

TOWARDS AN OPTIMAL FEATURE SET FOR ENVIRONMENTAL SOUND RECOGNITION

Dalibor Mitrovic, Matthias Zeppelzauer, 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

Feature selection for audio retrieval is a non-trivial task. In this paper we aim at identifying an optimal feature combination for environmental sound recognition. The feature combination is constructed from a broad set of features. Additionally to state-of-the-art features, we evaluate the quality of audio features we previously introduced for another domain. We examine the properties of features by quantitative data analysis (factor analysis) and identify candidates for feature combinations. We verify the quality of the combination by retrieval experiments. The optimal solution yields Recall and Precision values of 87% and 88%, respectively.

1. INTRODUCTION

Environmental sound recognition (ESR) is a research field that addresses the recognition of non-speech and non-music sounds. The goal of ESR is to distinguish between different classes of environmental sounds. The domain of environmental sounds is nearly infinite in size. Hence, most state-of-the-art research focuses on a limited domain of sounds. The variety of applications of ESR ranges from automatic surveillance to life logging [1]. Recent research focuses on multimodal retrieval that combines visual and auditory information to improve retrieval quality and semantic understanding of multimedia objects [2]. Furthermore, automatic annotation of audio and video data gains importance. Section 4 presents related work in the field of ESR.

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 (AD) in the domain of general purpose environmental sounds. Originally, we introduced the AD for animal sound recognition [3]. We examine the redundancy of the features by a quantitative data analysis (Principal Components Analysis, PCA). Moreover, we evaluate the retrieval quality of the features by a set of classifiers. Finally, we discuss retrieval results in context of the data analysis. The goal of these investigations is the identification of an optimal feature set for environmental sound recognition.

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. A brief survey of related work is given in Section 4.

2. EXPERIMENTS

In order to identify the optimal feature set for classification of environmental sounds, we perform a series of experiments. 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.

A vast number of audio features exists that were 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. Furthermore, we employ Perceptual Linear Prediction (PLP) and Relative Spectral Perceptual Linear Prediction (RASTA-PLP). PLP and RASTA-PLP are derivatives of LPC, optimized for speaker-independent speech recognition.

Pitch and Loudness are two widely used perceptual features (perceptual features try to imitate the human auditory sense). Pitch is the perceptual counterpart of frequency. It measures the perceived frequency of a signal. Beside Pitch, we employ the perceptual loudness measure Sone. Spectral Flux (SF) is a frequency domain feature. It describes the fluctuations in the spectrum of the signal.

Mel Frequency Cepstral Coefficients (MFCCs) and Bark Frequency Cepstral Coefficients (BFCCs) are two similar and powerful audio features usually applied in speech recognition. Since they only differ in the applied psychoacoustic scale we expect a similar behavior of both features. Cepstral Coefficients offer a compact and accurate high order representation of audio signals.

The Constant Q-Transform, is a music analysis feature. 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-frequency transforms (DFT, DCT, DWT).

Time domain features employed include Zero Crossing Rate (ZCR) and the Amplitude Descriptor (AD). The ZCR is a measure for the fundamental frequency of an audio signal. The AD is a set of novel features that describe the characteristics of the signal waveform [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.

Feature	Dimensions	Feature	Dimensions
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

Table 1. Investigated feature groups and their corresponding dimensions.

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 with an Euclidean distance measure.

The database for the experiments contains 557 samples (105 cars, 127 crowds, 118 footsteps, 105 signals, and 102 thunder sounds). The signal class contains two subclasses: horns and sirens. These subclasses differ on the technical level, while they represent the same concept of *sounds indicating danger*. The sample database is split into training sets and test sets. Each training set comprises of 12 randomly chosen samples. The training set of the signal class contains 24 samples. The remaining samples form the test sets.

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. The Varimax rotation rotates the eigenvectors to a position that maximizes the variances of the loadings of the feature components.

