Singing Voice Classification in Commercial Music Productions

Faiz MAAZOUZI
LabGED Laboratory
BP. 12
University of Annaba – Algeria
+213 38 87 29 04
mazouzi@labged.net

Halima BAHI
Computer Science Department
BP. 12
University of Annaba - Algeria
+213 38 87 29 04
bahi@labged.net

***Abstract*—Instead of the expansion of the information retrieval systems, the music information retrieval domain is still an open one. In this context, the singing voice classification is a promised trend. In this paper, we shall present our experiments concerning the classification of singers according to their voice type, and their voice quality. Some experiments were carried in which two sets of parameters are used in addition to the use of two classification approaches: The GMM (Gaussian Mixture Models) and the VQ (Vector quantization). The obtained results were compared to those provided by the related state-of-the-art approaches.**

KEYWORDS- Music Retrieval; Features Extraction; Voice Singer Classification; Gaussian Mixture Model; Vector Quantization.

I. INTRODUCTION

Due to the progress of the unlimited data storage capabilities and the proliferation use of the Internet, information retrieval systems encountered a large interest. Much of this data is in the form of speech from various sources. So, it becomes important to develop the necessary technologies for indexing and browsing such audio data. We are particularly interested of the music information retrieval domain.

However, the major part of the research dealing with the human voice information retrieval focuses on the speaking, not the sung voice, and the one dealing with sung voice focuses on the fundamental properties such as the pitch or the rhythm and not the quality of the voice itself. The purpose of this study is an attempt to define two classification categories: The first deals with the skills of a singer (quality), and the second with the kind of the sung voice (Type).

Concerning, the sung voice type, a categorization has already existed related to the field of the opera, and there are some ideas on how could (or should) be the classification of voices in the current music [1].

For the assessment of the singer voice quality, we consulted human expert, and we defined three categories: “good”, “medium” and “bad”. Then, we defined two sets of parameters and we pursued the experiments using two classifiers. The first one is based on the Gaussian Mixture Models [2] and the second one on Vector quantization [3]. The Obtained results are compared to those provided by similar studies [4] [5]. Pawel Zwan and al. [4] consider the two classification categories, while in [5] only the evaluation of singing voices is considered.

The rest of the paper is organized as follows: The next section will describe data used for experiments. Section 3 will present the separation of voice and music, this stage is critical for the rest of the study. In Section 4, we shall present the two sets of features we extract from the signal. In section 5, we shall describe the principles of GMM classifier and in section 6, those one of Vector quantization. In the 7th section, we shall present evaluation and experiments. Finally, we shall make our conclusion in section 8.

The figure below shows the general scheme of classification.

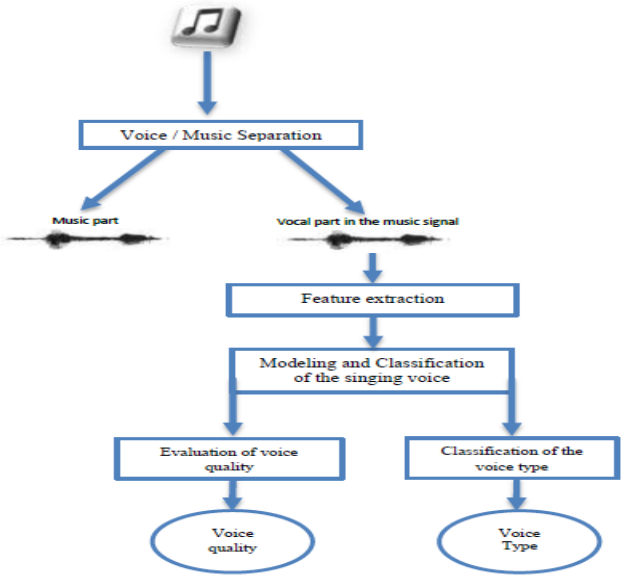

Figure 1. General scheme of classification

II. DATA COLLECTION

Zikdalgerie.com is a music website which contains thousands of Algerian songs. The songs are grouped in albums; each album represented a singer [6]. In our study, we consider a part of the provided songs, including songs of 50 singers from the website. The learning data base contains 240 audio samples (120 for the classification type and 120 for evaluating the voice quality).

