On Feature Selection in Environmental Sound Recognition

Dalibor Mitrović, Matthias Zeppelzauer, Horst Eidenberger
Vienna University of Technology
Institute of Software Technology and Interactive Systems
Favoritenstrasse 9-11, A-1040 Vienna, Austria
E-mail: *mitrovic@ims.tuwien.ac.at*

Abstract – *Given a broad set of content-based audio features, we employ principal component analysis for the composition of an optimal feature set for environmental sounds. We select features based on quantitative data analysis (factor analysis) and conduct retrieval experiments to evaluate the quality of the feature combinations. Retrieval results show that statistical data analysis gives useful hints for feature selection. The experiments show the importance of feature selection in environmental sound recognition.*

Keywords – *Feature Selection, Statistical Data Analysis, Environtmental Sound Recognition*

1. INTRODUCTION

Environmental sounds comprise all types of sound that are neither speech nor music making the domain nearly infinite in size. The goal of environmental sound recognition is to assign given input sounds to previously defined categories. The variety of applications ranges from automatic surveillance to multimodal video retrieval, e.g. of sports videos [1, 2].

Most audio features were originally designed for domains other than environmental sounds. Since feature design includes assumptions about the targeted domain, it is not a priori clear whether a feature performs sufficiently for another domain. In this paper we survey a large number of state-of-the-art features and investigate their quality for a set of environmental sounds. Furthermore, we evaluate the applicability of the Amplitude Descriptor [3] (originally devised for animal sounds) in the domain of general purpose environmental sounds.

In order to evaluate the data quality we examine the redundancy of the features by a quantitative data analysis method, namely Principal Components Analysis (PCA). Statistical data analysis allows for evaluation of the data quality of features [4] and supports feature selection [5]. Since high data quality is *necessary but not sufficient* for successful retrieval, we evaluate the retrieval quality of the features and their combinations by supervised classification in order to identify an optimal feature set for a given database. Finally, we evaluate and discuss the statistical properties of the resulting feature set.

The remainder of the paper is organized as follows. In Section 2 we present the structure of the experiments. Results of the experiments are discussed in Section 3 and summarize the paper in Section 4.

2. EXPERIMENTAL SETUP

We perform a series of experiments in order to identify the optimal feature set for classification of environmental sounds. The experiments are split into three steps:

1. Analysis of global redundancy of all features by PCA.
2. Construction of feature sets and optimization of retrieval quality (based on step 1).
3. Analysis of the data quality of the empirically optimized solution.

2.1. Audio Features

A vast number of audio features has been designed for specific application domains, such as speech recognition, audio segmentation, and music information retrieval. We examine features from all mentioned application domains. Popular features from speech recognition are Linear Predictive Coding (LPC) coefficients, Perceptual Linear Prediction (PLP) and Relative Spectral Perceptual Linear Prediction (RASTA-PLP). PLP and RASTA-PLP are psychoacoustically enhanced derivatives of LPC that represent the spectral envelope of a signal (timbral information).

Mel-Frequency Cepstral Coefficients (MFCCs) and Bark-Frequency Cepstral Coefficients (BFCCs) are two similar and powerful audio features usually applied in speech recognition. Cepstral Coefficients offer a compact and accurate high order representation of the spectral envelope similarly to PLP and RASTA-PLP. We expect a similar behavior of both cepstral features since they only differ in the applied psychoacoustic scale.

Pitch and Loudness are two popular perceptual features. Pitch is the perceptual counterpart of the fundamental frequency and represents the perceived frequency of a signal. Additionally, we employ a descriptor for the specific loudness sensation.

Spectral Flux (SF) is a frequency domain feature. It describes the temporal fluctuations in the spectral envelope of the signal.

The Constant Q-Transform (CQT) originates from music analysis. It is closely related to the Fourier Transform but yields frequency components that map efficiently to musical frequencies. Furthermore, we consider coefficients of unitary time-to-frequency transforms: Fourier-, Cosine-, and Wavelet transform (DFT, DCT, DWT).

Time domain features employed include Zero Crossing Rate (ZCR) and the Amplitude Descriptor (AD). The Zero Crossing Rate is a measure for the fundamental frequency of an audio signal. The AD is a set of novel features that characterize the waveform envelope in terms of low- and high-amplitude segments [3].

Table 1 lists the features (with their corresponding dimensions), applied in the experiments. All features add up to a feature vector of 230 dimensions.

Table 1. Features and their corresponding dimensions.

Feature	Dim.	Feature	Dim.
AD	7	BFCC	20
ZCR	1	MFCC	20
DFT	20	LPC	20
DCT	20	PLP	19
DWT	20	RASTA-PLP	19
CQT	20	Loudness	40
SF	2	Pitch	2

2.2. Classifiers

The retrieval quality of the features is usually evaluated by classification. We select a representative set of supervised classifiers. Support Vector Machines (SVM) are a sophisticated kernel-based machine learning technique. Furthermore, we apply Learning Vector Quantization (LVQ) and a K-Nearest Neighbor (K-NN) classifier based on a Euclidean metric.

