Animal Voice Recognition for Identification (ID) Detection System

Che Yong Yeo, S.A.R. Al-Haddad, Chee Kyun Ng
Department of Computer & Communication Systems Engineering,
Faculty of Engineering, Universiti Putra Malaysia, UPM Serdang,
43400 Selangor Darul Ehsan, Malaysia.
Email: ycy_1222@hotmail.com, {sar, mpnck}@eng.upm.edu.my

Abstract — **Voice recognition systems have become the important applications for speech recognition technology. In this paper, an animal identification (ID) detection system based on animal voice pattern recognition algorithm has been developed. The developed animal voice recognition 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 the end point detection of input voice such that the silence voice can be removed. MFCC is used for the process of feature extraction where a more compact and less redundant of the representative voice can be obtained from the input voice. While the voice pattern classification will be done by using DTW algorithm. The DTW voice pattern classification module is playing a very important role as it is used to get the optimal path between the input voice and the reference voice in the database. The obtained results show that the developed recognition system can be worked as expected.**

Keywords – Animal; Voice Recognition; ZCR; MFCC; DTW

I. Introduction

Animals represent a major group of the mostly multicellular and eukaryotic organisms of the area of Animalia or Metazoa. At the same time, as they develop, their body is already planned eventually becomes fixed, although some of them had undergone a process of metamorphosis later in their life. Most of the animals are dynamic, which means they can move with spontaneously and also independently. Besides, all the animals also are heterotroph, which means they must ingest the other organisms in order to get the nutrition [1].

Scientists had estimated that there may be 30 million species of animals on the earth and can be group into 6 basic groups, which are invertebrates, fishes, amphibians, reptiles, mammals, and birds. Among these groups, mammals are the vertebrates that evolved from the reptiles and there are about 5400 species of mammals alive today. The mammals display remarkable arrays of adaptions and these adaptions enable them to inhabit in a wide range of habitats. Some of the examples of the mammal animals are lion, tiger, cat, bear, dog, and so on [2].

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]. For example, the face recognition system can be used to replace the traditional key based system for the cars security. By using this face recognition system, the car owners can lets their face recognized by the car system in order to open the door. Therefore, the car owner can still open the car although they forget to bring or lose the key [3]. Nowadays, there are many types of recognition system, such as face recognition system [5], voice recognition system [6], iris recognition system [7], fingerprint recognition system [8], and so on.

Besides using the recognition system on the human and objects, the recognition system also can apply on the animal by using the same concept that applied to the human. There are many animal recognition systems had developed nowadays in order to apply in certain areas, such as security area, research area, and so on. Among them include the pattern recognition [9], vocal recognition, vision recognition [10], and animal fiber recognition system [11].

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 are having different vocal frequencies and hence the accurately of detecting the ID of a particular animal is higher. This means that the developed animal voice recognition system for detecting the ID of a certain species of animal is very useful especially for the security purpose in zoo. This system is also useful when it is applied in the hospital veterinary.

II. Literature Reviews

A. Dog's Vocalization

The common vocal communications of the dogs are barks, howls, growls, and whines. For the barks, the dogs' barking is mainly use to defend a territory and to demarcate their boundaries. There are various types of bark which two of them are the bark that direct to the human intruder and the bark that direct to the canine intruder. These two kinds of bark are different and unfortunately the dogs are more likely to bark in a response to the other dogs' bark rather than to the sound of a human intruder [12].

For the whining calls, it is a care-soliciting call of the dogs. This call is firstly used by the puppies to communicate with their mother, who provides the warmth and the nourishment. However, this kind of call is used by the mature dogs when

978-1-61284-413-8/11/$26.00 ©2011 IEEE

they want to relief from pain or are in the mildly frustrating situation. For example is when they want to escape outdoors. While for the howling calls, it is used by the canine and has not been deciphered well. The last type of call of the dog is the growling. This is an aggressive or distance-increasing call in dogs [12].

B. *Difference Between Animal Calls and Human Speech*

Recently, the audio data is gained importance in the field of content-based retrieval. The rising huge number of audio and video database states the needs for the efficient retrieval. The animal sounds are a domain of environmental sounds that has not been investigated yet in details [13].

The difference between the human speech and animal calls is in the human speech, each word means subject, concept, or action. But apart precise meanings, a person pronouncing a word puts in it his emotions and mood. 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 [14].

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 [14]

III. Methodology

A. *End Point Detection*

When there is an input voice signal, the initial step is to detect the beginning and ending point of the vocal signal. The reason of this process is to remove the silence sound such that the processing is only focusing on the main part of the sound. For this system, the end point detection algorithm will be using the zero-crossing rate, ZCR based algorithm.

The ZCR is known as the number of times the sound sequence change its sign per frame and it is given as

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

where:

$$\operatorname{sgn}[x(m)] = \begin{cases} +1 & x(m \geq 0) \\ -1 & x(m < 0) \end{cases} \qquad (2)$$

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 [15].

