Analysis of the Data Quality of Audio Descriptions of Environmental Sounds

Dalibor Mitrovic, Matthias Zeppelzauer, and Horst Eidenberger

Vienna University of Technology
Institute of Software Technology and Interactive Systems
Favoritenstrasse 9-11, A-1040 Vienna, Austria
{mitrovic, zeppelzauer, eidenberger}@ims.tuwien.ac.at

Abstract

In this paper we perform statistical data analysis of a broad set of state-of-the-art audio features and low-level MPEG-7 audio descriptors. The investigation comprises data analysis to reveal redundancies between state-of-the-art audio features and MPEG-7 audio descriptors. We introduce a novel measure to evaluate the information content of a descriptor in terms of variance. Statistical data analysis reveals the amount of variance contained in a feature. It enables identification of independent and redundant features. This approach assists in efficient selection of orthogonal features for content-based retrieval. We believe that a good feature should provide descriptions with high variance for the underlying data. Combinations of features should consist of decorrelated features in order to increase expressiveness of the descriptions. Although MPEG-7 is a popular and widely used standard for multimedia description, only few investigations do exist that address analysis of the data quality of low-level MPEG-7 descriptions.

1 Introduction

In the last decades a huge number of features was developed for the analysis of audio content. One of the first application domains of audio analysis was speech recognition [14]. With upcoming novel application areas the analysis of music and general purpose environmental sounds gained importance. Different research fields evolved, such as audio segmentation, music information retrieval (MIR), and environmental sound recognition (ESR). Each of these areas developed its specific description techniques (features). Currently, features are often employed in other domains than their original ones. A recent effort to standardize multimedia description tools led to the MPEG-7 standard. MPEG-7 is an ISO/IEC standard for multimedia content description [13]. The standard defines low-level descriptions techniques (including audio) as well as high-level tools for multimedia processing.

The huge number of existing features makes the selection of the most appropriate feature set for a task difficult. Statistical data analysis can help in the identification of independent features. In this paper, we perform a quantitative analysis of low-level MPEG-7 audio descriptors in the domain of environmental sounds. We compare MPEG-7 descriptors to a set of state-of-the-art audio features we previously analyzed in the domain of environmental sounds [19]. We investigate different description techniques by statistical data analysis in order to identify similarities and redundancies. Redundant features describe similar properties of the underlying data, while statistically independent features contain orthogonal information. The objective of feature selection is the combination of orthogonal features in order to maximize the amount of represented information. The method proposed in this paper supports the identification of independent and redundant features. Furthermore, we evaluate selected MPEG-7 high-level tools from this point of view. Additionally, we investigate the amount of information (entropy) contained in each feature. The information content of a feature is proportional to the variance of the feature values for a given dataset. We derive a measure that represents the information contained in a feature with respect to its variance in order to evaluate the expressiveness of a feature.

The remainder of the paper is organized as follows. In Section 2 we give background information about the low-level MPEG-7 descriptors. We present the structure of the experiments in Section 3. The results of the experiments are discussed in Section 4. A survey of related work is given in Section 5.

2 Background

The *MPEG-7 Audio* part specifies data structures and techniques for the description of audio content. It contains low-level descriptors (LLDs) as well as more sophisticated description techniques. In this section, we discuss the MPEG-7 LLDs relevant to this study.

2.1 Low-level MPEG-7 Audio Descriptors

The MPEG-7 Audio LLDs are a collection of low-level audio features that describe characteristic properties of sound such as harmonicity, sharpness, pitch, and timbre [16]. The descriptors are applied to short frames of the signal and are either scalars or vectors per frame. Aggregation of descriptions of several frames to a description of an entire media object is not a normative part of the MPEG-7 standard. Some LLDs (*TimbralSpectral* descriptors) support a single-valued summarization for entire media objects. For other LLDs the standard proposes mathematical operations, such as minimum, maximum, mean, and variance for summarization. The MPEG-7 Audio LLDs are organized in the six groups listed in Table 1.

Table 1: The low-level MPEG-7 audio descriptors

Group	Low-level descriptor	Abbreviation
Basic	AudioWaveform	AW
	AudioPower	AP
BasicSpectral	AudioSpectrumEnvelope	ASE
	AudioSpectrumCentroid	ASC
	AudioSpectrumSpread	ASS
	AudioSpectrumFlatness	ASF
SpectralBasis	AudioSpectrumBasis	ASB
	AudioSpectrumProjection	ASP
SignalParameters	AudioHarmonicity	AH
	AudioFundamentalFrequency	AFF
TimbralTemporal	LogAttackTime	LAT
	TemporalCentroid	TC
TimbralSpectral	SpectralCentroid	SC
	HarmonicSpectralCentroid	HSC
	HarmonicSpectralDeviation	HSD
	HarmonicSpectralSpread	HSS
	HarmonicSpectralVariation	HSV

In the following, we describe the MPEG-7 Audio LLDs together with their perceptual meaning and their application domain. A more detailed description can be found in [16].

