EFFECTIVE SINGING VOICE DETECTION IN POPULAR MUSIC USING ARMA FILTERING

Hanna Lukashevich, Matthias Gruhne and Christian Dittmar

Fraunhofer IDMT
Ilmenau, Germany
lkh@idmt.fraunhofer.de

ABSTRACT

Locating singing voice segments is essential for convenient indexing, browsing and retrieval large music archives and catalogues. Furthermore, it is beneficial for automatic music transcription and annotations. The approach described in this paper uses Mel-Frequency Cepstral Coefficients in conjunction with Gaussian Mixture Models for discriminating two classes of data (instrumental music and singing voice with music background). Due to imperfect classification behavior, the categorization without additional post-processing tends to alternate within a very short time span, whereas singing voice tends to be continuous for several frames. Thus, various tests have been performed to identify a suitable decision function and corresponding smoothing methods. Results are reported by comparing the performance of straightforward likelihood based classifications vs. postprocessing with an autoregressive moving average filtering method.

1. INTRODUCTION

The availability of digital music material to end users is continually increasing through new media and content distribution methods. As a result, there is a growing need to automatically categorize and annotate the large amount of data. This allows the user to locate music that fits his or her personal preferences. It's now common sense that semantically meaningful descriptions (e.g. genre, tempo and musical key) of audio content are a suitable means to achieve that goal.

Therefore, active research has been conducted in the field of Music Information Retrieval (MIR) during recent years. Discrimination between vocal and non-vocal parts of popular music has been identified as an importand base technology for further high-level analysis. This information can be used for example in artist identification [1] and singing language recognition [2]. It has furthermore much relevance in lyrics synchronization [3]. One of the early approaches of vocal/non-vocal detection in popular music has been derived from speech/music discrimination and introduced by Berenzweig and Ellis [4]. They performed experiments using several low-level descriptors and Hidden Markov Models (HMM) for discriminating between two classes of a previously annotated and trained database. The reported results vary between 55,2% and 81,2%, depending on the utilized features. Tzanetakis [5] performed experiments with different low-level features and a multitude of classifiers. The reported results range between 61% and 75%. Maddage et al. [6] introduced an approach for vocal/non-vocal detection without a previously trained classifier. They performed a Fourier transform on the subbands of the spectrum of the signal. Thereafter they decided if the signal is music or vocal based on simple thresholding. They reported an accuracy of 84%. Unfortunately all these approaches are not directly comparable, because all publications are based on a different test set, varying in musical content and size.

One of the base approaches that is relatively straightforward to implement uses Mel-Frequency Cepstral Coefficients (MFCCs) and a Gaussian Mixture Model (GMM) classifier. This technique has been used in artist detection, singing language detection and lyrics synchronization [1], [2], [3] and it exhibits performance comparable to more complex systems.

With the combination of MFCCs and GMMs one often encounters rapidly alternating output, that is semantically meaningless for the target application. Therefore, a smoothing function for decimation of outliers has been introduced in [7], where Tsai et al. accumulated the log likelihoods of single frames over a certain time span in order to achieve more reliable results. Thus, we decided to pursue this approach and concentrate on postprocessing of intermediate classification results. We identified that the instability in classifying depends on factors like model quality, generality of training data and complexity of test material. Since the influence of the above mentioned factors can only be reduced to a certain extent we investigated into finding a suitable smoothing algorithm. This paper introduces a novel method for deriving a bounded decision function and appropriate smoothing with an Autoregressive Moving Average (ARMA) filter [8].

The structure of the paper is organized as follows. The next two chapters describe feature vector extraction and GMMs. Section 4 presents our decision function, the subsequent ARMA filtering and additional smoothing. Thereafter the audio data set used in the evaluation is described. Section 6 depicts the details of the experiment and the corresponding results. Finally section 7 concludes this work and provides some perspectives for future directions.

2. FEATURE VECTOR EXTRACTION

From the multitude of features that have been suggested for MIR applications we have chosen to utilize MFCCs. MFCCs and derivates have found multiple successful applications in the field of speech recognition and speaker identification and has proved to be well-suited for MIR, for example in singer and artist identification [1], [2], [3]. The term cepstral originates from fundamental research of Bogert [9]. The main point is the implicit decomposition of a periodic signal into excitation and filter. The most straightforward way to compute MFCC is the summation of FFT bins weighted by the Mel-Filterbank passbands, taking the natural logarithm and subsequent discrete cosine transform.

