International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-11 September-2013

Identification & Detection System for Animals from their Vocalization

A. D. Mane[1], Rashmi R. A.[2], S. L. Tade[3]

Abstract

Until now little research has been done in the area of animal sound identification. Animal sound classification and retrieval is very helpful for bioacoustics and audio retrieval applications. This paper is a literature review of an animal identification and detection system based on animal voice pattern recognition. The system uses the Zero-Cross-Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) joint algorithms as the tools for recognizing the voice of the particular animal. ZCR is used for end point detection of input voice such that the silence voice can be removed. MFCCs are used as it is the most widely used features for speech recognition due to their ability to represent the speech spectrum in a compact form. DWT is used for voice pattern classification by getting the optimal path between input voice and referenced voice in database. The results are encouraging and motivate further research in this domain, particularly detecting the state of the animal like normal, hunger, sleep, heat etc.

Keywords

Zero-Cross-Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW), spectrum

1. Introduction

Recently, audio data gained importance in the field of content-based retrieval system. The rising number of audio and video databases states the need for efficient identification and detection. A very few investigation has been done till date on animal sounds which is a part of environmental sounds. [1]. In the daily life, we can hear a variety of creature's sounds, including human speech, dog barks, bird songs, cicada sounds, frog calls, and cricket calls, etc. Many animals generate sounds either for communication or as a by-product of their living activities such as eating, moving, or flying, etc [2]. Some of the practical issues researchers' faces include the difficulty of acquiring high quality acoustic data in adverse environments, imperfect labelling of data and inadequate knowledge about how animals produce and perceive sound. Recognition system is a very useful system that helps the users to recognize human, object, and animal. By having the recognition system, the security of some areas can be improved [3], [4]. The aim for this animal voice recognition system is to develop a system that can help the human to recognize the particular animals in order to know which animal are calling. Different animals have different vocal frequencies and hence the accuracy of detecting a particular animal is higher. This means that the developed animal voice recognition system for detecting a certain species of animal is very useful especially for the security purpose in zoos, national parks and sanctuaries. This system is also useful when it is applied in the hospital veterinary [5].

A. Background and Relevance

Identification of animals by their sounds is valuable for biological research and environmental monitoring applications, especially in detecting and locating animals. Furthermore, most of the animal vocalizations have evolved to be species-specific. Therefore, the utilization of animal vocalizations to automatically identify animal species is a natural and adequate way to ecological censuring, environment monitoring, biodiversity assessment, etc [2]. Some research obstacles when dealing with animal vocalizations are noisy data and label validity. The incorporation of noise models is important when dealing with animal vocalizations since the recording environment is usually poor with many interfering noise sources present. This noise can greatly decrease classification accuracy, especially if the characteristics of the noise vary across the dataset. Detecting state of animal is a challenging job since researchers can only guess as to what the animal is trying to communicate acoustically [6].

Section 2 covers the difference between human speech and animal calls. Section 3 covers the proposed methodology. The proposed method is divided into 3 modules. Under the first module, ZCR (Zero Crossing Rate) algorithm is used for end point detection. Second module consists of extracting the feature of the speech signal in the form of MFCC coefficients and in the third another module the nonlinear sequence alignment known as Dynamic Time Warping (DTW) introduced by Sakoe Chiba

has been used as features matching techniques. Experimental results and discussions are shown in section 4 while section 5 includes conclusion and future scope.

2. Difference between Human Speech and Animal Calls

The difference between the human speech and animal calls is that, in the human speech, each word means subject, concept, or action. But apart exact meanings, a person pronouncing a word adds his emotions and mood in it. If the word was pronounced with thin voice, we realize that it was a child, and if with bass, it was an adult man. In contrast to human words, although the animal calls do not represent the precise meanings, they are not senseless. This is because, similarly with human speech, the animal calls bear information about animal's mood and intentions. It is becoming clear from such call features as pitch (high or low), loudness, repetition rate, and many others. Besides that, most of the animal species possess by rich vocal repertoires: for example, they can select among growling, barking, howling, whining, whimpering, squealing, hooting, and sometimes have especially exotic sounds, such as echolocation clicks [7]. Both humans and most nonhuman mammals produce sounds using a couple of vocal folds, located in larynx and the vocal folds can vibrate with frequency of a few hundreds or thousands times per second. This frequency assumed the name of fundamental frequency and is measured in Hertz (Hz), where 1 Hz = 1 cycle per second [7].

