Statistical and Neural Classifiers: Application for Singer and Music Discrimination in Polyphonic Music Context

Hassan Ezzaidi[1] and Mohammed Bahoura[2]

[1] Département des Sciences Appliquées
Université du Québec à Chicoutimi,
550, boul. de l'Université, Chicoutimi, Qc, Canada, G7H 2B1
hezzaidi@uqac.ca
[2] Département de Mathématiques, d'Informatique et de Génie
Université du Québec à Rimouski,
300, allée des Ursulines, Rimouski, Qc, Canada, G5L 3A1
Mohammed_Bahoura@uqar.qc.ca

Abstract. The problem of identifying sections of singer voice and instruments is investigated in this paper. Three classification techniques: Linde-Buzo-Gray algorithm (LBG), Gaussian Mixture Models (GMM) and feed-forward Multi-Layer Perception (MLP) are presented and compared in this paper. All techniques are based on Mel frequency Cepstral Coefficients (MFCC), which commonly used in the speech and speaker recognition domains. All the proposed approaches yield a decision at every 125 ms only. Particularly, a large experimental data is extracted from the music genre database RWC including various style (68 pieces, 25 subcategories). The recognition scores are evaluated on data used in the training session and others never seen by proposed systems. The best results are obtained with the GMM (94% with train data and 80.5% with test data).

Keywords: music, song, artist, discrimination.

1 Introduction

The explosion of media, particularly the accessibility to digital information via Internet and the increasing production amount of the disk industry, creates many new needs as the browsing, archiving, maintenance, classification and authentication of data. Hence, the needs for an automatic processing of media documents are increasing fast and the challenge is becoming greater. Among these needs, there is the interest to classify music by style, singer, instrument or language. Automatic indexing is another need even more delicate and complex. Discrimination between music, silence, speech and singing is a crucial for the proper functioning of identification, recognition and segmentation systems. In other words, we want an automatic system to analyse, understand, track and mask like processed by the human auditory system. So, the challenge is very

A. Elmoataz et al. (Eds.): ICISP 2010, LNCS 6134, pp. 130–136, 2010.

strong despite the progress obtained in speech and speaker domains. More, the acoustical music signal is very complicated because it is composed from multiple sound sources including artist song. Many works were published in music information retrieval but in a very limited context. Indeed, the context depends on the used database, pre-processing and post-processing effectuated, parameter vectors, models recognition, learning approach and target task to be performed by the system. The global task is the speech and the music discrimination, genre recognition, singer's identification, instruments identification, tracking and detection of the singer, the tempo/beat/metric determination, etc [1,2]. In the case of the singer characterization, some works was previously published. All applied techniques are inspired from the speech and the speaker's domain. The Mel-frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) are known as the best acoustics features in speech, speaker and instrument recognition. Intuitively, they have been used in artist identification system [1,3,4,5].

Classification strategies including the discriminate functions, Gaussian Mixture Models (GMM), Nearest Neighbors (k-NN), Kullback-Leibler divergence and Kullback-Leibler distance measure were compared for singer identification in polyphonic music using a concept of vocal separation [1]. A Support Vector Machine (SVM) is a classifier estimating an optimal hyperplane to separate classes of data. Reporting good achieving results in genre classification, it was applied also in singer identification [1]. The artist classification was also performed by a Multi Layer Perceptron (MLP) neural network [4].

In the present work, we are interested in a system that performs artist detection from the beginning to the ending of his song at different time. Classical MFCC coefficients are used as parameters. We apply Linde-Buzo-Gray (LGB) [3], Multi-layer forwards perceptron (MLP) and Gaussian Mixture Models (GMM) as classifiers. The decision is determined at rhythm of 125 ms without overlapping shift windows.

2 Database

A 68 musical pieces of the RWC Music Genre Database [6] database is used. Each piece averages 4 minutes originating from more than 25 genres (Popular, Ballade, Rock, Heavy Metal, Rap/Hip-Hop, House, Techno, Funk, Soul/R&B, Big Band, Modern Jazz, Fusion, Bossa Nova, Samba, Reggae, Tango, Baroque, Classical, Romantic, Modern, Blues, Folk, Country, Gospel) of men and women. Musicals pieces used are numbered in the RWC database [1] by: G01_xx to G05_xx and G07_xx. Duration of whole database is approximately 5 hours with 3 hours of music only. All the musical pieces used in our experiments have been hand labeled. For each file, we make listen to a person segment by segment's of duration one second and asking him to type $\underline{1}$ if he hears an artist song or $\underline{0}$ if he hears only music. Each piece in the proposed database is split equally in two part: the first half segment and the last half segment. First half segments for all pieces are used as the set for training session. Also, all the second part are grouped together and they are never presented in training session in order to be used in testing session as validation.