2.1.1 Basic

The LLDs in the *Basic* group primarily enable a short description of the shape of an audio waveform. The *AudioWaveform* (AW) descriptor represents the waveform envelope and is mainly intended for economic display of a waveform in an audio editor. AW comprises the minimum and maximum values of a framed signal. The *AudioPower* descriptor computes the average square of the waveform samples in a frame. It describes the power of the signal over time.

2.1.2 Basic Spectral

The LLDs in the *BasicSpectral* group describe basic properties of the spectrum of an audio signal. The *AudioSpectrumEnvelope* (ASE) descriptor represents the short-term power spectrum of a signal with a logarithmic frequency scale in several frequency bands. The logarithmic frequency scale aims at imitating properties of the human ear. The ASE descriptor is the basis for the computation of the other descriptors in the *BasicSpectral* group.

The *AudioSpectrumCentroid* (ASC) is the center of gravity of the spectrum calculated by ASE. The ASC descriptor indicates whether high or low frequencies dominate the spectrum of the signal. The *AudioSpectrumSpread* (ASS) descriptor represents the deviation of the power spectrum from its centroid. ASS enables separation of tonal sounds from noise-like sounds.

The fourth descriptor of the *BasicSpectral* group is *AudioSpectrumFlatness* (ASF). ASF describes the deviation of the spectrum of an audio signal from a flat shape. A flat spectrum indicates a noise-like or impulse-like signal. According to the MPEG-7 standard ASF is designed to perform *fingerprinting*, which requires robust matching between pairs of audio signals.

2.1.3 Spectral Basis

The *SpectralBasis* descriptors *AudioSpectrumBasis* (ASB) and *AudioSpectrumProjection* (ASP) are techniques for general-purpose sound recognition. ASB transforms the spectrum of a signal to a much lower-dimensional representation under certain statistical constraints. The ASB descriptor is based on the power spectrum, similarly to the ASE descriptor and provides a compact representation of a spectrum, while preserving a maximum amount of information.

The ASP is used together with the ASB descriptor. ASP takes a decibel-scaled spectrum as input and projects it against spectral basis functions, previously computed by ASB.

2.1.4 Signal Parameters

The *SignalParameters* group contains the *AudioFundamentalFrequency* (AFF) descriptor and the *AudioHarmonicity* (AH) descriptor. The AFF descriptor represents the fundamental frequency of a sound. AFF may be applicable to sound segmentation of speech and music. AH is a measure for the degree of harmonicity in a signal. The descriptor comprises of two components: *harmonic ratio* and *upper limit of harmonicity*. The harmonic ratio is the proportion of harmonic components in a signal. A purely harmonic signal has a harmonic ratio of “1”, while the harmonic ratio of noise is “0”. The upper limit of harmonicity specifies the frequency beyond which the audio signal has no more significant harmonic components.

2.1.5 Timbral Temporal

Timbral descriptors are usually employed in MIR. Timbre is a sound property that is independent of pitch and loudness. The *LogAttackTime* (LAT) characterizes the attack of a sound. The attack time is the time from the beginning of a sound signal to a point in time where its amplitude reaches a maximum. LAT is the logarithm of the attack time. The attack characterizes the beginning of a sound, which can be smooth or sudden. LAT may be employed for classification of musical instruments. The *TemporalCentroid* (TC) is the point in time where most of the signal energy is located.

2.2 Timbral Spectral

Harmonic peaks in a spectrum correspond to frequencies that are a multiple of the fundamental frequency. They are appropriate to describe the timbre of a signal. The *TimbralSpectral* descriptors rely on harmonic peak estimation by the fundamental frequency of the audio signal. The *HarmonicSpectralCentroid* (HSC) is the amplitude-weighted average of the harmonic peaks in a spectrum. The *HarmonicSpectralSpread* (HSS) descriptor is the amplitude-weighted deviation of the harmonic peaks from the HSC.

The *HarmonicSpectralDeviation* (HSD) is the deviation of the harmonic peaks from the spectral envelope. The spectral envelope is the mean over a few neighboring harmonic peaks. *HarmonicSpectralVariation* (HSV) refers to the correlation of harmonic peaks in adjacent frames. The fifth *TimbralSpectral* descriptor is *SpectralCentroid* (SC), which is the power-weighted average of the frequencies in the power spectrum.

The timbre descriptors are usually applied to music information retrieval in which timbre plays an important role. We investigate the applicability of timbre descriptors in the domain of environmental sounds. The descriptors are expected to yield average results in the experiments.

