Music Retrieval in ICOR

Lutz Finsterle[1], Stephan Fischer[2], Ivica Rimac[2], and Ralf Steinmetz[2,3]

1
Institute of Communication Networks and Computer Engineering (IND)
University of Stuttgart
Pfaffenwaldring 47 • D-70569 Stuttgart • Germany
2
Industrial Process and System Communications
Department of Electrical Engineering and Information Technology
Technical University of Darmstadt
Merckstr. 25 • D-64283 Darmstadt • Germany
3
GMD IPSI
German National Research Center
for Information Technology
Dolivostr. 15 • D-64293 Darmstadt • Germany

email: finsterle@ind.uni-stuttgart.de, {fisch, rimac, rst}@kom.tu-darmstadt.de

Abstract

In this paper we describe music retrieval in ICOR (Intelligent COntent Retrieval), a project of Darmstadt TU. It is the goal of ICOR to find new interfaces to support applications of music video and music CDs. Although the project consists of audio and video analysis we concentrate on a description of the audio algorithms in this paper. We describe our MPEG-7 like data structure to store meta information for music pieces and explain which algorithms we use to analyze the content of music pieces automatically. We currently use an applause detection to distinguish live music from studio recordings, a genre classifier to distinguish pieces with beats (for example disco music) from classical music, and a singer recognition.

Keywords: Content processing, automatic audio classification, audio analysis, audio retrieval.

1 Introduction

In the last decade audio and video content processing became a major research issue. Early works dealt with the automatic recognition of video cuts [ZKS93] and with the automatic recognition of genres like newscast [ZGST94, FLE95] or sports games [GSCZS95]. Since that a great variety of different approaches to the automatic analysis of video and audio have been described. These can be categorized as follows:

- Fundamental research.
  Content-based research addresses a lot of fundamental problems, for example image similarity [SCA98], or a recognition of video cuts [ZKS93]. Also many other research areas are involved, for example image processing, neural networks or signal processing.
- General indexing and retrieval systems.
  Starting with QBIC [FSNA+95], systems such as Informedia [SC95], Virage [www.virage.com], and VideoQ [CCMS+97] which provide methods for retrieving digital images and videos by using visual examples and/or sketches for querying, and matching statical and dynamical visual cues, such as color, texture, shape, motion, and spatio-temporal composition attained the attention of the public. These systems are not restricted in terms of applicability for different purposes.

This work is supported in part by Volkswagen-Stiftung, Hannover, Germany.

- Specialized indexing and retrieval systems.
  Specialized systems have been described which focus on the implementation and solution of problems which are narrow in a sense that the solution cannot be adapted to other applications. An example is the image retrieval for petroleum applications described in [LSCB99].

Many systems which have been described follow a bottom-up approach, as the algorithms which have to be used to implement the specific parts of a system are developed and the system is then assembled from those.

A general problem which can be observed is that many projects implement their own set of algorithms. This can be explained by the fact that content processing algorithms are often application-specific and cannot be reused in another context. Another disadvantage is that many published algorithms tend to claim that their new algorithm is the best, but that neutral performance comparisons like [BR96,Lie99] are quite rare. Another problem is that many systems provide a complex user interface where the natural interaction in the form of text is neglected. It can be assumed that only those systems will be accepted widely which provide a straightforward look and feel and which hide the nature of algorithms from the user.

In this paper we present the ICOR (Intelligent COntent Retrieval) system currently being developed at the Technical University of Darmstadt. ICOR is used to index and retrieve music pieces from different genres (classical music, rock music, jazz and soul). The difference between other systems and ICOR is that we follow a top-down approach based on an MPEG-7 like model of music clips. Our approach hence starts with a user interface with a common look and feel. The needs of the user interface are then mapped onto our data model. Using the model we are able to derive the set of algorithms which have to be incorporated into the system, enabling us to reuse those algorithms which seem to be state of the art (for example cut detection). The paper hence explains the audio data model and highlights the audio algorithms necessary for our purpose. We understand ICOR as a first but promising step into the direction of developing powerful but natural user interfaces, using the emerging MPEG-7 standard to develop content retrieval systems which center around the needs of the end-user.

The paper is structured as follows: in Section 2 we explain the data model we use to structure the content of music pieces. We also derive the algorithms necessary to analyze the underlying content. In Section 3 we explain the algorithms we use to analyze the content of music pieces. Section 4 presents our experimental results. Section 5 concludes the paper and provides an outlook.

2 ICOR Data Model

ICOR aims at indexing and retrieving pieces of music, which have been recorded either from a video or from an audio CD. The functionality we offer to the user consists of the following aspects:
- retrieval of name of artist or director,
- retrieval of title of piece of music
- retrieval of genre (classical music, rock music, jazz, and soul music)
- search for pieces which are slower/faster/equally fast with regard to a reference clip
- search for similar pieces.

