Isolated Keyword Spotting in Multilingual Environment using ANN and MFCC

Brajen Kumar Deka, Pranab Das

Abstract: ***The performance and analysis of Keyword Spotting system (KWS) are applied when the training and testing in a multilingual environment. This paper exhibits an approach for building up a multilingual KWS framework for Assamese, English and Hindi language dependent on feed-forward neural system. Mel Frequency Cepstral Coefficient (MFCC) has been utilized for highlight extraction which gives a lot of highlight vectors from recorded sound examples. Neural Network backpropagation model is utilized to improve the acknowledgment execution on the recently made multilingual database utilizing the multi-layer feed-forward neural system classifier.***

Keywords : ANN, Backpropagation, Keyword Spotting, MFCC.

I. INTRODUCTION

Automatic speech recognition is a key part in applications, for instance, speech record recovery and human-PC association. The speech development is a rising advancement which means to take a speech signal as data and recognize the communicated word and made it as a yield [1]. KWS is a procedure which is utilized to interpret simply precise words from a relentless talk [4]. It is fabulously of language ASR framework which is displayed to out of language words. A good keyword spotter should recognize all keywords, removing false alarms, i.e. not reading non-keyword voice parts [5]. While some work has been done to identify multilingual keywords, most have been done using HMM [2].

We found also that KWS operates only in a one language environment. Multilingual KWS and linguistic recognition [6][7] are vital to making spoken exchange frameworks which can work in a multilingual context. Research on the influence of multiple languages on state-of-the-art KWS is critical for a highly multilingual nation like India. A significant number of the freely open KWS databases are worked from western material. In fact, the NE India linguistic situation is different from the rest of India. This is the area where two large Indo-European and Tibeto-Burman

linguistic families come together and fluently speak each other's language.

The structure of the paper is given below: Section II explains Keyword spotting, Section III clarifies the methodology approach, Section IV shows the result and Section V gives the conclusion.

Revised Manuscript Received on April 25, 2020.

***** Correspondence Author

Brajen Kumar Deka*, Research Scholar, Department of Computer Applications, Assam Don Bosco University, Guwahati, India. Email: brajendeka@gmail.com

Pranab Das, Assistant Professor (Sr.), Department of Computer Applications, Assam Don Bosco University, Guwahati, India. Email: Pranab.Das@dbuniversity.ac.in

II. KEYWORD SPOTTING SYSTEM

Speech Keyword Spotting (KWS) is magnificent structure in sound mining. It is the recovery of all cases of a given keyword in spoken expressions. Keyword Spotting is practically identical character to managing endeavors that procedure a more prominent than ordinary degree of talk like consistent keyword checking and to sound archive ordering. The technology has been used in a variety of applications ranging from telephone call centre system to covert surveillance applications. Keyword spotting plays a crucial role to identify the spoken words from the given speech inputs. A keyword spotting system that simply detects occurrences of relevant keywords will be more efficient than a fully-fledged LVCSR engine. KWS are often a strong and relevant technology and if it's used appropriately, it'll bring with reduced computational requirements increased scalability and potentially higher accuracies.

As a part of audio mining, Keyword spotting is used to automatically find the occurrences of keywords of interests in speech documents. The approach is also used to collect valuable information from vast quantities of records of speech. Keyword spotting also provides necessary statistical information is used in both online and offline applications [8][9][10].

III. METHODOLOGY

A. Database Generations

In this study, male and female voices are recorded in a noi sy, free environment, using a 16 kHz, mono channel, and a 16-bit laptop resolution microphone. With regard to English, Hindi and Assamese with 96-speaker consisting of 48 male and 48 female speakers who speaks 5 times each, isolated words are registered in three language primarily 7 days, 10 digits and 12 months. So we registered a total of 4640 separate samples of words. Among this 80% of the speech, data are used as training and the remaining 20% of data are used as testing dataset. Speaker dependent models are then produced for all the languages. The recorded datasets are:

Table-I: Speech Training dataset for days.

Serial No.	Spoken word in Days		
	Assamese	*English*	*Hindi*
1	সোমবাৰ	Monday	सोमवार
2	মঙ্গলবাৰ	Tuesday	मंगलवार
3	বুধবাৰ	Wednesday	बुधवार
4	বৃহস্পতিবাৰ	Thursday	गुरुवार
5	শুক্রবাৰ	Friday	शुक्रवार
6	শনিবাৰ	Saturday	शनिवार
7	দেওবাৰ	Sunday	रविवार

Table-II: Speech Training dataset for digits.

