SINGING VOICE TIMBRE CLASSIFICATION OF CHINESE POPULAR MUSIC

Cheng-Ya Sha,[1] *Yi-Hsuan Yang,*[2] *Yu-Ching Lin*[3] *and Homer H. Chen*[1]

[1]National Taiwan University, [2]Academia Sinica, [3]KKBOX

ABSTRACT

Singing voice plays an important role in the listening experience of music. In this paper, we propose to classify popular music by the timbre quality of the singing voice. Specifically, we adopt six singing voice timbre classes as the taxonomy and build a new data set, KKTIC, that contains the expert annotations of 387 Chinese popular songs. To build an automatic classifier, we resort to signal processing and machine learning techniques and extract a number of singing voice-related features such as vibrato and harmonic-to-noise ratio. We also propose the use of vocal segment detection and singing voice separation as preprocessing steps. Our evaluation identifies the relevant acoustic features and validates the importance of these preprocessing steps. The accuracy in timbre classification reaches 79.84% in a five-fold stratified cross validation.

Index Terms**—** Singing voice timbre, music information retrieval, vocal segment detection, singing voice separation.

1. INTRODUCTION

In order to organize and retrieve the growing collections of digital music, we need automatic systems that index each song with useful information such as genre. In response to this demand, the last decade has witnessed substantial progress towards the classification of music by genre, style, instrument, and emotion [1, 2]. However, singing voice has not yet been fully utilized as a music retrieval method, even though it is an important characteristic of music. Being one of the most versatile musical instruments, singing voice not only adds verbal components to the performance, but also allows a singer to express emotion. For example, a music piece with screaming and roaring voices usually expresses anger, whereas a song with sweet voices tends to evoke positive emotions.

In light of this, we propose a new scheme for music classification in this paper based on singing voice timbre. A data set is built specifically for this classification task. Moreover, we investigate several techniques to classify music by singing voice timber, including the extraction of singing voice-related acoustic features, the use of vocal segment detection to filter out irrelevant (non-vocal) temporal segments in a piece of music, and the enhancement (suppression) of the vocal part (instrument part) in the music piece by singing voice separation. Our study analyzes the contribution of different system components and identifies the setting that leads to the highest accuracy.

In summary, the major contributions of the paper include:

- We propose a new scheme to classify music and build a data set that is made publicly available for research purpose (http://mpac.ee.ntu.edu.tw/dataset/KKTIC).
- We demonstrate that singing voice detection and singing voice separation are helpful front-end operations for singing voice timbre classification.
- We compare different audio feature sets and find that voice-related features perform the best for this task.

2. RELATED WORKS

Research on singing voice has advanced significantly in the last few decades [3]. For instance, Berenzweig *et al.* [4] proposed a machine learning approach to locate singing voice segments. Regnier *et al.* [5] used vibrato and tremolo parameters to detect singing frames. Kim *et al.* [6] proposed the use of vocal segment detection and voice coding features to identify the singer in a song. This singer identification task has also been addressed by using vocal separation and MFCC [7] and by combining accompaniment sound reduction with reliable frame selection [8]. A great amount of efforts have also been made to separate singing voice from the accompanying instruments [9], as the two channels are mixed in most popular songs sold in the market.

Fujihara *et al.* [10] developed a retrieval system that searches for songs having vocal timbres similar to the query song. The system first suppresses the energy of the accompaniment (instrumental) sound, and then extracts speech features such as LPC-derived Mel-cepstral coefficient and delta F0 for song representation. This system differs from our system in that it retrieves songs based on similarity measurement, whereas ours is a learning-based system that captures the high-level semantics of voice timbre.

Turnbull *et al.* [11] built the CAL500 data set that is annotated by students. The data set is made of 502 songs with a tag lexicon of 135 musical semantic concepts, in which 22 are voice timbre-related. Though many studies have used this data set, none of them focuses on vocal timbre classification. In addition, our pilot study shows that the annotation of voice timbre is very sparse in this data set. Moreover, the annotations are noisy as they are entered by paid subjects. In contrast, the data set we develop is annotated by music experts and is specifically designed for singing voice timbre classification.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

Table 1. The singing voice timbre classes and the number of songs classified in each category by experts.

