Exploring Perceptual Based Timbre Feature for Singer Identification

Swe Zin Kalayar Khine, Tin Lay Nwe, and Haizhou Li

Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613
{zkkswe,tlnma,hli}@i2r.a-star.edu.sg
http://www.i2r.a-star.edu.sg

Abstract. Timbre can be defined as feature of an auditory stimulus that allows us to distinguish the sounds which have the same pitch and loudness. In this paper, we explore timbre based perceptual feature for singer identification. We start with a vocal detection process to extract the vocal segments from the sound. The cepstral coefficients, which reflect timbre characteristics, are then computed from the vocal segments. The cepstral coefficients of timbre are formulated by combining information of harmonic and the dynamic characteristics of the sound such as vibrato and the attack-decay envelope of the songs. Bandpass filters that spread according to the octave frequency scale are used to extract vibrato and harmonic information of sounds. The experiments are conducted on a database of 84 popular songs. The results show that the proposed timbre based perceptual feature is robust and effective. We achieve an average error rate of 12.2% in segment level singer identification.

Keywords: Timbre, Singing Voice Detection, Vibrato, Harmonic.

1 Introduction

The rapid evolution of the digital multimedia technologies in computer and Internet technology has enabled huge multimedia database. With these continually growing databases, automatic music information retrieval (MIR) has become increasingly important. Singer Identification (SingerID) is one of the important tasks in MIR. It is the process of identifying the singer of a song. In general, a SingerID process comprises three major steps.

The first step is detecting singing segments (vocals) in a song. Vocals can be either pure singing voice or a mixture of singing voice with background instrumentals (nonvocals). The second step is singer feature computation. Features are extracted from vocal segments. The last step is formulating singer classifier using feature parameters. In this paper, we propose new solutions for the second step of singer feature computation.

Earlier studies in SingerID use features such as Mel Frequency Cepstral Coefficients (MFCC) [6]. Recently, studies start looking into perceptually motivated features which are able to appreciate the aesthetic characteristics of singing voice

R. Kronland-Martinet, S. Ystad, and K. Jensen (Eds.): CMMR 2007, LNCS 4969, pp. 159–171, 2008.

for music content processing and analysis. For example, vibrato motivated acoustic features are used to identify singers in [1],[11]. Beside vibrato, harmonic is also a useful feature for SingerID. In fact, harmonics of soprano singer's voice are widely spaced in the spectrum in contrast to that of bass singer's voice [8]. Hence, harmonic spectrum is useful to differentiate between low and high pitch singers.

One of the basic elements of music is timbre or color. Timbre is the quality of sound which allows human ears to distinguish among different types of sounds [15]. Cleveland [3] states that an individual singer has a characteristic timbre that is a function of the laryngeal source and vocal tract resonances. Timbre is assumed to be invariant with an individual singer. On the other hand, Erickson [6] states that the traditional concept of an invariant timbre associated with a singer is inaccurate and that vocal timbre must be conceptualized in terms of transformations in perceived quality that occur across an individual singer's range and/or registers. In general, these studies suggest that timbre is invariant with an individual singer or there is a particular range of timbre quality associated to an individual singer. In this paper, we would like to study the use of timbre based features in SingerID task. Poli [9] measured the timbre quality from spectral envelope of MFCC features to identify singers. In [17], timber is characterized by the harmonic lines of the harmonic sound. In this paper, we propose determining timbre by the harmonic content of a sound and the dynamic characteristics of it such as vibrato and attack-decay envelope [15].

The rest of this paper is organized as follows. In section 2, we present the methods for vocal detection. In section 3, we study perceptually motivated acoustic features and their characteristics. In section 4, we describe the popular song database, experiment setup and results. Finally, we conclude our study in section 5.

2 Vocal Detection

We extract vocal segments from songs. Subband based Log Frequency Power Coefficient (LFPC) [10] is used as acoustic features. We train hidden Markov models (HMM) as vocal and nonvocal acoustic models to detect vocal segments. Vocal detection errors can affect SingerID performance. To reduce the vocal detection error, we formulate the vocal detection as a hypothesis test [12] based on confidence score. In this way, we only retain vocal segments, with which the acoustic models have high confidence. Vocal segments with confidence measure which are higher than a predetermined threshold are accepted for the SingerID experiment.

