Identifying the population of animals through pitch, formant, short time energy-A sound analysis

N.Raju, S.Mathini, T.Lakshmi priya, P.Preethi,M.Chandrasekar
Department of Electronics & Communication Engineering
School of Electrical & Electronics Engineering
SASTRA University
Thanjavur, India.

***Abstract*—For a long time humans have employed animal sounds to recognize and find them. This paper proposes an automatic sound classification system based on the features extracted from the animal sounds. This paper explores the use of sound analysis algorithms to extract distinct features of different animals. Pitch, formant and short time energy are the features extracted. Support Vector Machine (SVM), a recent classifier is used for classification of animal sounds. It also deals with the comparative performance analysis of pitch by means of Auto Correlation Function (ACF) and Average Magnitude Difference Function (AMDF). More specifically the objective of this paper is to classify animal sounds and to identify the location of animals.**

***Keywords*—features, pitch ,formant, short time energy, ACF,AMDF,SVM**

I. INTRODUCTION

Animals- The word animals derived from the Latin word Animalia, which means having breath. Animals are the major group of multi-cellular organisms of animal kingdom Animalia, which are found all over the world. They are differed in sizes, shapes and colors even if they fit to the same animal family. Each kind of animal has their own unique behavior and lifestyle. Sound is a versatile form of communication.

Acoustic signals have distinctiveness which is particularly suitable for communication and practically all animal groups have some means of communication by sound[1,2]. Sound travels reasonably long distances in air and water and even through the obstacles between the source and the recipient. The vehicle for the provision of this communication is called a signal. The signal may be a posture, touch, electrical discharge, movement, discharge of an odorant or a combination of these mediums. Each species has a distinct set of signals in its series since an extensive variety of sound signals are feasible. There are multiple approaches exists for classifying animal sounds based on the extracted features.

Features are the distinct characteristics of an audio signal, which took major part in deciding the class of audio signals. Animal researchers typically use different features than speech researchers to analyze animal vocalizations. In this paper our focus is towards the extraction of Pitch, formant and short time energy.

The pitch determination is an important facet of many sound processing algorithms [1-15]. Animals may produce sounds of varying pitch and loudness to communicate different information. The frequency of a sound is recognized as pitch. With the rise in frequency, higher is the pitch and shorter is its wavelength. A large variety of methods for determination of fundamental frequency of pitch have been proposed. Auto correlation function (ACF) and Average magnitude difference function (AMDF) are the methods used to extract pitch. Excellent reviews of these methods are available. A spectral peak of sound spectrum is defined as formant. Formants are extracted via Auto Regressive (AR) method. Short time energy computes the energy of audio signal, which plays a vital role in distinguishing between sound and unsound samples.

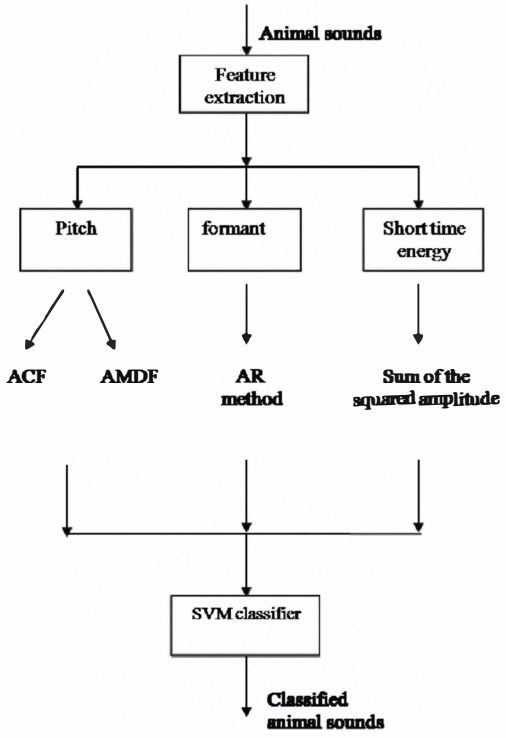

Fig 1: Flow diagram for animal's sound classification

The classification algorithm determines the animal identity based on the features extracted from the audio signal. Support Vector Machine (SVM)[5,12,13,14], a recent classifier is used for classification. In recent years certain species of animals are endangered, so there is a need to monitor the population of animals without letting them to extinction[3,7,9,10]. The above problem can be rectified by identifying the location of animals.