Serial No.	Spoken Word in Digits		
	Assamese	*English*	*Hindi*
1	০ (xuinno)	0 (zero)	० (shunya)
2	১ (ek)	1 (one)	१ (ek)
3	২ (dui)	2 (two)	२ (do)
4	৩ (tini)	3 (three)	३ (teen)
5	৪ (sari)	4 (four)	४ (chaar)
6	৫ (pas)	5 (five)	५ (paach)
7	৬ (soy)	6 (six)	६ (chhe)
8	৭ (xat)	7 (seven)	७ (saat)
9	৮ (ath)	8 (eight)	८ (aath)
10	৯ (no)	9 (nine)	९ (nau)

Table-III: Speech Training dataset for months.

Serial No.	Spoken word in Months		
	Assamese	*English*	*Hindi*
1	বহাগ	January	जनवरी
2	জেঠ	February	फ़रवरी
3	আহাৰ	March	मार्च
4	শাওন	April	अप्रैल
5	ভাদ	May	मई
6	আহিন	June	जून
7	কাতি	July	जुलाई
8	আঘোণ	August	अगस्त
9	পুহ	September	सितंबर
10	মাঘ	October	अक्टूबर
11	ফাগুন	November	नवंबर
12	চ'ত	December	दिसंबर

B. MFCC Feature Extraction Process

MFC Coefficients is widely accepted in most of the application. Feature extraction means to distinguish the part of that is good for recognizing the content and removal of all background noise. The key distinction between MFCC and cepstral coefficients lies within the process concerned once characterized the speech signals [3][11]. We use MFC coefficients because it is analogous to human hearing mechanisms and the complete process for obtaining MFC coefficients is shown in the following figure [12][13]:

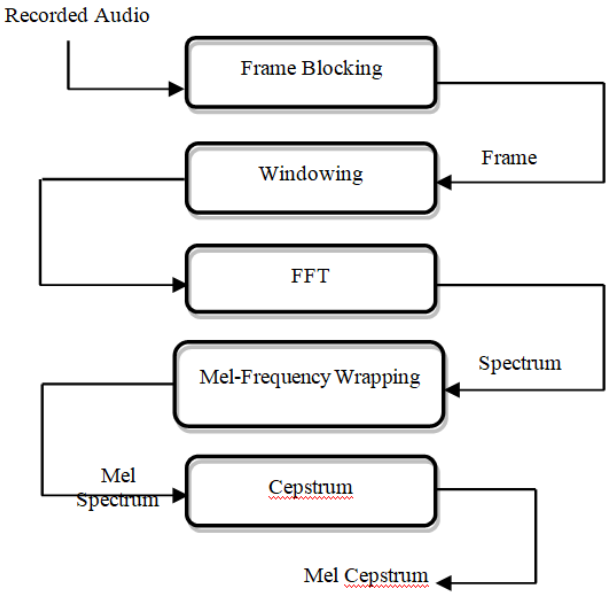

Fig.1: MFCC feature extraction process.

C. ANN Classifier

For isolated words which are set a classification threshold, we use classifier to differentiate between the languages. The experiment uses ANN, because this method of classification gives high precision values. Neural network training is done to connect a particular input with the same output goals. As a training set, 3712 audio samples are used, and 928 audio samples are used for testing purposes. There are various sorts of neural frameworks that are dependably used for structure accreditation structures. The most conventionally used neural structure is a multi-layer feed-forward framework including three layers where each layer offers examination to each layer that follows. The tremendous idea is to spread the information from layer to layer. A trade combine is joined, which gives additional information. The backpropagation model is used to improve the confirmation execution [14][15][16].

D. Backpropagation Neural Network

The neural network [14][17] contains interconnected layers of nodes. Form vectors are used as neural network data sets, with unique samples of the MFC coefficients.

The vector functiondata set is split into three neural network data sets which are train data set, validation data set and test data set. Weights are evenly balanced where necessary.

The hidden layer is connected to the output layer which shows the result. We have used ANN in this paper to describe the audio classification. The neural network is equipped using the backpropagation algorithm in a supervised way. There are many differences in the training algorithm when it comes to classifying new data and this data set is then applied to the network based on the test results.

IV. RESULTS AND DISCUSSION

In this experimental work, we have produced the confusion matrix as the output. The following table shows the experimental results based on the isolated word of days, digits and months considering English, Hindi and Assamese language.

Table-IV: Accuracy rate for days, digits and months.