3 Acoustic Features

We next study several perceptually motivated features, namely harmonic, vibrato and timber features, to characterize song segments. We propose to use subband filters on octave frequency scale in formulating these acoustic features.

3.1 Harmonic

Sopranos have higher fundamental frequency than bass singers. Hence, harmonics of soprano's voice is widely spaced in contrast to that of bass singing. Upper panels of Fig. 1 (a) and (b) show the examples of harmonic spectrum of soprano and bass singing respectively. To capture this information, we implement the harmonic filters with the centre frequencies located at each of the musical notes as in middle panel of Fig. 1 (a) and (b). The list of the frequencies of the musical notes can be found in [7]. Subband filters span up to 8 octaves (16 kHz). Each octave has 12 subbands and there are 96 subbands in total. Output of subband filtering is given in lower panels of Fig. 1. For soprano, lower panel of Fig. 1 (a) shows widely spaced peaks. However, the peaks are narrowly spaced in lower panel of Fig. 1 (b) for bass singers.

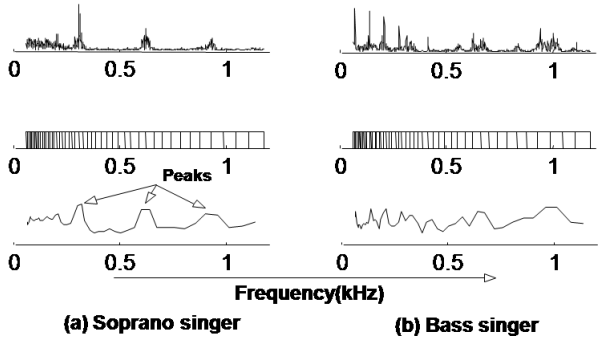

Fig. 1. Harmonics and harmonic filtering

3.2 Vibrato

Vibrato is a periodic, rather sinusoidal, modulation of pitch and amplitude of a musical tone [14]. Vocal vibrato can be seen as a function of the style of singing associated to a particular singer [11].

Vibrato is characterized by two parameters: the extent or excursion and the rate as illustrated in Fig. 2 (a). Female singers tend to have a slightly faster mean vibrato rate than male singers [4]. Vibrato excursions occurring at the tone D6 for three different singers are shown in Fig. 2 (a), (b) and (c). In Fig. 2(c), vibrato excursions to the up and down from the note is balanced. However, unbalanced vibrato excursion can be seen in Fig. 2 (a) and (b). According to [2], such irregular vibrato excursions are very common in most of the tones. In [5], vibrato extent is categorized into two different types, 'wobble' and 'bleat'. Wobble has a wider pitch fluctuation and slower rate of vibrato as in Singer-A. However, bleat has a narrower pitch fluctuation and faster rate as in Singer-C. Hence, the information such as 1) regularity or irregularity in vibrato excursion, 2) two different vibrato types of 'wobble' and 'bleat', and 3) vibrato rate is integrated into acoustic feature.