3 Experiments

3.1 Test Setup

The investigations in this paper employ a database containing of 940 sound samples from 9 classes of environmental sounds. The sounds can be categorized into noise-like and tonal sounds. This distinction is of interest since several MPEG-7 descriptors model these properties (see Section 2). Table 2 summarizes the 9 different classes, the respective number of sound samples and predominant sound characteristics.

Table 2: The class names, the respective number of samples and the sound characteristics

Class name	# of samples	Sound characteristics	Class name	# of samples	Sound characteristics
bird	99	tonal	dog	84	noise-like
cat	110	tonal	footsteps	118	noise-like
car	105	noise-like	thunder	102	noise-like
cow	90	noise-like	signal	105	tonal
crowd	127	noise-like			

The audio data in the experiments are sampled at 11025 Hz and quantized to 8 bits. MPEG-7 descriptors are computed with the LLD extractor provided by the TU Berlin [17]. The investigations include 12 previously analyzed audio descriptors from different application domains listed in Table 3 [19].

Table 3: Non-MPEG-7 descriptors and their abbreviations.

Descriptor	Abbreviation	Descriptor	Abbreviation
Mel-Frequency Cepstral	MFCC	Constant Q Transform	CQT
Bark-Frequency Cepstral	BFCC	Spectral Flux	SF
Linear Predictive Coding	LPC	Zero Crossing Rate	ZCR
Perceptual Linear Prediction	PLP	Loudness	Sone
Relative Spectral - Perceptual	RASTA-PLP	Amplitude Descriptor	AD
Discrete Wavelet Transform	DWT	Pitch	Pitch

Combination of MPEG-7 and non-MPEG-7 descriptors yields a 391-dimensional feature vector. Frame-based features are summarized by statistical moments (mean and variance) in order to obtain descriptions of entire media objects.

The analysis steps are as follows: Firstly features are extracted from the raw sound samples. In the second step the feature vectors of all sample files are combined into a matrix of 391 columns and 940 rows. The third step is a dimension reduction via Principal Components Analysis (PCA). The resulting *factor loading matrix* is the basis for statistical data analysis.

3.2 A novel measure for expressiveness

The quality of a feature may be measured by the amount of variance of its numeric values. A good feature should provide descriptions with high variance for the underlying data. Statistical data analysis reveals the amount of variance of a feature. The PCA decorrelates the second statistical moments (variances) of the feature data. We select only the principal components (PCs) with an Eigenvalue greater or equal than “1” for data analysis. PCs with lower Eigenvalues are not considered since they explain less variance than the original data. From the PCA we yield the factor loading matrix that describes the influence of PCs on the particular feature components and vice versa. The factor loading matrix has the PCs in its columns (ordered by descending Eigenvalues) and the features form the rows of the matrix. Each entry (factor loading) represents the influence of a feature on a PC. The factor loadings are in the interval [-1, 1]. A high absolute factor loading indicates a high degree of correlation between a feature and a PC. Features that load the same PCs are correlated. A Varimax rotation simplifies the interpretation of factor loadings by maximizing their variances.

In order to quantify the information contained in a feature we sum the absolute factor loadings weighted by the corresponding percentage of explained overall variance. The result is normalized by the sum of variances of all PCs. We call this measure *Weighted Average LoaDing Indicator* (WALDI). It is computed as defined in Equation 1.

$$WALDI(f_j) = \frac{1}{\sum_{i=1}^{C} \sigma^2(c_i)} \sum_{i=1}^{C} \left| L(c_i, f_j) \right| \sigma^2(c_i) \,, \tag{1}$$

Where f_j is the *j*-th feature from a set of F features and the c_i are the C principal components. The variance of the *i*-th PC is denoted by $\sigma^2(c_i)$. The factor loading matrix L is a R^{CxF} matrix with C columns and F rows.

The WALDI of a feature is a measure of its information content in terms of orthogonal variances in the data. This value is proportional to the expressiveness of a feature. However, it does not contain information about redundancies among different features. This information can be derived from the factor loading matrix.

We build a graph based on the WALDI (see Figure 1). Peaks in the WALDI graph indicate high expressiveness, which may have two reasons: either the corresponding feature loads a large number of less important PCs or it highly loads the few most important PCs.

Additionally, we compare WALDI with the information entropy, introduced by Shannon and Weaver [18]. We determine the entropy as defined in Equation 2.

$$H(f) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \tag{2}$$

We consider features as random variables and quantize the observed values of the features into n=256 bins. Then we compute the probability $p(i)$ of the *i*-th bin for a feature *f*. For n=256 the unit of entropy is *bit per byte*. The entropy is a measure for the information content of a variable. The uncertainty of a random variable is proportional to its entropy. A powerful feature should have high entropy. We discuss the entropy of MPEG-7 features in Section 4.1.

4 Results

Quantitative data analysis discloses the data quality of numerical features. The basis of the investigation is the factor loading matrix that shows the mapping of features to PCs. In the first step of the analysis we investigate the expressiveness of features.