Having performed a search the user gets a scrollable list which contains the possible answers. Clicking on one of the answers opens a player which plays back a piece of music. To be able to map the requirements of the user interface to the necessary algorithms we use an MPEG-7-like music model. As MPEG-7 is an emerging standard, no obligatory model can be used. However, the MPEG group uses a set of proposals in the standardization process. Our model has been developed based on these proposals. We currently use the following model where we highlight the content similar to the language C instead of XML for the sake of clarity.

data_struct(genre_type) {
type=list;
values='blues','rock','soul','classic','jazz';
}
	data_struct(song_type) {
	type=list;
	values='studio','live';
	}
		data_struct(title) {
		type=string;
		}

cd(ID=string) {
title=string
type=string
artist=string // name of artist
session_type=song_type
genre=genre_type
label=string
company=string
date=DATE
link=string
song=list(song_type)
}
	data_struct(song_type) {
	title=string
	type=song_type
	genre=genre_type
	label=string
	company=string
	date=DATE
	info {string}
	description=descriptive_data
	vocals=list(lead_vocals)
	musicians = list(musician)
	applause = applause_type
	}
		data_struct(artist) {
		name=string
		type=genre
		link=string
		}

data_struct(applause) {
type=string
position(){
start=TIME
end=TIME
duration=TIME
}
extracted_by(){see above}
link=string
}
	data_struct(musician) {
	vocals() {
	list(artist)
	}
	instruments(){
	list(artist, instrument_type)
	}
	}
		data_struct(lead_vocals) {
		name=string
		extracted_by() {
		name= matrix|vector|raw
		}
		link=string
		}

data_struct(bpm_type){
value=integer
extracted_by() {
name= matrix|vector|raw
}
}
	data_struct(descriptive_data) {
	beats_per_minute=bpm_type
	link=string
	}
		data_struct(instrument_type) {
		name=string
		link=string
		}

3 Audio Analysis

To fill the proposed MPEG-7 model with meta data audio features have to be extracted automatically. The algorithms we apply include:

- an algorithm to detect whether the examined piece is a live or a studio production. To achieve this goal we search for appearances of applause or massive hand clappings within the audio data, considering that applause is represented as a statistical energy distribution over some specific frequency band.
- an algorithm to calculate the beats per minute of a music piece to be able to take the decision what kind of genre the examined data belongs to.
- The last and most complex analysis is needed for a singer recognition. The algorithm we use grounds on the extraction of the unchangeable vocal track characteristics every artist has.

The mechanisms and algorithms for the automatic classification described above will now be described in detail, starting with the detection of applause within the audio data stream.

3.1 Applause Detection

Applause can be represented as a frequency band where the energy is distributed statistically. This is due to the fact that applause can be modeled as peaks which are distributed statistically. In Figure 1 a close relation between the synthesized applause signal and gaussian noise filtered by a bandpass digital filter can be observed. This phenomenon can be taken for common with slight parameter changes for different recordings. These parameters are mainly the center frequency of the frequency distribution and the width of the frequency band.

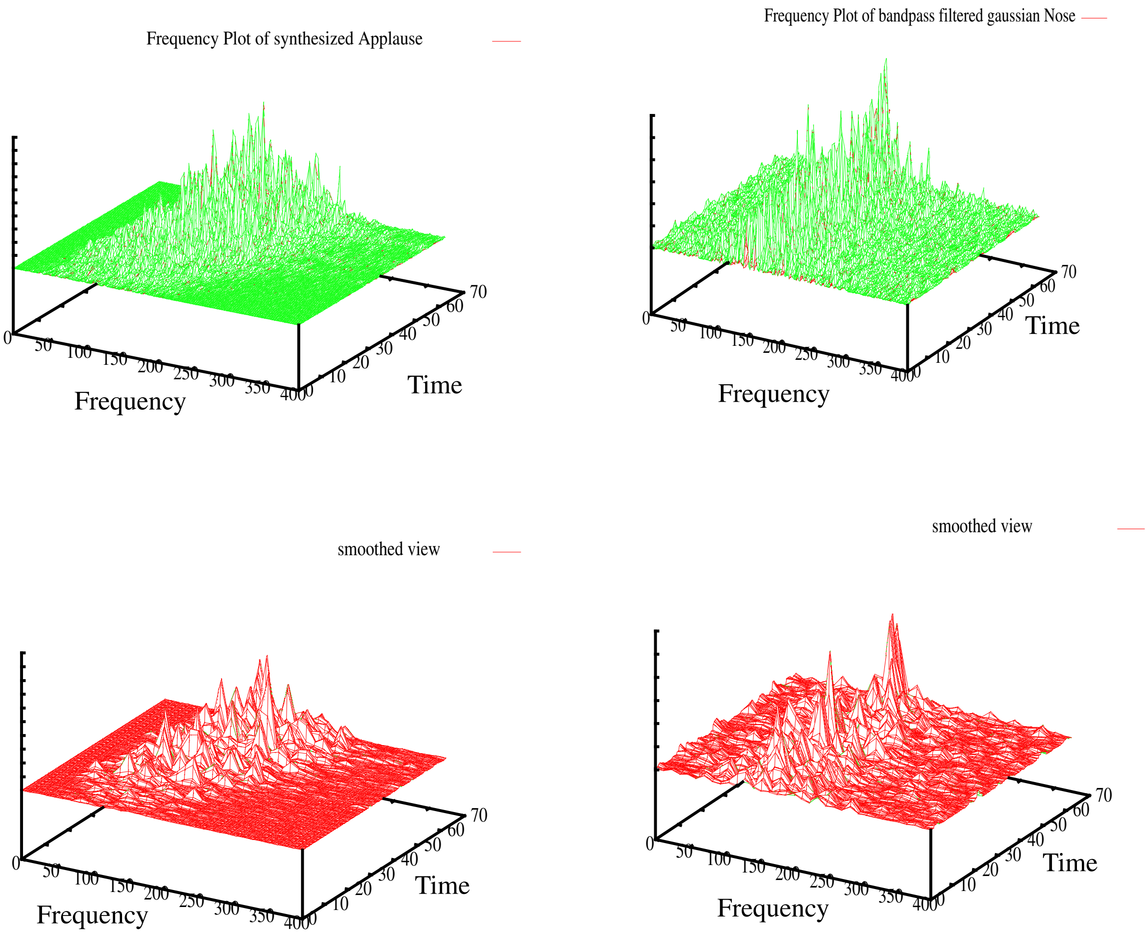


Figure 1: Spectral plots of synthesized applause and filtered noise

To be able to recognize applause we propose to calculate the gravity point and the variance of the energy/ frequency distribution of every frame. If the difference of subsequent values is close to constant during some interval Δ_t, we observed that it is very likely that there is an occurrence of applause.