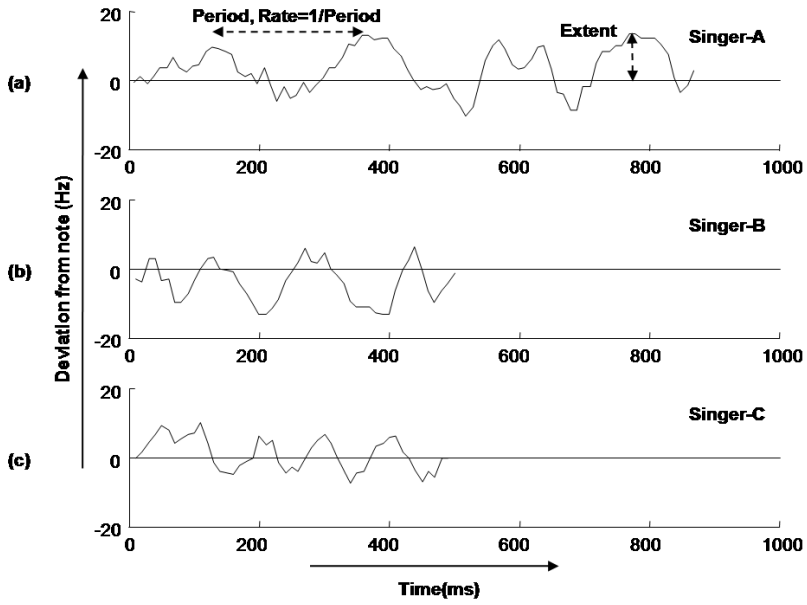

Fig. 2. Vibrato waveforms of 3 singers at note D6, 1174.6Hz

As a result of the modulation of pitch, the frequencies of all the overtones vary regularly and in sync with the pitch frequency modulation [13]. Therefore, we implement the subband filters with the center frequencies located at each of the musical notes to characterize the vibrato. The list of the frequencies of the musical notes can be found in [7]. Due to the fact that singing voice contains high frequency harmonics [16], our subband filters span up to 8 octaves (16kHz). Our subband filters are implemented with a bandwidth of ± 1.5 semitone from each note since vibrato extent can increase more than ± 1 semitone when a singer raises his/her vocal loudness [13]. We employ cascaded subband filters (referred to as vibrato filter) [11] to capture vibrato information from acoustic signal.

In Fig. 4, the upper panel shows spectrum partial. The middle panel presents the frequency response of the vibrato filter. The lower panel demonstrates the instantaneous amplitude output of the vibrato filter. With the output from the vibrato filter, we are able to track the local maxima to derive the vibrato extent [11]. We illustrate the vibrato filter [11] and subband outputs in Fig. 3 and Fig. 4. The filter has two cascaded layers of subbands. The first layer has overlapped trapezoidal filters. The second layer has 5 non-overlapped rectangular filters of equal bandwidths for each trapezoidal subband. Trapezoidal filters are tapered between ± 0.5 semitone to ± 1.5 semitone. The vibrato fluctuations are observed by tracking the local maxima in the instantaneous amplitude output of the subbands in the second layer as shown in the lower panel of Fig. 4. Local maxima indicate the position of the vibrato. The distance between the center frequency of the corresponding filter and the local maxima informs the vibrato

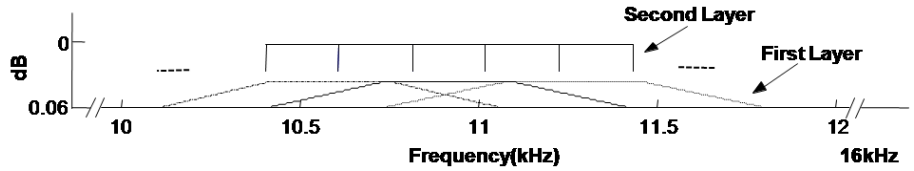

Fig. 3. A bank of cascaded subband filters

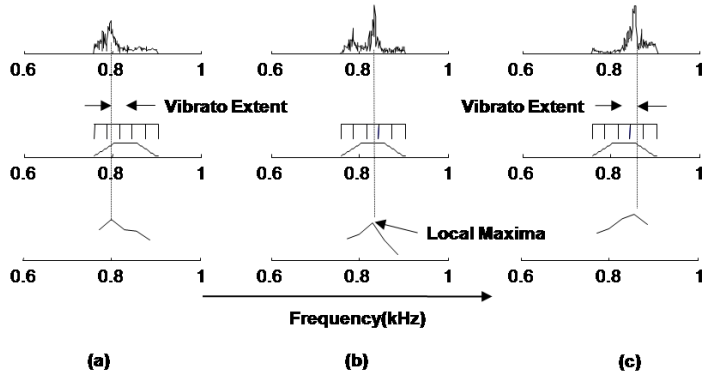

Fig. 4. Vibrato fluctuations and vibrato filtering observed at the note G#5, 830.6Hz. (a) vibrato fluctuates left (b) no fluctuation (c) vibrato fluctuates right.

extent. The tapered and overlapped trapezoidal filters in the first layer allow vibrato fluctuations of adjacent notes observed at the output of the subbands in the second layer to be 'continuous'. The vibrato filter captures irregularities or regularities in vibrato excursion and 'wobble' or 'bleat' vibrato types.