Timbre	Description	#Song
Deep	Low-pitched voice that is usually sexy, confident, and friendly	74
Gravelly	Voice that is hoarse, husky, croaky, mature and mellow	57
Powerful	Full and powerful sound	70
Sweet	Medium-pitched voice that is bright and pleasant, silvery and clear	54
Ethereal	Voice that is clear, healing and distinctively delicate	63
High-pitched	High-pitched voice that is sonorous, passionate, and penetrating	81

3. DATA SET

As there is no standard taxonomy of singing voice timbre, we consult three professional and experienced music editors of KKBOX (http://tw.kkbox.com), a leading cloud-based music service provider in East Asia, to determine a proper taxonomy. We opt for a small number of categories which are representative enough of the universe of singing voice timbre and also easy to understand. The editors must agree upon the definition of the singing voice timbre classes, such that they can provide exemplar songs for each class in a consistent manner. Moreover, to reduce cultural bias, we focus on only Chinese pop songs released in Taiwan and Hong Kong. This selection process results in six classes, which are shown in Table 1 along with verbal descriptions provided by the editors. Each class is associated with 54 to 81 exemplar songs, giving rise to 399 exemplars in total. The six classes are by nature not mutually exclusive, so 11 songs are found to be adequate for two classes. We refer to this data set as KKBOX TImbre Chinese, or KKTIC.

As a singer may perform songs with different styles and singing timbres, the annotation is made per song instead of per artist. To avoid possible bias, we select at most five songs from a singer. In the end, the 387 unique songs are from 91 singers (28 male, 56 female, and 7 group singers).

4. SYSTEM

The flow chart of the proposed system is shown in Fig. 1. Given an input song, we first identify the segments with singing voice using a pre-trained vocal/non-vocal classifier, and then separate the voice signal from the accompaniment using a singing voice separation algorithm. Finally, singing voice-related features are extracted.

4.1. Vocal segment detection (VD)

Some temporal segments of a song are purely instrumental (e.g., intro, bridge, and outro) and are therefore less relevant to singing voice timbre. Instead of dealing with the whole song, it makes sense to neglect these non-vocal segments via a VD algorithm. To this end, a vocal/non-vocal classifier is built by using an in-house collection of 1,019 Karaoke songs. Each song has two channels: one is accompaniment-only $\mathbf{m}$, and the other is a hybrid of vocal signal $\mathbf{s}$ and the accom-paniment $\mathbf{m}'$. We use this data set because it is easy to differentiate the vocal segments from the non-vocal ones from Karaoke songs and use these segments as training data for the vocal/non-vocal classifier.

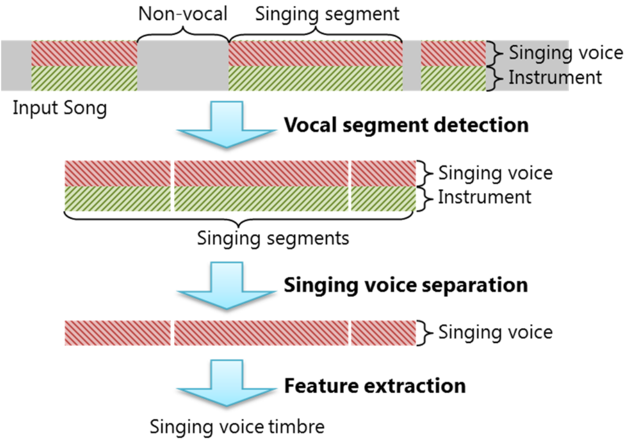

Fig. 1. System flowchart.

Specifically, we identify the vocal segments in these Karaoke songs by exploiting the fact that $\mathbf{m}$ and $\mathbf{m}'$ are usually similar. We adopt the least-mean-square algorithm proposed in [12] to isolate the voice channel $\mathbf{s}'$ from the hybrid $\mathbf{m}+\mathbf{s}'$. Moreover, since the frequency range of human singing voice is seldom lower than 80 Hz [13], the frequency components lower than 100 Hz is removed from $\mathbf{s}$. We consider the segments in $\mathbf{s}$ whose energy is smaller than the mean value plus 1/5 standard deviation of the energy values of all frames in that song as vocal segments. We then re-combine the two channels and use the vocal/non-vocal information estimated from the above procedure as ground truth label and then train a support vector machine (SVM) binary classifier.

