APPLICATION OF SPEECH RECOGNITION FOR SWIFTLET VOCALIZATIONS

Siti Nurzalikha Zaini Husni Zaini[1], Sunardi[2], Kamarul Hawari Ghazali[3], Saiful Nizam Tajuddin[4]

Faculty of Electrical and Electronic Engineering, Universiti Malaysia Pahang, MALAYSIA
[1]*snzhz.ct89@gmail.com,*
[2]*sunardi@ump.edu.my,*
[3]*kamarul@ump.edu.my*

Faculty of Science and Technology, Universiti Malaysia Pahang, MALAYSIA
[4]*saifulnizam@ump.edu.my*

Abstract

This research is about speech recognition technique are used for swiftlet vocalization application. Swiftlet vocalization need a system for recognize because there are many types of swiftlet sounds use in industry only can inspection by human expert. This research use speech recognition by using Mel Frequency Cepstral Coefficient (MFCC) for feature extraction and Distance Time Warping (DTW) for classification to calculate accuracy and efficiency combination both techniques.

Keywords: swiftlet, sound, attraction, MFCC

1. Introduction

Health care is very important to human life. Because of that, we might need to choose health nutrition such as bird's nest from swiftlets. The nests of some species are built entirely from threads of their saliva, and are collected for the famous Chinese delicacy bird's nest soup.

Swiftlets's nest make skin whiting agent and also good for eye's health. For asthma sufferer, it also became best agent restore respiratory system and strengthens lungs. The bird's nest benefiting all age level such as collagen nutrient which include in every swiftlets's nest can launch blood vessel increase appetite and improve alimentary canal.

The nests can give high potential and also benefiting for health although the value added. Within more this a decade, entrepreneurs explored various methods and new technology to increase production. There are a few factors to make swiftlets attract such as aroma, light, temperature, humidity and sound.

The swiftlets character is sensitive toward sound. Previously, sound that produced at swiftlets husbandry premise actually is produced from recording audio sound bird voice. Therefore, the research and development about sound characteristic for swiftlets attraction needed to develop swiftlets industry. This is used for industry to attract swiftlets enter and build their nests in man-made house. The income can give benefits for good economic and healthy.

Nowadays, bird house for swiftlets farming usually developed and equipped with recorded sound of chirping and mating from cave (natural habitat) to attract swiftlets to enter and build nest. These sounds just taken using trial and error method without analysis of sound involve in signal to attract the swiftlets. This method is sometimes successful to attract the swiftlets, but certainly these sounds contains noisy and disturb by another sound.

The objective of this research is to identify the sound of characteristic for swiftlets attraction. There are 10 samples of sound have placed at external location in swiftlets's house to be analyzed for frequency and magnitude.

2. Swiftlet Attraction

2.1 Sound Attraction

Swiftlets are birds contained within the four genus Aerodramus, Hydrochous, Schoutedenapus and Collocalia. They form the Collocaliini tribe within the swift family Apodidae. This group contains around 30 species which is mostly confined to southern Asia, south Pacific islands, and northeastern Australia. All of them are within the tropical and subtropical regions. They are in many respects typical members of the Apodidae, having narrow wings for fast flight, with a wide gap and small reduced beak surrounded by bristles for catching insects in flight.

A small-sized swift (*Family Apodidae*) have 24 species worldwide. The main producers of edible nest are White-nest Swiftlets (*Aerodramus fuciphagus*) and Black-nest Swiftlets (*A. maximus*). Two unique characters are salivary gland to build nest and Echolocation [1]. The distinguishes are many but not all species from other swifts and indeed almost all other bird is their ability to use a simple but effective form of echolocation to navigate in total darkness through the chasms and shafts of the caves where they roost at night and breed. The nests of some species are built entirely from threads of their saliva, and are collected for the famous Chinese delicacy bird's nest soup.

There are environmental factors such as temperature, light intensity, humidity and sound is the key of successful place for swiftlets [2]. Sound is the main attraction for swiftlets for place in their house. The most interesting feature of swiftlets is that many species utilize a sonar-like system [3]. The swiftlets's voice proven very effective attracts swiftlets to be nested in bird house for swiftlets farming. This is shown that swiftlets very sensitive on sound.

The previous research state that swiftlets hearing responses to the frequency 1 - 16 kHz [2] and which most energy on 2 - 5 kHz [4]. This frequency falls into normal hearing. This statement is shown that in general, the animals generate sounds to communicate with members of the same species [5]. In 1990, technique for swiftlets attraction by using recording began to be expended but recording quality that is adverse. Through technology development, swiftlets recording voice that produced with quality, clear and similar authentic swiftlets voice. This swiftlets's recording voice usable to increase swiftlets population to build nest. There are two locations to attract the swiftlets entered the swiftlets farming house which are puller and external.

