Discrimination and Retrieval of Animal Sounds

Dalibor Mitrovic, Matthias Zeppelzauer and Christian Breiteneder
Vienna University of Technology
Institute of Software Technology and Interactive Systems
Favoritenstrasse 9-11, A-1040 Vienna, Austria
{mitrovic, zeppelzauer, breiteneder}@ims.tuwien.ac.at

Abstract

Until recently few research has been performed in the area of animal sound retrieval. The authors identify state-of-the-art techniques in general purpose sound recognition by a broad survey of literature. Based on the findings, this paper gives a thorough investigation of audio features and classifiers and their applicability in the domain of animal sounds. We introduce a set of novel audio descriptors and compare their quality to other popular features. The results are encouraging and motivate further research in this domain.

1. Introduction

Recently, audio data gained importance in the field of content-based retrieval. The rising number of audio and video databases states the need for efficient retrieval. The Quality of retrieval depends on the features that represent the signal, and on the classifiers that discriminate between classes of signals. Animal sounds are a domain of environmental sounds that has not been investigated yet in detail. Some investigations consider animal sounds among other classes of sound [14], [12]. To the authors' knowledge there is no prior work analyzing the discrimination of animal sounds from each other. Our contribution to this research field is represented by a thorough investigation of the applicability of state-of-the-art audio features in the domain of animal sound recognition. Additionally we introduce a set of novel features and compare their performance with popular audio features. Besides, we present an extensive survey of state-of-the-art features and classifiers.

In this paper the authors try to identify an efficient method for automatically distinguishing between sounds of different animals. Such a technique could be part of a supporting system for the deaf, providing information about the surrounding environment. Automatic surveillance and annotation of time-dependent media may employ animal sound recognition as well. Additionally, life logging applications could take advantage of such a technique, imagine a visit to the zoo.

Audio data may be coarsely divided into three classes: speech, music, and environmental sounds. Speech recognition has a long tradition and is extensively surveyed by Rabiner and Juang in [23]. Music analysis deals with the identification of music genre, artist, instruments and structure [9].

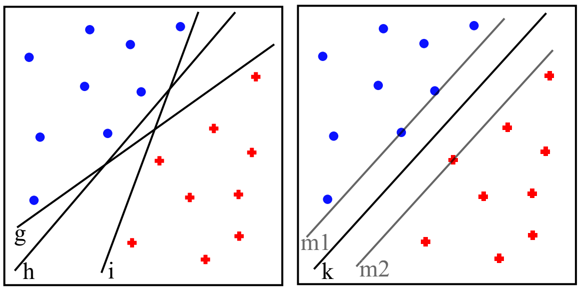
The remainder of this paper is organized as follows: In Section 2 the principles of Support Vector Machines (SVM) are given. Section 3 addresses the methodology considered. Results are discussed in Section 4. A survey of related work is performed in Section 5. Finally in Section 6 conclusions and future work are presented.

2. Background

Classification is an important step in content-based retrieval. The process of classification tries to correctly predict the class of a sample. A recent classifier is the Support Vector Machine (SVM) [3][30]. SVMs are supervised, statistical learning methods applicable for classification and regression. They are also known as maximum-margin classifiers.

Given two separable clouds of points $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_k, y_k)$ where $\mathbf{x}_i \in R^n$ and $y_i \in \{-1, +1\}$, an SVM constructs an optimal separating hyperplane $\mathbf{wx} + b = 0$, that maximizes the distance between the hyperplane and the nearest data point of each cloud (these points are the support vectors). The distance between the support vectors and the hyperplane is called margin. Figure 1 depicts the difference between a suboptimal and an optimal separating hyperplane.

The hyperplane is not constructed in feature space, instead the saddle point of the following Lagrange functional is cal-



(a) (b)

Figure 1. Optimal Separating Hyperplanes (OSH): (a) g, h, i are valid but not optimal separating hyperplanes. (b) k is the OSH, the distance between k and m1 respectively m2 is equal and maximal.

culated:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\left\|\mathbf{w}\right\|^2 - \sum_{i=1}^{l} \alpha_i \left\{ y_i \left[(\mathbf{w} \cdot \mathbf{x}) + b \right] - 1 \right\}, \tag{1}$$

where α_i are the Lagrange multipliers. Equation (1) may be transformed into problem (2) which is easier to solve.

