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 [7], [6]. 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 a 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 [13]. Music analysis deals with the identification of music genre, artist, instruments and structure [5].

The remainder of this paper is organized as follows: Section 2 addresses the methodology considered in our experiments. Results are discussed in Section 3. A survey of related work is performed in Section 4. Finally, in Section 5 conclusions and future work are presented.

2. 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

1-4244-0028-7/06/$20.00 ©2006 IEEE

result output options. This enables efficient and consistent tests of various features and classifiers.

2.1. Feature Extraction

The survey in this paper considers multiple state-of-the-art features applied in speech recognition, music analysis 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 Zero Crossing Rate (ZCR) and Short Time Energy (STE). The following spectral features are investigated: Linear Predictive Coding (LPC) coefficients, Relative Spectral Predictive Linear Coding (RASTA PLP) [4], Pitch [15], Sone [12], Spectral Flux (SF) and coefficients from basic time to 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 novel time-based features that describe characteristics 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).

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 [17]. The goal of LPC is to extract 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 minimize 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 3).

Cepstral Coefficients (CCs) are a popular feature in audio retrieval [10], [21]. The authors of [18] 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 [2]. 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 high and low amplitude. 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 segments with high energy from segments with low engery. 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 lengths of high energy segments in the signal. Figure 1(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 lengths of low energy segments in the signal. Details are depicted in Figure 1(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 high amplitude sequence. In other words the AHA feature represents the extent of high energy segments in the signal. Figure 1(c) illustrates this concept.

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 seven-dimensional feature vector which is used for classification.

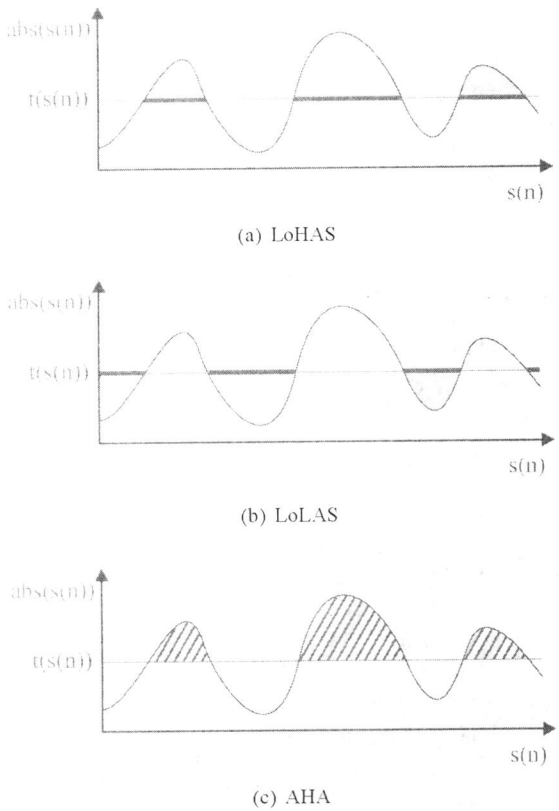

Figure 1. LoHAS, LoLAS and AHA for signal s(n) with threshold t(s(n)): (a) Length of High Amplitude Sequence (LoHAS). (b) Length of Low Amplitude Sequence (LoLAS). (c) Area of High Amplitude (AHA).

2.2. Classification

This section offers a brief discussion of the classification methods and the parameters used. We employ three supervised classifiers: SVM is a sophisticated kernel based machine learning technique introduced by Vapnik in [19]. The SVM is applied with a linear kernel and an RBF kernel. Furthermore, we apply the MATLAB implementation of Linear Vector Quantization (LVQ) by Kohonen [9]. LVQ is a classification method closely related to Self Organizing Maps (SOMs) [8]. The third classifier is Nearest Neighbor (NN) with Euclidean distance measure. NN is considered to indicate the quality of the features. Features that discriminate classes well, provide disjoint partitions of the feature space. Satisfactory results with the NN algorithm imply such a partitioning in the feature space.

3. 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 low-dimensional features, such as ZCR, SF, and Pitch is below that of high-dimensional features. Low-dimensional features usually 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 the Amplitude Descriptor (AD). 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 [2]. 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 [2], [4]. 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.

For each feature we compute recall and precision per class. Calculations of recall and precision depend on the number of retrieved documents for a given query. In our case recall and precision are computed for the complete test set. Table 1 shows mean recall and mean precision over all classes for selected features. More detailed results are presented in [22].

MFCCs and BFCCs perform nearly identically. This is due to 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 in 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 shown in

Table 1. 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.

	SVM		K-NN		LVQ	
	R	P	R	P	R	P
MFCC	0.79	0.81	0.81	0.83	0.77	0.77
BFCC	0.80	0.81	0.82	0.82	0.77	0.78
LPC	0.80	0.82	0.72	0.71	0.73	0.73
AD	0.79	0.79	0.75	0.75	0.71	0.74
Combin.	*0.90*	*0.91*	*0.85*	*0.86*	*0.83*	*0.84*

Table 1. Recall (R) and precision (P) values for selected features and classifiers. The last row summarizes the results of the combined feature.

In contrast to MFCC, BFCC, and LPC, AD is a time-based feature. Classification with the SVM and a linear kernel yields a recall and a precision of 0.79. This is comparable to the other features. 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. AD performs comparably to the more complex spectral features mentioned above.

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. The SVM is able to maintain higher precision and recall values than LVQ and NN for the selected features.

Up to now we concentrated on individual features. In order to improve retrieval quality, we combine several features into one 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. Table 1 lists recall and precision values of the combined feature vector for different classifiers.

4. Related Work

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. Pioneering work in environmental sound recognition is performed in [20]. 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.

A popular research field is audio recognition in broadcasted video. In [11] 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 neural networks. The authors of [16] 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 in [21]. The authors analyze keywords in commentator speech and audience which are relevant to important actions of the game. They apply a Hidden Markov Model trained with low level features (Energy and MFCCs including delta and double delta features) to recognize the keywords. Investigations presented in [14] 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 [3].

In [23] 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 [7] 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 similarly to MFCC. SVMs are applied successfully to environmental sound recognition in [6].

A challenging area of retrieval 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 & 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 outperforms NN and LVQ.

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

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

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

[5] 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.

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

[7] 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*, 14:716–725, 2004.

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

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

[10] 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.

[11] 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.

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

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

[14] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. *In Proceedings of the ACM International Conference on Multimedia*, pages 105–115, 2000.

[15] 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.

[16] 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.

[17] T. Tremain. The government standard linear predictive coding algorithm: Lpc-10. *In Speech Technology Magazine*, 1:40–49, April 1982.

[18] 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.

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

[20] 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.

[21] 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.

[22] M. Zeppelzauer. Discrimination and retrieval of animal sounds. *Technical Report TR-188-2-2005-06, http://www.ims.tuwien.ac.at/publication_master.php*, 2005.

[23] 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.