The coefficients computed by that method can be thought of as

weighting factors for different periodic characteristics in the logarithmic distribution of energy in the Mel-bands. The very first coefficient equals the overall energy and should be omitted for classification purposes to be prone against different amplification factors. The succeeding coefficients represent a more detailed description of the energy distribution in Mel-bands. Therefore, the number of coefficients is limited to D in order to generalize the properties of the current audio frame whilst omitting subtile dynamic aspects. Furthermore, the implicit orthogonality of MFCCs simplifies the theoretical background of statistical modeling.

3. GAUSSIAN MIXTURE MODELS

Our main interest is targeted towards discrimination of the two classes: music and music plus singing voice (further denoted as MUS and VOX respectively). For each of the above mentioned classes one particular Gaussian mixture model represents the distribution of the raw data in D-dimensional feature space as linear combination of several D-dimensional Gaussian probability density functions (PDF). These two Gaussian mixture models are further denoted as MUS GMM and VOX GMM. The parameters of the component densities are estimated with the well-known expectation maximization (EM) algorithm [10] [11]. The linearly weighted combination of Gaussian basis functions is expected to generalize the collected features forming smooth approximations of their arbitrarily shaped PDFs. Equation 1 gives the definition of a GMM defined as a weighted sum of M component PDFs according to [12]

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i g_i(\mathbf{x}) \tag{1}$$

where $g_i(\mathbf{x})$, $i = 1, ..., M$ represent the component PDFs, $\mathbf{x}$ is a D-dimensional observed feature vector and p_i the individual mixture weights or priors. Each component is defined as a D-variate Gaussian PDF

$$g(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \, exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\} \tag{2}$$

with empirically estimated mean vector μ and covariance matrix $\mathbf{\Sigma}$. This way, a particular mixture PDF is completely parameterized by the tuple $\lambda_i = \{p_i, \mu_i, \mathbf{\Sigma}_i\}$. The training process is constituted by the maximum likelihood (ML) estimation of the model parameters that maximize the likelihood of the GMM given the training data consisting of feature vectors for one class. The ML optimization is actually carried out by the expectation maximization (EM) algorithm, iteratively refining the initial estimation of parameters [12].

The initial estimation of parameters is computed per model by choosing an appropriate M and partitioning the classes in feature space using k-means clustering. The preceding clustering step guarantees convergence to invariant ML estimates and is therefore favoured in contrast to random model initialisation.

4. DECISION FUNCTION

The usage of VOX GMM and MUS GMM allows us to calculate likelihoods of both models for every input frame of data. Let $L(\lambda_v|\mathbf{x})$ and $L(\mathbf{x}|\lambda_m)$ denote the likelihoods of feature vector $\mathbf{x}$, to belong to VOX and MUS classes respectively.

In previous works [7], a decision function was derived as a simple difference between log-likelihood values for VOX and MUS classes as given in equation (3).

$$f_1(\mathbf{x}) = log(L(\lambda_v|\mathbf{x})) - log(L(\lambda_m|\mathbf{x})) \tag{3}$$

If the value of the decision function is above the theoretical threshold of 0 then the corresponding frame is considered to belong to the VOX class while values below 0 indicate MUS class.

In this work we propose a novel approach for computing a decision function as given in equation (4).

$$f_2(\mathbf{x}) = \frac{L(\lambda_v|\mathbf{x})}{L(\lambda_v|\mathbf{x}) + L(\lambda_m|\mathbf{x})} - 0.5 \tag{4}$$

The theoretical threshold of the proposed decision function $f_2(\mathbf{x})$ is also arranged to 0. It should be noted that whithout further post-processing both decision functions essentially produce the same results, when it comes to a binary threshold based decision (i.e. indicate if $L(\lambda_v|\mathbf{x})$ is higher than $L(\lambda_m|\mathbf{x})$). Since both decision functions exhibit a very noisy slope, they are not directly suited for utilization in real-world applications. It is not beneficial to make a decision for audio excerpts that are too short to provide semantically meaningfull interpretations. Therefore, the decision functions need an additional smoothing and/or filtering.