3.3 Timbre

Sounds may be generally characterized by pitch, loudness and quality. For sounds that have the same pitch and loudness, sound quality or timbre describes the characteristics which allow human ears to distinguish among them. Timbre is a general term for the distinguishable characteristics of a tone. Timbre is mainly determined by the harmonic content of a sound and the dynamic characteristics of the sound such as vibrato and attack-decay envelope of the sound [15]. Tone quality or timbre seems most strongly related to the physical phenomena of unfolding partials in the spectrum of a sound or the spectral envelope which distinguishes between two different instruments playing the same note at the same amplitude. The spectral envelop consists of the basis for our tonal judgement [18]. Attack-decay processes of two different singers are shown in Fig. 5 (a) and (b). Singer-1's voice takes more time to develop to its peak than that of Singer-2. And, the decay process of Singer-1 is more gradual than that of Singer-2.

Studies found that it takes a duration of about 60ms to recognize the timbre of a tone. If a tone is shorter than 4ms, it is perceived as an atonal click [15].

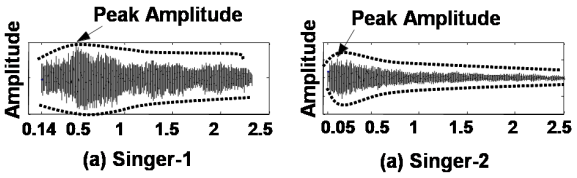

Fig. 5. Attack-decay envelopes

3.4 Cepstral Coefficient Computation

We first divide a music signal into frames of 20ms with 13ms overlapping and apply Hamming window to each frame to minimize signal discontinuities at the end of each frame. Each audio frame is passed through harmonic filters for harmonic content analysis to derive log energy of each band. Finally, we compute a total of 13 Octave Frequency Cepstral Coefficients (OFCC_{har}) from the log energies using Discrete Cosine Transform. We then replace the harmonic filters with vibrato filters to compute the OFCC_{vib} coefficients. To account for timbre characteristics, output log energy of vibrato filter is augmented by that of harmonic and Mel-scale filters. Then, we compute 13 TimBre Cepstral Coefficients (TBCC). We augment the feature coefficients with time derivatives or delta parameters from two neighbouring frames to capture temporal information. For example, delta parameters take care of vibrato rate and attack-decay envelope in OFCC_{vib} and TBCC respectively.

3.5 Formulating Singer Identification as Verification

Traditionally, singer identification system is formulated as a pattern classification problem,in which we find singer model that has the highest likelihood score. The likelihood score measures the similarity between model and test samples. However, it does not take discriminative information between singers into accounts. Here, we propose formulating the identification as a verification task [19] in which we use likelihood ratio instead of likelihood score for decision making. Let O be the sequence of input feature vectors representing a vocal segment from a song of target singer group, otherwise known as the observation. Following the statistical hypothesis test theory, we define two hypothesis. The true hypothesis H_1 is that O belongs to the target singer model λ^m. And the false hypothesis H_0 is that O belongs to non-target model λ^k. We have the following decision rule.

$$p(H_1) \underset{\text{Reject}}{\overset{\text{Accept}}{\gtrless}} p(H_0) \tag{1}$$

When it is applied in singer verification, equation (1) can be rewritten in the form of posterior probability:

$$p(\lambda^m|O) \underset{\text{Reject}}{\overset{\text{Accept}}{\gtrless}} p(\lambda^k|O) \tag{2}$$

where $p(\lambda^m|O)$ is the posterior probability of target singer model λ^m given the song segment O. We apply Bayesian decision rule, equation (2) can be formulated as follows.