There are different voice types' classifications. However, most of these types fall under six major different voice categories known in all the major voice classification systems. Women' voice is typically divided into three groups: soprano, mezzo-soprano, and contralto. Men' voice is usually divided into three groups: tenor, baritone, and bass.

For the evaluation of the voice quality, we define three vocal qualities, "good voice", "bad voice" and "medium voice ".

III. VOICE / MUSIC SEPARATION

The audio files often contain multiple sound sources (singers, instruments, noises) mixed with live recording or studio. The source separation aims at reconstruct the source signals to listen to them individually.

There are several ways for voice and music separation; these ways are all based on approaches known as "blind" of extracting "generic" audio signal descriptors and using them to learn the two classes of segments "sung" and "non-sung". In this context, we adopt the method proposed by Bonada and all in [7].

Figure 2 presents the diagram of a system Voice / Music Separation.

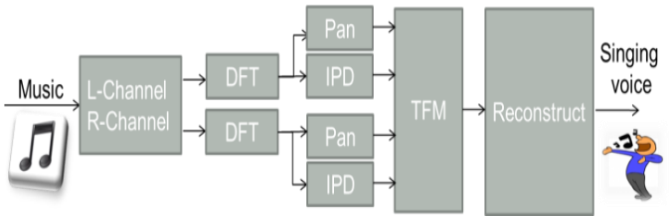

Figure 2. Voice / Music Separation

In the case of stereo mixes, we used in the separation phase, the time-frequency masking method to calculate the Discrete Fourier Transform (DFT) for left and right channels.

$DFT_0(out_L)[0] \cdots DFT_0(out_L)[N/2]$
$DFT_0(out_R)[0] \cdots DFT_0(out_R)[N/2]$
. . .
$DFT_{P-1}(out_L)[0] \cdots DFT_{P-1}(out_L)[N/2]$
$DFT_{P-1}(out_R)[0] \cdots DFT_{P-1}(out_R)[N/2]$

The DFT coefficients are grouped by adjacent pan and IPD (Inter-channel Phase Difference) to choose between the candidate solutions generated by TFM.

Where: $\forall f \in [0 \ldots N/2]$

Pan:

$$pi = \arctan \left| \frac{\mathrm{DFTp}(S_i^R)}{\mathrm{DFTp}(S_i^L)} \right| .2\pi \quad (1)$$

IPD (Inter - Channel Phase-Difference):

$$|Arg(DFTp\ (S_i^L)f) - Arg((DFTp\ (S_i^R)f)| = 0 \quad (2)$$

We can see the difference between the original signal (voice + music) and the singing voice signal in Figure 3**.**

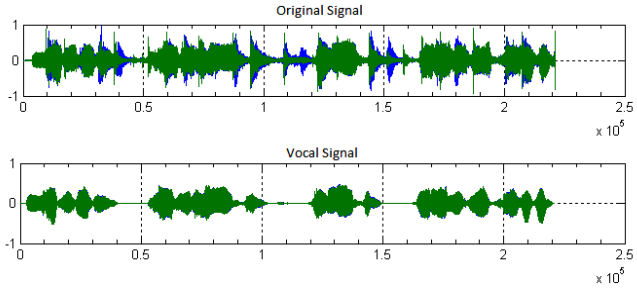

Figure 3.The difference between the original signal (voice + music) and the sung voice.

Systematic evaluation shows that, despite its simplicity, the proposed system achieves a competitive level of performance [7].

IV. FEATURES EXTRACTION

The features extraction stage aims at providing a discriminate representation of the signal.

The first parameter set (FV1) we define, contains 12 cepstral coefficients. The audio signal is sampled at 16 kHz and coefficients MFCC (Mel Frequency Cepstral Coefficients-) are calculated from a bank of 24 Mel filters applied every 10 ms windows of 30ms.

Another more general way to determine the parameters of the sung voice is to use the description of signals such as audio descriptors of the MPEG-7 standard.

