Singer Identification in Rembetiko Music

Andre Holzapfel, Yannis Stylianou
Institute of Computer Science, FORTH, Greece, and
Multimedia Informatics Laboratory, Dep. of Computer Science, University of Crete, Greece
{hannover,yannis}@csd.uoc.gr

Abstract— **In this paper, the problem of the automatic identification of a singer is investigated using methods known from speaker identification. Ways for using world models are presented and the usage of Cepstral Mean Subtraction (CMS) is evaluated. In order to minimize the difference due to musical style we use a novel data set, consisting of samples from greek Rembetiko music, being very similar in style. The data set also explores for the first time the influence of the recording quality, by including many historical gramophone recordings. Experimental evaluations show the benefits of world models for frame selection and CMS, resulting in an average classification accuracy of about 81% among 21 different singers.**

Keywords- **Artist Identification, Music Information Retrieval, Gaussian Mixture Model**

I. Introduction

The research activity in the field of music information retrieval has increased continously throughout the last years. This increase has been motivated by the growing amount of music exchanged using the internet, and the simultaneous interest of the music industry to find proper means to deal with this new way of distribution. A major part of popular music is characterized by the performing artist. As most kind of popular music include vocals, the use of systems to recognize the singer is obvious. Such systems have been presented in the past, and the efforts have been outlined in [1]. Common to the approaches is to use techniques from speaker identification. As such, most commonly used features are Cepstral coefficients, and these features are usually modeled for a specific speaker/singer by using mixtures of Gaussians (GMM), see [2]. Furthermore, in [3] and [1], world models for the instrumental and vocal frames from all classes have been used to decide if a particular frame in a test song contain vocals or not. To our knowledge, these world models have not been evaluated for the use in score normalization as described in [4]. Also, ways to optimize the vocal/non-vocal decision based on world models have not been evaluated. Furthermore, we would like to examine a straightforward approach to eliminate some influences of the mastering of the audio CD. As in this procedure an equalization is applied to a piece of music, we would like to evaluate the influence of Cepstral Mean Subtraction (CMS) [5] applied to the MFCC features.

A disadvantage in all the presented approaches is the choice of musical pieces from popular artists, with the artists belonging to different musical styles or even genres. Furthermore, some of the best results in singer identification have been achieved on very small datasets [3]. Because of this it is unclear, if the performance of the system can be assigned to the correct recognition of the singer, or if this decision has been simplified by different styles of the artists. Thus a major effort of this publication is the compilation of a big database, consisting of music of one particular style, with a detailed analysis of the data characteristics. The authors' choice was Greek Rembetiko music, where a huge amount of singers is available that perform a musical style that has been preserved for over eighty years. Apart from representing a homogeneous dataset, it also contains a lot of noisy recordings. This is due to the fact that the oldest pieces have been recorded in the 1930's. These pieces have been digitized from old gramophone disks and contain a significant amount of noise. This way, the evaluation of the sensitivity of music information retrieval systems to bad recording conditions is possible.

II. Description of the data set

The compiled data set consists of 290 songs from 21 Rembetiko singers. The number of songs per singer ranges from eight to 18. Details of the data set are depicted in Table I. The numbers for musical activity list the decades in which the artist recorded music. It was tried to cover a wide range of this period with the contained pieces of music. Because of that, for some singers, as Sotiria Bellou, the singer's voice varies strongly. Note that the artist Xarmas represents a male/female duo, that throughout the given period performed together. It is interesting to see, if a system aimed to describe a single singer's voice, can cope with the mix of two voices.

TABLE I
Data set description

Singer	male/female	activity	songs	ID
Agathonas	m	70-now	11	S1
Batis	m	30	13	S2
Bellou	f	40-80	18	S3
Dalkas	m	30-50	14	S4
Delias	m	30-40	8	S5
Genitzaris	m	40-90	9	S6
Gkoles	m	70-now	11	S7
Glykeria	f	70-now	12	S8
Marika (Papangika)	f	20-30	18	S9
Mario	f	70-now	13	S10
Markos Bambakaris	m	30-60	15	S11
Menidiatis	m	60-now	17	S12
Nikolaidis	m	60-now	11	S13
Rita Ampatzi	f	30-50	13	S14
Roukounas	m	30-50	14	S15
Roza Eskenazi	f	30-60	15	S16
Stellakis Perpiniadis	m	30-60	18	S17
Stratos Pagiumtzis	m	30-60	16	S18
Tsaousakis	m	50-70	16	S19
Tsitsanis	m	30-70	13	S20
Xarmas	m+f	40-50	12	S21

From each singer four songs have been hand labelled with the following labels:

- INSTR : instrumental sounds without any voice
- VOICE : voice of target singer without second voice
- MIXED : voice of target singer with second voice
- OTHER : interjections

For singer S21 all vocal frames have been labelled as VOICE, as we want to recognize this particular singer duo. Three of the labeled songs have been put into the training set, and one labeled song for each singer was kept in the test set, in order to evaluate the automatic identification of vocal frames later on.

