MFCC – Based Speech Recognition using Artificial Neural Networks

Engr. Ranil M. Montaril, MSECE[2], Dela Cruz, Tonichi Paul M.[1], Castañeda, Joshua Robert B.[1], Gecarane, Jay Leonarth F.[1], Lim, Kenno John S.[1], Marzan, Eadrian N.[1], Militar, Jessica Justine B.[1], Sembrano, Jessie James V.[1]

[1]Bachelor of Computer and Information Sciences, College of Computer and Information Sciences, Polytechnic University of the Philippines, Sta. Mesa, Manila, Philippines
[2]Faculty Member, Department of Computer, Science, Polytechnic University of the Philippines

Abstract – Speech recognition is an important role in emerging technology especially in devices which uses speech. In this project, the researchers applied MFCC which gives 26 coefficients and fed in Artificial Neural Networks. The main focus of the study is to recognize isolated words which have been spoken by different speakers. In testing the data, the tools got 80.95% accuracy rate which got 12 hit and 3 miss from 15 samples.

Key Terms – ANN, MFCC, periodogram, signal framing, mel – filter bank, speech recognition

1. Background of the Study

1.1 Introduction

With the advancement of automated system, speech recognition is an important and emerging technology with great potential. It has significant role in operating devices which uses speech. It can be used in several applications like household appliances, mobile phones, security devices and computers [1]. In addition, it eases the communication barrier in helping computer understand human language. The communication has extent in providing access for anybody who has handicap and hearing loss. There are people who cannot use keyboard or cannot hear sound videos they have watched [2]. In accordance, speech recognition could potentially make their lives easier [3].

The complexity and recognition problem in processing speech is increasing. It found that problem is being more complex when processing on randomly analog signals like speech signals [1]. Different researches have various methods that being proposed for efficient extraction of speech and developing model in recognizing speech.

In this paper, the proponents developed a tool that will recognize speech of each member using Artificial Neural Networks. Each member recorded a voice or speech signal using Matlab. Then, the raw data exported in a sheet. It is used to generate 26 MFCC and pre – train it in ANN. The ANN embedded in Excel and applied speech processing and recognition. The outputs are waveform of the speech signal with corresponding recognized words. In this works, it is limited to recognized 7 words/names spoken by 7 speakers.

1.2 Problem Statement

The study aims to develop a tool that will recognize speech of each member using Artificial Neural Networks. Specifically, it aims to answer the question:

1. What is the accuracy rate of the tool in recognizing words?

1.3 Applicable Related Studies

In a research of Kamble [4], it provides a comprehensive study of Artificial Neural Network (ANN) in speech recognition. It focuses on the different neural network related methods that can be used for speech recognition including its advantages and disadvantages. The researchers experiment different neural network such as feed – forward network, recurrent neural network, modular neural network and Kohonen self – organizing maps. In the results, ANN proves that it can be very useful in speech signal classification. It concluded that RNN have achieved better speech recognition rates than MLP. But the training algorithm of RNN is more complex and dynamically sensitive which can cause problems.

The study of M. A. Raba Bah, A. Al – Marghilan and M. A. Eyad proposed the application of wavelet transform for reduction of the value of artificial neural networks for speech recognition tasks. In their study, artificial neural networks as a model for speech recognition are notable, resistant and effective. In training the neural network, genetic algorithm is being implemented. It suggests focusing in advantages of wavelet transforms target values in order to reduce likelihood of lower learning in local minimum. It includes of convergence acceleration, improvement of accuracy, reduction in number of iterations and decreasing the probability of falling into a local minimum [5].

Base on the study of Ki-Seung Lee [6], it is well known that a strong relationship exists between human voices and the movement of articulatory facial muscles. In their paper, the researchers utilize this knowledge to implement an automatic speech recognition scheme which uses solely surface electromyogram (EMG) signals. The sequence of EMG signals for each word is modelled by a hidden Markov model (HMM) framework. The proposed model reflects the dependencies between each of the EMG signals, which are described by introducing a global control variable. EMG signals were acquired from three articulatory facial muscles. The findings indicate that such a system may have the capacity to recognize speech signals with an accuracy of up to 87.07%, which is superior to the independent probabilistic model.