The second vector contains a number of MPEG-7 descriptors and other descriptors. The descriptors are divided into the following groups:

- Energy: Audio Power (AP).
- Harmonic: Audio Fundamental Frequency (AFF)
- Spectral: Audio Spectrum Spread (ASS), Mel-Frequency Cepstral Coefficients (MFCC).
- Temporal: Log-Attaque Time (LAT), Temporal Centroid (TP).
- Various: Audio Spectrum Flatness (ASF).

From these descriptors, a feature vector (FV2) with 130 parameters was established; the parameters are divided into the following groups:

- Parameter of Energy: 4 parameters.
- Parameter of harmony: 4 parameters.
- Spectral parameter: 54 parameters.
- Temporal parameters: 36 parameters.
- Various parameters: 32 parameters.

In this research, we investigated 2 types of feature vectors, (FV1) with 12 dimensions and (FV2) with 130 dimensions.

V. Modeling and Classification with Vector Quantization (VQ)

The motivation in the use of Gaussian mixture densities is the intuitive notion that the individual component densities of a multi-modal density, like the GMM may model some underlying set of acoustic classes. These acoustic classes reflect some general singer vocal tract configurations that are useful for characterizing his voice type and quality. The spectral shape of the ith acoustic class can be represented by the mean μi of the ith component density, and variations of the average spectral shape can be represented by the covariance matrix i.

In the classification of type, we used 24 mixtures of Gaussian distributions to model each sample in the training base types that contains 20 samples for each type "Tenor, Baritone, Bass, Soprano, Mezzo-Soprano, Contralto".

For the evaluation of voice quality, we used 24 Gaussian mixture distributions to model each sample in the training set of vocal qualities that contains 20 samples for each voice quality (3 male and 3 female).

The three qualities of male and female voices are "good voice", "Medium voice" and "Bad voice". At the training stage, the singers' voices were manually classified by an expert.

Figure 4 presents the diagram of system modeling and classification based on Gaussian mixture model.

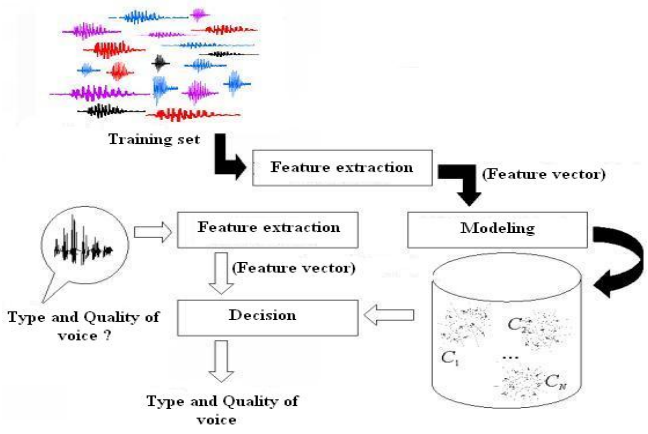

Figure 4 diagram of system modeling and classification based on Gaussian mixture model.

A. GMM Modeling

A Gaussian mixture density is a weighted sum of M component densities, given by the equation:

$$P(x|\lambda) = \sum_{i=1}^{M} pigi(x) \qquad (3)$$

Where x is a D-dimentional random vector, gi(x) are the component densities and pi are the mixture weights. Each component density is a D-variate Gaussian function of the form:

$$g(x) = \frac{1}{(2\pi)^{D/2}|\sum 1/2|} e^{-1/2\,(x-\mu)T\Sigma-1(x-\mu)} \qquad (4)$$

Where μ represents the estimated mean and covariance matrix. The parameters of the Gaussian probability densities are estimated by the EM algorithm (Expectation-Maximization), whose description can be found in [8].

The complete Gaussian mixture density is then parameterized by the mean vextors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation: = {pi, i, i}.

For the voice singer classification and assessment, each class is represented by a GMM and is referred to its model λ.

For the evaluation of voice quality: The GMM of the "Good" (λg) class is driven on descriptors extracted sections annotated as good voice. The GMM of the "Medium" (λm) class: descriptors extracted sections annotated as Medium voice. The GMM of the "Bad" (λb) class: descriptors extracted sections annotated as bad voice.

At the same time of the voice type classification, the following models were built: λte , λbr , λba, for male voice and λso, λms and λal, for female voice.