Another peculiarity of the data set is that some of the artists take part in the others' recordings. As such the artists Markos Bambakaris, Anestis Delias, Stratos Pagioumtzis, Giorgos Batis and Stellakis Perpiniadis formed a group for many years. Because of that, in many songs of the target singer, another singer, who is part of the data set, is featured as second singer. The same holds for Vasilis Tsitsanis, who wrote many songs for Bellou and Tsaousakis, and sings the second voice in some songs of Bellou. Similar relations exist for the currently performing artists Gkoles, Glykeria and Agathonas. In the experimental section it will be observed if these relations influenced the results.

III. System and Parameters

A. Signal Representation

For representing the signals Mel Frequency Cepstral Coefficients (MFCC) have been chosen. Twenty coefficients are used, the zero coefficient is neglected. This parametrization is motivated by the findings of [6], where it has been found the optimum in a similar task. The used window length is 20 ms with 50 % overlap. As we suppose that the piece of music under consideration has passed through a filter bank in the mastering process, we can eliminate the effect of this filtering by using Cepstral mean subtraction [5]. This is because the filter bank represents a linear filter, and as such the MFCC, $\tilde{\mathbf{c}}$, of the signal can be considered the sum of the cepstrum of the input music and the cepstrum of the filter bank. So we can compute the cepstrum $c_i(n)$ of the i-th coefficient at the n-th window as

$$c_i(n) = \tilde{c}_i(n) - \sum_{m=1}^{M} \tilde{c}_i(m) \qquad (1)$$

with M being the number of coefficients computed over the full length of the song. Note that this approach does not address mastering effects as dynamic compression, as these are non-linear.

B. Signal Modeling

Singer Models: As described in Section II the songs in the training set have all been hand labelled. As depicted in Figure 1 this gives the meaningful choices to train the singer models either on only solo vocal, all vocal, or instrumental frames, resulting in models for the i-th singer $\theta_i^{solo}, \theta_i^{voice}, \theta_i^{instr}$ respectively. The difference between the first two choices gives information about how sensitive the system is to the existence of another singer's voice. Comparing the classification incorporating models trained on instrumental frames, with the accuracy achieved with models trained on vocal frames, will give information about the homogeneity of the musical style. If the assumption of the data set being homogeneous holds, we would expect a very low performance, when training the singer models on instrumental frames. In this paper all singer models contained 20 components with full covariance matrices, Expectation Maximization was used for training.

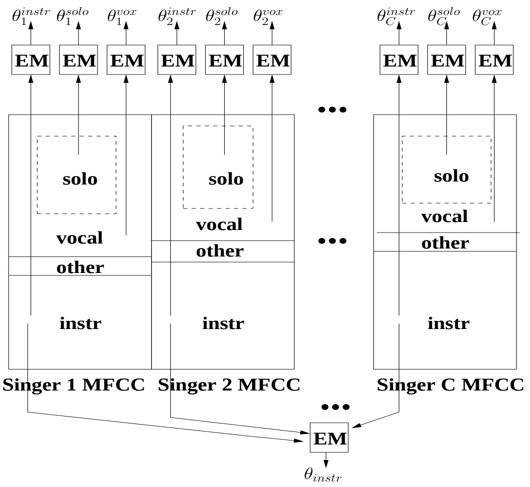

Fig. 1. Possible feature sets for training statistical models

World Models: In speaker recognition a procedure called score normalization uses a world model for all speakers, *i.e.*, a GMM is trained on all the training data frames of all classes. In classification the score $S(\mathbf{c}, \theta_k)$ of the k-th singer's GMM θ_k for a test vector $\mathbf{c}$ is then computed as

$$S(\mathbf{c}, \theta_k) = \log p(\mathbf{c}|\theta_k) - \log p(\mathbf{c}|\theta_{wld}) \qquad (2)$$

with $p(\mathbf{c}|\theta_{wld})$ being the likelihood of the vector given the world model. In our case two choices for the world model are possible: a world model for the instrumental frames (θ_{instr} as shown in Figure 1) and another for all the vocal frames, which will be denoted as θ_{vox}. Apart from the use for score normalization we can use world models also for the automatic selection of vocal frames in test songs, as will be described in Section III-C. In this paper all world models contained 128 components with full covariance matrices, Expectation Maximization was used for training.

