SINGER IDENTIFICATION IN POLYPHONIC MUSIC USING VOCAL SEPARATION AND PATTERN RECOGNITION METHODS

Annamaria Mesaros, Tuomas Virtanen, Anssi Klapuri
Tampere University of Technology
Institute of Signal Processing

ABSTRACT

This paper evaluates methods for singer identification in polyphonic music, based on pattern classification together with an algorithm for vocal separation. Classification strategies include the discriminant functions, Gaussian mixture model (GMM)-based maximum likelihood classifier and nearest neighbour classifiers using Kullback-Leibler divergence between the GMMs. A novel method of estimating the symmetric Kullback-Leibler distance between two GMMs is proposed. Two different approaches to singer identification were studied: one where the acoustic features were extracted directly from the polyphonic signal and one where the vocal line was first separated from the mixture using a predominant melody transcription system. The methods are evaluated using a database of songs where the level difference between the singing and the accompaniment varies. It was found that vocal line separation enables robust singer identification down to 0dB and -5dB singer-to-accompaniment ratios.

1 INTRODUCTION

Singing voice is the main focus of attention in musical pieces with a vocal part; most people use the singers voice as the primary cue for identifying a song. Also, a natural classification of music, besides genre, is the artist name (often equivalent to singers name). A singer identification system would be useful for MIR (music information retrieval) systems in case of identifying singers for songs. 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.

Two main approaches to singer identification have been studied: one where features are computed directly from the polyphonic signal and another using separation and analysis of the vocal source. Treating the polyphonic mix directly and extracting the features for classification relies on the assumption that the singing voice is sufficiently dominating in the feature values. As preprocessing, the authors of [9, 10] located the time segments where vocals are present. After endpoint detection, in [10] the author used a fixed-length segment of 25 s to compute the features. Reported results were 82% on a number of 45 songs from 8 singers, using MFCCs as features and GMM models and maximum likelihood classification.

The second approach is the separation of vocals from the polyphonic mixture. A statistical approach to vocals separation is presented in [5]. Another method to accomplish vocals separation is extracting the harmonic components of the predominant melody from the sound mixture and then resynthesizing the melody by using a sinusoidal model [1, 8]. In addition, the authors of [1] selected reliable frames of the obtained melody to get classification between the vocal and non-vocal frames. Reported results are 95% correct classification on a number of 40 songs from 10 singers, using 15 linear prediction mel cepstral coefficients and 64 components GMM maximum likelihood classification.

The question that arises is which of the former methods is more robust to accompaniment influences, and to which degree. This paper 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. 65 songs from 13 singers were mixed at levels starting with clean voice to 0dB and -5dB singing-to-accompaniment ratio (SAR). 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. The acoustic material was produced so that the accompaniment does not provide any information about the singer's identity. This ensures that the evaluation is based on singer identification and not on the accompaniment.

The paper is organised as follows. Section 2 gives general guidelines about the features and the classification methods, including a detailed description of the proposed Kullback-Leibler divergence between GMMs. Section 3 explains the vocal separation algorithm, then in section 4 the organization of the different classification tasks is described. The experimental results are presented in the same section, then conclusions and future directions are pointed out.

2 FEATURES AND MODELS

The MFCCs (Mel-frequency cepstral coefficients) have been the most successful acoustic features in speech and speaker recognition systems. They have also been successfully used in artist identification [4] and instrument identification. A bank of filters equally spaced in Mel-frequency scale resamples the frequency axis. A discrete cosine transform (DCT) is applied to the mel-resolution power spectrum, and the lower coefficients of the DCT are used to represent a rough shape of the spectrum. The features used for classification are vectors of 12 MFCCs, computed on 34 ms frames. The zeroth order coefficient was used to detect the voiced frames and was discarded in the classification. Delta-MFCCs are not used.

2.1 Linear and quadratic discriminant functions

Discriminant analysis is a simple technique for classifying a set of observations into predefined classes. Based on training data, the technique constructs a set of discriminant functions

$$L_i = \mathbf{x}^T \mathbf{a}_i + c_i \tag{1}$$

where $\mathbf{a}_i$ is a vector of discriminant coefficients of class i, $\mathbf{x}$ is a feature vector and c is a constant. Given a new observation, the discriminant functions are evaluated and the observation is assigned to the class having the highest value of the discriminant function. After individual frames classification, the entire signal is assigned to the class where the majority of the frames were assigned. By allowing cross terms, we obtain quadratic discriminant functions of the form $\mathbf{x}^T \mathbf{A}_i \mathbf{x} + c_i$ ($\mathbf{A}_i$ being a matrix) that can model more complex boundaries between classes.