Due to the complexity inherent to training two GMMs covering the entire body of real-world music, the absolute values of $L(\lambda_v|\mathbf{x})$ and $L(\lambda_m|\mathbf{x})$ tend to be relatively small. Moreover, absolute values of likelihoods for MUS and VOX parts even within a particular song may exhibit significant differences.

The statistical properties of the above mentioned decision functions have been examined in-depth in order to benefit from their peculiarities. We investigated the PDFs of the values returned by each of the desicion functions separately for MUS and VOX classes of input data. Since manually segmented songs from our audio data test set (see section 5) were available, we had the possibility to split the set of observed input feature vectors $\mathbf{x}$ in two subclasses: VOX and MUS frames contained in the song. For each of these subclasses the PDFs of the decision functions were estimated. Exemplary results for a representative song are shown in Figure 1. It can clearly be seen that although the experimental results for both decision functions proved the liability of the theoretical threshold, the PDFs do exhibit distinct properties. In the upper plot (results for $f_1(\mathbf{x})$), the overlapping region of the PDFs covers a large amount of observations. Thus, even small changes of the thresholding could have significant impact on the classification results. In contrast, the lower plot ($f_2(\mathbf{x})$) depicts overlapping in less critical regions. A well-established technique to improve correct classification rate is defining a so-called uncertain zone around the threshold. One can see that for the decision function $f_1(\mathbf{x})$ it will yield a high amount of uncertain frames. In addition, the borders of uncertain zone for $f_2(\mathbf{x})$ must be given in absolute values which tend to vary depending on the song.

Moreover, our experiments proved that the decision function $f_2(\mathbf{x})$ is the most suitable for filtering and smoothing. It is ranged between -0.5 and 0.5, and it is symmetrical around the threshold. As it will be shown below, $f_2(\mathbf{x})$ can be successfully filtered using ARMA filtering [8].

4.1. Autoregressive Moving Average Filtering

As singing voice generally tends to be continuous for multiple consecutive frames, we assumed that the instantaneous value of decision function of frame i is partly determined by k previous frames,

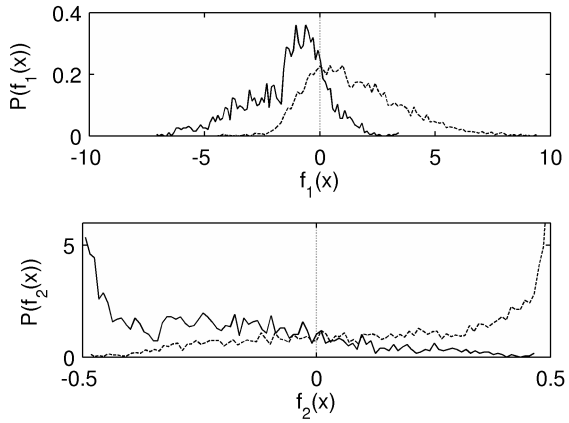

Figure 1: *Comparison of PDFs of the decision functions for a representative song. The solid line in both plots corresponds to MUS frames of the song, and the dashed line corresponds to VOX frames of the song. The upper plot shows the PDFs recieved for $f_1(\mathbf{x})$ and the lower plot shows the PDFs recieved for the decision function $f_2(\mathbf{x})$.*

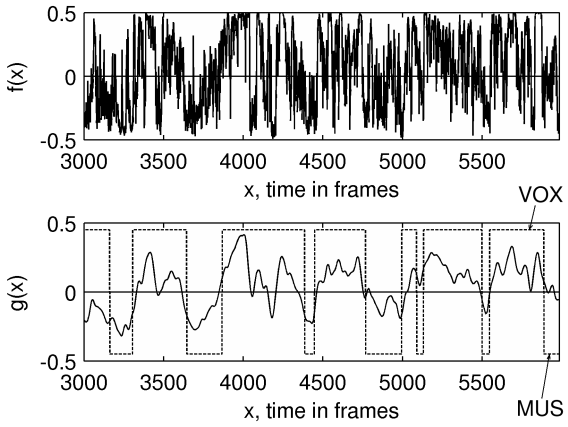

Figure 2: *The upper plot shows decision function $f_2(\mathbf{x})$ for an exemplary excerpt of a representative song. The lower plot depicts the decision function $g(\mathbf{x})$ after ARMA filtering and smoothing. For comparison, the dashed function represents the manual segmentation for this audio excerpt.*