For smoothness and efficiency, we consider 3-second as the basic unit for VD. We randomly pick 1/2 of the 30,363 3-second segments from the monauralized Karaoke data set and extract MFCC for feature representation [14]. Cross-validation on the remaining 1/10 segments obtains accuracy of 73.62%, which is close to the state-of-the-art [14].

4.2. Singing voice separation (SP)

Unlike Karaoke music, the singing voice and accompany-ment are mixed in both channels for consumer available music. Therefore, although VD helps identify segments with

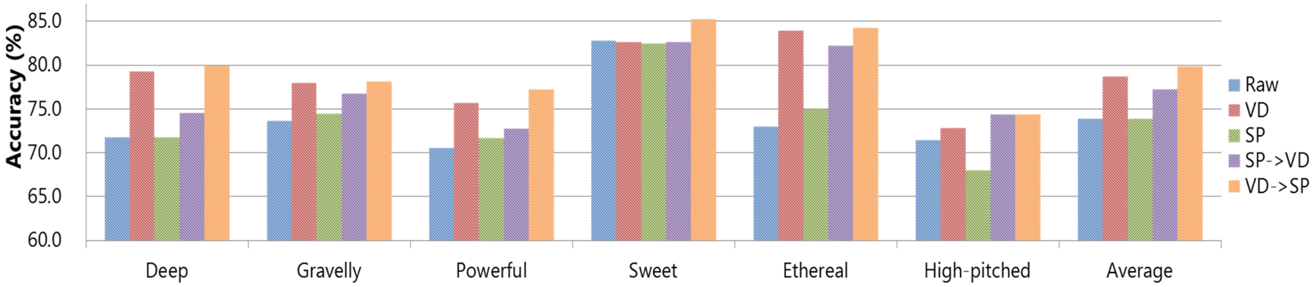

Fig. 2. Classification accuracy of the system using different preprocessing steps.

singing voice, the identified segments (e.g., for KKTIC) still contain instrumental sounds.

Because we are concerned about singing voice timbre, we investigate whether we can obtain better result by enhancing (suppressing) the vocal (instrument) part in the music piece by singing voice separation algorithms. To this end, the robust principle component analysis (RPCA) is adopted, for its excellent performance shown in recent work [9]. Given an input song, we first compute its *N*-point short-time Fourier Transform (STFT) to obtain the spectrogram $X = Me^{jP}$, where $M \in \mathbb{R}^{f \times t}$ is the magnitude and $P \in \mathbb{R}^{f \times t}$ is the phase. We then apply RPCA to decompose M into a low-rank matrix L and a sparse matrix S by solving the following optimization problem,

$$\min_{M=L+S} \|L\|_* + \lambda \|S\|_1, \qquad (1)$$

where $\|\cdot\|_*$ and $\|\cdot\|_1$ denote the nuclear norm (sum of the matrix singular values) and the l_1 norm (sum of absolute values of the matrix entries), respectively. The parameter λ is set to $\lambda_0 = 1/\sqrt{\max(\mathrm{f},\mathrm{t})}$ as recommended in [15] for a balance between the two terms. It has been found that after this decomposition, L and S, respectively, corresponds well to the background accompaniment and the singing voice, possibly because of the repetitive nature of the instruments and the sparseness of singing voice in the time-frequency domain [9]. Therefore, we can recover the time domain signal of the singing voice via the inverse-STFT of Ee^{jP}. Due to space constraint, we refer the reader to [9] and the reference therein for the details of solving Eq. (1).

4.3. Feature extraction

To better catch the singing voice traits, we represent the audio content with features that are designed specifically for voice signals [16, 17]. Specifically, we use the Opensmile toolkit [18] to extract audio features such as jitter and shimmer (both are related to vibrato) from the 30th second to the 60th second segment of each song and pool the extracted frame-level feature vectors into a song-level vector by taking the first four moments. For comparison, we also use the MIRToolbox [19] to extract music-specific features such as tempo and chromagram. See Table 2 for dimensions and brief descriptions of the extracted features.

Table 2. The feature sets. The number in the parenthesis after each feature set denotes the dimension.