B. GMM classification

The classification criterion adopted for the classification is the criterion of maximum a posteriori. According to this criterion, any observation vector x is assigned to the class that maximizes the probability of observing the model λ partner knowing.

a) For the evaluation of voice quality:

Class (x) = {i | p (x | λi)} = max {p (x | λg), p (x | λm), p (x | λb)}

b) For the classification of male type:

Class (x) = {i | p (x | λi)} = max {p (x | λte), p (x | br), p (x | ba)}

c) For the classification of female type:

Class (x) = {i | p (x | λi)} = max {p (x | λso), p (x | ms), p (x | λal)}

VI. Modeling and Classification with Vector Quantization (VQ)

The second system uses vector quantization (VQ) classifier. The classification of the sung voice is done by calculating the minimum distance vector (distortion) between the training sequence and the wave file as a singer was tested.

The modeling stage:

1- Calculate the feature vector (FV)

2- Using the algorithm of quantification LBGVQ (Linde Buzo-Gray Vector Quantizer), algorithm proposed by [3] to compute the dictionary or codebook

The classification stage:

1- Calculate the vector feature for the file to test (FVt)

2- Calculate the distance between the vector feature for the file to test (FVt) and the entries of the codebook.

Figure 5 presents the diagram of a system modeling and classification based on vector quantization.

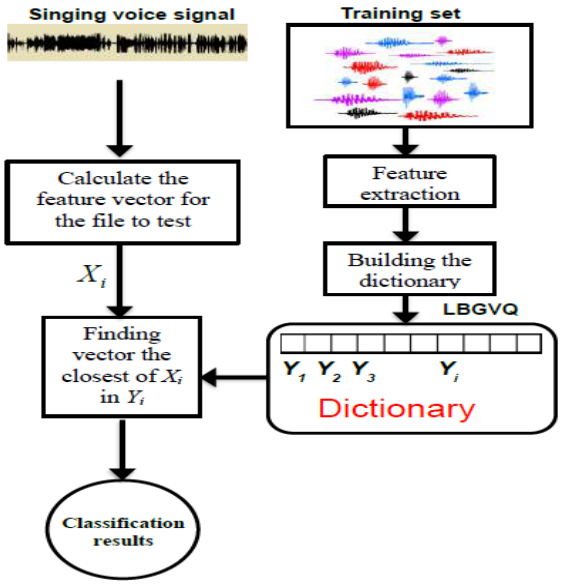

Figure 5. Diagram of a system modeling and classification based on vector quantization.

VII. Modeling and Classification with Vector Quantization (VQ).

In order to make a detailed comparison, between the GMM (Gaussian Mixture Model) and the QV (Vector Quantization) in the singing voice classification field. We made the following experiments.

The results of the singing voice classification and the voice quality assessment are presented in Table 1 and 2.

The confusion matrix or contingency table is used to assess the quality of classification.

Table1. The confusion matrix for the classification with GMM and FV2 vector feature (a) voice quality (VoiceQ) and (b) Voice type (VoiceT).

a.

VoiceQ	Good	Medium	Bad
Good	95	3	2
Medium	4	94	2
Bad	0	1	99

b.

VoiceT	bass	Baritone	tenor	alto	Mezzo-soprano	soprano
bass	97	1	2	0	0	0
Baritone	7	91	2	0	0	0
tenor	0	2	98	0	0	0
alto	0	0	0	96	2	2
Mezzo-soprano	0	0	0	0	99	1
soprano	0	0	0	0	4	96

Table2. The confusion matrix for classification with vector quantization (VQ) and FV2 feature vector (a) voice quality (VoiceQ) and (b) Voice type (VoiceT).

a.

VoiceQ	Good	Medium	Bad
Good	91	6	2
Medium	4	90	6
Bad	1	4	95

b.

VoiceT	bass	Baritone	tenor	alto	Mezzo-soprano	soprano
bass	95	5	0	0	0	0
Baritone	6	91	3	0	0	0
tenor	1	6	93	0	0	0
alto	0	0	0	92	3	3
Mezzo-soprano	0	0	0	3	88	9
soprano	0	0	0	1	2	97

Another evaluation method is to calculate the recall and precision for the two classification methods (GMM and VQ).

