledge of song structure. The proposed framework starts with a vocal detection process followed by a hypothesis test for the vocal/non-vocal verification. This method allows us to select vocal segments of high confidence for singer identification. From the selected vocal segments, the cepstral coefficients which reflect the vibrato characteristics are computed using the parabola bandpass filters spread according to the music frequency scale. The strategy in our classifier formulation is to utilize the high level musical knowledge of song structure in singer modeling. The proposed framework is validated on a database containing 84 popular songs of commercially available CD records from 12 singers. We achieve an average error rate of 17.9% in segment level identification.

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 information retrieval. Singer identification (SingerID) is one of the important tasks of great interest. Singer ID is essentially the task of automatically identifying the singer from a song in commercially available CD records. SingerID is much more complex than the speaker identification task due to the fact that singing voice is present in an adverse acoustic environment of music accompaniment. Although research efforts in speaker identification have made major advances, the speaker identification techniques can’t be directly applied to SingerID in a simple way.

The basic procedure for any singer identification system includes the detection of vocal segments in a song and the extraction of feature parameters from vocal segments with a time resolution constrained by the analysis window length. Then, the singer of each vocal segment is identified using a statistical classifier.

A large number of features have been explored to represent audio signals for SingerID. Some of the methods originate from efforts in speech recognition. They include Mel Frequency Cepstral Coefficients (MFCC) [1] and Linear Prediction Coefficients (LPC) [2]. Others benefit from the research in music analysis. They are robust estimates of spectral envelopes [3], Octave Frequency Scale based cepstral coefficients [4] and features derived from Composite Transfer Function estimates [5]. However, none of the studies has explored the use of vibrato information, which represents the exclusive vocal characteristics of a singer [6]. Howes [7] observed that some elements of vibrato may affect listeners’ perception of the voice, which affects their preference for a particular singer. Eric [8] found that vibrato characteristics such as rate and extent seem clearly relevant to perception. Melody [9] stated that temporal patterns in the vocal passage such as vibrato and note transitions are likely cues to singer identity and vocal quality. Inspired by these studies, we find that it is important to include vibrato characteristics in the acoustic features of singing vocals for SingerID. The method in [4] uses Octave Frequency Scale to characterize the singer’s harmonic structure. Apart from harmonic analysis, we find that this Octaves Frequency Scale can be used to investigate singer’s vibrato characteristics as well.

Common pattern classifiers such as HMM and Neural Networks (NN) have been proposed for SingerID [1,2]. In those studies, a statistical model for each of the singer was created using only acoustic cues. We find that singer modelling can be more effective if we integrate the information of pop song structure into statistical modelling.

Taking the existing research a step forward, we would like to formulate the acoustic feature parameters to capture the singer’s vibrato characteristics and to construct a statistical classifier. We employ the multi-model HMM (MM-HMM) to describe the singer’s vibrato characteristics present in different compositions of a song. Our approach consists of three steps. First, the test song is segmented and classified into vocal and non-vocal segments. Then, we compute the confidence scores of the classification decision to confirm the decision with a hypothesis test. Finally, we extract singer features from the vocal frames and evaluate them with the MM-HMM singer model.

The rest of the paper is organized as follows. In Section II, we present a method for vocal/non-vocal detection. In Section III, we discuss in details the procedures for singer feature extraction from an audio signal. In Section IV, we discuss the process of singer modelling. In Section V, we present the pop song database, experiment set-up and results. Finally, we conclude our study in Section VI.

2. Vocal/Non-vocal segmentation

2.1. Vocal detection

The vocal detector detects the presence of vocals in the musical signal. We assume that the spectral characteristics of different segments (pure vocal, vocal with instruments and pure instruments) are different. Based on this assumption, we extract feature parameters based on the distribution of energy in different frequency bands from audio signal to differentiate vocal from non-vocal segments. We compute a subband based Log Frequency Power Coefficients (LPC) [10] to form our feature vectors. For segmentation, we employ Multi-Model HMM (MM-HMM) training method [11] in which multiple vocal and non-vocal models are created based on the structure of the song. This process integrates song structure information into vocal/non-vocal modeling. Then, we employ an automatic bootstrapping process [11] which adapts the test song’s own
models for increased classification accuracy of vocal
detection.

2.2. Vocal verification

It is possible that the vocal detector may wrongly detect non-
vocal segments as vocal segments. Therefore, we evaluate the
reliability of the classification decision. The process has two
steps. First, a hypothesis test [12] is performed for each
detected vocal or non-vocal segment. Then, the accept/reject
decision for a given segment is made by comparing the
confidence measure with a predetermined threshold.

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 neighbourhood information in HMM
model space [13] to determine the properties of the possible
competing source distribution of the target model. A
hypothesis test [12] is then performed for each segment of
audio to obtain a confidence score for its current classification
as obtained from MM-HMM. This confidence score is
compared with a predetermined threshold to retain only the
frames that have a high confidence of vocal/non-vocal
decision.

