Singer Identification Using Gaussian Mixture Model (GMM)

Su Lin Wai
University of Computer Studies, Yangon
sulinwai@gmail.com

Abstract

*In this paper, identifying the artist of a query song from the audio database is considered. To build the model of a specific signer, only the vocal segments of a song is employed. Mel-Frequency Cepstral Coefficient (**MFCC**) is used for extracting salient features of each artist. Classification among a group of artists is performed by Gaussian Mixture Model (**GMM** classifier. The desired "singer" may be defined as an individual or group who records or performs popular songs under a particular identification name. After extracting the vocal segments, they are fed into the singer identification system that has been trained on data taken from songs of other album by the same artist. Experiments are carried out a group of singers where the songs are in three different genres.*

1. Introduction

Human auditory physiology and perceptual apparatus have evolved to a high level of sensitivity to the human voice because singing voice is the oldest musical instrument. Singing voice is the main focus of the attention in music pieces with a vocal part; the most people use the singers' voice as the primary cue information identifying a song. Also, a natural classification of music, besides genre, is the singer's name. 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. The recent rise of the field of MIR (Music Information Retrieval) has spawned several works on the topic of singer identification for popular songs.

Singer identification system would be useful for MIR systems, the interdisciplinary science of retrieving information from music, in case of identifying singers for songs. In MIR system, there are three main audiences that are identified as the beneficiaries of MIR: industry bodies engaged in recording, aggregating and disseminating music; end users who want to find music and use it in a personalized way; and professionals: music performers, teachers, musicologists, copyright lawyers, and music producers. Many popular and great singers have voices that are particularly unique and thus often instantly recognizable. The singer's information is essential in organizing, browsing, searching, exploring music data and retrieving music collections. Because the human voice is a personal tool and no two voices are quite the same. The inherent difficulties lie in the nature of the problem: the voice is usually accompanied by other musical instruments and even though humans are extremely skilful in recognizing sounds in acoustic mixtures, interfering sounds usually make the automatic recognition very difficult. Most of the singer identification systems are combined two stages; vocal/non-vocal separation and identification. Identification stage includes feature extraction and classification. In this presented method, the separation of vocal and instrumental-only parts is conducted manually. So the separated vocal parts are mixed with background music.

2. Related Work

In [1], the music is segmented into sub-frames according to the inter-beat-interval. Vocal and instrumental part is separated by the SVM is trained with the 10^{th} order **Octave Scale Cepstral** Coefficients (**OSCCs).** Singer names are identified through **Linear Prediction Cepstral Coefficients (LPCC)** and **Mixture Model** (**GMM**) classifier. The work in [2] is based on the idea of using only the vocal segments of a song to build the model of a particular singer. The borders between vocal and instrumental parts are first detected with the Bayesian Information Criterion (**BIC)**. Then, each segment is classified as vocal or instrumental by a decision tree based on MFCCs.

Having vocal segments located, by training a GMM for each singer. HSI (**Hybrid Singer Identification**)[3], for large databases. Preprocessing module separates music into vocal and non-vocal segments by using SVM. For singer modeling module, GMM is used to model statistic characteristics, where vocal segments are for singer feature and non-vocal segments for music structure. At the end of this system, the classification result is

enhanced by the HSI with a decision model further reducing the misclassification using a neural network.

Li et al. [4] developed some new acoustic features for singer identification that extracted information about the singer's vibrato. Applying several banks of filters (triangular, parabolic and cascaded), and transforming the resulting energies into the cepstral domain, they extracted the **Octave Frequency Cepstral Coefficients** (**OFCC**) and experimented on a 12-singers database. Klapuri et. [5] gives an evaluation of different classification methods in polyphonic case and also separation of the vocal line. Mixtures with various relative levels of the singing and accompaniment were used in order to evaluate the robustness of the methods. Classification strategies include linear and quadratic discriminant functions, GMM based maximum likelihood classifier and nearest neighbor classifiers using Kullback-Leibler divergence between GMMs of the song under analysis and the singers.

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. P. Annesi, etc. al. [7] analyzed the average and standard deviation of 6-dimentional vectors such as volume, bests, spectral, energy, centroid, pitch and 5-MFCC, over the entire song. Then SVM classifier was used on two different data sets to classify.

W. H. Tsai, etc. al. [8] evaluated several feature measurements, including MFCC and Perceptual Linear Prediction (PLP), both with and without their first-order derivatives and Cepstral mean normalization (CMN) was applied to minimize channel-induced perturbations. Then vocal/instrumental discrimination was performed by using MFCC and GMM and the accuracy of the result was obtained from the comparison of manual segmentation and automatic segmentation on test data. M. A. Bartsch [9] used PESCE (Peak, Edge, Strand, and Complex Extractor" as fundamental frequency estimation algorithm, that takes a short audio signal as input and produces fundamental frequency estimates of "voice-like" sources from the signal. First, PESCE was used for the detection of the singing voice within a polyphonic mixture at the time of PESCE returns a fundamental frequency estimate and then the fundamental frequency estimate allowed one to extract time-varying amplitudes for the partials of the voice signal from a time-frequency distribution.