Automatic animal sound classification is very helpful for bioacoustics and audio retrieval applications[1,8].The automatic animals sound classification system used to monitor respiratory diseases in commercial animal houses [4].Recent researches illustrates that animals activity would be infrequent prior to the occurrence of natural disasters. These evidences demonstrate that animals have the capability to identify the awaiting natural disasters [5].The methodology to extract above features are explained in following sections and their results were tabulated.

II. FEATURE EXTRACTION

Feature extraction is the methodology of the conversion of an audio signal into a succession of feature vectors that bear the distinguishing information about the signal. They are also used as basis for numerous types of audio analysis algorithms [6,11]. It is typical that the features are computed based on the window basis. These window based features are considered as portrayal of the signal for that particular moment in the short duration of time.

A. pitch

The fundamental or first harmonic of any tone is perceived as its pitch. There are several methods which have been proposed to extract pitch information from an audio signal. In general, these methods are based on the average magnitude difference function (AMDF) and the autocorrelation function (ACF). They are calculated for each frame with fixed length.

1) ACF

Auto correlation function is a time domain method which reckons the similarity between a frame s (i) and its delayed version by the auto-correlation function:

$$acf\ (t) = \sum_{i=0}^{n-1-t} s(i)s(i+t) \qquad (1)$$

t -time lag in terms of sample points.

The maximum value of acf (t) over a specified time period is selected as the pitch.

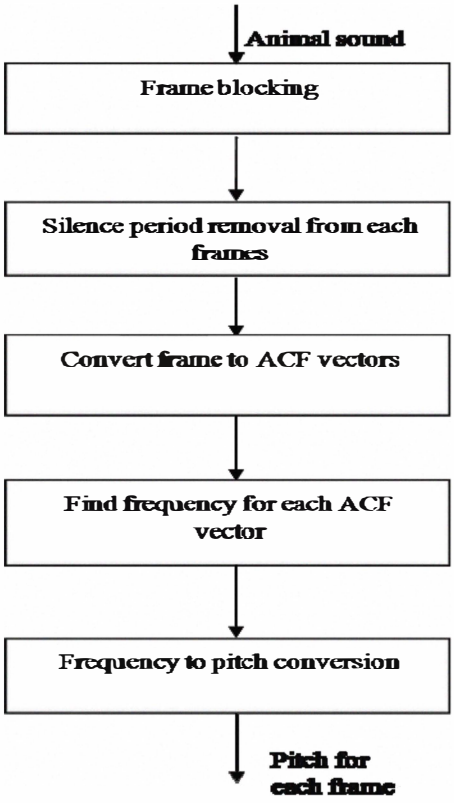

Fig 2 :Pitch detection using ACF method

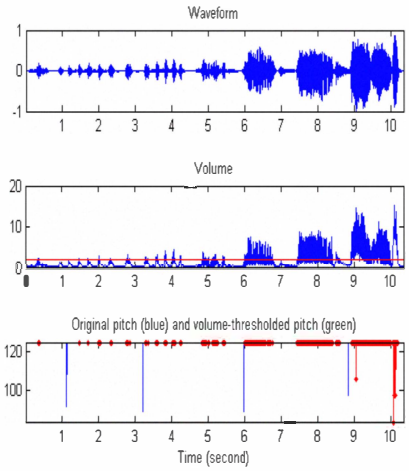

Fig 3: Pitch plot for tiger vocalization using ACF method

2) AMDF

The concept of AMDF (average magnitude difference function) is very close to ACF with the exception of distance estimation instead of similarity measurement.

The short-time AMDF is defined as:

$$amdf\ (t) = \sum_{i=0}^{n-1-t} s(i) - s(i+t) \qquad (2)$$

t - Time lag in terms of sample points.