B. *Feature Extraction*

For the feature extraction section, the used algorithm is calculating the Mel-Frequency Cepstral Coefficients (MFCC). The aim of this feature extraction process is to obtain a new voice representation which is more compact, less redundant, and more suitable for statistical modeling. The MFCC is based on the known variation of the human ear's critical bandwidths with frequency, where it filters the space linearly at low frequencies and logarithmically at high frequencies. It is used in order to capture the phonetically important characteristics of the voice. There are several steps in order to implement the MFCC as shown in the Figure 1 [16].

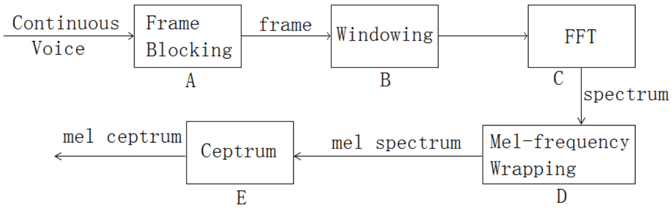


Figure 1. Computation of Mel Frequency Cepstral Coefficients (MFCC).

Process A: Frame Blocking

The framing process is firstly applied to the voice signal of the producer. This signal is partitioned or blocked into N segments (frames).

Process B: Windowing

The second process of the processing is to window each of the individual frame such to minimize the signal discontinuities at the beginning and end of each frame.

Process C: Fast Fourier Transform

The next process is the Fast Fourier Transform (FFT) where each frame of N samples is converted from time domain to frequency domain.

Process D: Mel-Frequency Wrapping

The obtained spectrum from the FFT process is then Mel Frequency Wrapped. The major aim of this process is to convert the frequency spectrum to the Mel spectrum.

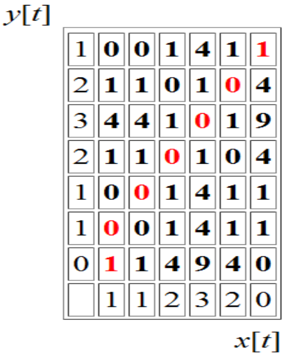
Process E: Cepstrum

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

C. *Pattern Classification*

After the feature extraction process, the next process is matching two signals in order to undergo the verification process. However, the voice input signal may be vary in term of speed or time when compare with the reference voice signal. Therefore, these two signals must be aligned in order to get the optimal match between them. This process is called as the Dynamic Time Warping (DTW) algorithm.

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.



$y[t]$

1	0	0	1	4	1	**1**
2	1	1	0	1	**0**	4
3	4	4	1	**0**	1	9
2	1	1	**0**	1	0	4
1	0	**0**	1	4	1	1
1	**0**	0	1	4	1	1
0	**1**	1	4	9	4	0
	1	1	2	3	2	0

$x[t]$

Figure 2. Difference between x[t] and y[t].

The value inside each of the cells is calculated as

$$(x[t]-y[t])^2 \tag{3}$$

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 I. THE BEST PATH FROM CELL TO CELL.

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.

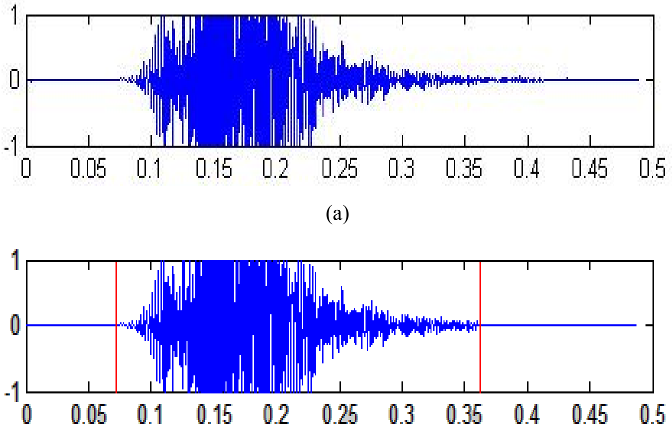
By using the DTW, the calculated path through the matrix represents the details about the similarity between the samples of the test signal and the samples in the reference. Besides that, it also gives the information about how much the two signals differ in the best alignment. Hence, the process of verification can be executed with more accurate.

IV. RESULTS AND DISCUSSION

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

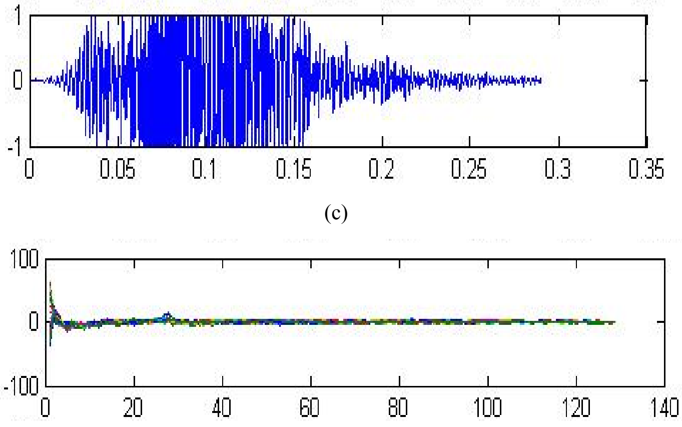
Figure 3b shows the silence sound of the dog is detected and marked using the red line. After that, this silence sound is removed and left the main part of the dog sound, which is shown in Figure 3c.