3 Feature Extraction

The conventional parameters Mel-Frequency Cepstral Coefficients (MFCC) often used in systems for recognition and identification (speech or music) were considered. In deed, all musical pieces are first down sampled from 44.4 kHz to 16 kHz. Then, the musical signal is divided into frames of 2000 samples. Theses frames are shifted at a frame rate of 2000 samples (no overlapping). In fact, the musical signal is more stable and quasi-stationary than the speech signal (co-articulation phenomena). For each frame, a Hanning window is applied without pre-emphasis. The Fourier power spectrum is computed and coefficients are extracted from 32 triangular Mel-filters. After the application of log operator for each output filters, a discrete cosine transform is applied to extract 24 coefficients in cepstral space.

4 Proposed Classification Techniques

The problem of discrimination of singing and music can be naively seen as a categorization into two classes. A first class is associated with musical sequences and a second one is associated with the song signal. In fact, the modeling with just two classes is far from ideal realty related to the rich information conveyed by the singer signal (melody, form, prosody, sex, age, feeling, etc.) and the musical signal (instruments, style, etc.). However, in the field of speaker identification several contributions have been proposed to track and recognize one or some people engaged in a conversation [3]. Inspired by the progress realized last years in this field, we propose to consider the following statistical approaches.

4.1 Non-parametric Model

Linde-Buzo-Gray algorithm (LBG) [3] clustering is a method that aims to partition N observations into K clusters (prototypes or centers) in which each observation belongs to the cluster with the nearest distances. Strategy based on vector quantization using LBG algorithm was investigated to calculate respectively 32, 64 and 128 centers (one codebooks related music and second codebooks related song). Same MFCC coefficients described previously were used as vectors parameters.

4.2 Multi-Layer Feed-Forwards Perceptron (MLP)

Neural networks (NN) are known able to perform complex tasks in various fields of application as pattern recognition, identification, classification, prediction and control systems. The NN designs are inspired by biological cells of neurons systems. Among the networks most used is the multi-layer feed-forwards networks. Each layer consists of simple units operating in parallel process. Unit receives their input from units from a layer directly below and sends their output to the unit layer above the unit. Each unit is weighted with an appropriate coefficient

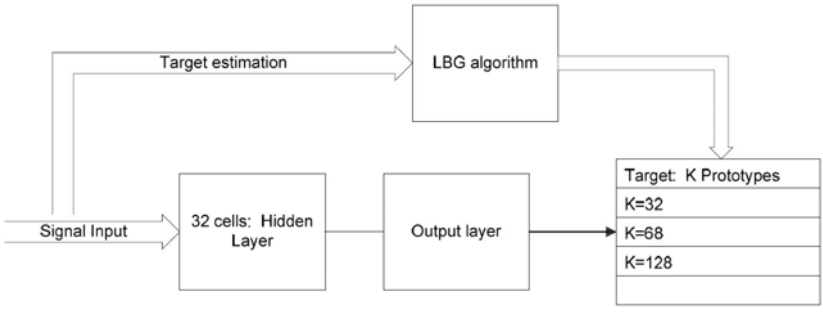

Fig. 1. Neural network structure system

wi (synapses). Commonly, NN is trained by adjusting the weights and biases, so that each input leads yields in output the specific target. The auditory toolbox for Matlab is used for this experiment [7].

Based on labeling data manually performed in music and song, we used conventional multi-layer feed-forwards network. Output is a binary encoding; the input is the vector of MFCC coefficients with a hidden layer of 32, 62 and 128 cells respectively.

In principle, we should use audio signal segmentation based on phonemes, for the segments song. Thus, several alternative coding can take place to establish the targets and to train the network. This is a feasible task, despite the variability of the speaker. However, for segments music this strategy becomes very complicated and difficult due to the number of instruments and musical notes combinations. To bypass this problem, we applied the LBG algorithm to perform segmentation by using a simple centroid representation. The architecture system is illustrated in Fig. 1. Specifically, we separated the learning data into two subgroups respectively for music and song, based on the manual labelling. Then, for each sub-group it applied the LBG algorithm to estimate prototypes cookbooks. Three sets of prototypes were estimated for each subgroup: 32, 64 and 128 codebooks. Relying on the similarity measure, the target for each entry is coded in this case as the nearest prototype. So we are left with 32 prototypes (24 of dim.) respectively for music and singing