$$p(O|\lambda^m)P(\lambda^m) \underset{\text{Reject}}{\overset{\text{Accept}}{\gtrless}} p(O|\lambda^k)P(\lambda^k) \tag{3}$$

where $P(\lambda^m)$ and $P(\lambda^k)$ are the priori probabilities of O to be target singer group or non-target singer group respectively. In our system, $P(\lambda^m)$ and , $P(\lambda^k)$ are assumed equiprobable. Equation (3) can be rewritten in terms of likelihood ratio as in equation (4).

$$\frac{p(O|\lambda^m)}{p(O|\lambda^k)} \underset{\text{Reject}}{\overset{\text{Accept}}{\gtrless}} \frac{P(\lambda^k)}{P(\lambda^m)} \tag{4}$$

In this way, we use a likelihood ratio instead of a likelihood score for decision making.

Now, let us re-examine equation (4) by considering the total cost of a singer identification system. The total cost is defined as follows [19]:

$$C = C_{\lambda^k|\lambda^m}.P(\lambda^m).p(\lambda^k|\lambda^m) + C_{\lambda^m|\lambda^k}.P(\lambda^k).p(\lambda^m|\lambda^k) \tag{5}$$

where $C_{\lambda^k|\lambda^m}$ and $C_{\lambda^m|\lambda^k}$ are the costs of a false rejection and a false acceptance respectively. $p(\lambda^k|\lambda^m)$ and $p(\lambda^m|\lambda^k)$ are the probabilities of a false rejection and a false acceptance yielded by the system. Assuming the costs of $C_{\lambda^k|\lambda^m}$ and $C_{\lambda^m|\lambda^k}$ are the same, equation (5) reflects the error rate averaged over the test database. We find that the decision strategy of equation (4) leads to minimization of this total cost in equation (5).

4 Experiments and Discussion

We compile a database of 84 popular songs from commercially available CD albums of 12 solo English and Chinese singers. The titles of the albums are shown in Table 1. A total of 7 songs are extracted from each album. Four songs of each singer are allocated to TrainDB and the remaining 3 songs to TestDB. Vocal and nonvocal segments of each song are manually annotated to provide the ground truth. The sampling frequency of the song is 44.1 kHz and 16 bit per samples. We define error rate (ER) as the number of errors divided by the total number of test trials.

We first segment a song into 1 second fixed-length segments. Then, each segment is classified as vocal or nonvocal class using the method mentioned in section 2. The average vocal/nonvocal classification error rate of vocal detection system is reported at 8.7%.

Table 1. Singers and Album Titles for 84 Popular Songs

Number	Singer	Album
1	Michael Bolton	Vintage
2	Richard Marx	My own Best Enemy
3	Adu	Tian Hei
4	Ou De Yang	Ocean
5	Jay Chou	Qi Li Xiang
6	Kathryn Williams	Relations
7	Agnetha Faltskog	My colouring book
8	Jennifer Lopez	J. Lo
9	Shania Twain	Come on over
10	Gabrielle	Play to win
11	Madonna	Like a virgin
12	Dido	Life for rent

Using the vocal segments from vocal detector, SingerID experiments are further conducted. We use the continuous density HMM with four states and two Gaussian mixtures per state for all HMM models in our experiments. Using the TrainDB, we train a singer model, λ_s, for each of 12 singers. To identify singer for a vocal segment O, we estimate the likelihood score of O being generated by each of 12 singer models. The model with the highest likelihood suggests the singer identity. We conduct experiments to compare the SingerID performance of five feature types, namely, TBCC,OFCC$_{vib}$, OFCC$_{har}$, MFCC, and LPCC. Window size is 20ms and frame shift is 7ms in all tests.