Accuracy in (%)				
Languages	***Gender***	***Days***	***Digits***	***Months***
English	Male	80.4	93.5	87.5
	Female	83.9	94.1	86.5
	Male and Female	82.15	93.8	87
Hindi	Male	80.4	93.7	80.2
	Female	83.9	92.5	84.4
	Male and Female	79.5	93.1	82.3
Assamese	Male	80.4	93.9	84.1
	Female	78.6	94.9	85.7
	Male and Female	79.5	94.4	84.9

From the above table, we have found that the levels of accuracy of digits are higher than those of days and months. The test result reveals an accuracy rate of 93.8%, 93.1% and 94.4% since the accuracy classification of ten classes of both male and female speakers with regard to English, Hindi and Assamese language. The highest isolated Assamese digits recognition efficiency is found to be 94.4%.

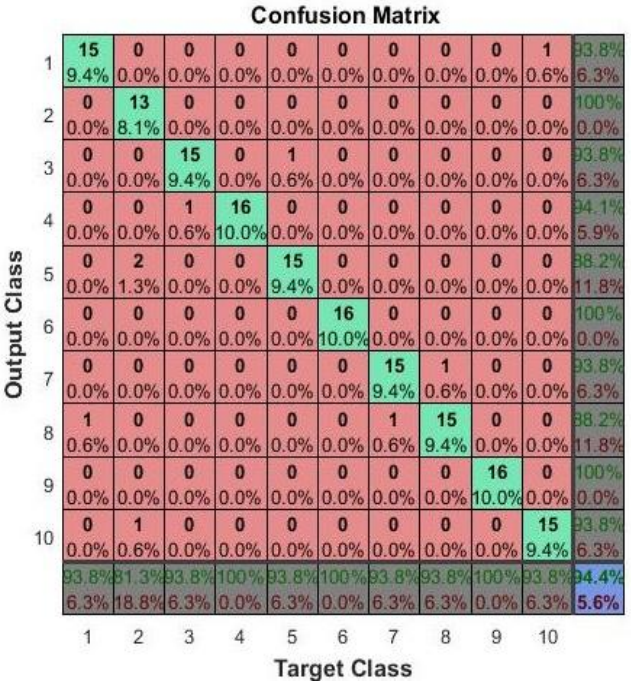

Fig. 2. Confusion matrix of Assamese digit for both male and female.

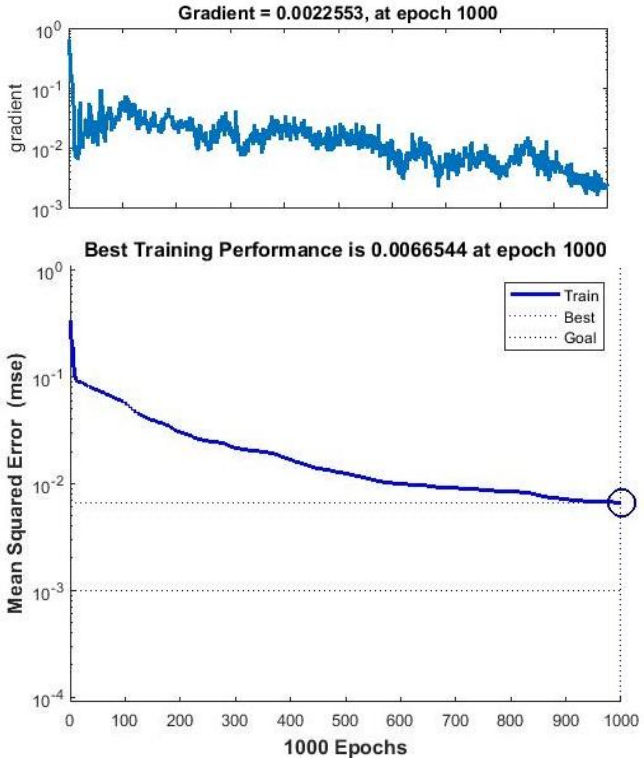

Fig. 3. Performance Curve in terms of mse versus epochs.

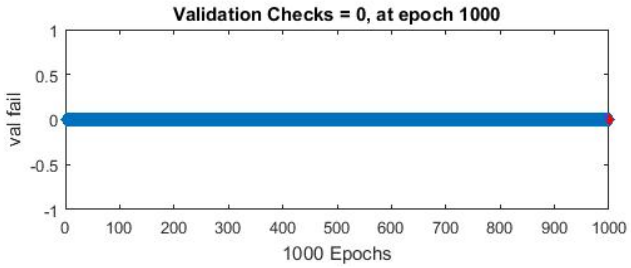

Fig. 4. Plot Training State.

V. CONCLUSION

Speech processing is an essential area of machine learning technology and multi-lingual keyword spotting is an exciting development that has attracted us today. This technology can be used to combine various technologies. An essential tool in this method is a neural network. We have tried to use MFCC and ANN in this paper to implement a system in Assamese, English and Hindi language and it has successfully spotted the words according to their language.