C. Vocal frame selection

In the next step, the likelihoods of the MFCC's to be produced by world models are computed. It is assumed that frames containing vocals have a high likelihood on a vocal world model, while having a low likelihood on a instrumental world model. An approach to use these likelihoods for automatically choosing voice like frames for classification was presented in [1]. There, for each song to be classified, a segment length L was accepted as vocal, only if

$$\sum_{i=1}^{L} \log p(\mathbf{c}_{kL+i}|\theta_{vox}) > \sum_{i=1}^{L} \log p(\mathbf{c}_{kL+i}|\theta_{instr}) \qquad (3)$$

with k the segment index. In this paper, this method will be referred to as maximum method. Another possibility is computing these two likelihood sums for all $k = 1...N_k$ segments as in (3), but then just keeping those segments, which have a big likelihood sum for the vocal model (left hand term) and a small sum on the instrumental model (right hand term). This was done by keeping the biggest half of the first and the smallest half of the latter, and then finding those segments, contained in both sets, resulting in a set k_{bst} of $N_{bst} \leq N_k/2$ segments. This method will be referred to as intersection method. The length of the frames was set to one second. This is due to the experience gained while hand labeling data. It is not possible to segment data more exactly particularly when the singing voice is dying away in a continuous instrumental accompaniment.

D. Classifying a song

After the vocal frame selection, the classification to one of the $j = 1...C$ singers is performed by maximum likelihood taking in account all the N_{bst} frames from the segments chosen above as depicted in (4).

$$S_{max} = \underset{j}{\operatorname{argmax}} \sum_{k_{bst}} \sum_{i=1}^{L} \log p(\mathbf{c}_{k_{bst}L+i}|\theta_j) \qquad (4)$$

IV. Experiments

In the experiments we examined the truth of the following **hypothesis**:

In order to give the best mean classification accuracy, a system has to use features that have been pre-processed by CMS as described in (1). The singer model should be trained on solo vocal frames, world models should be used for frame selection.

In this section this hypothesis will be examined and the influence of a score normalization will be shown. The errors of the system are featured as well, because it is assumed that by explaining and interpreting these errors, conclusions can be drawn concerning the character of the data set and the used statistical models. Values for the system accuracy will be given as the mean percentage of correct classified songs throughout all classes. First we will decide for a proper way to select frames in the testing phase.

A. Vocal frame selection

The performance of using the maximum method following (3) for vocal frame selection, will be compared with the proposed intersection method, as described in Section III-C. The 21 hand labeled songs from the test set have been used to evaluate the accuracies of the methods. The results of this experiment are depicted in Table II. It shows that the intersection method almost never failed. From the 375 frames labeled as vocal, only 5 frames were instrumental frames. The maximum method has a higher false acceptance rate, resulting in an accuracy of 83%, compared to 99% with the proposed method. The obtained number of segments marked as vocal was smaller for the intersection method, but with at least ten seconds per song enough for a reliable decision. Motivated by these results, for the following experiments, the labeling has been performed using the intersection method.

TABLE II
CLASSIFICATION ACCURACIES OF HYPOTHESIS SYSTEM

method	accuracy	correct frames	false accepted frames
intersection	99%	370	5
maximum	83%	1970	343

B. Verification of Hypothesis

The classification accuracy of the hypothesis system for each class and the confusion matrix are shown in Tables III and IV, respectively. Each row of the confusion matrix shows, how many songs of this artist have been assigned to which class.

For 14 classes the percentage of correctly classified songs is 80% or more. Next we have four singers that are recognized with a percentage over 50%. But it is remarkable that three classes are far below 50% (classes S3, S17 and S18). This characteristic result remained the same throughout all experiments. An analysis of the content of these classes will be performed in Section IV-C. It is interesting to point out here, that singer S21 was perfectly recognized, as this class contains a singer duo.

In Table V, we report the mean classification accuracies for all tested changes in the feature set and the trained models. For convenience, the first row reports the mean value of the accuracies from the hypothesis setup as listed in detail in Table III. The results in the next row (θ^{vox}) have been achieved by using all vocal frames of a class for training. The accuracy has slightly decreased, so concentration on the frames sung only by the target singer results in a more descriptive model for his/her voice. The next row, θ^{instr}, delivers the proof for the similarity of the music throughout the classes. The value of 36% result from most classes having very low accuracies and three that remained the same at a high level. These are singers S6, S20 and S21. This indicates a small variance of the songs contained in this class. Indeed this can be confirmed, all songs from S6 and most from S20 are from specific concert recordings, all songs from S21 are from the only album of the artist available to the authors. Note that for the classification with the θ^{instr} models the world model segment selection has not been used.