i.e. it can be interpreted as autoregressive (AR) process. In addition, smoothing of the decision function for removing short term outliers can be efficiently performed by means of moving average (MA) processing. The combination of the above mentioned post-processing steps can be interpreted as an ARMA(p,q) process. This process can be approximated by a rational transfer function [13] given by the linear difference equation:

$$x_i = \sum_{l=1}^{q} b_l n_{i-l} - \sum_{k=0}^{p} a_k x_{i-k}. \tag{5}$$

The system transfer function $H(z)$ between the input (n_i) and the output (x_n) for the described ARMA process is the rational function $H(z) = B(z)/A(z)$, where $A(z)$ and $B(z)$ represent the z-transforms of the AR and MA branches respectively. We calculate the coefficients b_l and a_k of the ARMA filter via Prony's method [13], [14]. Prony's method is an algorithm for finding an IIR filter with a prescribed time domain impulse response. This filter can recover the coefficients b_l and a_k exactly if the data sequence is truly an ARMA process of the correct order. The order of the ARMA filter was determined experimentally. The best results were received for $p = q = 10$. An examplary result of ARMA filtering applied to the decision function $f_2(\mathbf{x})$ is shown in the lower plot of Figure 2. In that plot, additional smoothing via convolution with a Hamming window was applied.

5. AUDIO DATA TEST SET

To assess the performance of the proposed method, we had to define a proper evaluation test bed. Due to the fact that there exists no well established database for that particular task, we decided to set up a proprietary test set by ourselves. Our test database consists of 84 PCM WAV-files. All files are downsampled to 16 bit, 22050 kHz, mono. The database contains 10 singers: 5 male and 5 female (see Table 1). The songs of every singer were randomly separated into training set (38 songs, 3-5 songs for every singer) and test set (46 songs, 4-5 songs for every singer). Every record in the database was manually labeled with regard to instrumental and vocal parts using the open source tool Wavesurfer. The total duration of the training set is 5815.47 sec, which equals more than 1.5 hours of music. The total duration of the test set is about 3000 sec, or 50 minutes, whereas only 1 minute excerpts of every song were considered (from 20 sec to 80 sec).

Male Singers	**Female Singers**
Brian Adams	Barbara Streisand
Eros Ramazotti	Anna Netrebko
Frank Sinatra	Nelly Furtado
Ozzi Osbourne	Anne Clark
Sting	LeAnn Rimes

Table 1: Singers in the Database

6. EVALUATION AND RESULTS

At the stage of feature extraction we used $D = 13$ Mel-frequency coefficients computed with 30 ms framesize and 10 ms hopsize. As our system is considered to constitute a front-end for further singer identification and lyrics alignment, we focussed on minimizing the error in identification of MUS frames, thus errors for VOX frames were considered to be less critical. For that reason, the number of mixtures for MUS GMM was set to 20, and the number of mixtures for VOX GMM was 13. These optimum parameters had been identified experimentally, the search was performed in an interval from 4 to 52 mixtures per model. The covariance matrices Σ were assumed as diagonal, considering the fact that they describe uncorrelated MFCCs.

The criterion F used to describe the classification rate has been defined as the harmonic mean (8) of V (6) and M (7).

$$V = \frac{number\ of\ voice\ frames\ detected\ correctly}{total\ number\ of\ voice\ frames} \tag{6}$$