The number of misidentified singing voice errors for 12 singers are listed in Fig. 6. Each feature type is extracted from 1s segments and there are 5089

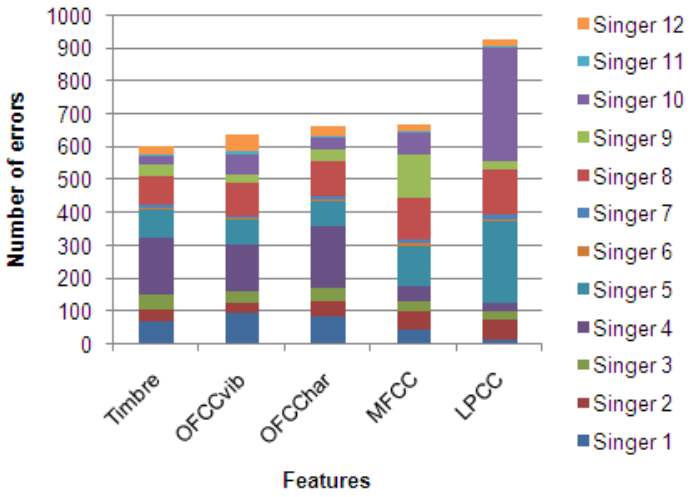

Fig. 6. Number of misidentified test samples on TestDB by 5 feature types

total segments in TestDB. From Fig. 6, we observed that the performance of the singer identification depends on the effect of song content and the singer. In general, the same gender of singers can be confused and the singers with light instrumental background achieve higher accuracy than the singers with strong instrumental background.The minimum error rate of 0.48% (Kathryn Williams) and the maximum error rate of 28.85% (Ou De Yang) are obtained based on the characteristics of the songs.

Table 2. Error rate (ER%) of SingerID on TestDB

TBCC	$OFCC_{vib}$	$OFCC_{har}$	MFCC	LPCC
12.2	12.8	13.6	12.9	21.9

Table 2 shows that the TBCC feature, with an average error rate of 12.2%, outperforms all other features. It is observed that timbre based features capture the singer characteristics well by 5% and 10.3% relative error reduction over $OFCC_{vib}$ and $OFCC_{har}$ features. Furthermore, TBCC feature perform well by 5.4% to 44.3% relative error reduction over traditional features such as MFCC and LPCC. Perceptual based acoustic features (TBCC, $OFCC_{vib}$ and $OFCC_{har}$) in general give better results than the traditional features.

A false alarm rate (FAR) is calculated to be aware of the discriminating ability of our singer models. FAR is defined as the number of false alarms to singer divided by the singer's total negative test trials. All the negative test samples for each singer are collected and so a negative sample for one singer can be a negative sample for another singer in two independent tests. From the 5089 test

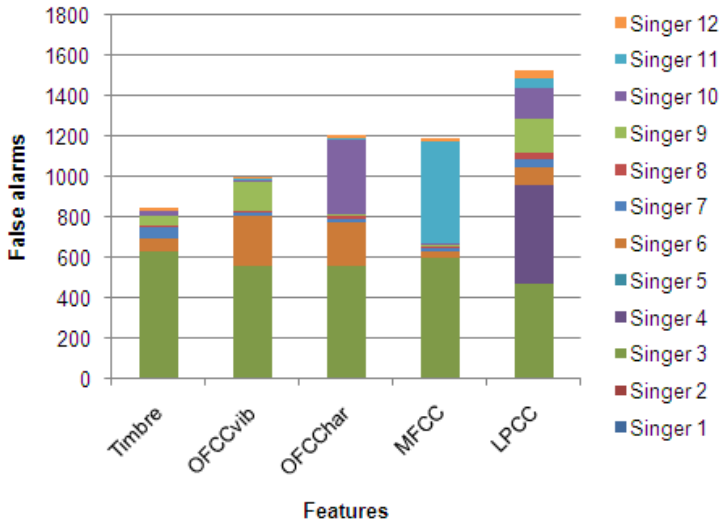

Fig. 7. Number of false alarms on TestDB by five feature types

Table 3. Average error on TestDB by five feature types

TBCC	$OFCC_{vib}$	$OFCC_{har}$	MFCC	LPCC
1.5	1.78	2.15	2.13	2.72

Table 4. Error rate (ER%) on TestDB with different window sizes of TBCC

Window size (ms)	ER (%)
59	12.1
60	15.3
61	15
63	11.9
64	15.4
65	13.6

samples, we come up with 55979 negative samples in total. The accumulated counts of false alarms are illustrated in Fig. 7.