4.1 Information content analysis

We evaluate the audio descriptors based on their expressiveness and compute therefore the WALDI for the descriptor components (see Section 3.2 for details). We average over all WALDIs of the descriptor components to gain a more compact representation. That is equivalent to the average amount of information contained in each descriptor component. Figure 1 depicts the resulting graph.

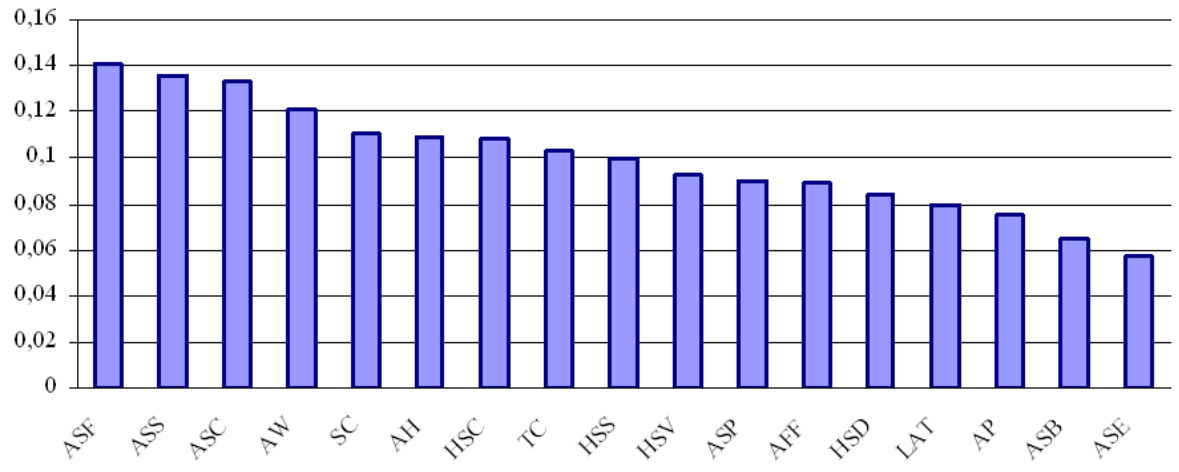

Figure 1: The WALDI graph of the MPEG-7 descriptors. High values indicate high expressiveness.

The *BasicSpectral* descriptors ASS, ASC, and ASF yield high WALDIs, because these descriptors have high loadings for the first few and most important PCs. In contrast to this, the ASB descriptor has a small WALDI value. The reason for this is that the components of ASB do not load any of the important PCs. The factor loading matrix reveals a similar situation for ASE. This can as well be observed in the WALDI graph (Figure 1). The timbral descriptors yield average values for the environmental sounds in the experiments. This is an unexpectedly good result since they mainly represent characteristics of musical sounds.

We believe that high entropy is a necessary property of a good feature. In the following, we compare the results obtained from the WALDI graph with the entropy of the descriptors. Figure 2 illustrates the average entropy of all components of a descriptor.

AH is the descriptor with the highest entropy (7.2 bit per byte), followed by the *BasicSpectral* descriptors, TC, and SC. Generally, the entropy of the LLDs is high with an average of 6 bit per byte. We observe that the MPEG-7 audio LLDs have higher entropy than the visual descriptors of the standard [15].

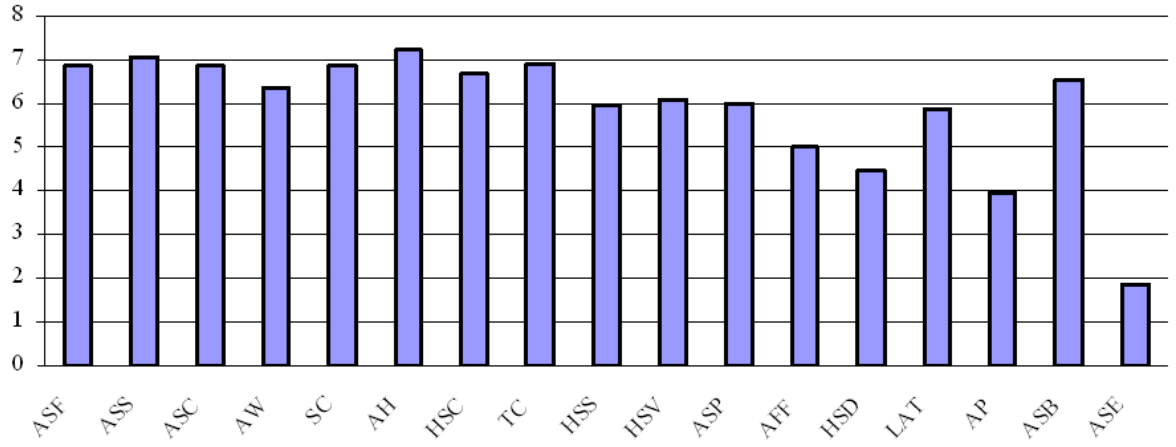

Figure 2: The entropy graph of the MPEG-7 descriptors.