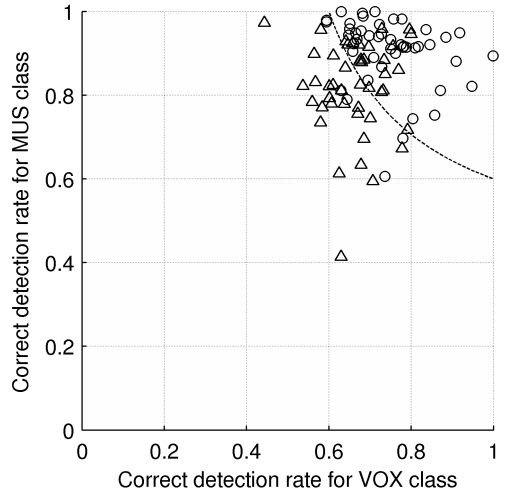

Figure 3: *Correct detection rates for all 46 songs of the test set. Dotted line corresponds to $F = 0.75$. Triangles represent baseline classification results without post-processing. Circles depict classification results achieved with the proposed method.*

$$M = \frac{number\ of\ music\ frames\ detected\ correctly}{total\ number\ of\ music\ frames} \quad (7)$$

$$F = \frac{2VM}{V+M} \quad (8)$$

Figure 3 shows the achieved results for each of the 46 songs of the test set with and without post-processing. Application of the proposed approach resulted in an average increase of the F-score from 72,7% to 81,3%. With our approach, the average result for MUS class is 90,5% while the average result for VOX class is 75,0%. Classification performance can be observed to increase significantly for the VOX class. This is due to fact that the VOX GMM contains less mixtures than its counterpart. So the possibility of spurious thresholding becomes higher in the raw unsmoothed detection function. As we mentioned before, the mistakes in the MUS class are considered more critical and therefore the outcomes correspond to our target. Besides relatively high correct detection rate, the usage of the suggested approach allows to retrieve semantically meaningful consecutive song segments of MUS and VOX. These can be effectively used for further applications e.g. lyrics alignment.

7. CONCLUSION

This paper described our approach towards automatic detection of singing parts in popular music. We used the well established methods of combining MFCCs and GMMs as a front-end. We showed that comparably straightforward methods of post-processing produce significant increase in classification results. Moreover, the application of the proposed decision function in conjunction with subsequent ARMA filtering explicitly enhances the perceptual quality of the achievable song segmentation. The properties of the described decision function can presumably be exploited in systems using further audio features and additional classification techniques such as HMMs, Support Vector Machines or Neural Networks. The information that can be derived from statistical analysis of the decision function allows for additional refinement stage based on heuristics. In addition, the filtered and smoothed decision function carries valuable information that can be interpreted in a semantically meaningful manner. For instance, its local minima indicate borders of phrases apparent while singing. These peculiarities will be studied more in-depth in future work.

8. REFERENCES

[1] A. Berenzweig et al., "Using voice segments to improve artist classifiaction of music," in *Proc. AES 22 International Conference on Virtual, Synthetic and Entertainment Audio, Espoo, Finland*, June 2002.

[2] W. H. Tsai and H.-M. Wang, "Towards automatic identification of singing language in popular music recordings," in *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR)*, 2004.

[3] S. G. Kai Chen, "Popular song and lyrics synchronisation and ist application to music," in *Proc. of the Thirteenth Annual Conference on Multimedia Computing and Networking (MMCN)*, 2006.

[4] A. Berenzweig and D. Ellis, "Locating singing voice segments within music signals," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2001.

[5] G. Tzanetakis, "Song-specific bootstrapping of singing voice structure," in *Proc. of the IEEE International Conference on Multimedia & Expo (ICME)*. 2004, pp. 2027–2030, IEEE.

[6] N. Ch. Maddage, K. Wan., Ch. Xu, and Y. Wang, "Singing voice detection using twice-iterated composite fourier transform," in *Proc. of the IEEE International Conference on Multimedia & Expo (ICME)*. 2004, pp. 1347–1350, IEEE.

[7] 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 & Language Processing*, vol. 14, no. 1, pp. 330–341, 2006.

[8] S. L. Marple, Jr., *Digital spectral analysis with applications*, Prentice Hall, Englewood Cliffs, 1987.

[9] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Proc. of the Symposium on Time Series Analysis, Ed.: M. Rosenblatt, John Wiley, New York*, 1963, pp. 209–243.

[10] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, November 1996.

[11] N. M. Dempster, A. P. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, pp. 185–197, 1977.

[12] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[13] S. L. Marple, "A tutorial overview of modern spectral estimation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 2152–2157.

[14] T. W. Parks and C. S. Burrus, Eds., *Digital Filter Design*, J. Wiley & Sons. Inc., 1987.