According to Dr. Philip Jackson, human neural networks are consists of neurons in the brain in inter – connectivity. It has different types of neurons with network topologies that can be applied automatic speech recognition through its discriminative features. In using ANNs for speech recognition, it innate connectivity and learnt patterns and difficult to interpret hidden weights. But it seems have successful applications like isolated words (ISR) and phoneme classification. On the other hand, it encountered problems like segmentation needed for training and poor modelling of time – scale variation [7].

2. Software Design Architecture

2.1 Functional Block Diagram

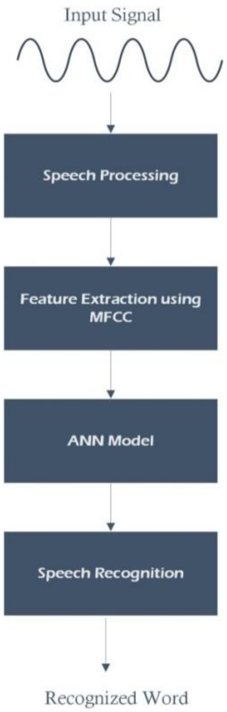

Figure 1 Block Diagram for Speech Recognition

An audio signal is the input that obviously constantly changing on even short time scales. In signal processing, framing the signal into 20 – 40ms frames is necessary. The next step is to calculate the power spectrum of each frame. The periodogram estimate performs of identifying which frequencies are present in the frame.

In extracting features, the filtering gives indication of energy exists near 0 Hertz. The Mel scale tells exactly in making space and making wide our filter banks. It relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Then, the system will take the logarithm of filter bank energies and compute its discrete cosine transform (DCT). It decorrelates the energies and used to model with artificial neural network. The ANN finds pattern match utterance and makes feature matching and recognizing speech. Then, it display speech waveform signal and corresponding recognized word.

2.2 System Architecture

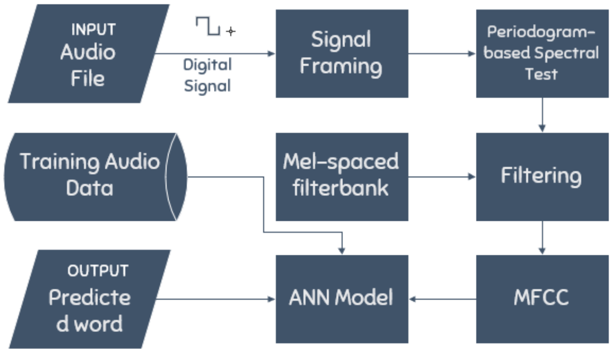

Figure 1 System Architecture

The audio file is the input form for the tool. It is in *m4a* file type with less than 2 seconds speech recorded by each member. It only consists of one word audio recorded. Then, the signal was framed into short frames. For each frame, the calculation of periodogram used to estimate of power spectrum. In filtering, the Mel filter bank applied to the power spectra and sum the energy in each filter. By taking the logarithm and its DCT of all filter bank energies, the system can get the MFCC computation to extract the features. The artificial neural network will get the features pattern and modeled to recognize words based on the training data. The training data is consists of audio file recorded from each member. The output will be predicted word and also display a speech waveform signal.

3. Simulation/Test Results

Table 1 Results of accuracy

Words	Sample numbers	Hit	Miss	Accuracy
eadrian	2	2	0	100%
jaja	2	1	1	50%
jay	2	1	1	50%
jessie	2	2	0	100%
jocas	2	2	0	100%
kenno	2	2	0	100%
tonichi	3	2	1	66.66667%
Total	**15**	**12**	**3**	**80.95238%**

In the results, the tool hit 12 correct recognized and miss 3 correct recognized. The tool also displays the recognized word and the corresponding signal waveform. The word “eadrian”, “jessie”, “jocas” and “kenno” got the highest accuracy rate with 100 percent. Second, the word “tonichi” got 66.67 percent and the word “jay” and “jaja” got the low accuracy rate with 50 percent. In overall, the tool got 80.95 percent accuracy rate.