3.2 Detection of Applause in the Audio Data Stream

Our algorithm starts with the computation of a time frequency representation of the audio data using the Fourier Time Transformation (FTT) [Ter72, TGS82] using a similar algorithm as described in [HeKe92].The Fourier Time Transformation is described by the following formula:

$$P(t_i, f_i) = \int_0^T p(t) e^{-i\omega t} e^{-a(t-t_i)} dt$$

The frequency distribution has to be computed with a varying time scale to obtain an acceptable performance, reducing the distance between two analysis points when a match has been detected. Hence, the recognition of an applause segment present during the time when the pattern can be found can be done in a reasonable amount of time. To locate the applause pattern examine the statistical features of a time interval. The power spectrum of applause is illustrated in Figure 2. This dense kind of spectrum is generated by overlaying a sequence of 15 frames.

Having calculated the energy distribution of the signal over all frequencies at a given time t, the mass center and the distribution of the energy around the center have to be calculated. If these data meet a statistical distribution, we consider this as a matching pattern for applause. We then refine the time scale to verify the occurrence and to detect the begin and the end of the applause sequence to be stored in our MPEG-7 data model. This approach also allows for a manual verification of the recognition result. If applause has been located within the audio data the piece is marked as a “live”-recording.

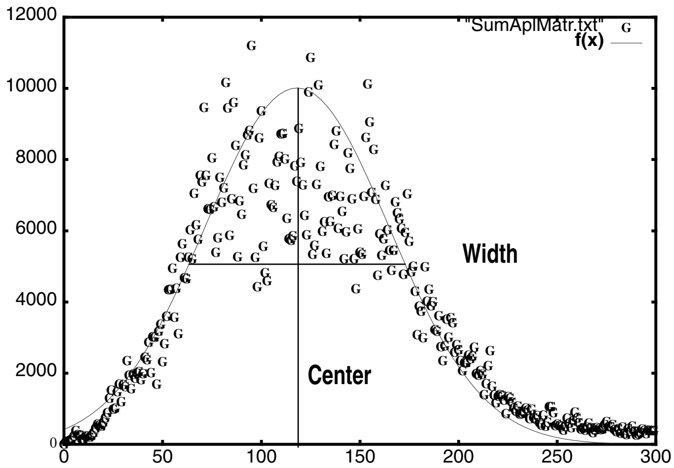


Figure 2: Mean energy over frequency with Gauss-fitted function

3.3 Recognition of Drum and Bass

The human perception of rhythm mostly uses the structure of the lower frequencies of sound, as well as steep increases of volume/loudness to synchronize the occurrences. Thus we consider these items as the key parts of the detection process [SPE96a, Pfei99]. Music pieces often use rhythm instruments similar to drums or bass-guitars.

In Figure 3 the temporal signal and the resulting energy distribution of a drum and a bass note are displayed. It is can be observed that the steep increase of loudness can be detected easily. Our approach focuses on this observation, using a low-frequency band for detection. An advantage of that approach is also the lower computational cost.

Assuming that the beats per minute and the rhythm structure do not vary to a large extent in the given audio data, and that the audio data only contains a single piece of music we use two different methods to recognize the beat of a music piece:

- a method to count the beats per minute and
- a method to determine the genre of that piece.

The overall quality of the information about the beats per minutes and about the rhythm structure depends on the genre of the examined piece. For pop and rock, and especially for dance music these data can be calculated extremely well. However, for jazz or classical music it is very difficult to detect an overall rhythm information. The strength of the resulting data is hence an important hint towards the genre of a music piece.

3.4 Recognition and Verification of Vocals

For the vocals recognition and verification we propose the application of the commonly used Linear Prediction Coding (LPC), refining it as described in [Ter74]. A problem is that normally the vocals do not appear as a single sound within the music track, and are always surrounded by other instrumental music. A refined filter technique has to be used to isolate the vocals from the rest of the audio track.

3.4.1 Digital - Voice Filter and LPC

A key issue, as described above, is the separation of vocals and what we would call “background noise” with respect to the vocals. The question arises how to separate these two information channels with respect to the fact that both interleave on some part of the frequency band. The frequency band that can be reproduced from CD-quality data spans from 20 Hz to

20.000 Hz. Normally speech can be found within 100 Hz and 6.000 Hz [Pfei99], however, musical instruments cover almost the whole range.

Speech and also vocals are produced in a special way, that can be physically modeled as a “lossless tube” filter. The model is also the basic idea of all LPC related speech coding systems. In the LPC model speech is produced in the following way (see Figure 4).