In [10], the authors applied the standard text-independent speaker identification techniques or a singer identification task. He manually collected vocal segments from several music recordings; extracted Linear Prediction derived Cepstral Coefficients (LPCCs) and modeled each singer with a GMM. In this method, the beginnings of the vocal sections were detected using simple threshold settings which were calculated from extracted features i.e. 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.

3. Background

This system is performed by two steps, feature extraction and classification. Feature extraction is the process of converting an audio signal into a sequence of feature vectors carrying characteristics information about the signal. These vectors are used by classification algorithms. In speech and music processing, the time-domain signal is often of substantially less interest than any number of frequency-domain representations. This is primarily because many visible "features" in the time domain are dependent upon the relative phase of slowly-varying sinusoids, which is generally imperceptible to human listeners. However, a variety of time-domain statistics have been proposed.

Mel-frequency Cepstral Coefficient (MFCC) is the short-term spectral decomposition of an audio signal that conveys the general frequency characteristics important to human hearing. In classification stage, previously unknown input data is assigned to a class, such assignment is made by decision rule. In this paper, Gaussian Mixture Model (GMM) will be used as classification method. GMM classifier is a type of classifier which combines the advantages of parametric and non parametric methods. GMM does not require storage of entire training vectors to make a classification. It is a very flexible model that can adapt to encompass almost any distribution of data.

3.1. Mel-Frequency Cepstral Coefficients (MFCCs)

The MFCCs are well known compact forms that can represent speeches. They are the most common representation used for *Spectra* in Music Information Retrieval (MIR). The first step in this process is to block a continuous audio signal into frames. The second step is to use a window function on each individual frame in order to minimize discontinuities at the beginning and end of each frame. The third step is the process of converting each frame from the time domain to the frequency domain. The fourth step is the transformation of the real frequency scale to the Mel frequency scale (unit of measure of perceived pitch or frequency of a tone). The final step is the log Mel spectrum is converted back to the time domain and the result is the Mel Frequency

Cepstral Coefficients. Step by step calculation of MFCC is illustrated with Figure 1. To calculate the Mel frequency scale, the following formula has been used.

$$\mathrm{Mel}(f) = 2595 \log_{10}(1 + f/700) \quad (1)$$

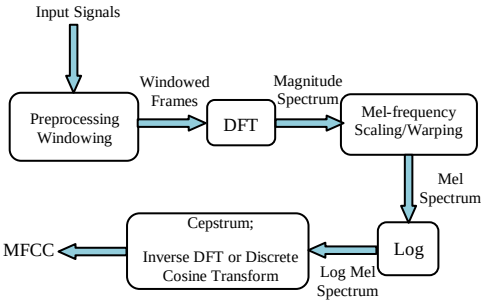

Figure 1. Block diagram for computing MFCCs.

3.2. Gaussian Mixture Model (GMM)

The GMM classifiers combine some of the benefits of both the k-NN and quadratic classifiers. Like quadratic classifiers, they employ a trainable model that done not require all of the training data to make a classification. Like kNN classifiers, GMM classifiers with sufficiently high order can approximate any distribution with arbitrary accuracy. GMM is a very flexible model that can adapt to encompass almost any distribution of data. The Gaussian means were first initialized by using the k-means clustering and then the model is refined using the Expectation Maximization algorithm.

GMM is a model of the probability density function for given set of data and has the form of a sum of individual Gaussian component, each possessing its own mean and covariance. The Gaussian Probability (pdf) of x for the i^{th} state:

$$f(x) = \sum_{j=1}^{J} \frac{P(j)}{(2\pi)^{p/2} \left| \sum_j \right|^{1/2}} e^{-1/2(x-\mu_j)^T \Sigma_j^{-1} (x-\mu_j)} \quad (2)$$

To calculate mean $\hat{\mu}$ and covariance $\hat{\Sigma}$, the following equations are used.

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad (3)$$

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^T (x_i - \mu) \quad (4)$$

The usage number of GMM classifiers can be calculated by the following equation;

$$number \quad of \quad GMM = \frac{n(n-1)}{2} \quad (5)$$

where n = number of classes.

Expectation Maximization (EM) Algorithm

EM is an iterative algorithm that converges on parameters that are locally optimal according to the log-likelihood function. This algorithm includes two main steps; estimation and maximization steps.