We calculate these values and we compare them to those obtained in [4] and [5].

The results are shown in Table 3 and 4, where PM means our proposed method.

Table 3. Comparison of the results of PM and studies [4] and [5] for evaluating the quality of singing voices.

project	Descriptors	Classification	Base	recall	precision
[4]	Vector for 331 parameters	The artificial neural networks (ANN)	1700 ECH 42 singers	93.6	93.4
[4]	Vector for 331 parameters	rough set-based (RS)	1700 ECH 42 singers	90.2	95.6
[5]	FBANK 12 dimensions	GMM (Gaussian Mixture Model)	10 singers	92.1	93.3
PM	MFCC 12 dimensions	GMM (Gaussian Mixture Model)	100 ECH 50 singers	92.96	93.57
PM	Vector for 130 parameters	GMM (Gaussian Mixture Model)	100 ECH 50 singers	95.99	96
PM	MFCC	VQ (Vector Quantization)	100 ECH 50 singers	90.11	87
PM	Vector for 130 parameters	VQ (Vector Quantization)	100 ECH 50 singers	92.4	92

Table 4. Comparison between the PM results and those of [4] for the classification type of sung voices.

project	Descriptors	Classification	Base	recall	precision
[4]	Vector for 331 parameters	The artificial neural networks (ANN)	1700 ECH 42 singers	89.5	89.6
[4]	Vector for 331 parameters	rough set-based (RS)	1700 ECH 42 singers	64	67.6
PM	MFCC 12 dimensions	GMM (Gaussian Mixture Model)	100 ECH 50 singers	91.96	92.92
PM	Vector for 130 parameters	GMM (Gaussian Mixture Model)	100 ECH 50 singers	95.96	96.11
PM	MFCC 12 dimensions	VQ (Vector Quantization)	100 ECH 50 singers	91	89.11
PM	Vector for 130 parameters	VQ (Vector Quantization)	100 ECH 50 singers	91.57	92.66

In this section we have compared the performance of PM with [5] and [4].

The results of both tables show that the complete vector of 130 is more efficient than the MFCC vector for the two classification methods (GMM and VQ).

From the obtained results of the two tables we can observe that the PM gives good results for evaluating the quality of singing voice and the classification of voices type.

VIII. CONCLUSION

The performed work in this study aims at making an automatic classification of sung voice.

Two feature vectors were formed (FV1, FV2), Then, we've compared the results of two methods of modeling and classification (GMM, VQ), through this evaluation we've found out that the use of vector FV2 (130 parameter) and Gaussian mixture model outcomes better results for the automatic classification of sung voices.

The classification of the sung voice is subjective and its research is very difficult to perform without a large amount of data. The field is also new and there are a lot of problems waiting to be challenged.

REFERENCES

[1] Sundberg, J., and Rossing, T.D. The Science of Singing Voice, the Journal of the Acoustical Society of America, 87:468, 1990.

[2] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussian Mixture Speaker Models", IEEE Transaction on speech and audio processing, Vol. 3. N°1, 1995.

[3] Linde, Y., Buzo, A., Gray, R., An Algorithm for Vector Quantizer Design, IEEE Transactions on Communications, vol. 28, pp. 84–94, 1980.

[4] Pawel Zwan, Piotr Szczuko, Bozena Kostek, Andrzej Czyzewski: Automatic Singing Voice Recognition Employing Neural Networks and Rough Sets. T. Rough Sets 9: 455-473, 2008.

[5] Prasertvithyakarn Prasert, Koji Iwano, Sadaoki Furui: An Automatic Singing Voice Evaluation Method for Voice Training Systems. No. 2-5-12 pp. 911-912, 2008.

[6] www.zikdalgeire.com, last visit, 20 Dec. 2010.

[7] Bonada, J., Loscos, A., Vinyes Raso, M.: Demixing commercial music productions via human-assisted time-frequency masking. In: Proc. AES, 2006.

[8] R.McAulay and T. Quatieri. Speech analysis /synthesis based on a sinusoidal representation, IEEE Transactions on Acoustics, Speech, and Signal Processing, 34(4):744–754, 1986.