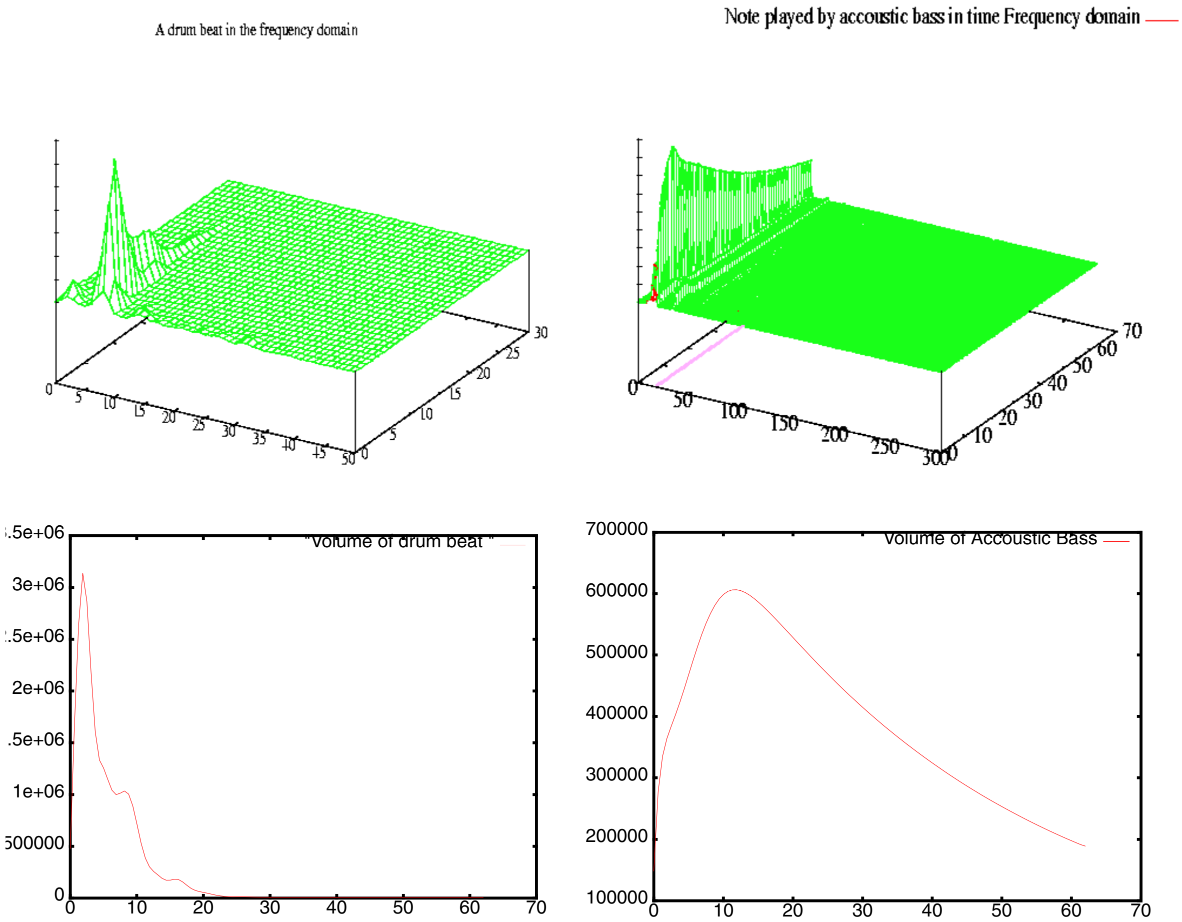


Figure 3: Frequency spectra and volume of drum and bass

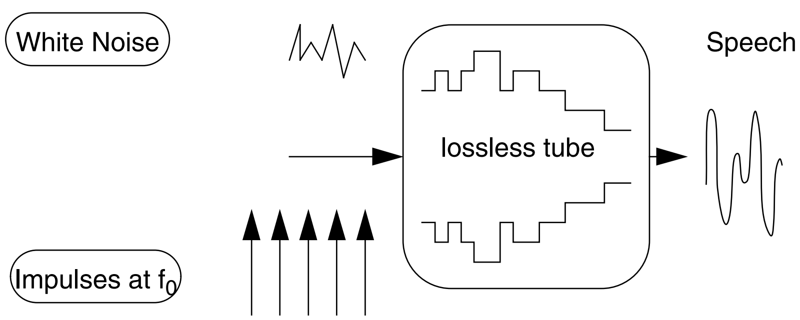


Figure 4: Physical model of LPC

The vocal tract as well as the throat and other parts of the human vocal system form some kind of a tube with wider and narrower sections. Considering this as lossless, hence no nasal cavity, this can be modeled mathematically as an IIR-filter (see Figure 5). The coefficients for the filter are calculated using the LPC.

$$y_n = \sum_{i=1}^{p} a_i y_{n-i}$$

The parameter estimation for Linear-Prediction-Filter coefficients can be found in [Ter74] and[Goo97].

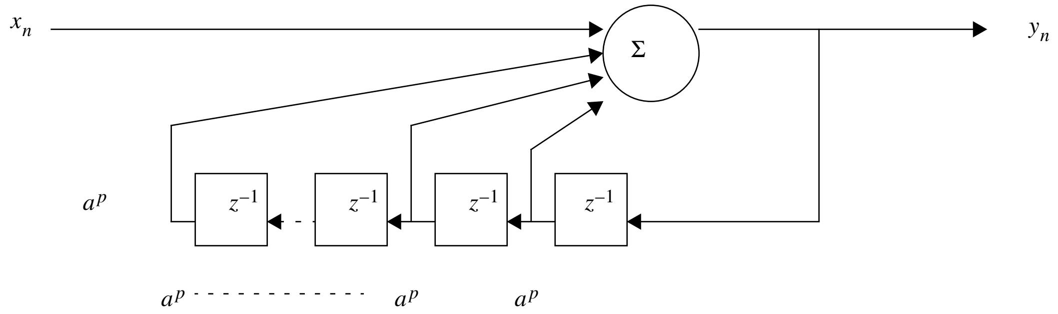


Figure 5: IIR filter

To recognize the vocals of a music piece we are only interested in the voiced parts of the audio data, as for all nasal (unvoiced) sounds the assumptions of the model are not correct. The parameters may be used for resynthesis, but the estimated parameters are not robust. The voiced/unvoiced decision within our framework is based on the Center Clipping Pitch detector (CLIP) method [Rab76]. The operations to be performed on each frame are (see also Figure 6):