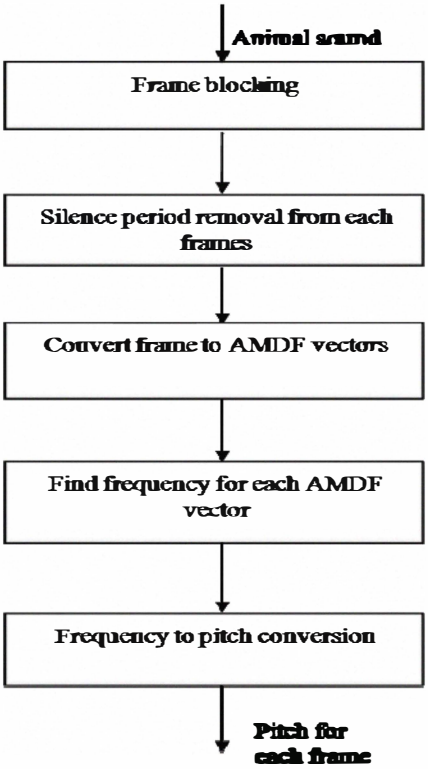

Fig 4: pitch Detection using AMDF method

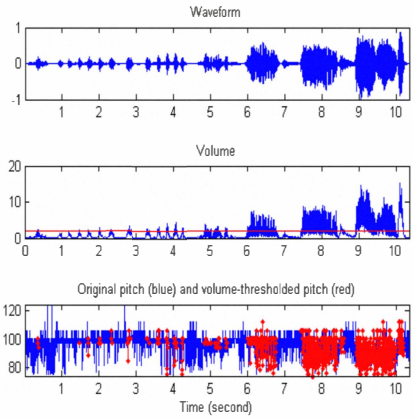

Fig 5: Pitch plot for tiger vocalization using AMDF method

FRAME BLOCKING:

The input audio signals are not stationary in nature, it is more strategic to fragment signals into frames. The audio signals tend to be stationary, when the frame duration is between 20- 35ms with the optional overlap of 1/3~1/2 of the frame size. In the process of frame blocking, overlap between neighboring frames was allowed to reduce discontinuity between them. Here the frame duration of 32 and an overlap of 10ms are taken for analysis. Usually the frame size is equal to power of two in order to facilitate the use of FFT.

SILENCE PERIOD REMOVAL:

In order to achieve better feature extraction there is a need to remove the silence/unvoiced periods from the audio signal. Removal of silence or unvoiced sounds from each frame can be carried by simple volume threshold (where a threshold equals one eighth the maximum volume) i.e., the pitch is set to zero if a frame is said to have a volume less than 1/8 of the maximum volume.

CONVERTING FRAME TO ACF:

This conversion provides ACF of a given frame, which is primarily needed for pitch tracking. ACF vectors can be obtained by performing dot product of frame and its shifted version. It is necessary that, maximum shift should be equal to frame size.

CONVERTING FRAME TO AMDF VECTORS:

This conversion provides AMDF of a given frame, which is primarily needed for pitch tracking. AMDF vectors can be obtained by performing summation of frame and its shifted version.

FREQUENCY TO PITCH CONVERSION:

After computing frequency (Hz) of each vector, it is converted to pitch (semitone), so that pitch information for each frame can be obtained.

Frequency to pitch conversion can be done by

$$\text{Pitch} = 69+12*\log2\ (\text{freq}/440) \qquad (3)$$

B. Formants detection via AR method

A formant is a component or a resonant frequency of the audio signal that does not change despite a change of the pitch. Formant frequencies is detected by means of Auto regressive model .The auto regressive model is a type of linear predictors, whose output depends upon the previous output.

The order of AR model was determined by using the formula

$$n = \text{round}\ (fs/1000) + 2 \qquad (4)$$

Where n is the order of the AR model, fess is the sampling frequency of the audio signal.

It is clear that the order of AR model depends upon the sampling frequency of the audio signal.

