Kadazandusun Speech Recognition: A Case Study

Mohd Shamrie Sainin[#], Mohd Hanafie Haris
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, MALAYSIA.
#Corresponding author. E-Mail: shamrie@ums.edu.my.

ABSTRACT Currently, there is no existing system that provides common information and utilities for Kadazandusun's speech recognition since Kadazandusun speech has different features that are not available in other languages. This paper presents a preliminary experiment using one of the famous feature extraction methods which is Linear Prediction Cepstral Coefficients (LPCC). Further investigation on the speech data is using several classifier algorithms to investigate the recognition rate of Kadazandusun words. There are 6 words of Kadazandusun collected as an individual speech to test the feature extraction and the classifiers. The objectives of this study are to investigate LPCC feature extraction and to propose a suitable classifier algorithm for Kadazandusun speech data.

KEYWORDS: Kadazandusun; Feature Extraction; Speech Recognition; Speech; LPCC
Received 25 March 2021 **Revised** 22 April 2021 **Accepted** 25 May 2021 **Online** 2 December 2021

Original Article

INTRODUCTION

Kadazandusun is one of the largest ethnicities in Sabah, Malaysia. Kadazandusun is one of Malaysia's most populous ethnic groups. The Kadazandusun is used to describe the union of the two ethnic groups in Sabah, Kadazan and Dusun, where Kadazandusun community is the widest in Malaysia's Sabah state. There are 13 distinct different dialects of the Kadazandusun language, with over 300,000 people which speak the language residing in the districts of Tambunan, Ranau, Penampang, Tuaran, Kota Belud, Papar, and parts of Kota Kinabalu. Kadazandusun community who still speaks can also be found in the districts of Labuk-Sugut, Kinabatangan, Beaufort, Keningau, Tenom as well as in the districts Tawau (Sullivan & Albert, 1988). Since 2004, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) has recognized Kadazandusun as a Borneo indigenous community with various documented heritage.

This is proof that this language is needed to be preserved for this language to be maintained for alternative learning in the future. Kadazandusun language is taught in several primary and secondary schools, teacher institutes, and universities such as Universiti Malaysia Sabah (UMS) and Universiti Pendidikan Sultan Idris (UPSI). This language is offered to UMS students in three levels of difficulty which Kadazandusun level 1, Kadazandusun Level 2, and Kadazandusun Level 3. Students can register for this course for every level, one level for one semester.

Speech recognition is a significant tool in human-computer interaction, especially in this information age, and becoming more important with the support of modern technologies. It mainly consists of two modules which are the engine module and the feature extraction module (Juan *et al.*, 2012). Speech Recognition is a variety of fields, including psychology, acoustics, signal treatment, pattern recognition, and linguistics. The complexity of understanding speech recognition comes from several aspects of those environments (Sloane & Silva, 2020). Kadazandusun's speech recognition has not been fully investigated specifically on feature extraction. The purpose of the research is to maintain the endangered language to provide alternative language learning in the future. While there have been many languages that have been implemented in the speech recognition system, Kadazandusun is not one of them since there is a little bit of research about Kadazandusun using speech recognition tools.

Many speech recognitions have been researched in recent years especially with new recognition tools such as Artificial Neural Network, Hidden Markov Model (HMM), and Support Vector Machine (SVM). However, there is no research has been done on Kadazandusun speech. Kadazandusun speech has different features that are not available in other languages. Thus, this research is to confront the important subject of Kadazandusun speech recognition. Mainly this work is to investigate the capability of the recognition to perform Kadazandusun speech-language.

BACKGROUND THEORY

In recent years, the accuracy of speech recognition (SR) has been one of the most active areas of research. The goal of voice recognition is to improve human-computer interaction by using a certain medium of an interface, to tackle challenging problems like speech-to-text or speech-to-speech translations, and even to create applications that can understand spoken language, as good as a human can perform (Këpuska & Elharati, 2015). Speech Recognition is the process of automatically recognizing spoken words of speakers based on information in a speech signal. With the use of a microphone, an automatic speech recognition application detects the words spoken by a person. The automatic speech recognizer is then identifying these words, and the system will display the identified words on the screen. Automatic Speech Recognition (ASR), often known as speech recognition, is a method for converting an audio signal (mainly utterance of speech) into a series of words or other language representation using computer software with a special recognition algorithm.