A comparison of WALDI and entropy shows that both measures correlate to a certain degree. Features with high WALDI tend to have high entropy. This is evident for the *BasicSpectral* descriptors (ASF, ASS, and ASC). Analogously, AP and ASE have low values for both measures. The correlation between WALDI and entropy shows that WALDI is a valid measure for information content and expressiveness.

4.2 Redundancy analysis

Identification of redundancies between features is an important pre-processing step of feature selection. The factor loading matrix is the basis for redundancy analysis. High absolute loadings indicate a high degree of correlation with the corresponding PC. Features that load the same PCs are dependent and contain redundant information.

The total redundancy of MPEG-7 descriptors is low, compared to the non-MPEG-7 features. In order to explain 85% of the overall variance in the data the non-MPEG-7 descriptors require 40 out of 228 PCs (~17.5%). MPEG-7 descriptors require 56 out of 163 PCs (~34.4%) to achieve the same result. Consequently, the set of MPEG-7 descriptors is less redundant.

In the following, we discuss the quality of the MPEG-7 Audio LLDs in more detail. The MPEG-7 descriptors of the *Basic* group (AP and AW) are highly correlated. The reason is that the computations of these two descriptors are similar. The descriptors represent related properties of the audio waveform. As expected, the expressiveness of these descriptors is limited. Both descriptors do not describe independent information with respect to the other features in the investigation. Most information explained by AP and AW is also contained in the Sone feature that measures the loudness of a signal.

The components of ASE are independent to a high degree. They load several less important PCs. However, ASE correlates with some Sone components. ASC, ASS, and ASF are highly correlated descriptors. The dependency of ASC and ASS is difficult to interpret since the descriptors describe different statistical moments. This may be a side effect introduced by the underlying data. In contrast to this, the redundancy of ASS and ASF is easier to explain since both descriptors model the same sound characteristics (noise-likeness and tonality). ASC, ASS, and ASF are highly redundant with several non-MPEG-7 features such as PLP, MFCC, and Sone.

The components of the *SpectralBasis* descriptor ASB are decorrelated. This is due to the SVD that yields orthogonal basis functions. The ASP descriptor bases on ASB. Thus, it correlates with ASB to some degree. ASB and ASP should be applied in conjunction according to the MPEG-7 standard. The factor loading matrix reveals the independence of ASB and ASP from all other features in this survey. These properties qualify ASB and ASP for combination with other features. Due to the similar computation of ASB and ASP, the two descriptors have comparable loadings in the factor loading matrix.

AH describes the harmonic structure of a signal. In the experiments AH is highly correlated with ASC, ASS and ASF because these descriptors depend on harmonic properties as well. Pitch, ZCR, and AD encode similar information as AH. A group of dependent MPEG-7 descriptors are AFF, ASC, SC, and HSC. They are correlated since they are all sensitive to the fundamental frequency of a signal. Consequently, the ZCR, which approximates the fundamental frequency, highly correlates with these features.

The timbral temporal descriptors LAT and TC are highly correlated with SF for the data in the experiments, even though they are computed entirely different. The reason for the high correlation is that these features do not model the characteristic of the environmental sounds very well. This is evident in the analysis of the expressiveness of these features (see Section 4.1). According to the WALDI graph in Figure 1, TC and LAT have average and low expressiveness, respectively. This is due to the fact that the shape of environmental sounds is generally not structured in attack, decay, sustain and release as is the case for musical sounds. The expressiveness of SF with a WALDI of 0.7 is low as well because of the complexity and noise-like structure of the environmental sounds. However, LAT, TC, and SF describe information that is not captured by the other features in the study. That qualifies them for combination with other features.

The timbral spectral descriptors are completely redundant for the environmental sounds in the experiments. However, they describe unique information that is not captured by any other feature in the experiment. Since environmental sounds contain only little timbral characteristics the expressiveness of the timbral features is limited. Nevertheless timbral descriptors may be applicable to separate certain sound classes, such as bird sounds from other environmental sounds.