The main part of the sound that shown in the Figure 3d is then analyzed using the MFCC in order to get a more accurate data. The output of the MFCC is shown in the Figure 3d and its data will be kept with different Mat files.



(a)

(b)



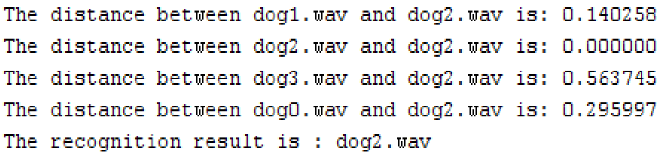
(c)

(d)

Figure 3. Dog bark sound, (a) original waveform, (b) end point detection, (c) filtered silence sound, (d) calculated MFCC.

A new Mat file, which represents 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
```

Figure 4. Final result of the developed recognition system.

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.

V. CONCLUSION

In this paper, an animal identification (ID) detection system based on animal voice pattern recognition algorithm has been developed. Based on the experimental results, the used ZCR algorithm for the end point detection can accurately detect the silence voice of the input voice before is removed. Besides, a more compact and less redundant of the representative voice is successfully extracted from the main part of the input voice by using the MFCC algorithm. Finally, the experiment results show the DTW algorithm is work as expected in order to calculate the distance between the input voice and the voice inside the database in order to get the optimal path between them and used for the pattern recognition. As the conclusion, the overall experiment results show that the developed system can be worked as expected.

REFERENCES

[1] Wikipedia, “Animal,” http://en.wikipedia.org/wiki/Animal, Access on October 30, 2010

[2] About.com, “Animal Groups - The 6 Basic Animal Groups,” http://animals.about.com/od/zoologybasics/tp/sixbasicanimalgroups.htm, Access on October 30, 2010

[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] Ming Zhao and Tat-Seng Chua, “Markovian mixture face recognition with discriminative face alignment,” Automatic Face & Gesture Recognition. FG '08. 8th IEEE International Conference, pp. 1 – 6, 2008

[6] R.A. Rashid, N.H. Mahalin, M.A. Sarijari, and A.A. Abdul Aziz, “Security system using biometric technology: Design and implementation of Voice Recognition System (VRS),” Computer and Communication Engineering. ICCCE 2008. International Conference, pp. 898 – 902, 2008

[7] Chia-Te Chou, Sheng-Wen Shih, Wen-Shiung Chen, V.W. Cheng, and Duan-Yu Chen, “Non-Orthogonal View Iris Recognition System,” Circuits and Systems for Video Technology, IEEE Transactions, vol. 20, pp. 417 – 430, 2010

[8] Kaisheng Zhang, Jiao She, Mingxing Gao, and Wenbo Ma, “Study on the Embedded Fingerprint Image Recognition System,” Information Science and Management Engineering (ISME), International Conference, vol. 2, pp. 169 – 172, 2010

[9] Shou-Jue Wang, and Xu Chen, “Biomimetic (topological) pattern recognition - a new model of pattern recognition theory and its application,” Neural Networks. Proceedings of the International Joint Conference, vol. 3, pp. 2258 – 2262, 2003

[10] Liming Zhang, and Jianfeng Mei, “Theoretical confirmation of simple cell's receptive field of animal's visual systems and efficient navigation applications,” Neural Networks. Proceedings of the International Joint Conference, vol. 4, pp. 3179 – 3184, 2003

[11] L. X. Kong, F. H. She, S. Nahavandi and A. Z. Kouzani, “Fuzzy pattern recognition and classification of animal fibers,” IFSA World Congress and 20th NAFIPS International Conference, vol. 2, pp. 1050 – 1055, 2001

[12] A. Houpt Katherine, “Domestic Animal Behavior for Veterinarians and Animal Scientists,” John Wiley and Sons, 5th edition, 2010

[13] Mitrovic D., Zeppelzauer M., and Breiteneder C., “Discrimination and retrieval of animal sounds,” Multi-Media Modelling Conference Proceedings, 12th International, 2006

[14] Moscow Zoo, “Gallery of animals’ sounds,” http://www.moscowzoo.ru/get.asp?id=C130, Access on October 30, 2010

[15] 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

[16] Soo Yee Cheang, and A.M. Ahmad, “Malay language text-independent speaker verification using NN-MLP classifier with MFCC,” Electronic Design, ICED 2008, International Conference, pp. 1 – 5, 2008