4. Conclusion

In this project, the recognizing of words yielded good results. In the tests performed, we used names as words to recognize. It includes of “eadrian”, “jaja”, “jay”, “jocas”, “jessie”, “kenno” and “tonichi”. The word “eadrian”, “jessie”, “jocas” and “kenno” got the best performance among all other isolated words. While other words like “tonichi” got nearly similar performance with words “jaja” and “jay” which got poor performance. To the words which got low performance, there are some recorded audio files that have zero value or did not cover the two seconds range of audio file. There are also advantage in updating of ANN model particularly with the number of hidden layer and output layer which affects the training and recognition performance of the model.

5. Recommendations

In this project, the researchers focused only in 7 words and limited for isolation words. In this case, the accuracy only depends only for one – word recognition rather than in word recognition at sentence level. In speech recognition, accuracy and speed are common measurements. The future researchers can add speed accuracy aside from word recognition accuracy. Also, other language model can be used to train the features such as Hidden Markov Model.

References

[1] I. Patel and S. Rao, "Speech Recognition using HMM with MFCC - An Analysis using Frequency Spectral Decomposition Technique," *Signal & Image Processing : An International Journal(SIPIJ),* vol. 1, no. 2, 2010.

[2] B. C. Condino, J. A. A. Estinopo, E. L. Praico, J. L. G. Salac and R. M. Montaril, "GENOSUBS: An Automatic Subtitle Generator in Filipino Using Hidden Markov Model," 2015.

[3] S. Deshmukh and M. Bachute, "Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization," *International Journal of Engineering and Innovative Technology (IJEIT),* vol. 3, no. 1, 2013.

[4] B. C. Kamble, "Speech Recognition Using Artificial Neural Network - A Review," *Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE),* vol. 3, no. 1, 2016.

[5] M. Al - Raba Bah, A. Al - Marghilan and M. E. Eyad, "Artificial Intelligence Technique for

Speech Recognition based on Neural Networks," *Oriental Journal of Computer Science & Technology,* vol. 7, no. 3, pp. 331 - 336, 2014.

[6] K.-S. Lee, "EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables," *IEEE Transactions on Biomedical Engineering,* vol. 55, no. 3, 2008.

[7] D. P. Jackson, "Artificial Neural Networks in Speech Recognition," University of Surrey.

[8] S. Y. M.Q Wang, "Speech recognition using hidden Markov model decomposition and a general background speech model," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 5, 1992.

Curriculum Vitae

Joshua Robert B. Castañeda is a Bachelor of Science in Computer Science student at the Polytechnic University of the Philippines who has a vast knowledge on programming languages such as C, Java and Python. He is well-equipped with necessary knowledge and skills in the field of computing and sees his self of developing his own algorithm someday.

Tonichi Paul Dela Cruz lives in New Manila, Quezon City, Philippines. Currently taking a bachelor degree in Science in Computer Science at Polytechnic University of the Philippines. He has a high interest in computers and how everything works on it. This includes computational intelligence and internet of things.

Jay Leonarth F. Gecarane is a fourth year student, currently taking BS in Computer Science in Polytechnic University of the Phlippines. He has research interests in mobile, natural language processing and image processing. He is knowledgeable about web and mobile development, Java, Python, C and C#.

Kenno John Lim is taking a Bachelor of Science in Computer Science at Polytechnic University of the Philippines. Educated using the platform of HTML5, C, C#, MATLAB and has little knowledge of Python. Personal interests are music, tekken and movies. Hobbies are modeling and collecting stuffs like Trolls.

Eadrian Marzan lives in Caloocan City, Philippines. Currently taking a bachelor degree in Science in Computer Science at Polytechnic University of the Philippines. He is focusing in web developing as front end programmer and interests on robotics and AI's.

Jessica Justine B. Militar is a Computer Science student at Polytechnic University from the year 2014 up to the present. She enjoys experimenting different algorithms and is driven by challenge. She has knowledge in programming languages such as Java, C#, C.

Jessie James Sembrano is taking up Bachelor of Science in Computer Science at Polytechnic University of the Philippines. He has knowledge on some programming language such as C, C++, C#, Java, PHP and Matlab. Interested in playing chess and physical sports like basketball, volleyball and badminton. Also plays guitar and loves music a lot.