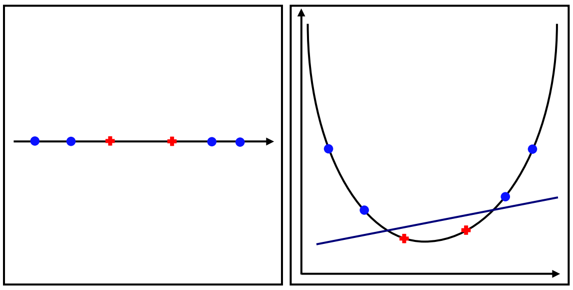
$$\bar{\mathbf{w}} = \sum_{i=1}^{l} \bar{\alpha}_i y_i \mathbf{x}_i, \qquad \bar{b} = -\frac{1}{2}\bar{\mathbf{w}} \cdot [\mathbf{x}_r + \mathbf{x}_s] \tag{2}$$

where $\mathbf{x}_r$ and $\mathbf{x}_s$ are two arbitrary support vectors with $\bar{\alpha}_r, \bar{\alpha}_s > 0, y_r = 1, y_s = -1$. Slack variables ζ_i and a penalty function $F(\zeta) = \sum_{i=1}^{l} \zeta_i$ are the means by which SVMs become applicable for the non-separable case [6]. The separating hyperplane is constructed in such a manner, that the number of falsely classified $\mathbf{x}_i$ is minimal. This consequently minimizes $F(\zeta)$. The slack variables only influence the Lagrange multipliers α_i, hence the solution for the optimization problem stays the same as for the separable case.

In practice most problems are not linearly separable. Instead of identifying a non linear separating function, the data points are transformed into a higher order space in which they become linearly separable. This is achieved by the use of kernels. Figure 2 illustrates the effect of a polynomial kernel that maps the input space into a feature space of higher order. Equation (3) describes the SVM classifier, where $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel used.

$$f(x) = sign\left(\sum_{support\ vectors} \bar{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + \bar{b} \right) \tag{3}$$

There are three typical kernel functions:



(a) not linearly separable (b) linearly separable

Figure 2. The Kernel maps the one dimensional input space (a) into a feature space of higher dimensionality, where the inputs become linearly separable (b).

1. polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = \left[(\mathbf{x}_i \cdot \mathbf{x}_j) + 1 \right]^d$,
2. Radial Basis Function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-(\mathbf{x}_i - \mathbf{x}_j)^2 / 2\gamma^2 \right)$, and
3. sigmoid: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(scl \cdot (\mathbf{x}_i \cdot \mathbf{x}_j) - off)$,

where *scl* (scale) and *off* (offset) are parameters that have to be chosen with care. The kernel becomes invalid for certain parameter values. Kernel functions are not limited to the ones mentioned above. Any symmetric function that satisfies the conditions in Mercer's Theorem is a valid kernel function [2].

3. Experiments

Distinction of animal sounds has not been investigated yet. In this paper we examine ways to distinguish between animal sounds. We choose four animals, namely birds, cats, cows, and dogs. Sounds by birds and cats respectively by cows and dogs show significant similarity on a perceptual level. That qualifies them to measure the quality of features and classifiers.

There is no publicly available reference database of animal sounds. The authors built a custom database of sound samples from an internet search. The database contains 383 samples (99 birds, 110 cats, 90 cows, 84 dogs). The data have a sample rate of 11025 Hz, are quantized to 16 bit and are single channel. A sound sample contains one or more repeated sounds of an animal (e.g. repeated barks of a dog). Additionally some samples contain background noise of other animals. File lengths and loudness levels vary over the samples.

All experiments are conducted in MATLAB using an extensible framework. Our framework supports the definition of experiment setups by configuration files. Configuration files specify ground-truth, test data, features, classifiers, and result output options. This enables efficient and consistent tests of various features and classifiers.

3.1. Feature Extraction

The survey in this paper considers multiple state-of-the-art features applied in speech recognition, music information retrieval and environmental sound recognition. The goal is to identify suitable features for the domain of animal sounds. The authors examine different types of features. Time domain features include Linear Predictive Coding (LPC) coefficients, Zero Crossing Rate (ZCR), Periodicity Histogram, Sone and Short Time Energy (STE). The following spectral features are investigated: Relative Spectral Predictive Linear Coding (RASTA PLP), Pitch, Spectral Flux (SF) and coefficients from basic time-frequency transforms (FFT, DCT, DWT, CWT and Constant Q-Transform). Cepstral domain features are Mel Frequency Cepstral Coefficients (MFCC) and Bark Frequency Cepstral Coefficients (BFCC). Additionally we introduce a set of new time-based features that describe the shape of the waveform of the signal. We call them Length of High Amplitude Sequence (LoHAS), Length of Low Amplitude Sequence (LoLAS) and Area of High Amplitude (AHA).