Estimation step: compute the probability for each data point to belong to the class and so this step represents a soft classification, since a point can belong.

Maximization step: update the means, update the variances and finally update the priors.

The EM algorithm is best suited for fitting Gaussian clusters and it is an easy task to guess the parameters for initialization. With the EM algorithm the discriminate surfaces have the form of (hyper) parabola. In Figure 2, the general flow of classification algorithm using GMM can be seen.

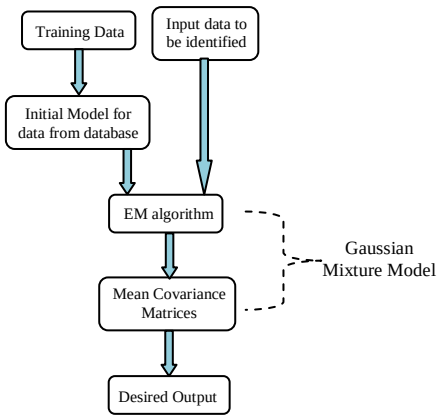

Figure 2. GMM classification algorithm.

4. Singer Identification Method

This signer identification method is composed with two main procedures: training and testing. In training stage, the following steps are processed.

- The input music is manually separated into vocal and instrumental parts by manually.
- The extracted input vocal segments for training are represented with coefficients using 13 MFCC coefficients from 20 ms frames.
- These coefficients are filtered EM algorithm and finally, the initial model for this system is saved in the database.

Illustration of training stage is given in Figure 3.

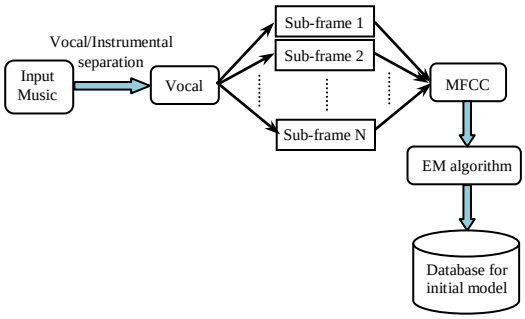

Figure 3. Training stage.

For testing, four steps are needed. The diagram of testing phase is shown in Figure 4.

- To know the artist of a song, vocal music excerpts are manually extracted as in the training phase.
- The vocal parts are then altered into required coefficients by using MFCC.
- The coefficients, the result of MFCC, are classified together with the initial model result from training data by using GMM classifier.
- After classification is finished, this identification method produces the desired output; the singer's name.

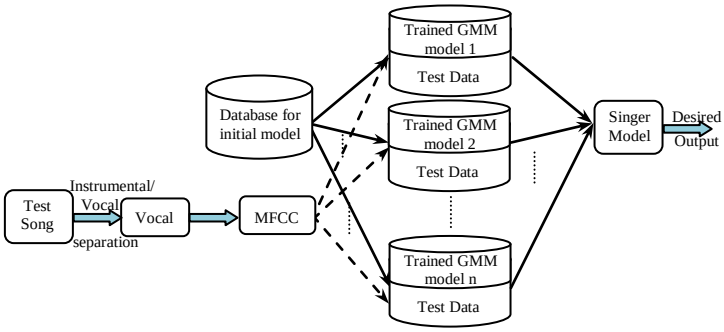

Figure 4. Testing stage.

5. Experimental Result

5.1. Data

This singer identification system is tested with a music database that contained the songs from 18 albums by 8 different singers; 4 male singers, 2 female singers and 2 groups. The genres involved in this experiment are pop, rock and country. All the music data used in these experiments are collected from commercial audio CDs at a 44.1 kHz sample rate, and 16 bits per sample in stereo format.

A singer has a unique structure and required to have shorter time resolution. In order to determine the unique structure of each song, manually annotated vocal fragments are used. The feature vectors are then extracted from 20 ms Hamming-windowed frame with 50% overlapping. One album of each singer is processed as the training data and the songs from the other albums are used as test data. In average, eight to fourteen songs are trained for each artist. In this experimental study, test songs for artists Don Williams and Rod Stewart are drawn from 3 music albums while identification test of the other singers are performed from 2 music albums.

5.2. Result

A confusion matrix contains information about actual and predicted classifications done by a classification system. The performance of this singer identification system is summarized with a confusion matrix as given in Table 1. The number of test songs for each artist is ten. For some of the test songs, it is found that singers are erroneously identified as Bee Gees. The reason behind it could be that the identification of an artist from a song is performed involving all vocal fragments with varying intervals obtained in manual separation stage.