REFERENCES

1. G. Hemakumar and P. Punitha, "Speech Recognition Technology: A survey on Indian Languages," International Journal of Information Science and Intelligent System, vol.2, no.4, 2013.
2. Li. Weifeng, A. Billard, and H. Bourlard, "Keyword Detection for Spontaneous Speech," Image and Signal Processing, 2009. CISP'09, 2nd International Congress on, IEEE, 2009.
3. M. Picheny, D. Nahamoo, V. Goel, B. Kingsbury, B. Ramabhadran, S. J. Rennie, G. Saon, "Global Trends and Advances in Speech recognition," IBM Journal of Research and Development, Vol. No-5, PP. 1-18 sept-oct-2011.

4. Tejedor, Javier, and Jose Colas, "Spanish keyword spotting system based on filler models, pseudo N-gram language model and a confidence measure," Proceedings of IV Jornadas de TechnologiadelHabla, 2006, PP: 255-260.
5. P. Kumar and S.L. Lahudkar, "Automatic Speaker Recognition using LPCC and MFCC," International Journal on Recent and Information Trends in Computing and Communication, Vol. 3, Issue. 4, April 2015, ISSN: 2321-8169:2106-2109.
6. V.K. Jain and N. Tripathi, "Multilingual Speaker Identification using analysis of Pitch and Formant frequencies," Published in IJRITCC Journal, Vol. 4, Issue. 2, February 2016, ISSN: 2321-8169: 296-298.
7. H. Bahi, N. Benati, "A New Keyword Spotting Approach," in IEEE transaction on Speech and audio processing, 2009.
8. B. K. Deka, P. Das, "A Review of Keyword Spotting as an Audio Mining Technique," International Journal of Computer Science and Engineering, Vol.7, Issue. 1, 2019.
9. S. Mandal, "Environmental Natural sound detection and classification using content based retrieval (CBB) and MFCC," International journal of engineering research and application (IJERA) ISSN: 2248-9622, Vol. 2, Issue. 6, Nov-Dec 2012, PP. 123-129.
10. S. A. Ali, S. Burkiand, S. Hasan, "Performance Analysis of learning classifier for spoken digit under Noisy condition," in Journal of Emerging Trends in Computing and Information Sciences, vol.4, No.-3, March 2013.
11. F. Pachet and P. Roy, "Analytical Features: A Knowledge-Based Approach to Audio Feature Generation," Journal of Applied Signal Processing, 2009.
12. P. Roy and P. K. Das, "Language Identification of Indian Languages Based on Gaussian Mixture Models," International Journal of Applied Pattern Recognition Journal, vol. 1, no. 1, 2013.
13. C.S. Kumar and F. S. Wei, "A bilingual Speech Recognition System for English and Tamil," ICICS-PCM, 2003.
14. S. Nidhi, "Speech Recognition using Artificial Neural Network," IJEST, Vol. 3, 2014.
15. S. Tabibbianm, A. Shokri, A. Akbari and B. Nasersharif, "Performance evaluation for an HMM-based keyword spotter and a Large-margin based one in noisy environments," Procedia Computer-Science (ELSEVIER), 2011.
16. S. Shetty, K. K. Achary, "Audio Data Mining using Multi-perceptron Artificial Neural Network," IJCSNS International Journal of Computer Science and Network Security, 2008.
17. D. Li, I. K. Sethi, N. Dimitrova and T. Mc Gee, "Classification of General Audio Data for Content-Based Retrieval," Pattern Recognition Letters, vol.22, no.1, pp.533-544, 2001.

AUTHORS PROFILE

Mr Brajen Kumar Deka is a research scholar in the Department of Computer Applications at Assam Don Bosco University, Guwahati, Assam. He completed his B.Sc. in Mathematics Major from Gauhati University, Guwahati in the year 1999 and MCA from Indira Gandhi National Open University, New Delhi in the year 2006. He is currently working as an Assistant Professor in the Department of Computer Science, NERIM Group of Institutions, Guwahati, Assam. He has published more than 6 research papers in reputed international journals and conferences' including IEEE and it's also available online. His main research work focuses on Speech Processing, Machine Learning, and Data Mining. He has 10 years of teaching experience and 4 years of Research Experience.

Dr Pranab Das received his PhD in Computer Science & Engineering from Rajiv Gandhi University, Arunachal Pradesh in the year 2016. He is currently working as an Assistant Professor (Sr.) in Department of Computer Applications, Assam Don Bosco University, Guwahati, Assam. He is a member of IAENG & ICST. He has published many research papers in reputed international journals and conferences including IEEE, Springer and is available online. His main research work focuses on Speech Processing, Machine Learning, and Audio Mining. He has 13 years of teaching experience and 6 years of Research Experience.