# Singer Similarity

Bertrand Scherrer

March 16, 2007

## 1 Introduction

This paper presents the state of research on singer identification. After examining the context of singer identification and identifying the challenges it raises, we will follow and explain the usual structure of the singer identification systems found in literature:

1. Vocal/NonVocal Segmentation
2. Feature Extraction
3. Classification

### 1.1 Applications of Singer Identification

Singer identification can be used in a variety of situations. For example, it could allow to automatically label musical data for which no (or not much) information is available and still recognize the singer. This is especially interesting today with the proliferation of unlabeled data on the Internet for popular music. Singer identification could also be used by record companies to identify bootleg recordings of their artists. Finally, music recommendation systems could use singer identification to group singers with same voice characteristics.

### 1.2 Main strategies

Singer recognition can be defined as the group of tasks involved in distinguishing music data from a singer characteristic's database (Tsai and Wang, 2006). Here are some of the strategies available:

- **Singer Identification (SID):** Given a group of candidate singers determine who sang a given part of the song. It can be understood as a N-class decision task from a database of labelled voice data.
- **Target Singer Detection (TSD):** Decide whether or not a given singer performs in a given part of a song. This is a binary classification.
- **Target Singer Tracking (TST):** Determine where, in a given recording, the target singer is singing. TST is TSD as a function of time.

We will mostly encounter SID strategies in the rest of the article with the notable exception of (Tsai and Wang, 2006) which considers all three.

### 1.3 Challenges

One of the main challenges of singer identification comes from the very nature of the singing voice. Indeed, singing voice is in between speech and a musical instrument (Mesaros and Astola, 2005): singing consists mostly of sustained vowels with highly harmonic spectrum like that of an instrument and, at the same time, it requires articulatory techniques related to speech. Thus, it requires new analysis methods that are a combination of speech processing and musical instrument recognition.

The other main problem is the fact that, in popular music[1], it is usually impossible to get a pristine solo voice recording of a singer: there is always a background 'noise' present (Tsai and Wang, 2006). Hence the necessity to identify segments containing vocal information.

# 2 Feature Extraction

Most of the available singer identification methods use frequency domain features extracted from recordings (Kim and Whitman, 2002):

- The Mel-Frequency Cepstral Coefficients (MFCC) are used in (Tsai, Wang, Rogers, Cheng, and Yu, 2003; Tsai and Wang, 2004, 2006), (Fujihara, Kitahara, Goto, Komatani, Ogata, and Okuno, 2005).
- The Modified Discrete Cosine Transform (MDCT) is used in (Liu and Huang, 2002).
- Several variations on Linear Predictive Coding Coefficients were used in (Fujihara et al., 2005) (regular LPCC), (Kim and Whitman, 2002) (warped LPCC), (Zhang, 2003) (cepstral coefficient of the LPC spectrum), (Fujihara et al., 2005) (MFCC of the LPC spectrum).

It is interesting to note that in (Fujihara et al., 2005), the feature extraction is done on the resynthesized output. Another interest of this work resides in its investigation of the best feature to use for singer recognition. It seemed to reveal that LPMFCC (the feature they seem to introduce) are yielding the best results[2].

# 3 Detecting Vocal/NonVocal Regions

This distinction has been identified as a valuable element of singer identification systems in (Berenzweig, Ellis, and Lawrence, 2002). Berenzweig *et al.* trained a neural net to segment vocal/nonvocal regions of radio recordings. They observed that focusing on vocal segments improved the performance of their classification scheme.

The idea behind such a segmentation resides in the observation that the voice spectrum exhibits different characteristics than the accompaniement's spectrum. Indeed, the voice's spectrum has a far more harmonic repartition of energy than the accompaniement's.

## 3.1 GMM-based methods

In (Tsai et al., 2003; Tsai and Wang, 2004, 2006), the principle of the method is the assumption that the accompaniement of singing and instrumental-only portions are often very similar. Hence, it would be possible to devise an *a priori* model for the background and then estimate the solo voice from this model.

In practice, using a database including vocal and non vocal music, one trains a Vocal GMM [3] and a Non Vocal GMM [4] respectively. Once the training is done, the likelihood of each feature vector to be Vocal/NonVocal is computed for each frame. Based on these likelihoods, there are different ways to decide on the nature of the frame. They defer mainly in terms of the analysis interval (frame-based, fixed-length segment, homogeneous-segment). This segmentation best results was 82.3%. Errors were due to misidentification of Vocal segments as NonVocal due to high level of the background compared to the voice.

In (Fujihara et al., 2005), Fujihara *et al.* present a slight variation on this GMM approach where it is assumed that the accompaniement during a sung portion **is not** the same as during an instrumental portion. Hence in order to perform the segmentation, Fujihara introduces another step denoted *accompaniement sound reduction*. It is performed by successively estimating the baseline of the piece, extracting the harmonic structure corresponding to the melody and resynthesizing that signal via additive synthesis. The resulting

[1] which is the subject of most if not all studies
[2] bias ?
[3] order 64
[4] order 80

signal, including harmonic elements of the voice mixed with those of instruments, is then passed through a similar test based on two GMM models (one for vocal sounds the other for non-vocal sounds).