2.3. Database

The database for the experiments contains 557 samples (105 cars, 127 crowds, 118 footsteps, 105 alerts, and 102 thunder sounds). Note that the alerts class contains two structurally different types of sounds: horns and sirens. These types differ on the technical level, while they represent the same semantic concept of sounds indicating danger. Consequently, we consider them as one class of sounds. The sample database is split into training sets and test sets. We randomly select 12 sounds of each class for the training set (except for the alerts class where we select 24). The remaining 485 sounds form the test set.

2.4. Employed Software

Feature extraction and classification is performed in MATLAB. Features are computed for entire sample files. SPSS is employed for data analysis. After PCA a Varimax rotation is performed in order to maximize the variances of the loadings of the feature components.

3. RESULTS

The goal of the investigations is to evaluate the quality of features and feature combinations. Statistical data analysis reveals the variance contained in the features. Moreover, it shows redundancies between features and feature components.

3.1. Data analysis

In the first step we compute the Principal Components (PCs) of all features involved in the investigation for all data samples in the database. The PCA (with varimax rotation) results in 44 PCs with an eigenvalue > 1 that explain 86.7% of the entire variance contained in the feature data.

The factor loading matrix shows the amount of influence (loading) each feature component has on each principle component. From a "good" feature we expect that its components have high loadings for many different PCs. In feature selection, we aim at combining features that do not load the same principal components.

In the following we discuss the distribution of loadings for different features and analyze similarities among them. Table 2 lists redundancy of the features' components and similarities to other features. Note that the similarities are asymmetrical because of the different dimensionalities of the features.

Table 2. Redundancies and similarities of features.

Feature	Redundancy	Similarity
AD	high	None
DFT	high	DCT
DCT	high	DFT
DWT	low	None
CQT	high	Loudness
SF	high	DFT, Loudn., LPC
BFCC	low	MFCC, LPC
MFCC	low	BFCC, LPC
LPC	avg	MFCC, BFCC
PLP	low	MFCC, BFCC
RASTA-PLP	low	None
Loudness	high	DCT, CQT
Pitch	high	None
ZCR	n/a	LPC

For example ZCR has similar loadings to LPC, however the opposite does not hold, because LPC has in total more components and thus loadings than ZCR.

First, we analyze the unitary signal processing transforms such as Fourier-, Cosine-, and Wavelet transform. The Fourier coefficients show low loadings for most PCs. This means, they contain only little variance. Most variance is contained in the high-frequency coefficients. That indicates that high frequencies are important for recognition of environmental sounds. Cosine coefficients have higher loadings but only for a few PCs. The Wavelet coefficients are independent from each other but only load PCs with low eigenvalues. They cannot capture the major variances contained in the data.

Data analysis reveals that CQT coefficients are highly redundant. All 20 coefficients load the first PC, which represents the largest variance of the data. This fact indicates that the feature may be discriminative only to a certain degree.

ZCR is a one-dimensional feature. Data analysis shows that it does not load any PC significantly. The Zero Crossing Rate represents the fundamental frequency of a signal. In the case of environmental sounds, the fundamental frequency of sounds may be similar for different classes (e.g. thunder and an idle car engine). We conclude that ZCR is not applicable to classification in isolation. Spectral Flux performs similarly to ZCR. We employ mean and variance of the Spectral Flux of entire sound files in the experiments. Data analysis shows that mean and variance of the Spectral Flux are highly correlated. Both load the same PCs moderately.

Similarly to SF, we employ mean and variance of Pitch. Both features are highly redundant, but independent from all other features in the experiments. Pitch loads a PC that is not explained by any other feature. However, the expressiveness of Pitch is limited since this PC explains only 1% of the overall variance

Data analysis of speech features reveals that the LPC coefficients are decorrelated. The loadings are low compared to other features. However, LPC coefficients are distributed over various directions of variance.

PLP and RASTA-PLP are techniques derived from LPC. They take several properties of human perception into account. Surprisingly, both features have weaker loadings than LPC. Retrieval experiments show that the optimizations of (RASTA-) PLP for speech have a negative effect on the retrieval of the investigated data set.

Loudness is a highly redundant but expressive feature (it highly loads the first four PCs). Redundancy may result from correlation of loudness in adjacent frequency bands.

MFCCs and BFCCs are popular features in audio retrieval. While independence of the features' components is high, MFCCs and BFCCs load the same PCs.

The last feature in the investigation is the Amplitude Descriptor (AD). The components of the AD show partially redundant factor loadings. However, the AD defines a PC that is not loaded by any other feature. As we show below, the AD is necessary in order to obtain the optimal feature set, because it captures information neglected by all other features.

3.2. Feature combination

In the second step we empirically search for an optimal solution. This is achieved by the following strategy. Starting from a well performing feature we add other features that have shown to be independent in the data analysis. The combination is then evaluated by the selected classifiers. Features or feature components that do not improve retrieval quality are removed from the combination. For example, the data analysis reveals that the information of LPC coefficients is already captured by the more expressive MFCCs. Classification proves that LPC coefficients in combination with MFCCs have only little influence on retrieval performance. Hence, we do not select LPC coefficients for the combination. For highly redundant features we choose only individual representative components. For example, a few components of Loudness suffice to represent most information contained in this feature.