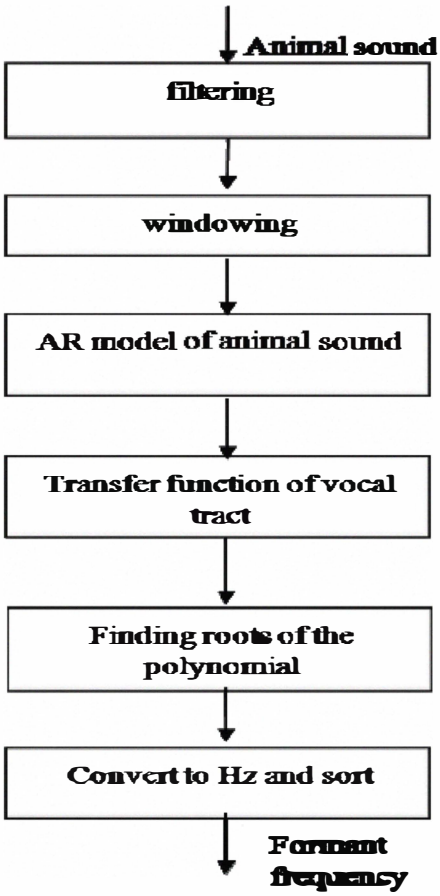

Fig 6: Formant detection using AR method

The input audio signal is pre -processed and further the parameters of AR model are computed. The roots of the resulted polynomial are computed and converted to Hz, which are sorted to provide the formant frequencies.

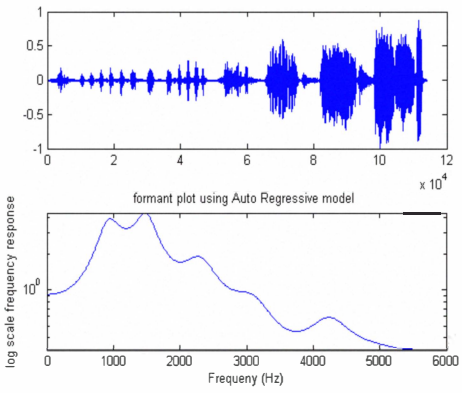

Fig 7: Formant plot for tiger vocalization using AR method

The above figure shows the formant plot of tiger vocalization using AR method.

C. Short time energy

Short time energy (STE) is a simple measure to distinguish voiced/silence signal. Short time energy can be defined as the sum of squared amplitude of samples in a frame. This measurement is used to distinguish between voiced and unvoiced sound segments. STE is based on the fact that energy in sound sample is greater when compared to the silent/unsound sample.

For a discrete-time signal x[n], the short-time energy measured at sample n is defined as

$$E_n = \sum_{m=n-N+1}^{n} (x[m])^2 \qquad (5)$$

i.e. sum of squared amplitude in an N -sample frame ending at n.

It can be used as the measurement to distinguish audible sounds from silence when the SNR is high. Its change in pattern over time may reveal the rhythm and periodicity properties of sound.

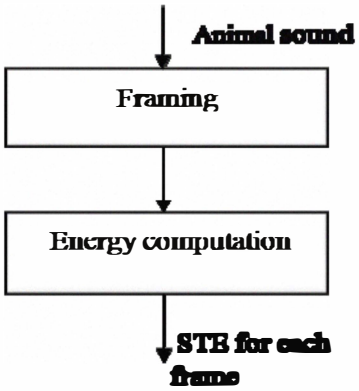

Fig 8: Method to compute short time energy

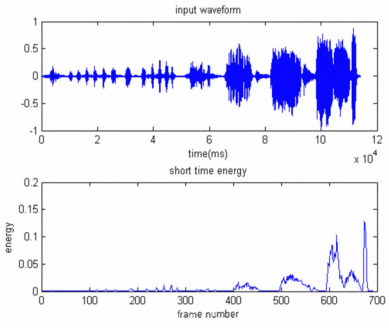

Fig 9: Short time energy plot for tiger vocalization

The above figure shows the energy plot for tiger vocalization. Thus the three distinct features were extracted from different animal samples. Pitch, formant and short time energy plots are shown for tiger vocalization.