The performance of TBCC feature consistently achieve better than the traditional spectral features in terms of both error rate and false alarm rate in Table 2 and Table 3. As mentioned in section 3.3, duration of about 60ms is necessary to recognize the timbre of tone. Hence, the error rate of 12.2% may not give the optimal performance for TBCC since window size is only 20ms. We further conduct several experiments with TBCC feature of different window sizes, with fixed frame shift of 7ms in all tests. In Table 4, the results show that performance peaks when window size is 63ms, giving the best error rate of 11.9%. It is observed that timbre based features capture the singer characteristics well by 0.9% and 1.7% (or 7% and 12.5% relative) error reduction over $OFCC_{vib}$ (reported earlier in [11]) and $OFCC_{har}$ features respectively. Furthermore, TBCC feature perform well by 5.4% to 44.3% relative error reduction over traditional features such as MFCC and LPCC. We believe that a window size of around 60ms is suitable to extract timbre characteristics from a music signal.

Both vocal and instrumental sounds have musical characteristics such as harmonic, vibrato, and timbre. This gives rise to a question as to whether the three features: TBCC, $OFCC_{vib}$ and $OFCC_{har}$, capture these musical characteristics from either vocal or instrumental sound. To look into this, we conduct SingerID experiments using manually annotated nonvocal segments. SingerID system performance using a) vocal segments, b) nonvocal segments are presented in Table 5.

Table 5. Average error rates using (a) vocal segments and (b) nonvocal segments

Features	Case (a)	Case (b)
TBCC	12.2	60.1
$OFCC_{vib}$	12.8	66.7
$OFCC_{har}$	13.6	60.6

Without surprise, in the presence of vocal timbre, vibrato and harmonic, the three features work the best as in Case (a). With absence of vocal timbre, vibrato and harmonic in Case (b), the error rate increases. This is because the singing voice usually stands out of the background musical accompaniments [13] and the three features are able to capture musical characteristics from vocal rather than from background instruments.

Inspired by speaker verification research [19], we conduct further singer verification experiments using likelihood ratio as in equation (5). We use TBCC feature with 20ms window size in this experiment. As mentioned above, we train each singer model for each of 12 singers. In addition, we train a Universal Background Model to represent non-target singer group using UBM-DB. During testing, each segment is evaluated against all the 12 models in the classifier, and is assigned to the model that gives the best match, as formulated in equation (4). Results in Table 6 show that the performance is improved by using likelihood ratio in a verification hypothesis test.

Table 6. Average Error Rate (ER) on TestDB with and without verification strategies

Method	ER(%)
without verification strategies	12.2
with verification strategies	11.4

In Fig. 8, we present the ER curve of the singer identification system after applying verification strategies. It is observed that the equal error rate (EER) is at 8.9%. It is worth noting that, with verification strategies, we have obtained average error rate of 11.4% (see Table 6) which is very close to the true optimum. With that, we believe that the proposed features and decision strategy are reliably effective.

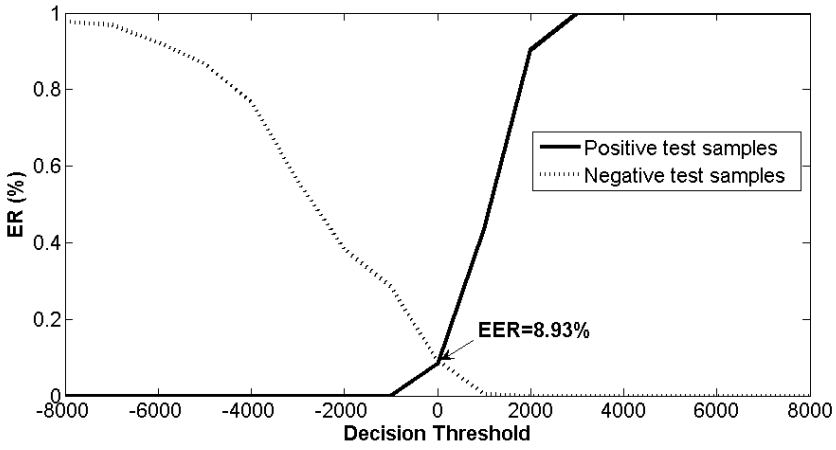

Fig. 8. Equal Error Rate (EER%) of singer identification using likelihood ratio scores