By this strategy we obtain a feature combination that contains the first 13 MFCCs, selected (hardly redundant) components of the AD, the mean SF, the first RASTA-PLP coefficient and the first Loudness component. This combination results in a 21-dimensional feature vector summarized in Fig. 1. Values in brackets give the corresponding dimensions.

FC = <**MFCC**(13), **RASTA-PLP**(1), **AD**(5), **SF**(1), **Loudness**(1)>

Fig. 1. The components of the performance-optimized feature vector (FC) and their dimensions.

Experiments show that this combination is able to discriminate the five classes of environmental sounds successfully. Table 3 summarizes the results of classification in terms of Recall and Precision values. Recall and Precision are computed for the entire test set. K-NN outperforms SVM because K-NN suits low-dimensional data better than SVM. To the authors' experience, SVM is better at classifying high-dimensional data (several hundred dimensions). Table 3 shows that SVM performs poorly for the alerts class. The alerts class contains two types of sounds (sirens and horns) that have similar semantic

meaning but different sound characteristics. That makes the decision boundary more complex than for the other classes. K-NN is better at modeling the complex decision boundary than the SVM.

Table 3. Results for the optimized feature combination.

Combi-nation	K-NN		LVQ		SVM	
	R	P	R	P	R	P
Cars	82%	85%	73%	69%	75%	89%
Crowds	99%	89%	3%	100%	94%	87%
Footsteps	90%	97%	77%	84%	93%	94%
Alerts	80%	94%	89%	33%	68%	96%
Thunder	87%	77%	62%	88%	90%	68%
mean	**88.1**	**88.7**	58.9	75.9	84.6	87.1

The linear SVM kernel used in the experiments is not able to model the complex boundary properly. LVQ is not able to discriminate successfully between the classes. The mean Recall over all classes obtained by K-NN is 88.1%, the mean Precision is 88.7%.

We performed more investigations than are presented in this paper. For example we compare the performance of the optimized solution with that of MFCCs because they usually achieve high recognition rates on their own (e.g. in speech recognition). The optimized feature vector outperforms MFCCs by 9.8% in recall and 8.0% in precision). Another experiment showed that the optimized solution also outperforms the feature set containing all features (by 12.8% in recall and 15.4% in precision). These results prove the importance of feature selection in this domain.

3.3. Data analysis of the optimal solution

The promising results of the feature combination are reflected by data analysis. Transformation of the combination by PCA yields 6 PCs with eigenvalues *>1* that represent 74.4% of the overall variance. This is a relatively large number of significant PCs indicating low redundancy in the feature set. The first PC covers 17.9% of the overall variance and is mainly loaded by the AD. The AD represents the direction of the highest variance in the data while MFCCs load the second to fifth PCs. Subsequent MFCCs show redundancies because of correlated information in adjacent frequency bands. The mean of Spectral Flux has a high loading for the fifth PC that explains 10.3% of the overall variance. The first RASTA-PLP loads the sixth PC higher than the other features. The Loudness feature loads the first five PCs moderately. However, it covers a significant amount of information and is beneficial for the optimal solution.

The performed investigation shows that factor analysis provides strong hints for feature selection. However, evaluation by classifiers is still necessary to fit the selection to the given data.

4. CONCLUSION

In this paper we evaluated the quality of a large number of features from various fields of audio retrieval. The goal has been the identification of an optimal feature set for the retrieval of environmental sounds. For this purpose, we have performed a quantitative data analysis in order to identify independent features. Data analysis reveals redundancies and dependencies between features. Information obtained by data analysis supports the selection of feature combinations. However, in general promising statistical properties do not guarantee satisfactory retrieval results. Due to this fact, we have empirically tested the identified feature combination for the environmental sounds in the database. Statistical data analysis of the optimal feature set reveals that the Amplitude Descriptor is independent from the other features and highly beneficial for retrieval.

ACKNOWLEDGEMENT

This work has received financial support from the Austrian Science Fund (FWF) under grant no. P16111-N05.

REFERENCES

[1] Y. Choi, K. Kim, J. Jung, S. Chun, K. Park, “Acoustic intruder detection system for home security,” *IEEE Transactions on Consumer Electronics*, Vol. 51, No. 1, February 2005, pp. 130-138

[2] S. Nepal, U. Srinivasan, G. Reynolds, "Detection of goal segments in basketball videos", *ACM Multimedia*, Ottawa, Canada, October 2001, pp. 261-269

[3] D. Mitrovic, M. Zeppelzauer, C. Breiteneder, "Discrimination and retrieval of animal sounds", *IEEE Multimedia Modelling Conference*, Bejing, China, January 2006, pp. 339-343

[4] D. Mitrovic, M. Zeppelzauer, H. Eidenberger, "Analysis of the Data Quality of Audio Descriptions of Environmental Sounds", *Fourth Special Workshop “Multimedia Semantics”*, Chania, Greece, June 2006, pp. 70-79

[5] L. Yu, H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy,” *The Journal of Machine Learning Research*, Vol. 5, 2004, pp. 1205-1224