Table 4: Dependencies between MPEG-7 and non-MPEG-7 features

	Correlated MPEG-7 descriptors	Correlated non-MPEG-7 features	Decorrelated MPEG-7 descriptors	Decorrelated non-MPEG-7 features
AW	AP, ASE, ASC, ASS	Sone	ASB, HSD, HSS, HSV	DWT
AP	AW, ASE	Sone	ASB, ASP, LAT, TC, HSC, HSD, HSS, HSV	BFCC, RASTA-PLP, DWT
ASE	AW, AP	Sone	ASB, ASP,TC, HSD, HSS, HSV	DWT
ASC	AW, ASC, ASS, ASF, AH, AFF	AD, ZCR, LPC, Pitch, Sone, PLP	HSD, HSS	DWT
ASS	AW, ASC, ASS, ASF, AH	Pitch, Sone, PLP	HSD	DWT
ASF	ASC, ASS, AH	BFCC, MFCC, Pitch, RASTA-PLP, Sone, PLP	HSD, HSS	DWT
ASB	-	-	AW, AP, ASE, AH, TC, SC, HSC, HSD, HSS HSV	AD, BFCC, SF, LPC, MFCC, Sone, DWT, CQT PLP
ASP	-	-	AP, ASE, AFF, TC, SC, HSC, HSD, HSS, HSV	BFCC, SF, MFCC, Sone, DWT, CQT
AH	ASC, ASS, AFF	AD, ZCR, Pitch	ASB, HSD, HSS	DWT
AFF	ASC, SC, HSC	ZRC	ASP, HSD, HSS	DWT, CQT
LAT	TC	SF	AP, SC, HSD, HSS, HSV	DWT, CQT
TC	LAT,	SF	AP, ASE, ASB, ASP, HSD, HSS, HSV	DWT, CQT, MFCC, BFCC
SC	AFF, HSC	ZCR, LPC	ASB, ASP, LAT, HSD, HSS, HSV	DWT
HSC	AFF, SC	ZCR	AP, ASB, ASP	DWT, CQT
HSD	HSS, HSV	-	all except HSC	all
HSS	HSD, HSV	-	all except ASS and HSC	all
HSV	HSD, HSV	-	AW, AP, ASE, ASB, ASP, LAT, TC, SC	all

A detailed summarization of redundant and independent MPEG-7 features is given in Table 4. For each MPEG-7 descriptor the table lists correlated and decorrelated features. For the sake of clarity MPEG-7 descriptors and other features are listed in separate columns.

4.3 Summary

The following major insights can be derived from the analysis:

1. Most of the groups of MPEG-7 descriptors contain highly correlated components (*Basic*, *BasicSpectral*, *SpectralBasis*, *TimbralTemporal*, and *TimbralSpectral*). One reason may be the characteristics of the environmental sounds, which contain only little timbre and harmonicity. Another reason is the similarity of the computations of the descriptors in particular groups.
2. The descriptors in the *SignalParameters* group are decorrelated to a higher degree than the components of the other groups. This is because AH and AFF describe different signal properties.
3. The different descriptor groups are mostly independent from each other. The exceptions are the *Basic* group and the *BasicSpectral* group, which are highly correlated. The reason is that the two groups describe a signal waveform in time and frequency domain without further processing. Since the time and frequency representations of a signal are equivalent, the descriptions correlate. Two other correlated groups are *BasicSpectral* and *SignalParameters*. Both have correlated components such as AH and ASF, which describe the tonality of a sound. Further correlated components are ASC and AFF, which both heavily depend on the fundamental frequency.
4. Descriptors of the *SpectralBasis* and *TimbralSpectral* groups are independent from all other features including the non-MPEG-7 ones. Hence, they complement all possible feature combinations and are potential candidates for feature selection.
5. Several non-MPEG-7 features correlate with MPEG-7 descriptors. The Sone feature is redundant with the entire *Basic* and *BasicSpectral* groups. Pitch and PLP highly correlate with the *BasicSpectral* group. Further, several popular speech recognition features such as MFCC, BFCC, RASTA-PLP, Sone, Pitch, and PLP encode the information captured by ASF. The reason may be that all these features describe properties necessary for audio fingerprinting.

4.4 MPEG-7 high-level tools

The MPEG-7 high-level tools contain application-specific description schemes that build upon LLDs. These description schemes are *AudioSignature*, *Timbre*, and *SoundModel*. Other high-level descriptions do not consist of the LLDs discussed above and are therefore out of scope for the data analysis in this study.

The *AudioSignature* description scheme contains statistical summarizations of the ASF low-level descriptor. Its addressed application area is fingerprinting. Data analysis reveals that the components, which represent the ASF in different frequency bands, are highly redundant. Hence, a subset of ASF components may be sufficient for retrieval applications. Furthermore, it satisfies the requirements for fingerprinting since it contains a significant amount of information (see the WALDI graph in Figure 1).

Timbre is the second description scheme considered. It contains LAT, HSC, HSD, HSS, and HSV for harmonic instrument identification and SC, TC, and LAT for percussive instrument identification. In the experiments, we evaluate the quality of the timbral descriptors to prove whether timbre is a discriminative characteristic of environmental sounds. Experiments show that the descriptors for percussive instruments model environmental sounds better. The harmonic instrument descriptors are highly redundant for ES. However, they describe unique information with respect to the other descriptors in the experiments. Thus, only a subset of components of the *Timbre* description scheme may be feasible for ESR.