III. CLASSIFICATION

The classification technique attempts to allot each input value to one of the given set of classes. The classification algorithm determines the animal identity based on the features extracted from the audio signal. Support Vector Machine (SVM), a recent classifier is used for classification. A support vector machine (SVM) is the supervised learning strategy that probes the data and recognizes patterns, which is valid for classification and regression analysis. The standard SVM considers a set of input data and envisages the possibility of two classes that comprises the input for each given input, making the SVM a non-probabilistic binary linear classifier. Feature vectors with the same number of components are needed to train SVM [3]. For SVM classification the LIBSVM software was employed. They are also known as maximum-margin classifiers.

Let the training vectors are: xi, i = {1, 2, 3 . . , l}, Y vector can be defined as

$$Yi = \begin{cases} 1 \text{ if } xi \text{ in class } 1 \\ -1 \text{ if } xi \text{ in class } 2 \end{cases} \quad (6)$$

Given two separable clouds of points (x1, y1),...,(xk , yk) where xi € Rn and yi € {-1, +1}. SVM constructs an finest separating hyper plane wx + b = 0, such that the the distance between the hyper plane and the nearest data point is maximum. The data point of each cloud are called the support vectors .The distance between the support vectors and the hyper plane is called margin [12,14].

Practically, most problems are not linearly separable. The data points are transformed into higher order space so that they become linearly separable. This is achieved by kernels [2].SVM classifier can be described as

$$f(x) = sign\left(\sum_{vec} \alpha y_i K(xi, xj) + b\right) \quad (7)$$

Where K (xi; xj) is the kernel used.

The three typical kernel functions can be given as

1. Polynomial:

K (xi,yj)=[(xi,yj)+1]^d

2.Radial Basis Function(RBF):

K (xi,yj)=exp(-(xi-yj)^2/2(y^2)

3.Sigmoid:

K (xi,xj)=tanh(scl(xi,xj)-off)

Where scl (scale) and off (offset) are parameters should be chosen with care, because the kernel becomes invalid for certain values.

The features are tested using the SVM classifier. The classifiers are trained by the training set which provides a model describing the data. The test data serve to validate the model. Discrimination of animal sounds is achieved using the SVM classifier [13].

IV. EXPERIMENTAL RESULT

This paper proposes methodology to extract features from audio signal. In this analysis various animal sounds are taken and their features were extracted using above methodology and their results were compared with praat tool. The features extracted from our tests can be used to classify animal sounds efficiently. The results were tabulated in Table I.

TABLE I. FEATURE EXTRACTION RESULT

Animal	Pitch (ACF)	Pitch (AMDF)	F1 (Hz)	F2 (Hz)	F3 (Hz)	Short time energy
Dog	120.12	97.74	953.7	1382.1	2461.3	0.0150
Cat	108.66	124.76	1139.3	1720.0	2642.1	
Chimp	119.37	98.62	1120.5	1740.3	2135.2	0.0213
Bear	123.27	96.63	800.5	1351.0	2126.2	0.0409
Bengal tiger	124.14	97.65	831.7	1485.6	2258.7	0.0164
Cattle	124.59	97.82	722.4	1329.8	1961.7	0.0019
Wolf	124.52	96.36	810.3	1098.1	1892.1	0.087
Crow	109.55	101.09	1293.8	1399.1	2091.7	0.0133
Parrot	110.80	107.66	1194.6	1772.8	2300.4	0.0122
Frog	98.170	119.83	1076.3	2171.8	2869.4	0.0025
Mountain lion	124.29	81.58	272.5	1328.1	2236.7	0.0015
Lion roar	124.63	88.59	467.6	1106.7	2193.6	0.0026
Tiger growl	109.55	101.09	1293.8	2407.4	2699.6	0.0133
Tiger	124.44	96.16	915.6	1497.4	2314.1	0.0078
Elephant angry	123.96	92.05	1003.2	1444.1	2083.1	0.0170
Gorilla	124.44	96.16	915.6	1497.4	2314.1	0.0078
Elephant	122.54	101.36	785.5	1315.5	2038.8	0.0416
Sheep	85.625	107.96	873.4	1877.1	2767.0	0.0049
Owl	124.70	90.506	528.4	901.6	1785.8	0.0073
Unknown animal	The above features will be computed					

The features extracted from our tests can be used to classify animal sounds efficiently. On classifying animal sounds based on the extracted features, population of certain animal species can be acquired.