The primary goal of the speech and voice recognizer is to be able to capture, comprehend, and respond to what is being said by the human (Jain & Rastogi, 2019). Automatic speech recognition analyzes and processes speech signals before transforming them into a sequence of individual words with the help of intelligent computer software. It is, in simple terms, the process is to convert spoken speech audio into correct text. The speech recognition process can generally be divided into many different components which are summarized in Figure 1.

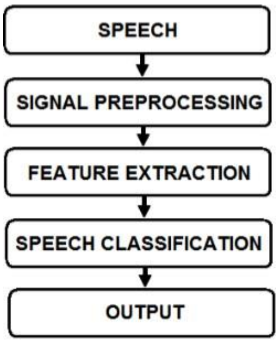


Figure 1. Speech recognition process.

Feature Extraction in General

The essential component of a speech recognition system is feature extraction. It is regarded as one of the most important functions of the system. This task entails extracting the valuable data from the input that aids the system in distinguishing the person. Extraction of features will reduce the magnitude of the speech audio signal without affecting the spoken signal's power. Many feature extraction approaches are available, including Probabilistic Linear Discriminate Analysis (PLDA), Relative Spectral (RASTA Filtering), Mel–Frequency Cepstrum (MFCC), Linear Predictive Coding, and many more. Every technique has its characteristics, advantage, and disadvantage (Kewal *et al.*, 2020). Figure 2 is an example of the basic process of feature extraction.

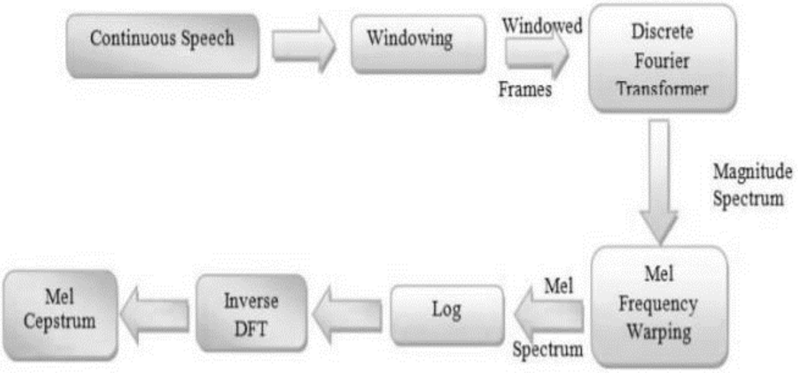


Figure 2. Feature extraction diagram (Kewal *et al.*, 2020).

LPCC is one of the primary low-bit-rate algorithms, and it was developed using the autocorrelation method in an attempt to replicate human speech signals. LPCC is among the well-known speech recognition approaches based on its performance and relative simplicity (Ooi *et al.*, 2012). Linear Predictive Cepstral Coefficients (LPCC) is the technique that we will be used to capture analog to digital sound. A set of cepstral coefficients produced from the spectral envelope determined by LPC will form an LPCC. This is acquired from the coefficients of the Fourier transform depiction of LPC's logarithmic magnitude spectrum. The cepstral analysis is widely used in the area of speech processing due to its capability to well represent speech waveforms and characteristics with a small number of parameters. The LPCC values are cepstral coefficients produced from linear prediction, where it is standardized between +1 and -1 and are obtained from LPC calculated spectral envelopes. The fundamental goal of linear prediction is to obtain the parameters of the vocal tract. A model, x(n) from a speech sample (x(n)) can be constructed as a linear mixture of the previous p speech samples at time n. Figure 3 shows a block diagram of the LPCC.

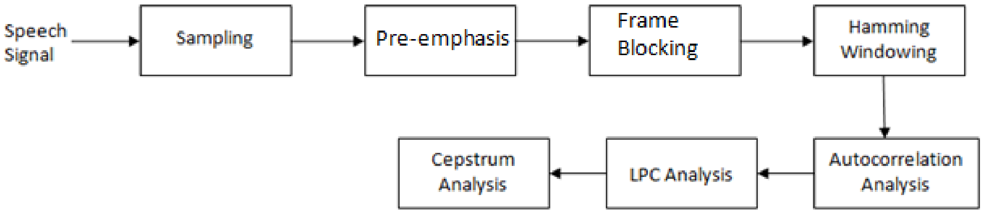


Figure 3. Block diagram of LPCC feature extraction method (Ooi *et al.*, 2012).