3. RESULTS

In this section we discuss the results of the quantitative data analysis and their relationship to retrieval quality. The goal of the investigations is to evaluate the quality of features and feature combinations. Data analysis reveals the variance contained in features. Moreover, it shows redundancies between

features and groups of features. Classification requires low variance inside classes and high variance between classes. The distribution of variances depends on the allocation of the data into classes. Data analysis does not consider class information. Hence, promising data analysis results are necessary but not sufficient for successful classification by a feature.

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 results in 44 PCs with an eigenvalue > 1 that explain 86.7% of the entire variance contained in the feature data.

In the following we discuss the distribution of loadings in the varimax-rotated factor loading matrix for different groups of features. Furthermore we analyze similarities among the groups of features. In Table 2 redundancy in and similarity of groups of features are listed.

Feature	Redundancy	Similarities
AD	high	none
DFT	high	DCT
DCT	high	DFT
DWT	low	none
CQT	high	Loudness
SF	high	FFT, Loudness, 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

Table 2. Redundancy inside groups of features and their similarities to other groups of features.

Firstly, we analyze the unitary signal processing transforms such as DFT, DCT, and DWT. The coefficients of the DFT 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. In contrast to DFT, DCT coefficients have high loadings. The coefficients are highly redundant, because they load the same PCs. Beside DFT and DCT, we analyze the expressiveness of the DWT. The DWT coefficients are independent from each other but only load PCs with low eigenvalues. The DWT 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 to a certain degree.

ZCR is a one-dimensional feature. Data analysis shows that it does not load any PC. ZCR 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). ZCR is not applicable to classification as a single feature. Spectral Flux performs similarly to ZCR. We employ mean and variance of the SF of entire sample files in the experiments. Data analysis shows that mean and variance of SF 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. Since this PC explains only 1% of the overall variance, the expressiveness of Pitch is limited.

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

PLP is a technique derived from LPC. It takes several properties of human perception into account and was developed for speaker-independent speech recognition. PLP has higher but fewer loadings than LPC. Data analysis of PLP is similar to that of MFCCs. Redundancy of PLP is low and loadings are distributed over several PCs. Experiments show that information described by PLP is not as discriminative as the information contained in MFCC. The optimizations of PLP for speech recognition have a negative effect on the retrieval of environmental sounds. We observe a similar behavior for RASTA-PLP.

Loudness is a highly redundant but expressive feature. Loudness is represented by some values of adjacent frequency bands. Redundancy may result from correlation of loudness in these frequency bands. The loadings of Loudness are especially high for the first four PCs.

MFCCs and BFCCs are popular features in audio retrieval. While independence inside the groups of features is high, MFCCs and BFCCs load the same PCs. MFCCs and BFCCs share all properties important for well performing features.

The last group of features in the investigation is the Amplitude Descriptor (AD). The features of the AD show partially redundant factor loadings. However, the AD defines a PC that is not loaded by any other feature in the survey. As will be shown below, the AD is necessary in order to obtain an optimal feature set for retrieval, 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 showed to be independent in the data analysis. The combina-

tion is evaluated by the selected classifiers. Features or groups of features 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 groups of features we choose only individual representative components. For example, a few components of Loudness suffice to represent most information contained in this group of features.

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 Figure 1.

$$FC = \langle \text{MFCC (13), RASTA-PLP (1), AD (5), SF (1), Loudness (1)} \rangle$$

Fig. 1. The elements 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 performs best, followed by SVM. LVQ is not able to discriminate sufficiently between the classes. The mean Recall over all classes obtained by K-NN is 87.4%, the mean Precision is 88.2%.

Combination	K-NN		LVQ		SVM	
	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.
cars	82%	85%	73%	69%	75%	89%
crowds	99%	89%	3%	100%	94%	87%
footsteps	90%	97%	77%	84%	93%	94%
signal	80%	94%	89%	33%	68%	96%
thunder	87%	77%	62%	88%	90%	68%

Table 3. Recall and Precision values obtained by the optimized feature combination.

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. 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. Finally, the Loudness feature does not yield high loads for any of the PCs. In contrast to the other features, it loads the first five PCs moderately. Due to this behavior, the Loudness feature covers a significant amount of information and is beneficial for an optimal solution.