The next row shows the result when leaving out the vocal frame selection in the testing phase for the θ_{solo}. This shows that focussing on vocal frames in the testing also improves results, even when the improvement is not very big. A major cause of the high accuracy of the presented system is the usage of Cepstral Mean Subtraction. As can be seen from Table V, the accuracy decreases from 81% to 66%, when not using CMS. This shows that the linear filtering applied in a CD mastering process affects the significance of a statistical model for the singer dramatically, at least for feature and data sets used here. The usage of score normalization (2) did not improve results. As shown the accuracy slightly decreased in

TABLE III

CLASSIFICATION ACCURACIES OF HYPOTHESIS SYSTEM

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
100	90	27	100	80	100	100	89	100	100	58	93	100	70	73	67	33	23	100	90	100

TABLE IV

CONFUSION MATRIX OF THE HYPOTHESIS SYSTEM

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
S1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S2	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
S3	0	0	4	3	0	1	0	2	1	0	0	0	0	0	0	1	1	0	1	0	1
S4	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S5	0	0	0	0	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
S6	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S7	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S8	0	0	0	0	0	1	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
S9	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0
S10	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0
S11	0	0	1	0	0	0	0	0	0	0	7	0	1	0	0	1	2	0	0	0	0
S12	0	0	0	0	0	0	1	0	0	0	0	13	0	0	0	0	0	0	0	0	0
S13	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0
S14	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	3	0	0	0	0	0
S15	0	0	0	1	0	0	1	0	0	0	0	0	0	0	8	0	1	0	0	0	0
S16	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	8	0	0	0	0	0
S17	0	0	0	9	0	0	0	0	0	0	0	0	0	1	0	0	5	0	0	0	0
S18	0	3	0	2	0	0	0	0	0	0	0	1	0	0	0	0	2	3	0	0	2
S19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0
S20	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0
S21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12

relation to the hypothesis system, to 79% when using the instrumental world model θ_{instr}, and to 78% when subtracting both world model likelihoods. This shows that score normalization is applicable when dealing with clean speech signals, but as our singer signals are strongly interweaved by music signals, it seems not applicable in the presented problem.

TABLE V

CLASSIFICATION ACCURACIES FOR DIFFERENT SETUPS

Setup	Accuracy
Hypothesis	81%
θ^{vox}	77%
θ^{instr}	36%
Hypothesis/No CMS	66%
Hypothesis/No frame selection	76%
Hypothesis/Score Normalization 1	79%
Hypothesis/Score Normalization 2	78%

C. Error analysis

We will focus on the worst classified singers S3, S17 and S18. Singer S3 (Sotiria Bellou) has been classified five times as female singer and six times as male, while the errors for the other female singers were almost always the assignment to another female singer. All the songs assigned to a male singer contain a male singer as second voice in many parts. Only five songs from this singer do not contain second voice, three of them are classified correctly, the two others to other female singers. Singer S17 (Perpiniadis) has been classified almost always as S4 (Dalkas). Both of these singers sung during the same period, both of them having a very powerful tenor voice. It seems like the model of S4 overlapped the model S17, making his classification impossible. The wrong classifications for S18 again have a musical reason. Half of the errors happen within the musical group, S18 was part of. Two others have strong second voice parts and lead to a classification as S21, the singing duo. And again, S4 also here gets 2 votes, indicating that the dominant voice of S4 indeed had its impact on his statistical model.

V. CONCLUSION

This paper introduced a new data set for singer identification and presented its characteristics. Using statistical models for the singers we were able to reach a high recognition performance, with the most significant improvement when using CMS. It was shown that a segment selection using world models lead to a choice of vocal frames almost without errors. The bad quality of the old recordings did not result in worse classification results than for the newer recordings.

The presented data base will be used for evaluating feature sets as introduced in [7] in future. All data along with the labeling can be obtained on request from the authors.

REFERENCES

[1] W.-H. Tsai and H.-M. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, 2006.

[2] T. Zhang, "Automatic singer identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2003.

[3] H. Fujihara, T. Kitahara, M. Goto, and K. Komatani, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. ISMIR Conference*, 2005.

[4] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *ICASSP 2003*, 2003.

[5] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, 1974.

[6] F. Pachet and J.-J. Aucouturier, "Improving timbre similarity: How high is the sky?" *Journal of negative results in speech and audio sciences*, vol. 1, no. 1, 2004.

[7] A. Holzapfel and Y. Stylianou, "A statistical approach to musical genre classification using non-negative matrix factorization," in *ICASSP 2007*, 2007.