Some of the above features do not perform sufficiently well and are not considered in detail. For example, the coefficients of the basic transforms (FFT, DCT, DWT, CWT and Constant Q-Transform) poorly discriminate the classes of animal sounds. The reason for this is that significant information in higher frequency bands is not considered in the first transform coefficients. RASTA PLP [8], ZCR, Pitch [26], Sone [22] and SF perform slightly better and are candidates for combinations with other features.

In the following we describe features that performed best for our data set. Linear predictive coding (LPC) represents a signal processing technique applied in signal compression, speech synthesis and speech recognition [28]. The goal of LPC is to separate formants from a speech signal. Formants describe the vocal tract (mouth, throat) of a speaker by its resonances. The formants are extracted by a linear predictor. The linear predictor tries to express the value of a sample by a linear combination of values of previous samples. LPC estimates coefficients using linear prediction, that minimizes the mean square error (MSE) between the original signal and the predicted signal. The coefficients of the linear predictor represent the formants of a speech signal. LPC coefficients are employed in speech recognition to distinguish between phonemes. It is beyond the authors knowledge that LPC coefficients have been introduced to environmental sound recognition. In this paper LPC features are successfully applied to animal sounds (see Section 4).

Cepstral Coefficients (CCs) are a popular feature in audio retrieval [18], [32]. The authors of [29] define the cepstrum as the Fourier Transform (FT) of the logarithm (log) of the spectrum of the original signal.

$$signal \rightarrow FT \rightarrow log \rightarrow FT \rightarrow cepstrum$$

In practice, CCs are derived from the FFT or DCT coefficients or linear predictive analysis [4]. CCs offer a compact and accurate high order representation of signals. Peaks in the cepstrum correspond to harmonics in the power spectrum.

Computation of MFCCs includes a conversion of the logarithmized Fourier coefficients to Mel scale. After conversion, the obtained vectors have to be decorrelated to remove redundant information. A DCT is applied to receive a decorrelated, more compact representation. MFCCs are an instance of CCs. In the following sequence the computation of MFCCs is illustrated.

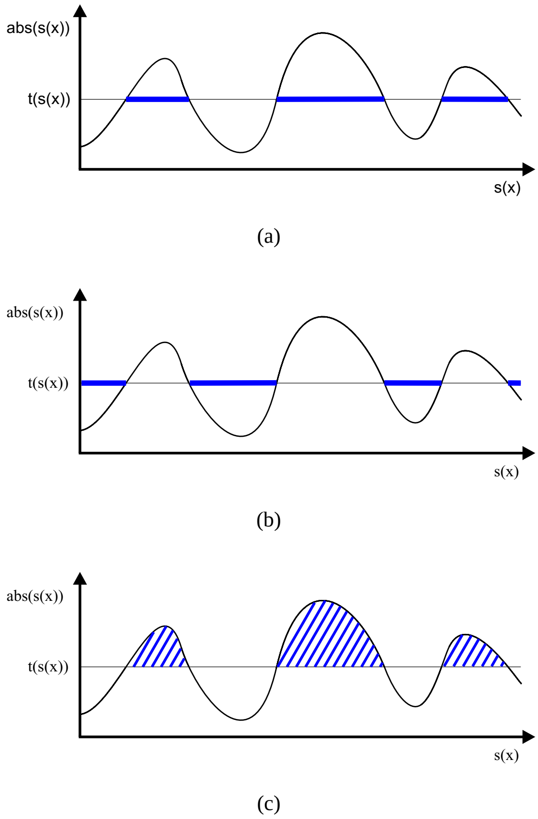
$$signal \rightarrow FT \rightarrow log \rightarrow Mel \rightarrow DCT \rightarrow MFCCs$$

A closely related group of features is BFCCs. BFCCs are similarly computed as MFCCs. They differ in the applied scale (Bark scale).

$$signal \rightarrow FT \rightarrow log \rightarrow Bark \rightarrow DCT \rightarrow BFCCs$$

Bark scale and Mel scale are perceptually motivated acoustical scales that nonlinearly map the signal frequency. Both nonlinear scales offer higher resolution for low frequencies than for high frequencies. MFCCs and BFCCs are expected to perform similarly.