One of the most essential aspects of speech recognition is the features applying the LPCC. Using finite numbers of the signal measured, the feature extraction seeks to highlight speech signals. The LPCC coefficients can be derived from LPC, which are then converted into cepstral coefficients. Autocorrelation analysis is then used to obtain the LPCC analysis. Furthermore, the signal's power spectrum is calculated by using the LPC, where it is a tool for analyzing formants. LPC is a prominent and powerful formant estimation approach in speech analysis methods. The following equation shows the LPC for s(n)

$$S[n] \;=\; \textstyle\sum_{i=1}^{p} a_i S(n-1) + Ge[n] \tag{1}$$

where a_i is the linear coefficient and $e[n]$ is the model error. According to the equation, after LPC spectral is computed, the cepstral coefficients using the following equation.

$$C_0 = \log_e p$$
$$C_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k}, for\ 1 < m < p\ and \tag{2}$$
$$C_m = \sum_{k=m-p}^{m-1} C_k a_{m-k}, for\ m > p$$

Classification Methods for Speech

There are various methods for speech recognition that uses machine learning algorithms. Previously, among the most used methods were Hidden Markov Model (HMM), Neural Network, K-Nearest Neighbor, and deep learning-based algorithms such as Recurrent Neural Network and Convolutional. Among the studies, Random Forest is used in speech recognition as discussed in Sincy *et al.* (2014) where the algorithm outperforms other algorithms in their study. Random Forest is classified as decision tree algorithms with other algorithms are Classification and Regression Trees (CART), Random Tree, and REPTree, Alternating Decision Tree (ADTree), and C4.5.

HMM has been used as one of the predominant methods in speech recognition since 1960. A study such as Rabiner (1989) is one of the early discussions of HMM in speech recognition. In this method, the components for speech recognition are speech, feature extraction, and a decoder that consists of an acoustic model, pronunciation, and language model. Therefore, relying on a feature extraction only after the extraction is not adequate. However, the study by Deshmukh (2020) that compares HMM with deep learning, shows that Recurrent Neural Networks could be better than HMM.

Similar to HMM, the Support Vector Method (SVM) is a traditional method applied to automatic speech recognition (Abbaschian *et al.*, 2021). The purpose of SVM is to find the best hyperplane in an N-dimensional space that distinctly classifies the data. Several potential hyperplanes could be chosen to divide the two groups of data points. A hyperplane is a "decision boundary" that separates tuples from one class from another. Studies by Nainan & Kulkarni (2020) and Swarna (2020) have shown that when feature extraction is done correctly, it can outperform other algorithms. The kNN algorithm in speech recognition has been widely used such as in Ooi *et al.* (2012) and Jena *et al.*, (2020), where kNN provides a better classification accuracy compared to the other tested algorithms.

METHODOLOGY

Sample Collection

When the study is carried out, there is no available data sample for Kadazandusun speech. Hence, the sample data is created. The data is important and can be a benefit for many projects that are followed by other researchers later. The details of the sample speech collection are summarized in Table 1.

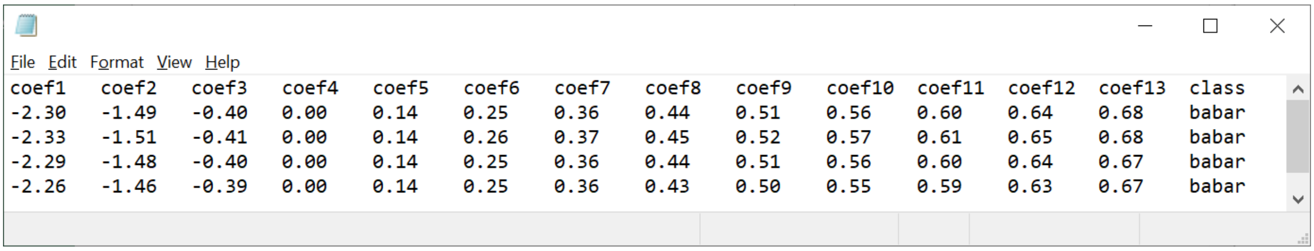
Table 1. The summary of sample collection.