2.2 GMM-based maximum likelihood classifier

A Gaussian mixture model (GMM) for the probability density function (pdf) of $\mathbf{x}$ is defined as a weighted sum of multivariate normal distributions:

$$p(\mathbf{x}) = \sum_{n=1}^{N} w_n \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n), \tag{2}$$

where w_n is the weight of the n-th component, N is the number of components and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ is the pdf of the multivariate normal distribution with mean vector $\boldsymbol{\mu}_n$ and diagonal covariance matrix $\boldsymbol{\Sigma}_n$. The weights w_n are nonnegative and sum up to unity. The standard procedure to train a GMM is the expectation-maximization (EM) algorithm, and the resulting parameters form an inherently discriminative model of the singer classes. The classification principle in the maximum likelihood classification is to find the class i which maximizes the likelihood L of the set of observations $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$:

$$L(X; \lambda_i) = \prod_{m=1}^{M} p_i(\mathbf{x}_m) \tag{3}$$

where λ_i denotes the i-th GMM and $p_i(x_m)$ the value of its pdf for observation x_m. The above criterion assumes that the observation probabilities in successive time frames are statistically independent.

2.3 Song-level nearest neighbour classifier

As an alternative to combining frame-level features, song-level features [4], where the classification is based on longer signal segments, have recently turned out to produce good results in artist classification. For example Mandel and Ellis [4] measured the similarity between two signals by the distance between their frame-level feature distributions.

In this paper we propose a similarity measure based on symmetric Kullback-Leibler divergence to be used in nearest-neighbor classification. We have a set of previously trained singer GMMs and the pdf of the observed features of a song is modeled with a GMM. The song is assigned to singer class having the smallest KL divergence value.

The symmetric Kullback-Leibler divergence between a singer pdf $p_1(\mathbf{x})$ and a song pdf $p_2(\mathbf{x})$ is given by

$$S(p_1(\mathbf{x})||p_2(\mathbf{x})) = D(p_1(\mathbf{x})||p_2(\mathbf{x})) + D(p_2(\mathbf{x})||p_1(\mathbf{x})), \tag{4}$$

where the Kullback-Leibler divergence D is given as

$$D(p_1(\mathbf{x})||p_2(\mathbf{x})) = \int p_1(\mathbf{x}) \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} dx, \tag{5}$$

where the integral denotes multiple integration over the whole feature space. When $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ are modeled with GMMs, the above integral can be solved only when a single Gaussian is used [3]. Some methods exist for approximating the divergence [3]. Monte-Carlo approximation [4] for multiple Gaussians calculates the divergence by using a set of samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$, drawn from the distribution $p_1(\mathbf{x})$:

$$D(p_1(\mathbf{x})||p_2(\mathbf{x})) \approx \sum_{m=1}^{M} \frac{1}{M} \log \frac{p_1(\mathbf{x}_m)}{p_2(\mathbf{x}_m)}. \tag{6}$$

When the dimensionality of $\mathbf{x}$ is large, an accurate approximation requires a huge amount of samples and is therefore not computationally practical.

Here we use the observations $X_1 = \mathbf{x}_1^1, \mathbf{x}_2^1, \ldots, \mathbf{x}_M^1$ that were used to train the distribution $p_1(\mathbf{x})$ as samples $\mathbf{x}_m$. They are the most representative samples of the distribution, since the distribution was trained using them. We observe that the resulting *empirical Kullback-Leibler divergence* can be written using the likelihoods (3) as

$$D_{\text{emp}}(p_1(\mathbf{x})||p_2(\mathbf{x})) = \frac{1}{M} \log \frac{L(X_1; \lambda_1)}{L(X_1; \lambda_2)}. \tag{7}$$

Since the term $L(X_1; \lambda_1)$ is fixed for each model λ_2, the empirical Kullback-Leibler divergence corresponds to the maximum likelihood classification [4].

In the symmetric empirical Kullback-Leibler divergence we include the empirical Kullback-Leibler divergence $D_{\mathrm{emp}}(p_2(\mathbf{x})||p_1(\mathbf{x}))$ obtained using the set of points $X_2 = \mathbf{x}_1^2, \mathbf{x}_2^2, \ldots, \mathbf{x}_N^2$ which are the observations used to train the distribution $p_2(\mathbf{x})$. The symmetric empirical Kullback-Leibler divergence can then be written as

$$S_{\mathrm{emp}}(p_1(\mathbf{x})||p_2(\mathbf{x})) = \frac{1}{MN} \log \frac{L(X_1;\lambda_1)L(X_2;\lambda_2)}{L(X_1;\lambda_2)L(X_2;\lambda_1)}) \tag{8}$$

The above measure is close to the cross-likelihood ratio [2, 7] with the exception that terms $L(X_1;\lambda_1)$ and $L(X_2;\lambda_2)$ are in [2, 7] replaced by $L(X_1;\lambda_{12})$ and $L(X_2;\lambda_{12})$, where the model λ_{12} is trained using both X_1 and X_2.