Additionally to the features above we introduce a set of time-based low-level features. The features describe characteristics of the waveform such as peaks and silence. The features are computed based on an adaptive threshold. The threshold for a particular sound sample is the sum of mean and standard deviation of the absolute sample values. This threshold separates peaks from silence in the waveform. Based on this threshold we compute the length of high amplitude sequences (LoHAS). The length of a high amplitude sequence represents the number of consecutive samples that have a value greater or equal to the threshold. LoHAS represents the distribution of the length of peaks in the signal. Figure 3(a) illustrates this feature. Analogously we define the length of a low amplitude sequence (LoLAS) as the number of consecutive samples that have a lower value than the threshold. LoLAS describes the distribution of length of the silent portions in the signal. Details are depicted in Figure 3(b). Sequences with high amplitude can be further characterized by the corresponding area below the waveform. We compute the area of high amplitudes (AHA) as area between the threshold and the signal in a LoHAS. In other words the AHA feature represents the extent of peaks in the signal. Figure 3(c) illustrates this concept.



(a)

(b)

(c)

Figure 3. LoHAS, LoLAS and AHA for signal s(x) with threshold t(s(x)): (a) Length of High Amplitude Sequence (LoHAS). (b) Length of Low Amplitude Sequence (LoLAS). (c) Area of High Amplitude (AHA).

The authors consider statistical properties of LoHAS, LoLAS, and AHA to build features that describe entire sample files. The final features comprise mean, standard deviation and median of LoHAS and LoLAS over the entire signal. Additionally we extract the mean of AHA. This results in a 7 dimensional feature vector which is used for classification.

3.2. Classification

This section offers a brief discussion of the classification methods and the parameters used. Three supervised classifiers are employed: SVM, described in Section 2, Nearest Neighbor (NN) with an Euclidean distance measure, and the MATLAB implementation of Linear Vector Quantization (LVQ). The SVM is applied with a linear kernel and an RBF kernel. NN is considered to test the quality of the features. Features that discriminate classes well, provide disjoint partitions of the feature space. Satisfactory results with the NN algorithm indicate such a partitioning in the feature space.

LVQ by Kohonen [16] is a classification method closely related to Self Organizing Maps (SOMs) [15]. LVQ tries to approximate the distribution of classes. The LVQ algorithm iteratively computes codebook vectors in a manner that the error rate is minimized. Each codebook vector represents a particular class. Results heavily depend on the class distribution in the training set and the initially chosen codebook vectors; i.e. two consecutive runs of the training algorithm do not necessarily yield the same results. We utilize a training set with evenly distributed classes and 200 epochs for training.