Description	Details
Candidates	10 persons of UMS Dusun ethnic student consists of 5 males and 5 females
Session	10 sessions for each person
Word to test	6 word: palad, babar, tanak, duo, korut, gayo
Total Data Sample	10 persons x 1 session x 6 words = 360 data (training data) 1 person x 1 session x 6 words = 60 data (testing data)

The experiment data is consists of 6 words which are “palad”, “babar”, “tanak”, “duo”, “korut”, and “gayo” that will be spoken by 10 Kadazandusun speakers consists of 5 males and 5 females which are among UMS student. Each speaker will speak 10 samples for each word. The recording was done in a quiet place environment. The reason for 5 females and 5 males in the data collection is to train the recognition engine using different gender.

Feature Extraction

The next step in the case study for speech recognition for Kadazandusun is the feature extraction to get the information from the speech data. The methods used which is LPCC due to 1) one of the popular methods and works well on most speech recognition applications, and 2) produce the important features as coefficient from the speech signal. The total number of features as coefficients extracted from the speech is 13, where it is considered as typically optimal for speech recognition. A sample of the features is shown in Figure 4 below.



```
coef1  coef2  coef3  coef4  coef5  coef6  coef7  coef8  coef9  coef10  coef11  coef12  coef13  class
-2.30  -1.49  -0.40  0.00   0.14   0.25   0.36   0.44   0.51   0.56    0.60    0.64    0.68    babar
-2.33  -1.51  -0.41  0.00   0.14   0.26   0.37   0.45   0.52   0.57    0.61    0.65    0.68    babar
-2.29  -1.48  -0.40  0.00   0.14   0.25   0.36   0.44   0.51   0.56    0.60    0.64    0.67    babar
-2.26  -1.46  -0.39  0.00   0.14   0.25   0.36   0.43   0.50   0.55    0.59    0.63    0.67    babar
```

Figure 4. Sample of feature extraction coefficients and class label.

Speech Recognition

In this phase, the feature data acquired from the feature extraction step is investigated through a machine learning approach using a single classifier and meta classifier algorithms adapted from the Weka machine learning tool. The single and meta classifier algorithms are listed in Table 2.

Table 2. Single and meta classifier used in Weka machine learning tool.

Classifier Type	Classifier Name
Single	Hidden Markov Model (HMM)
	Naïve Bayes (NB)
	Support Vector Machines (SVM or LivSVM in Weka)
	Multilayer Perceptron Neural Network (MLP)
	k-Nearest Neighbor (kNN)
	OneR
	Decision Tree (J48)
	Decision Tree (LMT)
	Random Forest (RF)
	Random Tree (RT)
	Decision Tree SimpleCart
Meta (Ensemble) With Random Forest as base classifier	AdaBoostM1
	Decorate
	Bagging
	END
	LogitBoost
	Rotation Forest

Each of the classifier algorithms is trained with the training data portion (360 samples) and then tested with test data (60 samples) to get the accuracy of the algorithm. In addition, the algorithms in Table 2 are implemented in Weka with their standard parameters.

RESULT AND DISCUSSION

In this section, the experiments are divided into two parts which are LPCC features with single classifiers and the second part is the features with meta classifiers. The results are then compared, and the best classifier algorithm is proposed to be implemented in web-based speech recognition. Table 3 is the results of single classifier algorithms on the LPCC features.

Table 3. Single classifier accuracy on the LPCC features.