3 VOCALS SEPARATION

For the separation of vocals from the accompaniment, we apply the melody transcription system [6] followed by sinusoidal modeling resynthesis. Within each frame, the melody transcriber estimates whether significant melody line is present, and estimates the MIDI note number of the melody line.

In the voice resynthesis, harmonic overtones are generated at integer multiples of the estimated fundamental frequency. Amplitudes and phases are estimated at every 20 ms from the polyphonic signal by calculating the cross-correlation between the signal and a complex exponential having the overtone frequency. Time-domain signal is obtained by interpolation of the parameters between successive frames

4 SIMULATION EXPERIMENTS

The database consists of 13 singers, containing both male and female perfomers with varying levels of singing skills. From each singer, 4-6 melodies with length of 20-30 seconds were recorded with sampling rate of 44100 Hz and 16 bit resolution. Each singer was given the same accompaniment. This ensures that the accompaniment and the mixing procedures are not singer specific. All the classification experiments were performed using 4-fold cross validation so that the training set contains all the data of a singer except the one song that is tested. The reported results are the average of the 4 experiments.

We used both artist-level and song-level GMMs, the latter resembling the modeling in [4]. The number of Gaussians in all the models was 10. The artist-level GMM is trained with all the songs from the training set, the resulting model being associated with the singer identity. For testing, the likelihood of the test song was calculated under each of the 13 GMMs representing singers, and the most likely singer was chosen. The song-level modelling constructs one GMM for each song, obtaining several GMMs associated to each singer, then the test song is classified according to the singer of the song which is closest to the one under analysis. The KL divergence distance was used with nearest neighbor classification, 1NN in artist-level GMM, 1NN and 3NN in song-level GMM. We also tested the symmetric KL divergence between artist-level single Gaussians and the Mahalanobis distance [4]. The acronyms used for the described classifiers are the following: LDF - linear discriminant functions; QDF - quadratic discriminant functions; GMM-A - artist-level GMMs, maximum likelihood classification; GMM-KL-A - artist-level GMMs and KL divergence; GMM-S - song-level GMMs, maximum likelihood classification; GMM-KL-S-1NN, GMM-KL-S-3NN - song level GMMs and KL divergence with one and with three nearest neighbors; G-KL-A - artist-level single Gaussian and KL divergence; G-Mah - artist-level Mahalanobis distance.

Each classification experiment was run for various SARs: -5dB, 0dB, 5dB, 10dB and 30dB, directly on the polyphonic mixture and also on the separated vocal line from each type of SAR mixture. The same SAR data was used both in training and testing. Also when separation was used, separation was applied also during the training.

In the first stage, the different classifiers were tested for the various SARs and the average classification rates are presented in Table 1. The linear discriminant function classifier is used to check the separability of the dataset; its classification performance and the two best classifiers are depicted in Figure 1, left.

With separation, the classification performance of the discussed classifiers shows visible improvement, as presented in Table 2 and in Figure 1, right. The identifica-

SAR [dB]	-5	0	5	10	30
LDF	28	42	55	61	63
QDF	42	53	57	69	75
GMM-A	38	36	53	65	71
GMM-KL-A	26	51	63	73	78
GMM-S	25	28	44	50	57
GMM-KL-S-1NN	21	32	32	55	59
GMM-KL-S-3NN	26	42	48	61	73
G-KL-A	13	25	36	40	38
G-Mah	25	34	48	57	65

Table 1. Classifiers performances on polyphonic mixtures at different SARs

SAR [dB]	-5	0	5	10	30
LDF	44	46	50	59	46
QDF	63	61	67	77	67
GMM-A	67	75	79	80	84
GMM-KL-A	63	69	82	78	75
GMM-S	51	59	71	73	76
GMM-KL-S-1NN	50	61	65	65	67
GMM-KL-S-3NN	51	61	59	65	69
G-KL-A	46	51	50	51	48
G-Mah	53	51	53	51	48

Table 2. Classifiers performances on vocals separated from polyphonic mixtures at different SARs