3. Proposed Methodology

A. End Point Detection

The initial step applied to input voice signal is to detect the starting and ending point of the vocal signal. This process is required to remove the silence part of sound so that the processing is only done on the main part of the sound. For this system, the zero-crossing rate (ZCR) algorithm will be used for detecting the end points. The ZCR is known as the number of times the sound sequence change its sign per frame and it is given as

$$Z(n) = \frac{1}{2}\sum_{m=1}^{N} \left|\text{sgn}[x(m+1)] - \text{sgn}[x(m)]\right|$$

where:

$$\text{sgn}[x(m)] = \begin{cases} +1 & x(m \geq 0) \\ -1 & x(m < 0) \end{cases}$$

This ZCR method is used in order to count the frequent of the signal that crosses over the zero axes. It is a very useful method for detecting the occurrence of silence sound [8].

B. Feature Extraction

The speech signal can be represented by a sequence of feature vectors which can help to apply mathematical tools without losing generality. The selection of peculiar features along with methods to extract them is known as feature selection and feature extraction [9].

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Few of the audio features that have been satisfactorily used for audio classification comprises Mel-frequency Cepstral Coefficients (MFCC), Local Discriminant Bases (LDB) and Linear Predictive Coding (LPC). Few techniques generate a pattern from the features and use it for classification by the degree of correlation. Some other techniques use the numerical values of the features parallel to statistical classification method. MFCC is used here.

The human peripheral auditory system forms the basis of MFCC. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each interval of sound with an actual frequency *f* measured in Hz, a intuitive pitch is measured on a scale called the 'Mel Scale' .The mel frequency scale occupies a linear frequency below 1000 Hz and logarithmic above 1kHz. The pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold is defined as 1000 Mels as a reference pint. A compact representation would be provided by a set of Mel-frequency Cepstral Coefficients (MFCCs), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. The MFCCs are proved to be more efficient. The MFCC is calculated by the following steps as shown in figure 1.

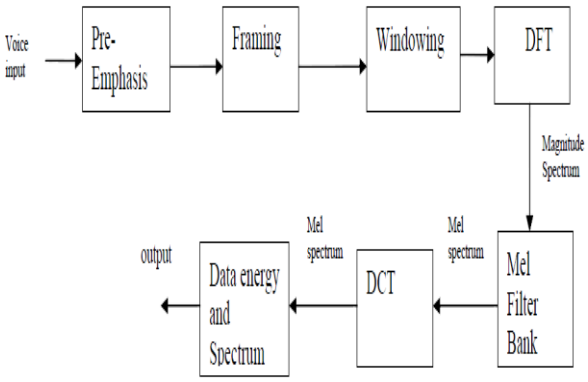

Fig.1:MFCC Block diagram [10]

Pre-emphasis
Pre-emphasis refers to a system process designed to increase, within a band of frequencies, the magnitude of some (usually higher) frequencies with respect to the magnitude of the others(usually lower) frequencies in order to improve the overall SNR. Hence, in this step the signal passed through a filter is processed which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

Framing
The process of segmenting the speech samples into a small frame with the length within the range of 20 to 30 msec is called as framing. The voice signal is divided into frames of say N samples. Adjacent frames are being separated by M where M<N. Typical values used are M = 100 and N= 256.
Speech is non-stationary but if we analyze it in frames it is stationary. Frame size depends on 2 factors viz. Average Speaking rate (frame size do not change within an utterance) and Data contained within the frame (frame size changes within an utterance). Short duration frames (wide-band analysis) are used to achieve high time resolution whereas, large frame sizes (narrow band) for high frequency resolution. Frames can be overlapped, and normally the overlapping region ranges from 0 to 75% of the frame size.