Classifier Name	Accuracy (%)
Hidden Markov Model (HMM)	16.67
Naïve Bayes (NB)	43.33
Support Vector Machines (SVM)	28.33
Multilayer Perceptron Neural Network (MLP)	51.67
k-Nearest Neighbor (kNN)	53.33
OneR	48.33
Decision Tree (J48)	43.33
Decision Tree (LMT)	41.67
Random Forest (RF)	**55.00**
Random Tree (RT)	45.00
Decision Tree SimpleCart	43.33

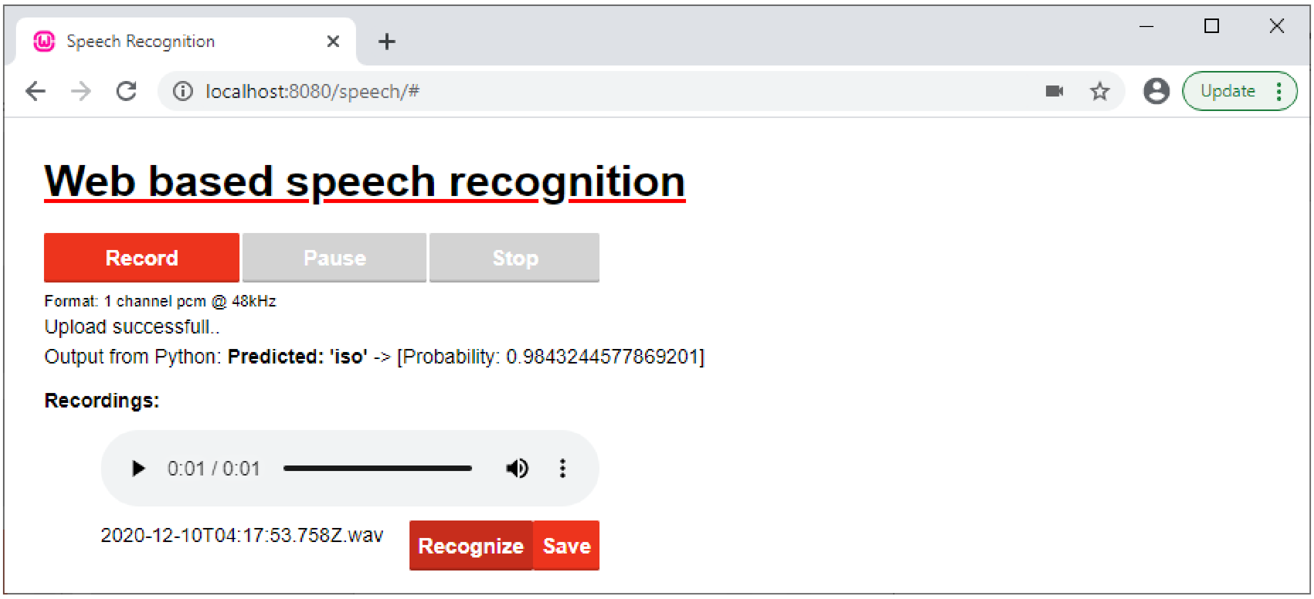
Based on the results in Table 3, LPCC features are not suitable to be used on HMM and SVM although that both of the classifiers are among the popular classifier in previous studies. Random Forest is the best algorithm although it only achieved 55% accuracy on the LPCC features. The next best classifier is the k-NN (using k=1) with 53.33% and then followed by MLP at 51.6%. Therefore, Random Forest is proposed to be used when LPCC is the feature extraction method on the Kadazandusun speech. Diving in detail on the class performance using Radom Forest, Table 4 shows the measures.

Table 4. Detailed accuracy by class using Random Forest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area	Class
	1	0	1	1	1	1	1	babar
	0.7	0.48	0.226	0.7	0.341	0.542	0.228	duo
	0.4	0.06	0.571	0.4	0.471	0.765	0.594	gayo
	0.5	0	1	0.5	0.667	0.658	0.587	korut
	0.5	0	1	0.5	0.667	0.605	0.583	palad
	0.2	0	1	0.2	0.333	0.639	0.376	tanak
Average	0.55	0.09	0.8	0.55	0.58	0.701	0.561	

According to Table 4, the class which contributes most of the accuracy is class 'babar' which is shown by the F-measure and other measures, 1, meaning that all of the samples in the test data (10 samples) were successfully recognized by Random Forest. In comparison, three other classes which are 'duo', 'gayo', and 'tanak' with lower than 0.5 weight, indicate that the word pronunciation may confuse the feature extraction and the classifiers. Further investigation is done on the meta classifiers to examine if there is any improvement in the accuracy. However, upon the completion of the experiments, all of the meta classifiers mentioned in Table 2 were not making any improvement from what Random Forest is achieved, where all of the algorithms produce 55% accuracy.