### 3.2 Harmonicity observation approach

The method proposed by Kim and Whitman in (Kim and Whitman, 2002), although it does not seem to perform very well (only 30% of vocal segments were detected at best ), has the merit to be different. It adpts a perceptual point of view, bandpassing the signal in the 200-2000Hz region (sensitive area for the ears) and then using a bank of comb filters with different delay values. From that bank of comb filters, ameasure of harmonicity is devised as the ratio of the signals energy and the minimum energy goign out of the comb filter bank. Then a simple test is done: if a segment of soud exhibits harmonicity above a certain value then it is classified as voice.

## 4 Classification Technique

The features extracted previously are used in a first step to train the classifiers. Three classifiers have been used in the present literature: GMM ((Tsai and Wang, 2004, 2006), (Zhang, 2003), (Fujihara et al., 2005), (Kim and Whitman, 2002)), SVM ((Kim and Whitman, 2002)) and k-NN ((Liu and Huang, 2002)). Since the GMM has been most widely used we are going to detail how it is used.

### 4.1 Gaussian Mixture Models

An example of this approach is developed by Tsai and Wang (2006). It can be summarized as follows:

- The observed feature vectors **on a vocal segment** ($\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2 \dots \mathbf{v}_T]$) can be thought of as an unknown combination of solo voice and background feature vectors ($\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2 \dots \mathbf{s}_T]$) and $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2 \dots \mathbf{b}_T]$ respectively). It is assumed that the solo voice feature vectors as well as the background feature vectors are distributed according to two GMMs: $\lambda_S$ and $\lambda_B$ respectively.
- Given certain assumptions [5], it is possible to derive an expression of the likelihood of the feature vectors grouped in $\mathbf{V}$ as a function of the likelihood of a given vocal segment feature vector given an underlying combination of solo and background models: $p(\mathbf{v}_t|\boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{b,j})$.
- A background GMM is created from the non vocal regions of the all the songs perfromed by one singer. The solo voice model is then determined using Maximum Likelihood Estimation: $\lambda_S^* = argmax(p(\mathbf{V}|\lambda_S, \lambda_B))$. This is done in the typical way using the EM algorithm.

This training is done for different singers to obtain different singer models. In the testing phase, for each frame inside a vocal segment, one will perform the computation of the likelihood of the feature vector for different solo voice model. The identified singer is the one for which the likelihood is the highest.

## 5 Experimentation Results

In most of the cases, the musical data used to test the different SID methods was made of real recordings of pop music coming from different sources: RWC database (Fujihara et al., 2005), the Internet (Kim and Whitman, 2002), Mandarin pop (Tsai et al., 2003; Tsai and Wang, 2004, 2006). The data usually featured, male and female singers, mostly in solo singing as well as instrumental pieces.

The results for singer identification range from a mere 45% (Kim and Whitman, 2002) to 80% (Liu and Huang, 2002) and 95% (Fujihara et al., 2005; Tsai and Wang, 2006).

[5] debatable independence of the vocal and background models for example

# 6 Conclusion

We have seen that the SID problem has been thoroughly studied and some very encouraging results have been presented. Nevertheless, only one study (Tsai and Wang, 2006) tackled the target singer detection and tracking inside a given song. Results were encouraging for duets although not perfect Moreover, most of the time, the data used for training and testing was specific to a style of music (pop) or a geographical location (mainly Asia). There is still room for improvement to provide a singer identification system that would work for other musical genres (many singers a cappela for example).

# References


Berenzweig, A., D. P. W. Ellis, and S. Lawrence. 2002. Using voice segments to improve artist classification of music. In *Proceedings of the AES 22nd International Conference on Virtual Synthetic and Entertainment Audio*.

Fujihara, H., T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. 2005. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proceedings of the International Conference on Music Information Retrieval*.

Kim, Y. E., and B. Whitman. 2002. Singer identification in popular music recordings using voice coding features. In *Proceedings of the International Conference on Music Information Retrieval*.

Liu, C. C., and C. S. Huang. 2002. A singer identification technique for content-based classification of MP3 music objects. In *Proceedings of the 11th International Conference on Information and Knowledge Management*.

Mesaros, A., and J. Astola. 2005. The mel-frequency cepstral coefficients in the context of singer identification. In *Proceedings of the International Conference on Music Information Retrieval*.

Tsai, W. H., and H. M. Wang. 2004. Automatic detection and tracking of target singer in multi-singer music recordings. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 221–224.

———. 2006. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Transactions on Audio, Speech and Language Processing* 14:330–41.

Tsai, W. H., H. M. Wang, D. Rogers, S. S. Cheng, and H. M. Yu. 2003. Blind clustering of popular music recordings based on singer voice characteristics. In *Proceedings of the International Conference on Music Information Retrieval*.

Zhang, T. 2003. Automatic singer identification. In *Proceedings of the 2003 International Conference on Multimedia and Expo*, vol. 1, 33–6.