Windowing
It is used to window each of the individual frame such as to minimize the signal discontinuities at the beginning and end of each frame (e.g.: Hamming Window). Window analysis is shifted by typically 10ms.

Fast Fourier Transform (FFT)
To convert each frame of N samples from time domain into frequency domain, FFT is being used. The Fourier Transform is used to convert the convolution of the glottal pulse x(t) and the vocal tract impulse response h(t) in the time domain into the frequency domain i.e. X(w) and H(w) respectively.

Mel-filter bank Processing
The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the Centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components.

Discrete Cosine Transform (DCT)
This is the process to convert the log Mel spectrum into time domain using DCT. The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

Delta energy and delta spectrum
The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. Each of the 13 delta features represents the change between frames corresponding to cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

MFCC's
In the final process, the log Mel spectrum is then converted back to time domain and the result is called the Mel- frequency Cepstral Coefficients (MFCC).

C. Pattern Classification
After the feature extraction process, the next process is matching two signals in order to undergo the verification process. Here, DTW (Dynamic Time Warping) technique is used. DTW algorithm is based on Dynamic Programming. This algorithm is used for measuring similarity between two time series which may vary in time or speed. This technique is also used to find the optimal alignment between two time series if one time series may be "warped" non-

linearly by stretching or shrinking it along its time axis [10].

In order to apply the DTW, two parameters that extracted from the voice signal are considered, where one of them is the input of test signal and the other one is the reference signal. For example, the input of test signal is x[t] = [1 1 2 3 2 0], and the reference signal is y[t] = [0 1 1 2 3 2 1]. These two signals are placed into a table in order to calculate the difference between them as shown in figure 2.

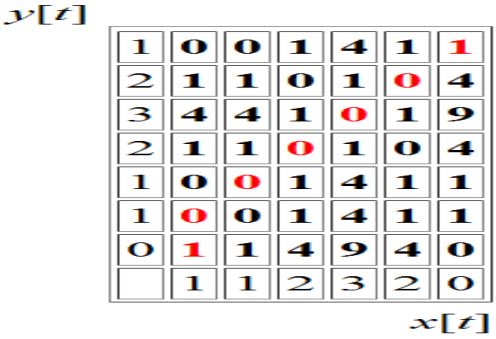

Fig.2: Matrix representing difference between x[t] and y[t] [5]

The value inside each of the cells is calculated by following expression,

$$(x[t] - y[t])^2$$

After calculate the difference between these two signals, the best or optimal path that move from the cell (1,1) to the cell (6,7) in this case is calculated and its result is shown in Table 1.

Table 1: The best path from cell to cell [5]

7	(0) B:7	(0) B: 5	(1) B: 3	(4) BL: 4	(1) BL: 2½	(1) BL: 1½
6	(1) B:7	(1) B: 5	(0) B: 2	(1) B: 2	(0) BL: 1	(4) BL: 4
5	(4) B: 6	(4) BL: 4	(1) B: 2 BL: 2	(0) BL: 1	(1) L: 2	(9) BL: 6½
4	(1) B: 2	(1) BL: 1½	(0) BL: 1	(1) L: 2 BL: 2	(0) L: 2	(4) L: 6
3	(0) B: 1	(0) B: 1, L: 1 BL: 1	(1) BL: 1½	(4) BL: 4	(1) L: 5	(1) L: 6
2	(0) B: 1	(0) L: 1, BL: 1	(1) L: 2	(4) L: 6	(1) L:7	(1) L: 8
1	(1) Start	(1) L: 2	(4) L: 6	(9) L: 15	(4) L: 19	(0) L: 19
	1	2	3	4	5	6

From the distance value shown in Figure 2, the optimal path from cell (1,1) to cell (6,7) can be calculated by taking the spent minimum cost when move along all the possible path. In order to find the best path from one cell to another cell, there are three ways to reach the destination; from left, bottom and bottom left. The used symbols to represent the spent cost that move from left, bottom and bottom left are L, B, and BL respectively.