The locations at puller swiflets's house is on the roof house. Mostly, this location will fit the swiftlets voice when they gather. The location at external swiflets's house is on the outer house. Mostly, this location will fit the adult swiftlets voice. External sound is focused in this research.

2.2 Mel-Frequency Ceptral Coefficient (MFCC) Feature Extraction

The Mel-Frequency Cepstral Coefficients (MFCC) is frequently used nowadays for feature extraction technique in speech processing. In this technique, the used of Mel scale in the derivation of cepstrum coefficients was introduced [6].As mentioned earlier, the main objectives of feature extraction is to extract the important characteristics from the speech signal, that are unique for each word, due to differentiate between a wide set of distinct words. MFCC is considered as the standard method for feature extraction in speech recognition and perhaps, the most popular feature extraction technique used nowadays [7]. MFCC able to obtain a better accuracy with a minor computational complexity, respect to alternative processing as compared to other feature extraction techniques [8].

The proposed method for feature extraction is given in figure 1 in methodology part. At this stage, it will emphasize on MFCC computational process, as the main algorithm for feature extraction analysis. Here, the feature extraction algorithm of MFCC has been used and applied to all collected of

speech samples to obtain the targeted output of features vector. There are certain parameters need to define first before the MFCC algorithm and coefficients value were estimated.

3. Methodology

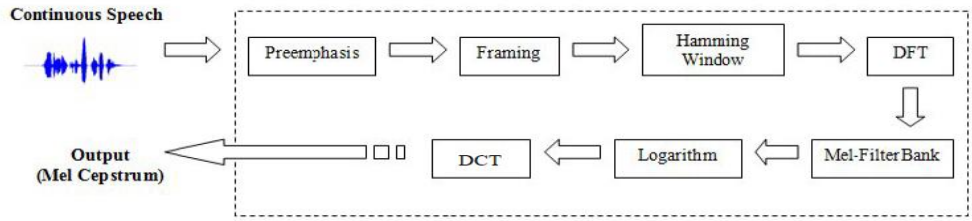

Figure 1: MFCC

3.1 Pre-emphasis

Pre-emphasis is considered as the first step of MFCC under the pre-processing stage in speech processing, which involved the signal conversion from analogue to digital signal. The size of the sample for digital signal is determined by the sampling frequency and the length of the speech signal in seconds. At the first stage in MFCC feature extraction, the amount of energy is used to boost into the high frequencies. It can be seen through the spectrum graph for speech segments like vowels, where there is more energy at the lower frequencies compared to the higher frequencies. This drop of energy across frequencies is caused by the nature of the glottal pulse [9]. When the frequency increases, pre-emphasis also increased the energy of the signal.

3.2 Framing

After pre-emphasis filtering process executed, the filtered of input speech will be framed. Here, the columns of data from the particular speech input will be determined. The Fourier Transform used here, only reliable when the signal is in a stationary position. In this case, speech or voice implementation holds within a short time interval only, less than 100 milliseconds of frame rate. Thus, the speech signal will be decomposed into a series of short segments and each of the frames will be analyzed, then any useful features will be extracted from it.

3.3 Hamming Window

Windowing is one of the important parts in MFCC feature extraction process. Here, each individual frame of speech signal is windowed, due to minimize the signal discontinuities at the beginning and at the end of each frame. The purpose of this action is to minimize the spectral distortion and to taper the signal to zero at the beginning and at the end of each frame.

3.4 Discrete Fourier Transform (DFT)

The Discrete Fourier Transform (DFT) normally computed via Fast Fourier Transform (FFT) algorithm. This algorithm is widely used for evaluating the frequency spectrum of the speech and converts each frame of N samples from the time domain into the frequency domain [10]. In this research, the windowed of speech segment is transformed into the frequency domain by using the Fourier Transform through the MATLAB.

3.5 Mel Filter Bank

Mel scale is applied due to place more emphasize on the low frequency components. It is because, the information carried by low frequency components of the speech signal is more important than the high frequency components. Mel scale is a unit of special measure or scale of perceived pitch of tone. Mel filter bank also known as Mel Frequency Warping, where it does not correspond linearly to the normal frequency, but behaves linearly below 1000Hz and a logarithmic spacing above 1000Hz.

In order to implement the filter banks, the magnitude coefficients of each Fourier transform of speech segment is binned by correlating them with triangular filter in the filter bank. In other hand, Mel-scaling is performed by using a number of triangular filters or filter banks [11].

3.6 Discrete Cosine Transform (DCT)

DCT is a Fourier transform, which is similar to the Discrete Fourier Transform (DFT), but using the real numbers only. DCT used to extract the Mel Frequency Cepstral Coefficients (MFCC) results, and it is often used to calculate the cepstrum instead of inverse FFT. In this research, this part was the final step of computing the MFCCs. It required computing the logarithm of the magnitude spectrum, in order to obtain the Mel-Frequency Cepstral Coefficients. The MFCCs at this stage are ready to be form in a vector format known as features vector. This features vector is then considered as an input for the next process, which is concerned with training the features vector for recognition purposes.