3. Acoustic features

3.1. Vibrato

Vibrato is a feature of the voice of professional singers that
occurs subconsciously [7]. The vibrato adds a special quality
to the tone and seems driven by a pulsation of subglottal
pressure [14]. Vibrato corresponds to a periodic, rather
sinusoidal, modulation of the phonation frequency [6]. The
rate of frequency modulation (vib rate) varies between the
tones and it tends to increase exponentially toward the end of
the tones [8, 15]. However, the mean vibrato rate is generally
considered to be a constant for a singer. In other words, a
singer is usually unable to change his/her vibrato rate in a
singing performance [14]. Female singers tend to have a
slightly faster mean vibrato rate than male singers [16].

![Figure 1: Vibrato waveforms of 3 different singers observed at
the notes of (a) G#5, 830.6Hz (b) D6, 1174.6Hz and (c) D6,
1174.6Hz.](image)

Fig. 1(a), (b) and (c) show examples of vibrato undulations
occurring at the tones G#5, D6 and D6 respectively for three
different singers. It is seen in Fig. 1(a) that vibrato cycles are
more or less similar to a sine wave and vibrato extents are
regular. However, considerable irregularity in vibrato extent
can be seen in Fig. 1(b) and (c). According to [15], such
irregular frequency excursions are very common in most
tones. In addition, the figures show that vibrato extent is
different from singer to singer. Some singers display a wider
pitch fluctuation and a slower rate of vibrato, which is referred
to as 'wobble'. Singers producing a vibrato with a narrower
pitch modulation and a faster rate may be considered to have a
'bleat' [17]. In summary, three different characteristics of
vibrato waveform should be integrated into acoustic feature
formulation. These are 1) regularity or irregularity in vibrato
excursion 2) two different vibrato types of 'wobble' and
'bleat' and 3) vibrato rate.

3.2. Subband filters

As a result of the modulation of phonation frequency, the
frequencies of all the overtones vary regularly and in
synchrony with the phonation frequency modulation [14].
Therefore, we implement the subband filters with the center
frequencies located at each of the musical notes [18] to
characterize the vibrato. Examples of musical notes G and G#
as across 5 octaves are listed in Table 1. Due to the fact that
singing voice contains harmonics in high frequency [2], our
subband filters span up to 7 octaves. We assume that the
amplitude of vibrato undulations is smaller than ±0.5
semitones. The bandwidth of each subband filter is ±0.5
semitone from each note.

![Figure 2: Parabola bandpass filters](image)

Table 1: Musical note frequencies (Hz) across 4 octaves

<table>
<thead>
<tr>
<th>Octave</th>
<th>C2 to B2</th>
<th>C3 to B3</th>
<th>C4 to B4</th>
<th>C5 to B5</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>98.0</td>
<td>196.0</td>
<td>392.0</td>
<td>784.0</td>
</tr>
<tr>
<td>G#</td>
<td>103.8</td>
<td>207.7</td>
<td>415.3</td>
<td>830.6</td>
</tr>
</tbody>
</table>

Figure 3: Vibrato fluctuations and parabola bandpass filtering
observed at the note G#5, 830.6Hz. (a) vibrato fluctuates
down (b) no fluctuation (c) vibrato fluctuates up. The upper
panel shows spectrum partial. The middle panel presents the
frequency response of the parabola bandpass filters. The lower
panel demonstrates the output of the parabola bandpass filters.

These parabola bandpass filters benefit from their
asymmetric structure. They capture the regularity or
irregularity in vibrato excursion and the ‘wobble’ and ‘bleat’ vibrato types.

Fig. 3 shows the frequency vibrato extracted from fifth octave of note G# of a song. This filter has less attenuation of the signal when the phonation frequency fluctuates down (Fig. 3(a)) and higher attenuation of the signal when the phonation frequency fluctuates up (Fig. 3(c)). In addition, the filter has the highest attenuation of the signal for the case of ‘wobble’ and the least attenuation of the signals for the case of ‘bleat’. The more the fluctuation occurs, the more the signal is attenuated.

3.3. Cepstral coefficient computation

A music signal is divided into frames of 20ms with a 13ms overlap. Each frame is multiplied with a hamming window to minimize signal discontinuities at the end of each frame. Then, the audio frame is passed through a bank of 96 parabola bandpass filters spaced in octave scale from 65Hz to 16kHz and the log energy of each band is calculated. Finally, a total of 9 Octave Frequency Cepstral Coefficients using parabola subband filters (OFCCPRA) are computed from log energies using Discrete Cosine Transform [19] for each audio frame. To integrate the information of vibrato rate, we augment the 9 coefficients with time derivatives or delta parameters from OFCCPRA of two neighboring frames.

4. Statistical singer modeling

A popular song is usually composed according to a common structure comprising of intro, verse, chorus, bridge and outro. Each composition section of a song has its own characteristics, which is uniquely interpreted by a singing voice and its music accompaniment [20]. We have good reason to believe that each composition of songs provides the singer cues in its own way. For this reason, we use the Multi-Model HMM (MM-HMM) training method [11] to integrate singer cues provided by each composition of a song into the singer models. We observe that the signal strength of the intro is relatively low compared to that of the verse or the chorus. The chorus section has stronger signal strength than the verse sections since it is usually accompanied by busier drums, additional percussion, a fuller string arrangement, and an additional melody line [20]. The bridge and the outro sections have similar signal strength as the chorus sections. To capture the intrinsic signal characteristics, we propose to model the singer models, according to the type of composition, with three different HMM models \( \Lambda_s = \{ \lambda_s^V, \lambda_s^I, \lambda_s^{CBO} \} \) for singer \( s \).