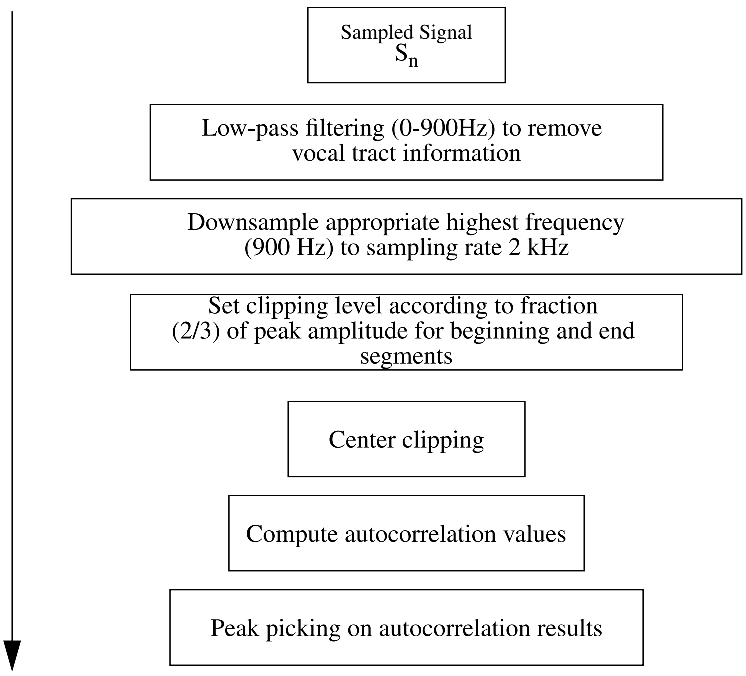


Figure 6: Algorithm to distinguish voiced from unvoiced segments

- pre-processing of a data frame to remove vocal tract information
- picking a clipping level and retain only those sample values that exceed the clipping level
- subtraction of clipping level for positive samples and adding of negative sample values
- computation of autocorrelation values of the center-clipped signal.

The voiced/unvoiced decision is then based on absence or presence of appropriate peaks. Also an estimation of the pitch period can be derived by this analysis. Figure 7 and Figure 8 show the basic steps of the algorithm.

Based on the assumption that every human being has some features of the vocal tract that do not change while speaking or singing and which are also independent from the language we use linear orthogonal parameters to recognize the vocals. These parameters are a subset of the LPC parameters of all voiced frames. The computation of linear orthogonal parameters has been described in [Goo97].

As mentioned above we try to find a set of parameters which are statistically uncorrelated. Linear prediction parameters are in general significantly redundant. To remove the redundancy we use a conventional eigenvector analysis [Cra51]. Taking into account that we only use those frames which we consider as voiced, the following calculation yields the orthogonal parameters where $\Omega(k, l)$ denotes the k-th orthogonal parameter in the l-th frame.

First the covariance matrix R of the linear prediction parameters of the voiced frames have to be calculated. If $x(k,l)$ denotes the k-th linear prediction parameter of the l-th voiced frame, the elements of R are given by the following equations:

$$r_{kl} = \frac{1}{F-1}\sum_{n=1}^{F}(x_{\ln} - \bar{x}_l)(x_{kn} - \bar{x}_k)$$

$$\bar{x}_i = \frac{1}{F}\sum_{n=1}^{F} x_{in}$$

Now these equations have to be solved to get the eigenvalue and the mutually orthogonal eigenvectors.

$$|R - \lambda \underline{1}| = 0$$

$$\lambda_l e_l = R e_l ; l = 1, 2, 3, \ldots, p$$

The output of the computation is a set of linearly orthogonal parameters $\Omega(k, l)$.

$$\Omega_{kl} = \sum_{n=1}^{p} e_{kn} x_{nl}$$

The output of the eigenvector analysis is a set of parameters which are linearly uncorrelated. These parameters reflect the unchangeable characteristic of the vocal tract of the vocals to be recognized. The mean values of these parameters are sufficient to recognize the vocals of a music piece.

4 Experimental Results

All experiments have to be done on high quality digital audio data (CD-Quality data) because of the higher time/frequency resolution. This is especially important to get the best representation of unchangeable vocal tract information for the singer recognition. The statistical data we obtained from our experiments were calculated based on a data set consisting of 300 pieces of 10 different artists.

4.1 Applause Detection

The performance of correct detection of applause within the audio data varies with the amount of noise present, as noise has a similar energy distribution compared to applause. The presence of a similar noise in the data is however not very likely.

Taking into account that most rhythm instruments which are responsible for what we call the “beat” of a piece of music have the most significant amount of their informational energy distributed in the frequency band from 20-150 Hz, a good recording equipment is needed to get these frequencies clean and strongly sampled.

Another vital issue is the correct mixture of the left and right stereo channels. For now we only use single channel analysis. Mixing the two channels is vital for the informational quality of the data. For future research it may be interesting to examine the two channels separately and later to correlate the gathered information to get more reliable classification results.