5 Conclusions

We have presented an approach for singer identification of popular songs. The proposed approach explores perceptually motivated timbre characteristics for SingerID. The main contributions of this paper are summarized as follows: 1) we propose using several perceptually motivated features using harmonic, vibrato and timbre information to represent the singer's characteristics. 2) With these features, we found that there is a strong correlation between singer characteristics and system performance. 3) We successfully apply speaker verification techniques into singer identification to achieve better system performance. We conclude that perceptually motivated features especially timbre features are effective in improving system performance.

References

1. Bartsch, M.A., Wakefield, G.H.: Singing Voice Identification Using Spectral Envelope Estimation. IEEE Transactions, Speech and Audio Processing 12, 100–109 (2004)
2. Bretos, J., Sundberg, J.: Measurements of Vibrato Parameters in Long Sustained Crescendo Notes As Sung by Ten Sopranos. Journal of Voice 17, 343–352 (2003)
3. Cleveland, T.F.: Acoustic Properties of Voice Timbre Types and Their Influence on Voice Classification. Journal of Acoustical Society of America 61, 1622–1629 (1977)
4. Dejonckere, P.H., Hirano, M., Sundberg, J.: Vibrato, ch. 2. Singular Pub., San Diego (1995)
5. Dromey, C., Carter, N., Hopkin, A.: Vibrato Rate Adjustment. Journal of Voice 17, 168–178 (2003)
6. Erickson, M., Perry, S., Handel, S.: Discrimination Functions: Can They Be Used to Classify Singing Voices? Journal of Voice 15, 492–502 (2001)
7. Everest, F.A.: Master Handbook of Acoustics. McGraw-Hill Professional, New York (2000)
8. Joliveau, E., Smith, J., Wolfe, J.: Vocal Tract Resonances in Singing: The Soprano Voice. Journal of Acoustical Society of America 116, 2434–2439 (2004)
9. Poli, G.D., Prandoni, P.: Sonological Models for Timber Characterization. Journal of New Music Research 26, 170–197
10. Nwe, T.L., Foo, S.W., De Silva, L.C.: Stress classification using subband based features. IEICE Trans. Information and Systems, Special Issue on Speech Information Processing E86-D(3), 565–573 (2003)
11. Nwe, T.L., Li, H.: Exploring Vibrato-Motivated Acoustic Features for Singer Identification. IEEE Transactions, Audio, Speech and Language Processing 15(2) (2007)
12. Sukkar, R.A., Lee, C.H.: Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. IEEE Trans. Speech and Audio Processing 4, 420–429 (1996)
13. Sundberg, J.: The Science of Singing Voice. Northern Illinois University Press (1987)
14. Timmers, R., Desain, P.: Vibrato: Questions and Answers from Musicians and Science. In: Proc. Int. Conf. On Music Perception And Cognition, England (2000)
15. Winckell, F.: Music, Sound and Sensation. Dover, NY (1967)

16. Zhang, T.: System and method for automatic singer identification. In: Proceedings IEEE International Conference Multimedia and Expo., Baltimore, MD (2003)
17. Zhang, T., Kuo, C.C.J.: Content-Based Audio Classification and Retrieval for Data Parsing. Kluwer Academic Publishers, USA (2001)
18. Helmholtz, H.: On the Sensation of Tone. Dover Publication, New York (1954)
19. Fredouille, C., Bonastre, J.-F., Merlin, T.: Bayesian approach based-decision in speaker verification, A Speaker Odyssey, Crete, Greece (2001)