When calculating the B and L, the possible way to the next cell is adding the previous cheapest cost with the reached destination cost, where L is rightward and B is upward. For example, in order to move from cell (1,1) to cell (2,1), the needed cost is adding the previous cheapest cost, which is 1 for this case, with the reached destination cost, which is 1 also. Therefore, the cost needed to move from cell (1,1) to cell (2,1) is 2.

While for the calculation of BL, the possible way to next cell is adding the previous cheapest cost with the half of the reached destination cost. For example, in order to move from cell (2,2) to cell (3,3), the needed cost is adding the previous cheapest cost, which is 1 for this case, with the half of the reached destination cost, which is 0.5. Therefore, the cost needed to move from cell (2,2) to cell (3,3) is 1.5. By using the DTW, the calculated path through the matrix represents the details about the similarity between the samples of the test signal and the samples in the reference. Besides that, it also gives the information about how much the two signals differ in the best alignment. Hence, the process of verification can be executed with more accuracy.

4. Results & Discussions

At the beginning of the system processing, a sound that produced by a dog is recorded and it acts as the input sound of the system. Figure 3a shows the original waveform of the sound that is recorded. In order to analyze the sound, the silence sound must firstly be removed and it is done by using the ZCR method as mention in the methodology section.

Figure 3b shows the silence sound of the dog which is detected and marked using the red line. After that, the silence sound is removed leaving the main part of the dog sound, which is shown in Figure 3c.

The main part of the sound shown in the Figure 3d is then analyzed using the MFCC in order to get a more precise data. The output of the MFCC is shown in the

Figure 3d and its data will be kept with different Mat files.

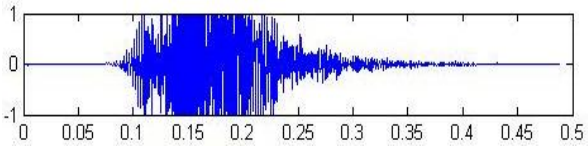

Fig 3a: Original waveform of the recorded sound

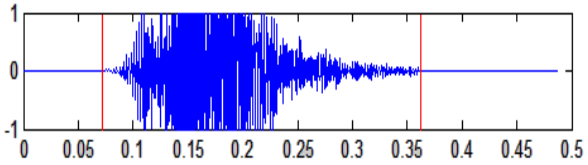

Fig 3b: End point detection

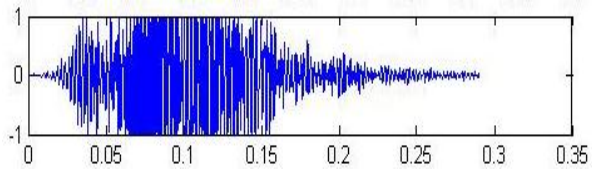

Fig 3c: Filtered silence part

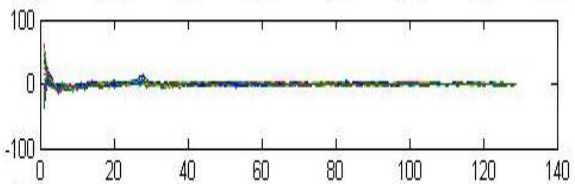

Fig 3d: Calculated MFCC

A new Mat file, which represent ts each of the sounds, will be generated in the database each time a new sound is saved. When an input is analyzed, its MFCC data will then be compared with all of the MFCC data inside the database.

Finally, the distance between the input sound and the reference sounds inside the database is calculated by using the DTW method. The final result of such experiment is shown in Figure 4.

```
The distance between dog1.wav and dog2.wav is: 0.140258
The distance between dog2.wav and dog2.wav is: 0.000000
The distance between dog3.wav and dog2.wav is: 0.563745
The distance between dog0.wav and dog2.wav is: 0.295997
The recognition result is : dog2.wav
```

Fig 4: Final result

From the given result, it is shown that the distance between the input sound and the sound dog3.wav has a maximum value; which is 0.563745, while the distance between the input sound and the sound dog2.wav has a minimum number; which is 0.00, after the comparison has been made. Therefore, the given input sound is recognized as the dog2.wav sound.