General audio features (166D)	
Dynamics (18D)	RMS energy (root-mean-square energy), loudness, ZCR (zero-crossing rate)
Spectral (148D)	Spectral centroid, roll-off, slope, flux, variance, skewness, kurtosis, MFCC
Singing voice features (428D)	
Pitch (50D)	F0 (fundamental frequency), jitter (deviations in pitch period), shimmer (deviations in pitch amplitude) [20]
Voice quality (VQ) (378D)	HNR (harmonic-to-noise ratio), LSP (line spectral pairs; speech coding used to represent linear prediction coefficients), voicing probability, and log Mel-frequency band
Music features (27D)	
Rhythm (5D)	Fluctuation, tempo, pulse clarity
Tonal (22D)	Chromagram, mode, and harmonic change detection function (HCDF) [21]

4.4. Voice timbre classification

We formulate this singing voice classification problem as a multi-label classification task, as each song can belong to more than one class. Specifically, for each of the six classes, we train a binary classifier predicting if a given song belongs to that class. We use SVM with radial basis function kernel as our classifier. In addition, because all of the classes have more negative examples than positive examples, we adopt the under-sampling method *EasyEnsemble* [22] to mitigate the class-imbalance problem.

5. EXPERIMENTS

The performance is evaluated on the KKTIC data set. For fair comparison, each song is converted to 22050 Hz, 16-bit PCM WAV, and mono-channel before any processing. We report the average result of 5-fold cross validation.

5.1. Singing voice detection and extraction

Fig. 2 shows the performance of the system using different preprocessing strategies for feature extraction. We use spectral and singing voice features described in Table 2 in this experiment. The following strategies are compared:

- Raw: use the raw input music directly
- SP: separated singing voice
- VD: segments of singing voice
- SP→VD: voice separation followed by vocal detection
- VD→SP: vocal detection followed by voice separation.

The following observations can be made from Fig. 2. First, voice separation (SP) contributes only minor improvement to the average accuracy. This may due to the interference introduced by the imperfect separation algorithm, especially for songs that have prominent singing voice and weak accompaniment. For example, for voice-prominent songs such as High-pitched songs, the removal of non-vocal segments is enough. On the other hand, features extracted from the remaining non-vocal segments bring negative effects. Second, the use of vocal detection (VD) greatly improves the average accuracy by about 5%. VD only does not work for Sweet songs. Third, the best result (79.84%) is attained by using vocal detection followed by voice separation (VD→ SP). The performance difference between VD→SP and Raw is significant (p-value<0.001) under the t-test. Finally, the accuracy degrades if we separate voice first and then detect the vocal segments (SP→VD), although we have expected that its result would be similar to VD→SP. This degradation is largely due to the errors in SP.

The parameter λ for SP can be adjust to accommodate the properties of the task in hand. The higher the value of λ, the sparser (but sometimes more distorted) the separated singing voice is. In the extreme case when $\lambda = 0$, RPCA acts like an identity function, i.e., it just outputs the input song. On the contrary, if the value of λ is too high, the separation algorithm introduces serious interference and hence leads to negative impact. The best result is obtained with $\lambda = 1.5\lambda_0$.

We note that these preprocessing steps and the parameters thereof can be used in a class-dependent manner, since for some singing timbre classes (e.g. Gravelly and Ethereal) the classification accuracy enhances little when additional SP is taken after VD.

5.2. Singing voice features

Next, we study the performance of different combinations of the feature sets using the VD→SP strategy. As shown in Fig. 3, using the general audio feature set (Dyn+Spec) already performs well, possibly because the inclusion of MFCC in the feature set Spec, which has been proved powerful as an audio signal representation.

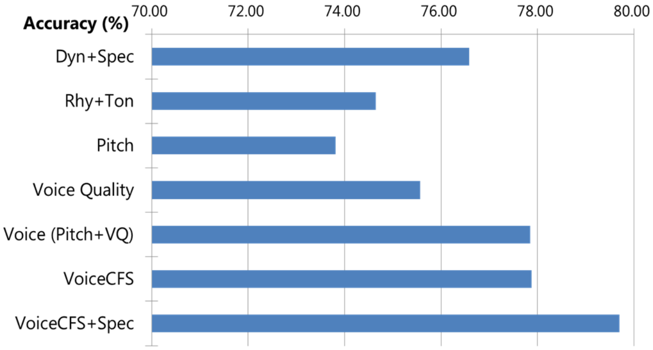

Fig. 3. Accuracy in singing timbre classification for different combinations of the feature sets.