![Figure 4: Creation of three HMM models for each singer](image)

Fig. 4 shows the general pop song structure and creation of HMM models based on the structure of the song. This modelling allows for more accurate singer modelling according to different acoustic environment settings that are characterized by the song composition type.

5. Experiments

Our experimental database consists of 7 popular songs from each of 12 solo singers. A total of 84 songs are included in the database. Songs are extracted from each album of the 12 singers. The songs are selected to cover a range of time period and artists. Four songs of each singer are allocated to the training set and the remaining 3 songs of each singer to the test set. We obtain a training set of 48 songs and a test set of 36 songs. There is no overlap between the two data sets.

5.1. Vocal/Non-vocal detection

We segment the song into 1 second segments and vocal/ non-vocal classification decisions are made on each segment. The error rate is reported as the number of error detection over the number of total segments that are present in the 36 songs in the test set. Two experiments are conducted. First, we carry out vocal/non-vocal classification. Second, we further examine each of the classification results with the additional hypothesis test as mentioned above. Both results are reported in Table 2. It is observed that the hypothesis test greatly improves the performance of vocal detection by 49.7% relative error reduction from 17.3% to 8.7%.

| Table 2: Error rates of vocal/ non-vocal detection |
|---|---|---|
| Vocals | N | V | Av |
| Vocal Detection | 27.2 | 7.5 | 17.3 |
| Vocal Detection + Hypothesis test | 7.7 | 9.7 | 8.7 |

5.2. Singer identification

Several experiments are conducted to evaluate the effectiveness of the MM-HMM based on vibrato features. We use the continuous density HMM with four states and two Gaussian mixtures per state for all HMM models in our experiments. Using training database, we train the three HMM models \( \Lambda_s = \{ \lambda_s^V, \lambda_s^I, \lambda_s^{CBO} \} \) for each singer \( s \), to form the MM-HMM classifier. The process of SingerID is carried out by using the vocal segments which are selected by the hypothesis testing method. Then, OFCCPRA features are calculated from 20ms windowing frames, with 13ms overlapping, from each vocal segment. Since we have 3 models for each singer, the MM-HMM classifier consists of 36 models for 12 singers. During identification, a SingerID decision is made on every 5-second subsegment. Each subsegment is matched with all the 36 models in the classifier, and is assigned to the model that gives the best match. The SingerID error rates are reported for all the 5-second subsegments in the test set.

| Table 3: Average SingerID error rates over 12 singers using 4 different feature sets. |
|---|---|
| Features | Error Rates (%) |
| OFCCPRA | 17.9 |
| OFCCTRI | 19.1 |
| LPCC | 21.1 |
| MFCC | 21.5 |

To compare the SingerID performance of OFCCPRA features with performance from traditional features,
experiments are conducted using Octave Frequency Cepstral Coefficients that employ triangular bandpass filters (OFCC_Trl), Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictive Cepstral Coefficient (LPCc). Results are shown in Table 3. The results show that the OFCCp feature, with an average error rate of 17.9%, outperforms all the traditional features. It is observed that employing parabola bandpass filters for vibrato characteristics formulation gives us 6.3% relative, or 1.2% absolute error reduction over triangular bandpass filters.

Next, we investigate the effectiveness of employing song structure information in singer modelling. First, we establish a baseline HMM, which only has one single HMM for each singer regardless of their song composition. With 2 mixtures per state, the results presented in Table 4 show that the MM-HMM training method outperforms the baseline HMM. Note that the MM-HMM has 3 times as many free parameters as the baseline HMM. For a fair comparison, we perform further experiment using baseline HMM with 6 mixtures per state. The results show that using more free parameters does not help the baseline HMM. The outstanding performance of MM-HMM can be attributed to the composition-based singer modelling.

| Table 4: Performance comparison between MM-HMM training and baseline HMM training |
|---------------------------------|-----------------|
| MM-HMM (2 mixtures/state)       | 17.9            |
| Baseline HMM (2 mixtures/state) | 19.2            |
| Baseline HMM (6 mixtures/state) | 19.8            |

6. Conclusion

We have presented an approach for automatic singer identification of pop songs. The proposed approach explores acoustic characteristics of singers present in the compositions of a song. Our contributions to the SingerID task can be summarized as follows: 1) we employ a statistical method to select vocal segment with high confidence score for singer identification; 2) we propose to use vibrato to represent singer’s characteristics; 3) We propose a method for multi-model HMM singer modeling based on the song composition type. For the first time, we have analytically and experimentally demonstrated the effectiveness of vibrato features in SingerID task. Experimental evaluations conducted on songs from commercially available CD records have shown the superiority of the proposed musical acoustic feature parameter formulation and singer modeling over conventional methods. We conclude that the vibrato-based singer feature is effective, and MM-HMM model has successfully made use of song structure information in SingerID.

7. References