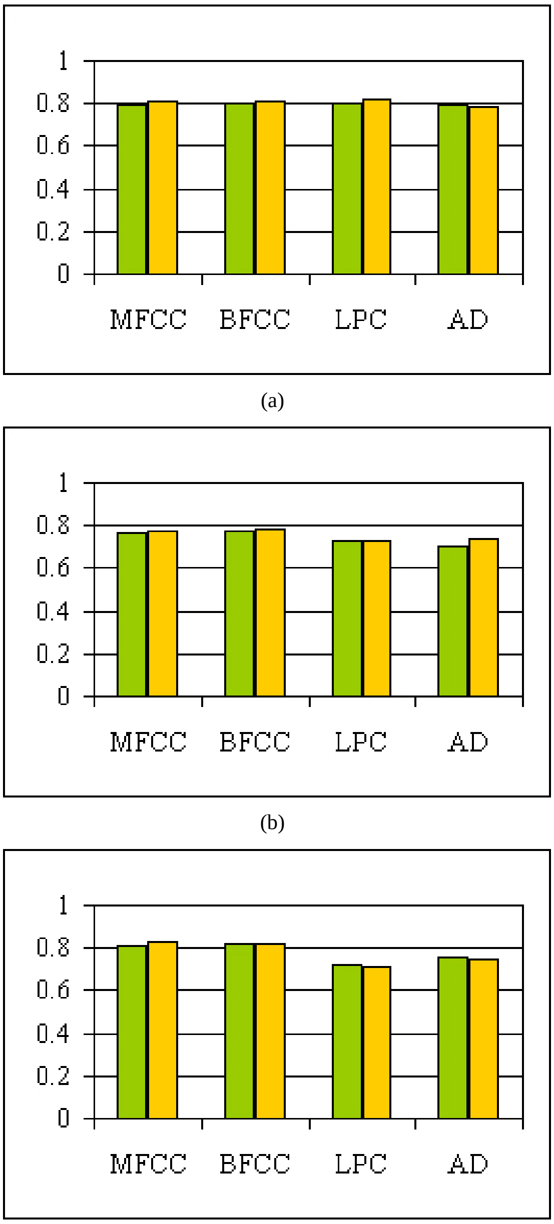
4. Results

In this section we present the results of our experiments. The sample database is split into a test set and a training set. The training set comprises 12 samples per class. The remaining samples form the test set: 87 bird samples, 98 cat samples, 78 cow samples, and 72 dog samples.

Multiple features performed poorly for our test data. The first few transform coefficients of FFT, DCT, DWT, CWT and Q-Transform insufficiently discriminate the animal sounds. The selected coefficients do not express the high frequencies well. In the case of animal sounds, high frequencies contain significant information (e.g. for cats and birds). Performance of one-dimensional features such as ZCR, SF, and Pitch is below that of multi-dimensional features, but low-dimensional features are not able to sufficiently represent the samples. In combination with other features ZCR, SF, and Pitch may improve results. STE is only useful in classification based on frames. When STE is computed for entire files, it represents the average energy of the sound sample, which does not provide meaningful information in our case.

In the following we consider the best performing features in detail, which are LPC, MFCC, BFCC, and an amplitude descriptor (AD). They are described in the rest of this section. The AD consists of LoHAS (mean, standard deviation, median), LoLAS (mean, standard deviation, median), and AHA (mean). LPC coefficients may be represented in many different ways [4]. For the data set used, the representation as impulse response is the best choice. 20 LPC coefficients are extracted from each sound sample. We consider the first 20 MFCCs and BFCCs [4], [8]. Delta and Double Delta Cepstrum features perform poorly and are not considered. At first the selected features are tested in isolation. Afterwards we try to identify an optimal solution to the recognition problem by combining features.

Each selected feature is tested with three classifiers: SVM, NN, and LVQ. The classifiers are trained by the training set



(a)

(b)

(c)

Figure 4. Recall (green) and precision (yellow) of the single features classified by: (a) SVM, (b) LVQ, (c) NN. AD can compete with the higher dimensional features.

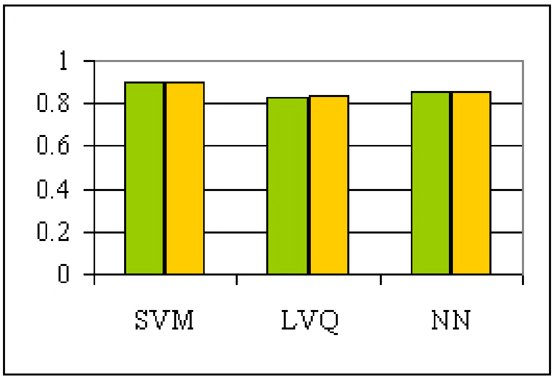


Figure 5. Recall (green) and precision (yellow) of the combined feature vector with different classifiers. The feature in combination with SVM performs best.

to construct a model that describes the data. The test data serve to validate the model. For each feature we compute recall and precision per class. Figure 4 shows mean recall and mean precision over all classes for the selected features. MFCCs and BFCCs perform nearly identically. This results from the fact that both are cepstral domain features that only differ in the psycho-acoustical scaling. MFCCs deliver the best results using the NN classifier (recall=0.81). That indicates that MFCCs cluster the feature space according to the classes. The SVM with a linear kernel yields similar results for MFCCs and BFCCs. LVQ provides slightly less performance for these features.

LPC coefficients discriminate the classes well. Best results are gained by the SVM with an RBF kernel illustrated in Figure 4(a). The NN classifier suboptimally explains the data. The distribution of the LPC coefficients appears to be too complex for the simple NN decision rule. LVQ demonstrates similar performance as NN for LPC. Figure 4(b) depicts the results obtained by LVQ.