Using only music feature set (Rhy+Ton) does not degrade the result much, partly because tonal features still represent the frequency content of the audio and partly because music and singing voice are usually correlated. For instance, love song singers often use a tender tone to express emotion, while we can expect to hear gravelly voices in heavy metal songs. The pitch feature set only captures the trend and variations of the fundamental frequency, so using pitch alone does not work well. Finally, we see that the voice feature set (Pitch+VQ) outperforms other feature sets, which is in line with our intuition that voice-related features should match the task well.

The dimension of the voice feature set might be too high (i.e., 428D). In view of this, we further use correlation-based feature subset selection (CFS) [23] and reduce the feature dimension to 58D. CFS evaluates the importance of a feature subset by awarding the individual feature predictive ability and penalizing the inter-correlation within the feature subset. Forward best-first selection strategy is used. The combination of Spec and the resultant feature set (VoiceCFS) leads to the best result in our evaluation.

6. CONCLUSION

In this paper, we have defined the problem of singing voice timbre classification and constructed a new data set for the task. We have empirically validated that the use of using vocal segment detection and singing voice separation improves the classification accuracy. In addition, voice features are remarkably effective. Future work will be directed towards using the vocal timbre classification results as an additional feature set for other tasks such as music emotion recognition and music recommendation.

7. ACKNOWLEDGMENT

This work was supported by the NSC of Taiwan under Grant NSC 100-2221-E-002-198-MY3, by KKBOX under Grant 100-S-C63 and by NTU under Grant 100R7609-2.

8. REFERENCES

[1] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *Signal Processing Magazine, IEEE,* vol. 23, pp. 133-141, 2006.

[2] Y. H. Yang and H. H. Chen, *Music Emotion Recognition*: CRC Press, 2011.

[3] M. Kob, N. Henrich, H. Herzel, D. Howard, I. Tokuda, and J. Wolfe, "Analysing and Understanding the Singing Voice: Recent Progress and Open Questions," *Current Bioinformatics,* vol. 6, pp. 362-374, 2011.

[4] A. L. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, 2001, pp. 119-122.

[5] L. Regnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," 2009, pp. 1685-1688.

[6] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," in *Proceedings of the 3rd International Conference on Music Information Retrieval*, 2002, p. 17.

[7] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proc. ISMIR*, 2007, pp. 375-378.

[8] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. ISMIR*, 2005, pp. 329-336.

[9] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-Voice Separation From Monaural Recordings Using Robust Principal Component Analysis," *ICASSP,* 2012.

[10] H. Fujihara and M. Goto, "A music information retrieval system based on singing voice timbre," in *Proc. ISMIR*, 2007, pp. 467-470.

[11] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 467-476, 2008.

[12] H. M. Yu, W. H. Tsai, and H. M. Wang, "A query-by-singing system for retrieving karaoke music," *Multimedia, IEEE Transactions on,* vol. 10, pp. 1626-1637, 2008.

[13] J. C. McKinney, *The diagnosis & correction of vocal faults*: Broadman Press, 1982.

[14] M. Rocamora and P. Herrera, "Comparing audio descriptors for singing voice detection in music audio files," in *Brazilian Symposium on Computer Music, 11th. San Pablo, Brazil*, 2007, p. 27.

[15] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Arxiv preprint ArXiv:0912.3599,* 2009.

[16] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, 2010, pp. 2794-2797.

[17] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Proc. Interspeech*, 2011, pp. 3201-3204.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, 2010, pp. 1459-1462.

[19] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects, Bordeaux, France*, 2007.

[20] L. Xi, T. Jidong, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and Emotion Classification using Jitter and Shimmer Features," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-1081-IV-1084.

[21] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," presented at the Proceedings of the 1st ACM workshop on Audio and music computing multimedia, Santa Barbara, California, USA, 2006.

[22] T. Y. Liu, "Easyensemble and feature selection for imbalance data sets," in *Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS'09. International Joint Conference on*, 2009, pp. 517-520.

[23] M. A. Hall, "Correlation-based feature selection for machine learning," The University of Waikato, 1999.