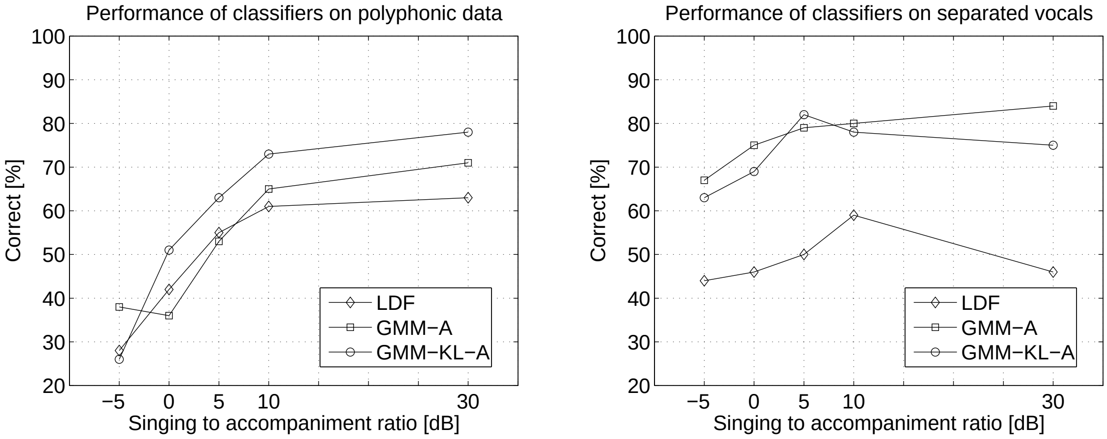

Figure 1. LDF baseline and the two best classifiers for polyphonic data (left) and separated vocals (right)

tion accuracy improves at 0dB SAR from 36% to 75% for GMM-A, and for GMM-KL-A it improves from 51% to 69%. One effect of the separation procedure is that the noisy sections of the melody, where no harmonic content is found, are reduced to silence.

The GMM-KL-A classifier seems to be more robust for the nonseparated case, and it performs comparable with the GMM-A classifier in the separated cases. The song level modeling and 3NN KL distance classification also shows robustness for the separated vocals case, but not as large improvement as the artist-level modeling. A simple explanation of this is the small number of training samples, this type of modeling being more appropriate to music classification in large databases where an artist GMM has a very large amount of data available for training.

5 CONCLUSIONS

In this paper we tested methods for singer identification in polyphonic music. Identification on both polyphonic music and separated vocals was tested. The simulation results show that singer identification down to realistic SARs (0dB, -5dB) is possible. The vocals separation improves the identification performance significantly at low SARs. The proposed method for approximating the Kullback-Leibler divergence produces comparable results with the best reference methods on separated vocals. On polyphonic data, it enables better average accuracy than the existing approaches. The future work includes different statistical models such as hidden Markov models and other classification methods such as support vector machines. [1]

6 REFERENCES

[1] Fujihara, H., Kitahara, T., Goto, M., et. al. ”Singer Identification Based on Accompaniment Sound Reduction and Reliable Frame Selection”, *Proc. of 6th ISMIR*, London, U.K., 2005.

[2] Gish, H., Siu, M-H., Rohlicek, R. ”Segregation of speakers for speech recognition and speaker identification”, *Proc. of ICASSP*, Toronto, Canada, 1991

[3] Hershey, J. and Olsen, P. ”Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models”, *Proc. of ICASSP*, Honolulu, USA, 2007

[4] Mandel, M. and Ellis, D. ”Song-Level Features and Support Vector Machines for Music Classification”, *Proc of .6th ISMIR*, London, U.K., 2005.

[5] Ozerov, A., Philippe, P., et. al. ”One Microphone Singing Voice Separation using Source-adapted Models”, *Proc. of 2005 IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics*, New York, USA, 2005

[6] Ryynänen, M. and Klapuri, A. ”Transcription of the Singing Melody in Polyphonic Music”, *Proc of .7th ISMIR*, Victoria, BC, Canada, 2006

[7] Tsai, W-H, Wang, H-M. ”Speech utterance clustering based on the maximization of within-cluster homogeneity of speaker voice characteristics”, *Journal of the Acoustical Society of America, no. 3, vol. 3*, 2006

[8] Yipeng, L. and Wang, D. ”Singing Voice Separation from Monaural Recordings”, *Proc. of 7th ISMIR*, Victoria, BC, Canada, 2006

[9] Youngmoo, E.K.and Whitman, B. ”Singer Identification in Popular Music using Warped Linear Prediction”, *Proc. of 3rd ISMIR*, Paris, France, 2002

[10] Zhang, T. ”System and Method for Automatic Singer Identification”, *IEEE International Conference on Multimedia and Expo*, Baltimore, MD, 2003.

[1] This work was supported by the Academy of Finland, project No. 5213462 (Finnish centre of Excellence program 2006-2011). The authors wish to thank Matti Ryynänen for providing the algorithm for melody transcription.