4. Experimental result

The samples of recording sound in 'mp3' format are converted into 'wav' format using mp3 to wav converter. There are 10 samples of sound have been measured. The measurement procedure is a sample taken to be cut at first, end and random audio in the middle by using easy audio cutter. Each sound deduction produced in period 120seconds.

Figure 2 shows that acquired MATLAB consist of continuous waveform and converted into its spectral. This waveform performed in time domain are contains certain peaks and valleys in the particular range. Based on the sounds profile is in 2 minute (120 seconds) has 20 MB memory and further. Thus, the sound can analysis maximum in 120 seconds. The waveform have irregular pattern with different amplitude and time it make difficult to pick the best sound.

Waveform performed two types of color which is blue and green after processing. The blue color representing the original sound while the green color representing the noise sounds.

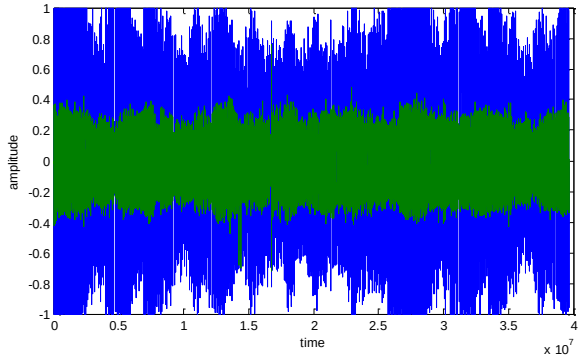

Figure 2: Swiflet sound

After apply step in MFCC, we get the last result as Figure 3. The figure shows the coefficient versus frame in the sound. The feature extraction is compute using DCT. For classification, we apply Distance Time Warping (DTW).

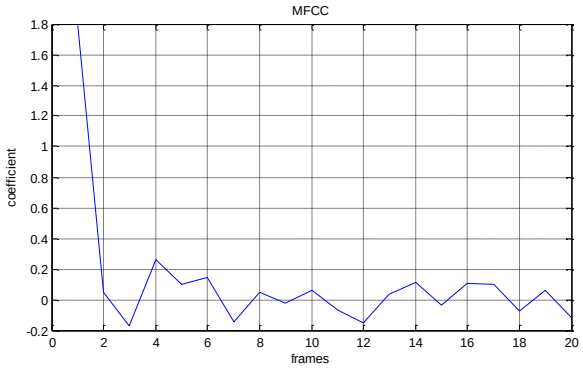

Figure 3: MFCC

5. Conclusion

For future work, this method can apply to other animal vocalization. Otherwise, the swiftlet vocalization can be extracted using other feature and classification to get better result. From the experiment, the feature extraction from swiftlet vocalization by using MFCC compares to standard and continues for classification after training the data.

References

[1] L.C. Koon. Opportunity and Sustainability of Swiftlet Farming in Malaysia, 2011.

[2] B.G. Coles, M. Konishi, J.D. Pettigrew, Hearing and Echolation in the Australian Grey Swiftlet, Collocalia Spodiopygia. Great Britain. The Company of Biologists Limited.1987.

[3] H. Thomassen. Swift as sound Design and evolution of the echolocation system in Swiftlets (Apodidae: Collocaliini). Print Partners Ipskamp B.V., Ensched, 2005.

[4] J.H. Fullard, R.M.R. Barclay, D.W. Thomas. Echolocation in free-flying Atiu Swiftlets (Aerodramus Sawtelli). Canada. Biotropica, 1993: 25: 334-339.

[5] C.H. Lee, C.H. Chou, C.C. Han, R.Z. Huang. Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. Pattern Recognation Letters. 31 August 2005.

[6] Levent, M.A., 1996, "Foreign Accent Classification in American English." Dissertation for Doctor of Philosophy in Department of Electrical & Computer Engineering, Graduate School of Duke University, Durham, USA.

[7] Ursin, M., 2002, 'Triphone Clustering in Finnish Continuous Speech Recognition', Master Thesis, Department of Computer Science, Helsinki University of Technology, Finland.

[8] Davis, S.B. & Mermelstein, P., 1980, 'Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences', IEEE

[9] Jurafsky, D. & Martin, J.H., 2007, Automatic Speech Recognition: Speech and Language Processing: An Introduction to natural language processing, computational linguistics, and speech recognition, Prentice Hall, New Jersey, USA.

[10] Owen, F.J., 1993, 'Signal Processing of Speech'. Macmillan Press Ltd., London, UK.

[11] Thomas, F.Q., 2002, 'Discrete Time Speech Signal Processing', Prentice Hall, New Jersey,USA.