5. Conclusion & Future Scope

In this paper, Identification and Detection of animals from their vocalizations has been developed. Firstly, the ZCR algorithm used for end point detection can efficiently remove the silence part. Secondly, MFCC algorithm is used for extracting the phonetically important characteristics of voice which is more compact and lastly DTW is used for comparing and classifying the accurate results.

These algorithms have been worked out for same speech signals as well as for different speech signals and it have been found that if both speech signals are same the cost will be 0 and if speech signal are of different voices then cost will definitely have some value which shows the mismatching of the signals.
So far, researches have been done for detecting the emotions of human beings [11], [12]. Animal calls also resemble their moods, intensions and the different states of their being such as normal, hunger, or heat state but they cannot convey or communicate through semantics. Therefore, it is a challenging job since researchers can only guess as to what the animal is trying to communicate acoustically.

References

[1] Mitrovic D., Zeppelzauer M., and Breiteneder C "Discrimination and retrieval of animal sounds", Multi-Media Modelling Conference Proceedings, 12th International, 2006.
[2] Chang-Hsing Lee, Yeuan-Kuen Lee, Ren-Zhuang Huang, "Automatic Recognition of Bird Songs Using Cepstral Coefficients", Deptartment of Computer Science and Information Engineering Chung Hua University, Taiwan, Journal of Information Technology and Applications Vol. 1 No. 1, May, 2006, pp.17-23.
[3] R. B. Chen and S. J. Zhang, "Video-based face recognition technology for automotive security," Mechanic Automation and Control Engineering (MACE), International Conference, pp. 2947 – 2950, 2010.
[4] Liying Lang and Hong Yue, "The Application of Face Recognition in Network Security,"

Computational Intelligence and Security. CIS'08, International Conference, vol. 2, pp. 395 – 398, 2008.

[5] Chi Yong Yeu, "Animal voice recognition for identification (ID) detection system", Dept of Computer & Communication Engineering, 2011 IEEE 7th International Colloquium on Signal Processing and its Applications.

[6] Deshmukh O., Rajput N., Singh Y., Lathwal S., "Vocalization patterns of dairy animals to detect animal state", 21st International Conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan.

[7] Moscow Zoo, "Gallery of animals' sounds," http://www.moscowzoo.ru/get.asp?id=C130, Access on October 30, 2010.

[8] Seman. Noraini, Bakar Zainab Abu, and Bakar Nordin Abu, "An Evaluation of Endpoint Detection Measures for Malay Speech Recognition of an Isolated Words", Information Technology (ITSim), International Symposium, vol. 3, pp. 1628 - 1635, 2010.

[9] Vibha Tiwari, "MFCC and its applications in speaker recognition", Deptt. of Electronics Engg., Gyan Ganga Institute of Technology and Management, Bhopal, (MP) INDIA, International Journal on Emerging Technologies 1(1): 19-22(2010).

[10] Anjali Bala, Abhijeet Kumar, Nidhika Birla, "Voice Command Recognition System Based on MFCC and DTW", International Journal of Engineering Science and TechnologyVol. 2 (12), 2010, 7335-7342.

[11] Dellaert, F., Polzin, Th., and Waibel, A., 1996 Recognizing emotions in speech. ICSLP 96.

[12] Hansen, L. and Salomon, P., 1990 Neural Network Ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence. 12: 993-1001.

[13] Ashwini D. Mane received her B.E. degree from MIT AOE, University of Pune in 2012. She is currently pursuing M.E in VLSI and Embedded Systems from P.C.C.O.E, Pune University.

Ashwini D. Mane received her B.E. degree from MIT AOE, University of Pune in 2012. She is currently pursuing M.E in VLSI and Embedded Systems from P.C.C.O.E, Pune University.

Rashmi R. A. received her B.E degree from Pune University in 2012. She is currently pursuing ME in VLSI and Embedded systems from P.C.C.O.E Pune University.

Prof. Sunil Tade is an assistant professor of electronics and telecommunication engineering at Pimpri Chinchwad College of Engineering, University of Pune, India. He received his B.E. degree from Amravati University and M.E. degree from University of Pune. His research interests include signal & image processing.