4.3 Parametric Model

We use a Gaussian Mixture Model (GMM) with a weighted sum of 32, 64 and 126 Gaussians. In Each case, GMM is defined for song features and music features extracted respectively from the training database. Expectation Maximization (EM) is used for estimating the means, covariances and weights parameters GMM models. Test features are classified by a maximum likelihood discriminate function, calculated by their distances from the multiple Gaussians of the class distributions. Recall that all the experimented data is hand labeled. It contains men and women artists, various genres and 5 hours recording. For all examined

Table 1. Score recognition (every 2000 samples): GMM method

Data set	Number of Gaussian		
	32	**64**	**128**
Train Music	86.5%	89.5%	94.1%
Train Song	89.5%	92.2%	93.4%
Test Music	77.7%	79.7%	81.8%
Test Song	81.2%	78.3%	79.1%

Table 2. Score recognition (every 2000 samples): LBG Method

Data set	Prototypes		
	32	**64**	**128**
Train Music	73%	76%	78%
Train Song	76%	78%	81%
Test Music	67%	69%	70%
Test Song	72%	73%	75%

Table 3. Score recognition (every 2000 samples): Multi-layer perceptron based LBG

Data set	Hidden layer units		
	32	**64**	**128**
Train Music	86%	86 %	88%
Train Song	54%	56%	58 %
Test Music	80%	80%	83 %
Test Song	54%	60%	59 %

Table 4. Score recognition (every 2000 samples): Multi-layer perceptron

Data set	Hidden layer units		
	32	**64**	**128**
Train Music	94%	92 %	92%
Train Song	60%	64%	66%
Test Music	90%	89%	88%
Test Song	35%	35%	37%

techniques we extract half of each file in the database for the training session and the last is used for testing session as unknown data.

5 Results and Discussion

The results for the three classifiers are shown in Tables 1 to 4. In the columns, we give the scores recognition for the 32, 64 and 128 prototypes for LBG, number

of weighted Gaussian of GMM and cells in hidden layer MLP. In line, we give the score recognition for train and test data respectively for segments music and song.

The recognition rate increases in general gradually dependently on the factor (32, 64 and 128). The results show that better score were obtained with training data for all systems. With test data, the rate recognition decrease for all strategies.

By comparing the results obtained by the GMM and the LBG, we find that the scores are lower for the LBG in all cases. In fact, quantization algorithm can be seen as a discrete version of GMM models. So, we expect lower performance and in best case the scores will be close. The difference is that the LBG algorithm is easy to implement and serves as an indicator.

By comparing the results obtained by the GMM and the proposed neural networks, we find that the neural networks not discriminate well segments of singing while GMM achieve a good performance. In the case of music discrimination, the two approaches give good results. However, it retains the best performance is obtained with the GMM in all cases.

It was concluded that the coding target by codebook is less effective for training the network compared to GMM system. Unsatisfactory results of the neural network can be improved by considering another way of coding and by balancing the amount of data between music and singing. The coding based GMM will be a good alternative.

6 Conclusion

In this works, the problem of identifying section of singer voice and musical instrument is explored. A parametric GMM models, a non parametric LBG model and MLP neural network were used as classifiers and are compared. Only, classical MFCC coefficients were used as parameters.

Best performance is obtained with GMM model until 94% score recognition with train data and 80.5% score recognition with test data. With the neural network results can still be improved but we must rework the part of coding using the only rang of codebooks elements or investigating other segmentation technique. The method of coding codebook remains a good idea for a quick learning the neural network in the case of music.

References

1. Mesaros, A., Virtanen, T., Klapuri, A.: Singer identification in polyphonic music using vocal separation and pattern recognition methods. In: Proc. ISMIR, Vienna, Austria (2007)
2. Tzanetaki, G., Essl, G., Cook, P.: Automatic musical genre classification of audio signals. In: Proc. ISMIR, Bloomington, Indiana (2001)
3. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. IEEE Trans. Comm. 28(1), 84–95 (1980)

4. Berenzweig, A., Ellis, D., Lawrence, S.: Using voice segments to improve artist classification of music. In: Proc. AES-22 Intl. Conf. on Virt., Synth., and Ent. Audio., Espoo, Finland (June 2002)
5. Kim, Y.E., Whitman, B.: Singer identification in popular music recordings using voice coding featuress. In: Proc. ISMIR, Paris, France (2002)
6. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: Rwc music database: Music genre database and musical instrument sound database. In: Proc. ISMIR, pp. 229–230 (2003)
7. The auditory toolbox for matlab, `http://cobweb.ecn.purdue.edu/~malcolm/interval/1998010/`