The *SoundModel* descriptor scheme is the third high-level tool in this investigation. It consists of ASB and ASP and addresses environmental sound recognition. While the components of ASB and ASP are highly decorrelated, their expressiveness is limited because they have only mediocre influence on the important PCs in the factor loading matrix. It may therefore be advantageous to combine the descriptors of *SoundModel* with other more expressive LLDs such as ASF.

5 Related Work

The MPEG-7 Audio standard provides a large set of low-level audio descriptors [13]. MPEG-7 audio descriptors are part of many state-of-the-art audio retrieval systems [6], [8], and [11]. MPEG-7 audio descriptors are applicable to different types of sound, such as music, speech and environmental sounds.

There are only few studies that address quantitative data analysis of content-based descriptions. We investigate low-level MPEG-7 visual descriptors from a statistical point of view in [15]. We employ statistical moments, factor analysis, and cluster analysis in the study. The investigation reveals that the visual descriptors are highly dependent on each other.

Most investigations apply features to a specific problem without a preceding data analysis. Such studies evaluate the quality of features empirically by performance measures such as recall and precision. A popular application domain is music information retrieval. In [8] the authors combine MPEG-7 descriptors with other common audio features for musical instrument classification. The investigation comprises of MPEG-7 *BasicSpectral* descriptors (ASE, ASC, ASS, and ASF) and *SpectralBasis* descriptors (ASB and ASP). Furthermore, SC and AH descriptors are incorporated. Additionally, the authors employ non-MPEG-7 features such as MFCCs, ZCR, and Spectral Rolloff Frequency [16]. After feature extraction different classifiers (HMM, GMM and non-negative matrix factorization) predict the class membership in terms of six different classes of instruments. MPEG-7 audio descriptors for instrument characterization are surveyed in [9] and [10] as well. The authors investigate the quality of high-level descriptors (*HarmonicInstrumentType* and *PercussiveInstrumentType*) for the distinction of instrument sounds. They show that the combination of the low-level descriptors contained in the high-level description schemes enable successful similarity matching in a musical sounds database. The authors of [2] present an MPEG-7 supported audio identification system that is robust with respect to common types of signal alterations. A flatness measure, similar to the MPEG-7 ASF descriptor is employed for similarity matching in a database of songs.

There are multiple investigations that deal with more general sounds such as environmental sounds. In [9] Casey presents the MPEG-7 sound recognition tools for general purpose sound recognition. The system employs the ASB and ASP descriptors. The author discusses the corresponding high-level description scheme (*SoundModel*) which includes continuous HMMs for classification. The authors of [1], [5], [6], and [7] extensively investigate the quality of the MPEG-7 sound recognition tools in different application domains. The authors perform speaker recognition and audio segmentation of speech and non-speech segments. Furthermore, they perform general sound classification of selected environmental sounds such as “Dog,” “Bell,” “Water,” and “Baby.” They compare the performance of MPEG-7 techniques with the traditional MFCC approach, originally developed for speech recognition. The MPEG-7 system employs ASB and ASP descriptors to represent the audio samples. Classification is performed by continuous HMMs. The investigations show that MPEG-7 descriptors perform comparably to MFCCs. However, MFCCs outperform ASB and ASP in some applications. A similar investigation is presented in [11]. The authors compare the widely used MFCC audio features to the low-level MPEG-7 descriptors designed for audio retrieval. Classification is performed with two types of HMMs (Maximum Likelihood HMM and Entropic Prior HMM). Again, MPEG-7 descriptors perform comparably to MFCCs. In [12] the authors present an MPEG-7-based retrieval system for environmental sounds. The system applies ASC, ASS and ASF to the description of audio samples. The samples are organized in classes, such as “doorbell,” “laughing,” “knock,” and “dog barks.” One HMM represents one particular class.

Investigations in [4] address highlight detection in sports videos. The authors analyze audio information from golf, soccer, and baseball games. They identify audio events that indicate highlights, such as cheering and applause. Again, MPEG-7 ASB and ASP descriptors serve as features. Classification is performed by HMMs.

The MPEG-7 standard defines a set of high-level description schemes in addition to the low-level description tools. An overview is given in [3]. The authors present MPEG-7 applications, such as Query by Humming, Audio Editing and application specific high-level description schemes (e.g. MusicalInstrumentTimbre, SpokenContent, and Melody).

6 Conclusions and future work

In this paper we perform an extensive statistical data analysis of low-level MPEG-7 descriptions in the domain of environmental sounds. We analyze correlations of MPEG-7 descriptors with other state-of-the-art audio features. Statistical data analysis allows us to identify features that are complementary to MPEG-7 LLDs. We extend data analysis by introducing a novel measure for information content.