Table 1: Confusion Matrix of Test Data

	Alan Jackson	Dio	Rod Stewart	Don Williams	Rush	Bee Gees	Colbie Caillat	Venessa Carlton
Alan Jackson	10	0	0	0	0	0	0	0
Dio	0	10	0	0	0	0	0	0
Rod Stewart	0	0	3	0	0	7	0	0
Don Williams	0	0	0	10	0	0	0	0
Rush	0	0	0	0	0	10	0	0
Bee Gees	0	0	0	0	0	10	0	0
Colbie Caillat	0	0	0	0	0	0	10	0
Venessa Carlton	0	0	0	0	0	0	0	10

The accuracy of the singer identification system is also shown in Table 2 with correct identification rate (accuracy) and incorrect identification rate (error). Correct identification rate for each singer is defined as the percentage of correctly identified songs to the total number of test songs for a particular singer. Similarly, error rate is defined as the percentage of incorrectly recognized songs to the total number of test songs for a specific singer. Based on this

experimental result, the identification of pop singer has higher accuracy rate than the others.

6. Discussion

In the vocal/instrumental separation, it is more difficult to extract vocal parts from the instrument part of the noisy songs such as rock. The country songs are easy to annotate vocal/non-vocal segments because the vocal part and non-vocal parts of these songs are clearly bounded. In some rock songs, the vocal parts of the rock songs are mixed with background instrumental part so that it is differentiate them. In testing with each singer, there is no overlap between test songs and train songs. If the songs have many vocal parts to be extracted or have too long vocal parts, the longer training time is needed to learn the model. The reason behind 0% accuracy for Rush could be difference between training album (heavy-rock) and testing album(soft-rock). On the whole, the classification results are far greater than previously reported paper such as [6] but still fall well short of expected human performance. The comparison with other methods could also be unfair due to the variety of the datasets used.

Table 2: Artist Identification Accuracy

Singer's Name	Gender	Genre	Accuracy	Error
Alan Jackson	Male	Country	100%	0%
Dio	Male	Country	100%	0%
Rod Stewart	Male	Pop	30%	70%
Don Williams	Male	Rock	100%	0%
Rush	Group	Rock	0%	100%
Bee Gees	Group	Pop	100%	0%
Colbie Caillat	Female	Pop	100%	0%
Venessa Carlton	Female	Pop	100%	0%

7. Conclusion

This presented method attempts to automatically establish the identity of a singer using acoustic features extracted from songs in a database of popular music. The unique qualities of a singer's voices make it relatively easy for us to identify a song as belonging to that particular artist. Genres used in this task are Pop, Country, and Rock. In this singer identification system, 8 albums of different singers with different genre are trained with GMM and 10 different songs from other albums of each singer are tested as test data. The accuracy of trained data is 100%. However in testing with all vocal parts from the entire song yields the correct identification rate up to 78.75%. With the review of other methods presented Section 2, the proposed method has reasonable accuracy with low complexity in feature extraction as well as classification stage. According to the results achieved, singer identification system can depend on the genres. With the accuracy rate of Table 2, this method can more accurately classify pop and country song singers. Improvements are still needed to validate the effectiveness of the method for larger database including non-English song singers.

8. References

[1] N. C. Maddage, C. Xu and Y. Wang, "Singer Identification Based on Vocal and Instrumental Models", *ICPR' 2004*.

[2] X. J. Mestres, "A BIC-based approach to Singer identification", *Master Thesis, Universitat Pompeu Fabra of Spain*, September 2007.

[3] J. Shen, J. Shepherd, K.-L. Tan and B. Cui; "HSI: A Novel Framework for Efficient Automated Singer Identification in Large Music Database"; *ACM Transactions on Information Systems*, Volume 27, Issue 3, 2009.

[4] H. Li and T. Lay; "Vibrato-motivated acoustic features for singer identification*"; IEEE Trans. Acoust. , May 2006.*

[5] A.Mesaros, T. Virtanen and A. Klapuri; "Singer Identification in Polyphonic Music Using Vocal Separation And Pattern Recognition Methods"; *ISMIR'07*.

[6] Y.E.Kim and B.Whitman; "Singer Identification in Popular Music Recordings Using Voice Coding Features"; *2002 IRCAM*.

[7] P.Annesi, R.Basili, R.Gitto and A.Moschitti; "Audio Feature Engineering for Automatic Music Genre Classification"; *Conference RIAO2007, Pittsburgh PA, U.S.A. May 30-June 1, 2007.*

[8] W.H.Tsai, H.M.Wang, and D. Rodgers, "Automatic Singer Identification of Popular Music Recordings via Estimation and Modelling of Solo Vocal Signal", *Institute of information Science, Academia Sinica, Taiwan, Republic of China,2001*.

[9] M.A.Bartsch, "Automatic Singer Identification in Polyphonic Music", *University of Michigan, 2004.*

[10] T. Zhang, "System and Method for Automatic singer identification*", ICME 2003.*

.