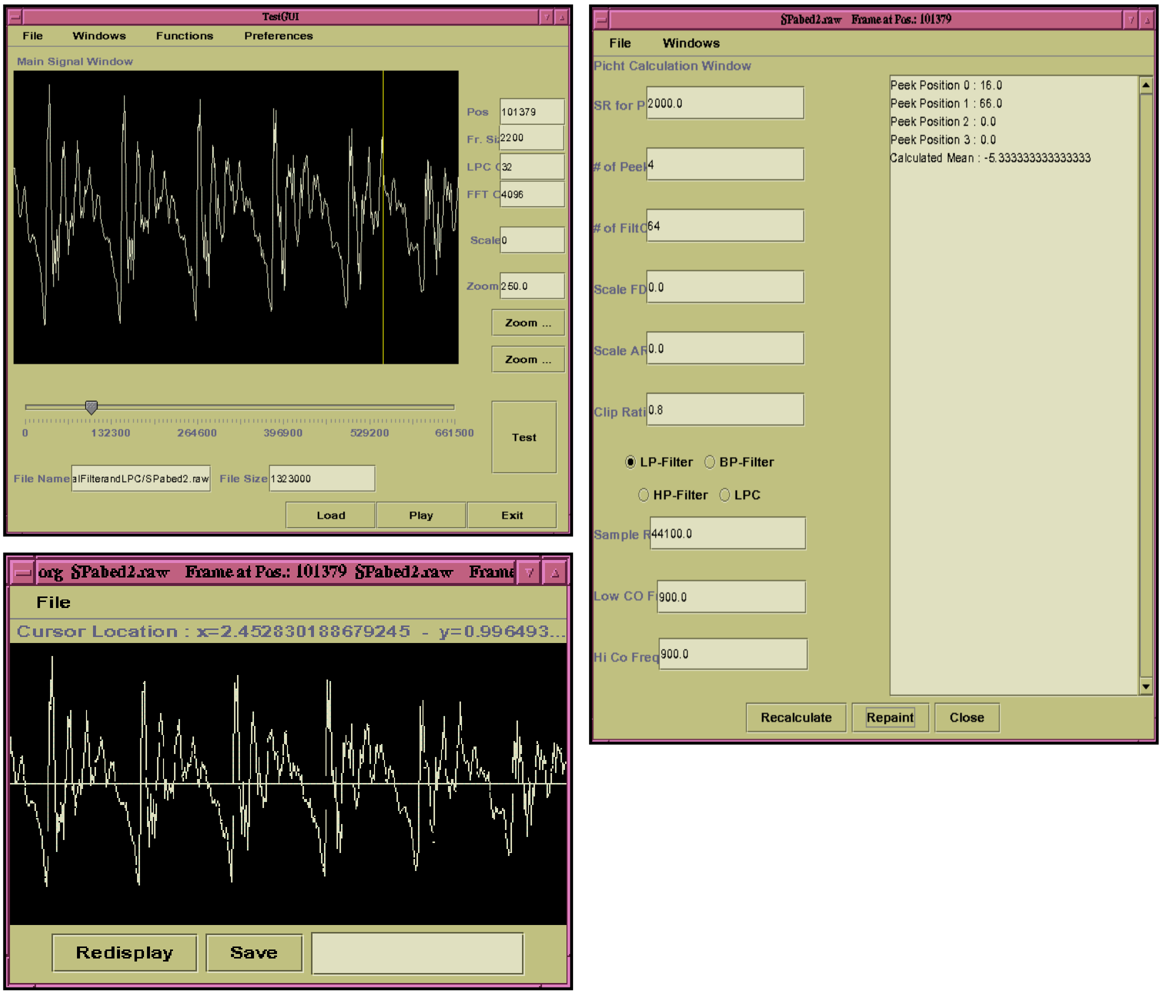


Figure 7: User interface for singer recognition analysis. Main window / analysis frame window / result window

4.2 Beat and Drum Recognition

Besides 300 music clips (see above) we also used a set of 150 sound files recorded from metronome and from the rhythm section of a synthesizer as reference rhythm structures. Using these reference data sets only consisting of beat data (metronome data) and of different rhythm data (walz, samba, foxtrott from synthesizer) together with the 300 music clips recorded from CDs we could classify app. 80 percent of the examined music pieces correctly. The remaining 20 percent where the algorithm failed consisted mostly of jazz and of classical pieces.

These results show that our genre recognition based on the idea of a rhythm detection works reliable for pop, rock or dance music. We classify all other genres as “unknown”; further work is needed to determine the genre of jazz or classical music.

4.3 Vocals Recognition

The retrieval and recognition performance of vocals which have been described by our parameters and stored in a database depends significantly on the metric used to compare the data. The linear prediction orthogonal parameters are by definition statistically uncorrelated. A dissimilarity feature is hence given by the following equation:

$$d = \sum_{i = p'}^{p} \left\langle \frac{\bar{\omega}_{i} - Z_i}{\sqrt{\lambda_i}} \right\rangle^2 \bar{J}$$

where

- $\bar{\omega}_i$: average value of orthogonal parameters of recognized vocals
- λ_i: reference eigenvalue of i-th parameter
- $\bar{J}$: average number of frames in set for recognized vocals
- Z_i: mean value of I-th parameter of unknown singer