The objective of the experiments is the analysis of the correlations in a large set of features. For this purpose, we employ Principal Components Analysis, which reveals low redundancy between most of the MPEG-7 descriptor groups. However, there is high redundancy within some groups of descriptors such as the *BasicSpectral* group and the *TimbralSpectral* group. Redundant features capture similar properties of the media objects and should not be used in conjunction. We discuss the quality of MPEG-7 high-level description tools based on the results of the analysis of the LLDs. The entropy of the LLDs is generally high (in average 6 bit per byte) which is significantly more than for the visual descriptors we surveyed in [15].

Future work in this area includes a comparison of results from data analysis with the results of actual retrieval experiments. This may answer the question to which degree results from data analysis support feature selection in real world retrieval tasks. Additionally, future research will focus on the combination of MPEG-7 descriptors with complementary features not in the standard in order to obtain richer descriptions.

Acknowledgements

The authors would like to express their gratitude to Professor Christian Breitenender for his guidance. We want to thank Professor Sikora from the Technical University of Berlin for providing us with the MPEG-7 LLD extractor. This work is supported by the Austrian Scientific Research Fund (FWF) under grant no. P16111-N05.

References

[1] Kim, H., Moreau, N., & Sikora, T. (2004). Audio classification based on MPEG-7 spectral basis representations. IEEE Transactions on Circuits and Systems for Video Technology. vol. 14. 716-725.

[2] Allamanche, E., Herre, J., Hellmuth, O., Fröba, B., Kastner, & T., Cremer, M. (2001). Content-based Identification of Audio Material Using MPEG-7 Low-level Description. Proceedings of the International Conference on Music Information Retrieval.

[3] Quackenbush, S., & Lindsay, A. (2001). Overview of MPEG-7 Audio. IEEE Transactions on Circuits Systems for Video Technology. vol. 11. 725-729.

[4] Xiong, Z., Radhakrishnan, R., Divakaran, A., & Huang, T. (2003). Audio-based highlights extraction from baseball, golf and soccer games in a unified framework. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). vol. 5. 628-631.

[5] Kim, H., Burred, J., & Sikora, T. (2005). How efficient is MPEG-7 for general sound recognition. Proceedings of AES 25th International Conference.

[6] Kim, H., & Sikora, T. (2004) Comparison of MPEG-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 5. 925-928

[7] Kim, H., Berdahl, E., & Sikora, T. (2003). Study of MPEG-7 Sound Classification and Retrieval. Submitted to the International ITG Conference on Source and Channel Coding.

[8] Benetos, E., Kotti, M., Kotropoulos, C., Burred, J., Eisenberg, G., Haller, M., & Sikora, T. (2005). Comparison of Subspace Analysis-Based and Statistical Model-Based Algorithms for Musical Instrument Classification. 2nd Workshop on Immersive Communication and Broadcast Systems (ICOB).

[9] Casey, M. (2001). MPEG-7 sound recognition tools. IEEE Transactions on Circuits and Systems for Video Technology. vol 11. 737-747.

[10] Peeters, G., McAdams, S., & Herrera, P. (2000). Instrument sound description in the context of MPEG-7. Proceedings of the 2000 International Computer Music Conference

[11] Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, S.T. (2003). Comparing MFCC and MPEG-7 audio features for feature extraction, Maximum Likelihood HMM and Entropic Prior HMM for sports audio classification. Proceedings of the International Conference on Multimedia and Expo. vol. 3. 397-400.

[12] Wang, J-F., Wagd, J-C, Huang, T-H., & Hsu, C-S. (2003**).** Home environmental sound recognition based on MPEG-7 features. Proceedings of the 46th IEEE International Midwest Symposium on Circuits and Systems. vol 2. 682-685.

[13] ISO/IEC 15938 (2002). Information Technology – Multimedia Content Description Interface. First Edition

[14] Rabiner, L. & Juang, B. (1993). Fundamentals of speech recognition. New York: Prentice-Hall.

[15] Eidenberger, H. (2004). Statistical analysis of content-based MPEG-7 descriptors for image retrieval. Multimedia Systems. vol. 10. 84-97.

[16] Kim, H-G., Moreau, N., & Sikora, T. (2005). MPEG-7 audio and beyond. West Sussex: Wiley.

[17] MPEG-7 Audio Analyzer Low Level Descriptors Extractor. http://mpeg7lld.nue.tu-berlin.de. (last visited 03/21/2006)

[18] Shannon, C., & Weaver, W. (1949). The Mathematical Theory of Communication. Urbana: University of Illinois Press.

[19] Mitrovic, D., Zeppelzauer, M., Eidenberger H. (2006). Towards an Optimal Feature Set for Environmental Sound Recognition. Technical Report TR-188-2-2006-03 (http://www.ims.tuwien.ac.at/publication_master.php)