In contrast to MFCC, BFCC, and LPC, each with 20 components, AD only consists of 7 components. Classification with the SVM and a linear kernel provides a recall and a precision of 0.78. This is comparable to the other features, illustrated in Figure 4(a). For the NN classifier recall and precision of AD lie between those of LPC and MFCC. The results of AD and LPC are similar using the LVQ classifier. In many cases AD performs equally or better than the other features, although it is of much lower dimension.

The features in our tests achieve satisfactory recall and precision with all classifiers (between 0.7 and 0.8). The classifiers do not perform equally well. While the SVM is able to maintain high precision and recall for all four features, LVQ and NN are not. The reason for this is, that NN and LVQ depend on the clustering of samples in feature space.

They deliver satisfactory results when the classes form disjoint clusters. In contrast, SVM constructs a more abstract model. As a consequence it depends less on the distribution of samples in feature space.

Up to now we concentrated on individual features. In order to improve retrieval quality, we combine several features to a feature vector. This makes sense because the combination aggregates information present in separate features. The feature vector comprises 26 components: 3 components (mean, standard deviation, median) of LoHAS respectively LoLAS, 4 LPC coefficients, 13 MFCCs, the mean SF, the mean Pitch, the first RASTA PLP coefficient and the mean of Sone. Classification based on this feature vector yields an average precision and recall above 0.9 using the SVM with a linear kernel. This is a significant improvement over results with the individual features. LVQ and NN profit from the combined feature vector as well. Figure 5 depicts recall and precision of the combined feature vector for different classifiers.

5. Related Work

There are different groups of audio retrieval techniques. Numerical representation of signals by features, is common to all methods. Approaches can be grouped by the way similarity among different signals is detected. A straight forward technique is to apply a distance measure directly to the features. Pioneering work in this area concerning audio is performed in [31]. The authors develop a content-based audio retrieval system (Muscle Fish) that distinguishes classes such as animals, machines, musical instruments, telephone, etc. They extract features such as loudness, pitch, brightness and bandwidth. Similarity is measured using a weighted Euclidean distance (Mahalanobis). Classification is accomplished by the nearest neighbor rule.

An alternative to directly measuring similarity is the use of artificial intelligence techniques such as Support Vector Machines (SVM) [6], Hidden Markov Models (HMM) or Artificial Neural Networks (ANN). An early example in the domain of audio processing is represented by [10]. The authors apply a self-organizing neural network to cluster similar sounds. Another way of classification is based on template matching [11]. The author extracts MFCC features from the audio signal and clusters the feature space into distinct cells with a quantization tree (Q tree). Histograms are considered as templates. They represent the distribution of feature vectors over the partitions of the tree. Templates are compared by distance measures (e.g. Euclidean distance or cosine distance).

Segmentation is an important preprocessing step of audio analysis. It is employed to discriminate different types of sound such as speech, music, environmental sounds and combinations of these. The authors of [25] separate music and speech with low level features. They apply Spectral Centroid, Spectral Flux (SF), Zero Crossing Rate (ZCR), Spectral Roll-off, and Percentage of Low Energy Frames to represent the audio signal. Different classification techniques such as Gaussian Mixture Model (GMM) and Nearest Neighbor (NN) are used to separate speech from music based on these features.

Based on successful segmentation of an audio stream, different audio types can be further analyzed. The most intensive research took place in the area of speech recognition. Beside classical recognition of speech [23] researchers focus on recognition of the spoken language [21]. Another field of research is classification of the speaker (e.g. for customization issues or authentication) [24]. In the area of multimodal dialog systems recognition of human emotions from audio gains focus [5]. The different areas of speech processing are a source to survey state-of-the-art audio features.

Beside speech recognition music information retrieval (MIR) gained importance through the availability of huge amounts of digital music. MIR consists of classification and structural analysis. Classification concerns recognition of instruments, artists and genres. Multiple speech recognition features are applicable to the classification of music. In [18] the authors distinguish between instruments (e.g. Brass, Keyboard, and String) by extracting features such as ZCR, Short Time Energy (STE), Bandwidth, Pitch, Formant Frequencies and Mel-Frequency Cepstral Coefficients (MFCC). These features are computed from short frames of the audio signal. The mean and standard deviations of the features over all frames add up to the final feature vector that represents the signal. Classification is performed by GMM and NN. Music genre classification is addressed in [13]. In this paper the authors propose the discrete wavelet packet decomposition transform to distinguish music genres.