The prototype of the web-based speech recognition is then developed using the proposed feature extraction and recognition model using Random Forest. Figure 5 shows the preliminary prototype of the application which will be used for learning the language pronunciation in future development. As mentioned before, the prototype of the system was developed using PHP and Python as backend which implements the learning and recognition.



Format: 1 channel pcm @ 48kHz
Upload successfull..
Output from Python: **Predicted: 'iso'** -> [Probability: 0.9843244577869201]

2020-12-10T04:17:53.758Z.wav

Figure 5. Prototype of the web-based Kadazandusun speech recognition for learning.

CONCLUSION

The learning of a language through an automated computer is an important alternative especially in the new norm which leveraging the use of the digital opportunity. Similarly, the Kadazandusun language which is mostly being taught in schools is becoming a priority to support the teaching environment and also the community that would like to learn the language and at the same time preserving the language. Therefore, this paper is one of the preliminary works that step up further to provide such support. In this paper, the discussion is directed from why this study is conducted, data collection, feature data construction, and then the investigation of what suitable recognition algorithm works. It is found that Random Forest outperformed the other classifiers, hence proposed to be used in the web-based application of the Kadazandusun speech recognition. However, to increase the performance of the Kadazandusun speech recognition, further studies need to be performed such as applying other feature extraction methods and other classifiers investigations.

ACKNOWLEDGEMENTS

This work is supported by the funding provided by Skim Geran Inovasi (UMS SGI0063-2018).

REFERENCES

[1] Abbaschian, B. J., Sierra-Sosa, D. & Elmaghraby, A. 2021. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors*, 21(4), 1249.

[2] Deshmukh, A. M. 2020. Comparison of Hidden Markov Model and Recurrent Neural Network in Automatic Speech Recognition. *European Journal of Engineering Research and Science*, 5(8), 1-8.

[3] Jain, N. & Rastogi, S. 2019. Speech Recognition Systems - A Comprehensive Study Of Concepts And Mechanism. *Acta Informatica Malaysia,* 3(1), 1-3.

[4] Jena, B., Mohanty, A. & Mohanty, S. K. 2020. Gender Recognition of Speech Signal using KNN and SVM. *Proceedings of the International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS 2020)*. 10-11 December 2020. Kerala, India.

[5] Juan, S. S., Besacier L., & Tan, T. 2012. Analysis of Malay Speech Recognition for Different Speaker Origins. *Proceedings of 2012 International Conference on Asian Language Processing*. 13-15 November 2012. Hanoi, Vietnam, pp. 229-232.

[6] Kewal, M., Amitesh, D., Rahul, K., Viraj, P. & Suvarna, P. 2020. Speech Recognition: General Idea and Overview. *International Research Journal of Engineering and Technology (IRJET)*, 7(10), 947-953.

[7] Këpuska, V. & Elharati, H. 2015. Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions. *Journal of Computer and Communications*, 3, 1-9.

[8] Nainan, S., & Kulkarni, V. 2020. Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN. *International Journal of Speech Technology*, 24, 809–822

[9] Ooi, C. A., Hariharan, M., Yaacob, S. & Lim, S. C. 2012. Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications*, 39(2), 2157-2165.

[10] Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) 257-286.

[11] Sincy, V. T., Sreekumar, K. T., Santhosh, K. C. & Reghu R. P. C. 2014. Random forest algorithm for improving the performance of speech/non-speech detection. *Proceedings of the First International Conference on Computational Systems and Communications (ICCSC)*. 17-18 December 2014. Trivandrum, India. pp 28-32.

[12] Sloane, E. B., & Silva, R. J. 2020. Artificial intelligence in medical devices and clinical decision support systems. *In:* Ernesto, I. (Ed.). *Clinical Engineering Handbook* (2nd Edition). Academic Press.

[13] Sullivan, A. G. & Albert, C. K. T. 1988. *SABAH, land of the sacred mountain*. Sabah Handicraft Centre.

[14] Swarna, R. N. 2020. Bangla Broadcast Speech Recognition Using Support Vector Machine. *Proceedings of the 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*. 21-22 December 2020. Bangladesh. pp 1-6.