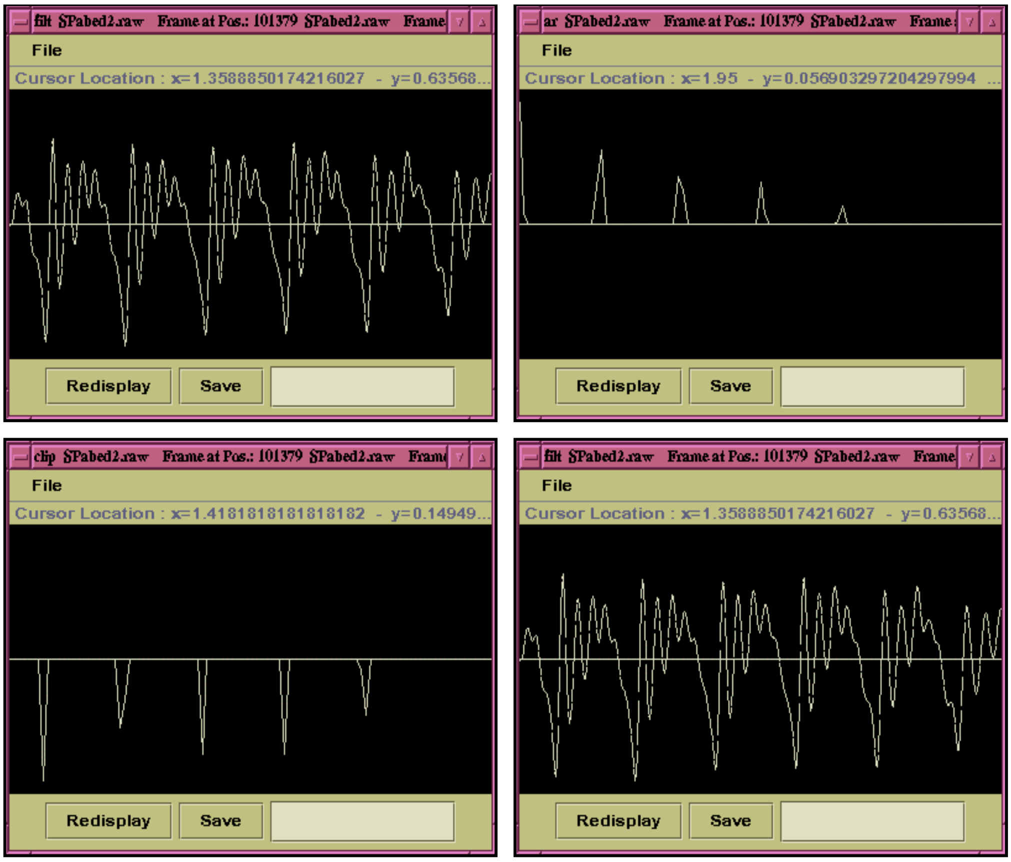
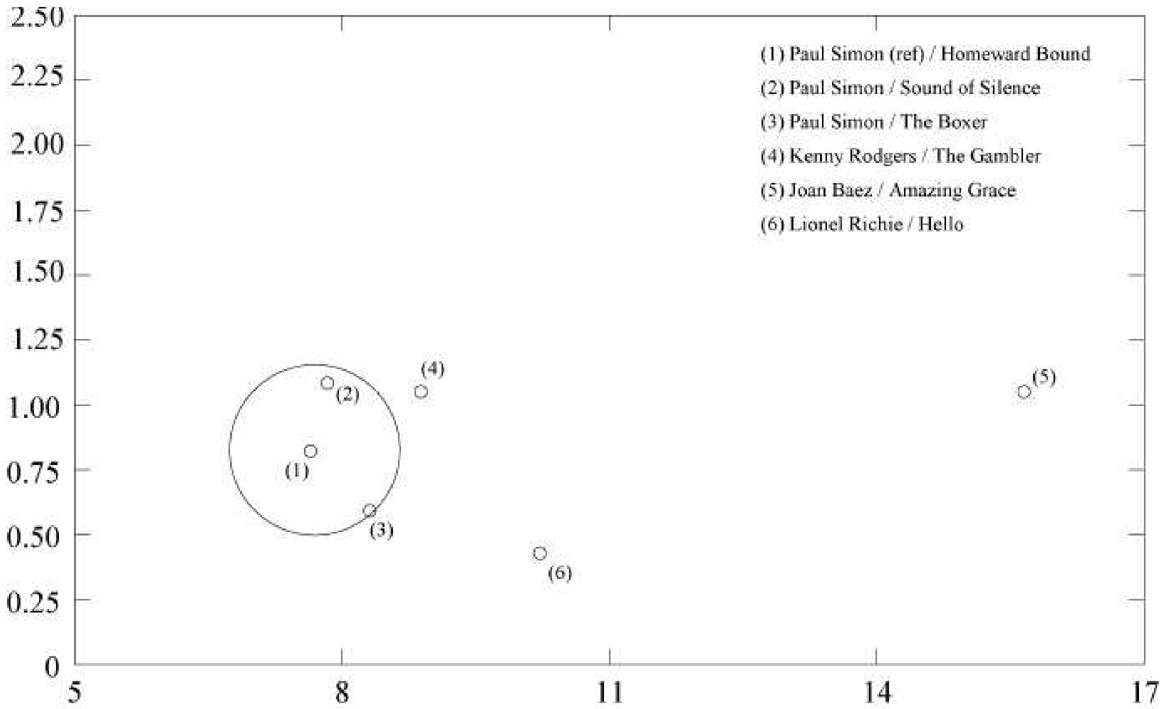


Figure 8: Interface for speaker recognition / Output of every stage of center clipping

The dissimilarity feature has a good performance (see Figure 9). Figure 9 shows that data points for different pieces of Paul Simon from different records are located in a close neighborhood. Data points for Lionel Richie, Kenny Rogers and Joan Baez are located in other sections. The overall performance of the linear prediction orthogonal parameters was around 90 percent for those pieces where only one artist was singing. However, for duets and chorus pieces the performance is much lower. In our future work we will have to think about mechanisms to deal with duets.



The reference Titles are chosen by the rate of voiced Frames detected in the song. Homeward Bound for Paul Simon is chosen because of this and as can be seen gives us a good performance in recognizing. Songs two and three lie within the bounds, both Paul Simon as singer, but Song Number 4 lies outside. So it can be seen that the recognition performance ist strongly dependent on a good reference.