Structural music analysis tries to extract similarities and recurrences in a piece of music. A comprehensive structural analysis is performed in [20]. Autocorrelation is computed to extract Rhythm from the wavelet-decomposed signal. Pitch Class Profiles in combination with HMM separate chords. Vocal and instrumental sections are characterized in terms of Octave Scaled Cepstral Coefficients (OSCC). An SVM trained with OSCC features separates vocal from instrumental sections.

Environmental sound recognition concerns 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 a restricted domain. A popular research field is audio recognition in broadcasted video. In [19] the authors recognize the scene content of TV programs (e.g. weather reports, advertisement, 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 sport videos. The authors of [27] 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. A similar approach is followed by [32]. The authors analyze keywords in commentator speech and audience which are relevant to important actions of the game. They apply an HMM trained with low level features (Energy and MFCCs including delta and double delta features) to recognize the keywords. Investigations presented in paper [33] address extraction of highlights in baseball games. Beside visual features the authors extract audio features (e.g. MFCC, Pitch, Entropy). An SVM detects excitement of the audience. Template matching is applied for baseball hit detection. These two audio cues are combined to improve quality of highlight detection. Another area of interest is surveillance and intruder detection. A broad survey of audio features and classification techniques, in context of automatic surveillance is given in [7].

In [34] multilevel classification is proposed. First the authors apply a coarse level segmentation to separate speech, music and environmental sound. In a second step an HMM is considered to analyze environmental sounds (e.g. footstep, laughter, rain, windstorm). The authors of [14] present an audio indexing system using MPEG-7 features. They apply Audio Spectrum Basis (ASB) and Audio Spectrum Projection (ASP) descriptors to distinguish classes such as ”Dog”, ”Bell”, ”Water”, and ”Baby” with HMMs. They show that MPEG-7 descriptors perform similar to MFCC. SVMs are successfully applied to environmental sound recognition in [12]. The authors compare and combine cepstral features (MFCCs) with perceptual features (Total Spectrum Power, Subband Powers, Brightness, Bandwidth, and Pitch). In [12] perceptual features outperform cepstral features. Best results are reached by a combination of both. In [12] SVM performs better than NN and k-NN.

A challenging area of environmental sound recognition is life logging. This research field is concerned with continuously analyzing the environmental sounds of a human user. From this information a diary is built where major events and the user’s activities are stored. Fundamental research in the domain of life logging is performed in the ”Forget-me-not” system [17]. ”Forget-me-not” is a mobile application that analyzes the activities of a user in his office. This includes monitoring the workstation, telephone, printer and the location of the user. In [1], Aizawa presents a life logging system that captures video and audio. Audio information is considered to detect human voice to recognize conversation scenes. The system supports GPS and provides inertial trackers to measure motion. Additionally it has access to documents, web pages, and emails. Applications discussed in this section prove the importance of environmental sound recognition for future information systems.

6. Conclusions & Future Work

Discrimination of animal sounds is a rarely considered area of environmental sound recognition. In this paper we presented a survey of widely used audio features and classifiers. Our research focus was the investigation of their applicability in the domain of animal sound recognition. We introduced a set of novel time-based audio features that are easy to compute. Despite their simplicity, they perform comparably to much more complex features, such as MFCC or LPC. We have shown that a combination of state-of-the-art features with our feature set is able to successfully classify more than 90% of the animal sounds in our database (using SVM). Beside SVM, we employed NN and LVQ classifiers in our experiments. All classifiers yielded satisfactory results. The SVM slightly outperformed NN and LVQ.

Further work will include comparison of the features discussed in this paper with MPEG-7 features for environmental sound recognition. Additionally we will examine context sensitive classifiers such as Hidden Markov Models and Artificial Neural Networks. Animal sound recognition will be incorporated into life logging applications. A future goal is the distinction of different sounds from the same species (”understanding animals”).

Acknowledgments

The authors are very grateful to Doris Divotkey and Horst Eidenberger for their guidance and advice. This work has received financial support from the Austrian Science Fund (FWF) under grant no. P16111-N05.

References

[1] K. Aizawa. Digitizing personal experiences: Capture and retrieval of life log. *In Proceedings of the 11th International Multimedia Modelling Conference*, pages 10–15, January 2005.