V. CONCLUSION AND FUTURE WORK

It is evident that ACF and AMDF are the basic pitch tracking algorithms, which have been widely used. The performance of ACF and AMDF are summarized by analyzing different dog samples and their results were tabulated. It was observed from the tabulation that ACF outperforms AMDF.

TABLE II. ANALYSIS OF PITCH FOR DIFFERENT DOG SAMPLES

Different dog samples	Pitch (ACF)	Pitch (AMDF)
Doberman	124.46	89.64
Dog growl	124.73	91.31
Dog barking1	120.67	90.44
Dog barking2	120.71	87.66
Dog barking3	119.45	101.38
Dog barking4	119.51	102.69
Dog barking5	119.56	102.98
Dog barking6	120.14	97.74

Our future work is to predict the location of animals. The short time energy of the signal decreases with the increasing distance. By using this concept the location of animals can be computed.

ACKNOWLEDGEMENT

Authors wish to thank Professor R.Amirtharajan SAP/ECE School of Electrical & Electronics Engineering PG Project coordinator for his constant encouragement and his time and support.

VI. REFERENCES

[1] S.Gunasekaran and K.Revathy," Automatic Recognition and Retrieval of Wild Animal Vocalizations", International Journal of Computer Theory and engineering, Vol.3, No.1, February, 2011 1793-8201

[2] Dalibor Mitrovic, Matthias Zeppelzauer and Christian Breiteneder . Discrimination and Retrieval of Animal Sounds. IEEE conference on Multimedia Modelling ,Beijing, China pp 339-343, 2006.

[3] Marius Vasile Ghiurcau, Corneliu Rusu . About Classifying Sounds in Protected Environments.3rd International symposium on Electrical And Electronics Engineering 2010,pp84-87.

[4] Vasileios Exadaktylos1, Mitchell Silva1, Sara Ferrari, Marcella Guarino2 and Daniel Berckmans . Sound localization in practice: An application in localization of sick animals in Piggeries. Advances in sound localization,576-588

[5] W. Astuti, A.M. Aibinu, M. J. E Salami, R. Akmelawati, and Asan G.A Muthalif.. Animal Sound Activity Detection Using Multi- Class SupportVector Machines. 4th International Conference on Mechatronics ,May ,2011.

[6] S.Gunasekaran, K.Revathy . Content-based Classification and Retrieval of Wild Animal Sounds using Feature Selection Algorithm . Second International Conference on Machine Learning and Computing. 2010

[7] Felix Weninger and Bj¨orn Schuller . Audio Recognition In The Wild: Static And Dynamic Classification On A Real-World Database Of Animal Vocalizations . 36th International Conference on Acoustics, Speech and Signal Processing . May, 2011

[8] 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.

[9] M. V. Ghiurcau, C. Rusu, and R. Bilcu, "Wildlife intruder detection using sounds captured by acoustic sensors," Proc. Of IEEE ICASSP 2010, Dallas, USA, 2010, pp. 297-300

[10] M. V. Ghiurcau, C. Rusu, and R. Bilcu, "A modified TESPAR algorithm for wildlife sound classification," Proc. of IEEE ISCAS 2010, Paris, France, 2010, pp. 2370-2373

[11] B. Han and E. Hwang, "Environmental sound classification based on feature collaboration," in Proc. IEEE ICME 2009, New York, USA, Jun 2009, pp. 542–545.

[12] V. Vapnick, The Nature of Statistical Learning Theory. Springer-Verlag, New-York, 1995

[13] 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

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

[15] www.animalsounds.org- animals' database