The performed investigation shows that factor analysis provides strong hints for choosing features combinations that capture a maximum of information of the data samples. The feature combination identified, discriminates the classes of environmental sounds well and classifies 87% of the samples correctly.

4. RELATED WORK

Environmental sound recognition concerns with the identification of sounds that do not originate from speech or music. The range of environmental sounds is extremely wide. Hence, most investigations concentrate on restricted domains. A popular research field is audio recognition in broadcasted video. In [4], the authors recognize the scene content of TV programs (e.g. weather reports, advertisements, basketball and football games) by analyzing the audio track of the video. They extract Pitch, Volume Distribution, Frequency Centroid and Bandwidth to characterize TV programs. Classification is performed by a separate neural network for each class. A well investigated problem is highlight detection in sports videos. The authors of [5] retrieve crucial scenes in soccer games by analyzing play-breaks. Whistles, that often refer to play-breaks in sports, are detected using Spectral Energy within an appropriate frequency band. Another indicator for highlights is the audience. Excitement is quantified by Loudness, Silence and Pitch. Another area of interest is surveillance and intruder detection. The authors of [6] detect intruders in a room by monitoring variations in a room-specific transfer function.

The authors of [7] compare and combine cepstral features (MFCCs) with perceptual features (Brightness, Bandwidth, Pitch, etc.). In [7], perceptual features outperform cepstral features. Best results are reached by a combination of both. In [7] SVM performs better than NN and K-NN.

A challenging area of environmental sound recognition is life logging [1]. This research field is concerned with continuously analyzing the environmental sounds surrounding a human user. From this information a diary is built where major events and the user's activities are stored.

5. CONCLUSIONS

In this paper we evaluated the quality of a large number of features from various fields of audio retrieval. The goal was the identification of an optimal feature set for the retrieval of environmental sounds. For this purpose, we 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. Since promising data analysis properties of features do not guarantee satisfactory retrieval results, we empirically tested and enhanced the identified feature combination. By this procedure we identified an optimal feature combination for the environmental sounds in the database. In the final step we prove low redundancy of the feature combination by data analysis. Data analysis reveals that the novel feature set (Amplitude Descriptor), we introduced in [3], is independent from the other features. Hence it is necessary to obtain an optimal solution for retrieval. Eventually, experiments show that features from related domains, such as speech recognition and music information retrieval, qualify for environmental sound recognition.

6. REFERENCES

- [1] K. Aizawa, "Digitizing personal experiences: Capture and retrieval of life log," *In Proceedings of the IEEE Multimedia Modelling Conference*, vol. 00, pp. 10–15, 2005.
- [2] S. Nepal, U. Srinivasan, and G. Reynolds, "Detection of goal segments in basketball videos," *In Proceedings of ACM Multimedia Conference*, 2001.
- [3] D. Mitrovic and M. Zeppelzauer, "Discrimination and retrieval of animal sounds," *In Proceedings of the IEEE Multimedia Modelling Conference (accepted)*, 2006 (www.ims.tuwien.ac.at/publication_master.php).
- [4] Z. Liu, J. Huang, Y. Wang, and T. Chuan, "Audio feature extraction and analysis for scene classification," *In IEEE Workshop on Multimedia Signal Processing*, vol. 20, pp. 343–348, 1997.
- [5] D. Tjondronegoro, Y. Chen, and B. Pham, "The power of play-break for automatic detection and browsing of self-consumable sport video highlights," *In Proceedings of the ACM Workshop on Multimedia Information Retrieval*, pp. 267–274, 2004.
- [6] Y. Choi, K. Kim, J. Jung, S. Chun, and K. Park, "Acoustic intruder detection system for home security," *In IEEE Transactions on Consumer Electronics*, vol. 51, pp. 130–138, 2005.
- [7] G. Guo and Z. Li, "Content-based classification and retrieval by support vector machines," *In IEEE Transactions on Neural Networks*, vol. 14, pp. 209–215, 2003.