[2] M. Aizerman, E. Braverman, and R. L. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[3] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152, 1992.

[4] M. Brookes. Voicebox is a matlab toolbox for speech processing. *http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, 2005.

[5] C. Chiu, Y. Chang, and Y. Lai. Analysis and recognition of human vocal emotions. *In Proceedings of the International Computer Symposium*, December 1994.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[7] M. Cowling. Non-speech environmental sound classification system for autonomous surveillance. *PhD Thesis*, Griffith University, Queensland, Australia, 2004.

[8] D. Ellis. Matlab audio processing examples. *http://www.ee.columbia.edu/∼dpwe/resources/matlab/*, 2005.

[9] S. Esmaili, S. Krishnan, and K. Raahemifar. Content based audio classification and retrieval using joint time-frequency analysis. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, 5(17-21):665–668, May 2004.

[10] B. Feiten and S. Gunzel. Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18(3):53–65, Summer 1994.

[11] J. Foote. Content-based retrieval of music and audio. *In Proceedings of the SPIE conference on Multimedia Storage and Archiving Systems II*, 3229:138–147, August 1997.

[12] G. G. and Z. Li. Content-based classification and retrieval by support vector machines. *In IEEE Transactions on Neural Networks*, 14:209–215, January 2003.

[13] M. Grimaldi, C. P., and A. Kokaram. A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques. *In Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 102–108, 2003.

[14] H. Kim, N. Moreau, and T. Sikora. Audio classification based on mpeg-7 spectral basis representations. *In IEEE Transactions on Circuits and Systems for Video Technology*, pages 716–725, 2004.

[15] T. Kohonen, editor. *Self-organizing maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1997.

[16] T. Kohonen. Learning vector quantization. pages 537–540, 1998.

[17] M. Lamming and M. Flynn. 'forget-me-not' intimate computing in support of human memory. *In Proceedings of FRIEND21 International Symposium on Next Generation Human Interface*, February 1994.

[18] M. Liu and C. Wan. Feature selection for automatic classification of musical instrument sounds. *In Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 247–248, 2001.

[19] Z. Liu, J. Huang, Y. Wang, and T. Chuan. Audio feature extraction and analysis for scene classification. *In IEEE Workshop on Multimedia Signal Processing*, pages 343–348, June 1997.

[20] N. Maddage, C. Xu, M. Kankanhalli, and X. Shao. Content-based music structure analysis with applications to music semantics understanding. *In Proceedings of the 12th annual ACM international conference on Multimedia*, pages 112–119, 2004.

[21] Y. Muthusamy, E. Barnard, and R. Cole. Reviewing automatic language recognition. *In IEEE Signal Processing Magazine*, pages 33–41, October 1994.

[22] E. Pampalk. A matlab toolbox to compute similarity from audio. *In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, 2004.

[23] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[24] D. Reynolds and R. Rose. Robust text-independent speaker identification using gaussianmixture speaker models. *In IEEE Transactions in Speech and Audio Processing*, 3:72–83, January 1995.

[25] E. Schreirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, pages 1331–1334, 1997.

[26] X. Sun. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, may 2002.

[27] D. Tjondronegoro, Y. Chen, and B. Pham. Applications ii: The power of play-break for automatic detection and browsing of self-consumable sport video highlights. *In Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 267–274, 2004.

[28] T. Tremain. The government standard linear predictive coding algorithm: Lpc-10. *In Speech Technology Magazine*, pages 40–49, April 1982.

[29] J. Tukey, B. Bogert, and M. Healy. The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe-cracking. *In Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed)*, pages 209–243, 1963.

[30] V. N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, November 1999.

[31] T. Wold, D. Blum, and J. Wheaton. Content-based classification, search, and retrieval of audio. *In Proceedings of the IEEE Multimedia*, 3(3):2736, 1996.

[32] M. Xu, L. Duan, L. Chia, and C. Xu. Audio keyword generation for sports video analysis. *In Proceedings of the 12th annual ACM international conference on Multimedia*, pages 758–759, 2004.

[33] R. Y., G. A., and A. Acero. Automatically extracting highlights for tv baseball programs. *in Proceedings of the ACM International Conference on Multimedia*, pages 105–115, 2000.

[34] T. Zhang and C. Kuo. Hierarchical classification of audio data for archiving and retrieving. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, 6:3001–3004, March 1999.