Figure 9: Similarity of music clips

coefficent #	(1) Paul Simon/ Homeward Bound	(2) Paul Simon/ Sound of Silence	(3) Paul Simon/ The Boxer	(4) Kenny Rodgers/ The Gambler	(5) Joan Baez/ Amazing Grace	(6) Lionel Richie/ Hello
1	6.98	7.97	8.32	8.60	16.49	10.23
2	0.78	1.07	0.56	1.08	1.10	0.49
3	0.124	0.098	0.105	0.137	0.071	0.113

Table 1: First three orthogonal linear prediction coefficients for different songs / artists

5 Conclusions and Outlook

In this paper we described the ICOR system which follows an MPEG-7 approach to index and retrieve music clips (audio and video). Unlike other systems ICOR is based on a scene model from which the necessary algorithms are derived. In this paper we described the detection of applause, the recognition of beats per minute and the recognition of the vocals (artist) of a music piece. However, these algorithms have to be understood as a first step. We currently try to improve the performance of the existing algorithms, especially the recognition of duets. We also develop new algorithms to improve the overall strength of the user interface offering the functionality to play with the content of our system. New algorithms include for example a feature to measure the similarity of music clips.

ACKNOWLEDGEMENTS

The authors would like to thank the Volkswagenstiftung Germany who partially funds the research project. We also owe thank to our colleague Frank Nack for his helpful comments.

REFERENCES

[BR96] Boreczky, J. and Rowe, L., *Comparison of Video Shot Boundary Detection Techniques*, Proc.SPIE Conference on Storage and Retrieval for Still Image and Video Databases IV, pp. 170-179, San Jose, CA, 1996.

[CCMS+97] S.F. Chang, W. Chen, J. Meng, H. Sundaram, and D. Zhong. *VideoQ: An Automated Content Based Video Search System Using Visual Cues*. ACM MM 1997, Seattle, 1997.

[Cra51] H. Cramer, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1951.

[FLE95] S. Fischer, R. Lienhart and W. Effelsberg. *Automatic Genre Recognition*. Proc. ACM MM 1995, San Francisco, 1995.

[FSNA+95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and B. Dom et. al. *Query by Image and Video Content: The QBIC System*. IEEE Computer, 28(9), 1995.

[Fas94] H. Fastl, *Psychoacoustics and noise evaluation*. In: NAM'94, (H.S. Olesen, ed.) Aarhus, Denmark, Danish Technol. Institute, 1-12, 1994.

[Goo97] M. M. Goodwin, *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*. Ph.D thesis, Univ. of Berkeley, 1997

[GSCZS95] Y. Gong, L. T. Sin, C. H. Chuan, H. J. Zhang, and M. Sakauchi. *Automatic Parsing of TV Soccer Programs*. ICMCS95, 1995.

[HeKe92]K.Heldmann, Dr. W.Keiper, *Merkmalsextraktion an technischen Geräuschen mittels Teiltonzeitmuster*, R. Bosch GmbH Stuttgart, Zentrale Forschung und TU München, Elektroakkustik, DaGa'92 (German only).

[Lie99] R. Lienhart. *Comparison of Automatic Shot Boundary Detection Algorithms*. Proc.SPIE Conference on Storage and Retrieval for Still Image and Video Databases IV, pp. 290-301, San Jose, CA, 1999.

[LSCB99] C.-S. Li, R. Smith, V. Castelli, L.D. Bergman, *Comparing texture feature sets for retrieving core images in petroleum applications*. Proc.SPIE Conference on Storage and Retrieval for Still Image and Video Databases IV, pp. 2-11, San Jose, CA, 1999.

[Pfei99] S. Pfeiffer, *Information Retrieval from digitized audio tracks*. PhD thesis, Univ. of Mannheim, 1999 (in German).

[QM86]T. F. Quatieri and R. J. McAulay. *Speech transformations based on a sinusoidal representation*. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-34:1449-1464, December 1986.

[Rab76] Rabiner et al., *A Comparative Performance Study of Several Pitch Detection Algorithms,*IEEE Trans. ASSP, Vol. ASSP-24, No. 5, October 1976.

[Rob98]Tony Robinson, *Speech Analysis*, Lent Term 1998, http://svr-www.eng.cam.ac.uk/~ajr/SpeechAnalysis

[SC95] M.A. Smith and M. Christel. *Automating the Creation of a Digital Video Library.* Proc. ACM MM 1995, pp. 357-358, San Francisco, 1995.

[SCA98]G. Sheikholeslami, W. Chang, and A. Zhang. *Semantic Clustering and Querying on Heterogeneous Features for Visual Data*. Proc. ACM MM 1998, pp. 3-12, Bristol, 1998.

[SPE96a] S. Fischer S. Pfeiffer and W. Effelsberg. *Automatic audio content analysis*. In Proceedings of Fourth ACM International Conference on Multimedia, page 21–30, November 1996.

[Ter72] E. Terhardt, *Zur Tonhöhenwahrnehmung von Klängen*, Acustica 26: 173-199, 1972 (in German only).

[Ter74] E. Terhardt, *Pitch, consonance, and harmony.* J. Acoust. Soc. Am. 55: 1061 - 1069, 1974.

[TGS82] E. Terhardt, G. Stoll, and M. Seewann, *Algorithm for Extraction of pitch and pitch salience from complex tone signals*. In Journal of the Acoustical Society of America, Volume 71, No 3, March 1982a, pp 679 - 688.

[ZGST94] H. Zhang, Y. Gong, S.W. Smoliar, and S.Y. Tan. *Automatic Parsing of News Video*. Proceedings of IEEE Conf. on Multimedia Computing and Systems, 1994.

[ZKS93] H. Zhang, A. Kankanhalli, and S.W. Smoliar. *Automatic partitioning of full-motion video*. Multimedia Systems, 1(1), pp. 10-28, 1993.