ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

Multilingual Blended Speech Recognition using Gaussian Mixture Model for Non-Dictionary Words

Seema Kumari[1], Gaurav Garg [2]

Perusing M. Tech, Department of CSE, AITM at Palwal, Haryana, India[1]

Assistant Professor, Department of CSE, AITM at Palwal, Haryana, India[2]

ABSTRACT: The area of automatic speech recognition is being discussed from past few decades, and significant advancement is being observed periodically on the automatic speech recognition (ASR) and multi-language spoken systems. However, there are many technological hurdles yet to reach flexible solutions that satisfy the user. This is because of many factors such as environmental noise, paucity of robustness to speech variations (foreign accents, sociolinguistics, gender, and speaking rate), spontaneous, or freestyle speech. To realize the ubiquitous adoption of speech technology, there is need to bridge the space between what speech recognition using technologies like N-Gram and GMM can convey and what human need from it. To make it up, technology must deliver robust and high-recognition accuracy near to man-like performance so it demands to focus on the challenges in speech technology for multilingual approach.

KEYWORDS: Automatic Speech Recognition, Gaussian Mixture Model, Hidden Markov Model, Phonetic Recognition, Language Modelling, Multilingual Blended Speech Recognition.

I. INTRODUCTION

The field of Speech recognition adds another dimension to the problem of natural language understanding. The problems include the background noise, changes in pronunciation according to whether words are spoken in isolation or in the sentence and in variations of accent between individuals. The biggest problem is that everyday speech is not grammatical and the words which are not a part of grammar or dictionary how they can be recognized, as attempts to transcribe a dialogue or conversation often reveal. In the domain of Computer Science, speech recognition is the translation of spoken words into text. It is also recognized as "automatic speech recognition", "ASR", "computer speech recognition", "speech to text", or just "STT". Speech recognition (SR) is a technology that can translate spoken words into text. Some SR systems use "training" where an individual speaker reads sections of text into the SR system. The systems analyze the person's specific voice and use it to fine-tune the recognition of that person's speech, resulting in more accurate transcription. These systems that do not use training are called "Speaker Independent" systems. The systems that use training are called "Speaker Dependent" systems. Speech recognition applications include voice user interfaces such as voice dialing, call routing, domestic appliance control, search, simple data entry, speech-to-text processing (e.g., word processors or emails), and aircraft (Direct Voice Input). The phrase voice recognition refers to finding the identity the words or phrases of words which is not a part of the dictionary or grammar.

The objective of this scheme to provide the new model of the Speech Recognition System for non-dictionary words. The main focus is on N-Gram and GMM to depict best results in pattern recognizing. The objective of the proposed work is:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

1. Study various methods used for Speech Recognizing systems.
2. Identifying issues in existing methods.
3. Proposed new method which is optimized that identify the words or phrases of words which are not a part of Grammar or Dictionary using the Multilingual Approach.

This scheme audits the measurable structure and shows an iterative methodology to choose an ideal number of Gaussian blends that display the most extreme exactness with regards to Hindi discourse acknowledgment framework. Subsequently, the below diagram depicts the general framework of the Automatic Speech Recognition System for perusal and ready reference.

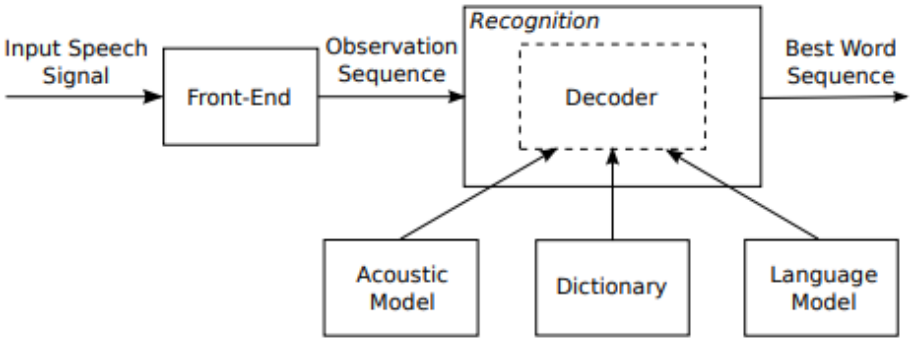

Figure 1. Automatic Speech Recognition System incorporating Input Speech Signal, Observation Sequence, Deoder based on Acoustic Model, Dictionary and Language Model.

Feature extraction: The information speech signal flag S is normally time-area tested speech waveform. Be that as it may, human hearing depends on the attributes of speech sounds in the recurrence space, in this way a phantom portrayal of speech flag is progressively helpful for speech acknowledgment. Since the speech flag is a period differing signal, which is stationary inside a brief timeframe yet changes over a more extended time [Rabiner and Juang, 1993]. While extricating highlights, we have to portion the information speech motion into little edges, at that point procedure each casing independently. The edge length is typically 25 msec. It is short enough to catch the fast changes in speech and adequate to accomplish adequate time-area goals. As the mel-scale approximates the human sound-related reaction better, the Gaussian Mixture Model (GMC) is a standout amongst the most well-known component portrayals in speech acknowledgment [Davis and Mermelstein, 1980].

Recognition: Following feature extraction, the recognition component decodes the most probable word sequence W from the observation sequence O. This recognition process can be represented by the following equation:

$$\begin{aligned}\hat{W} &= \arg\max_{W} P(W|O) \\ &= \arg\max_{W} \frac{P(W)P(O|W)}{P(O)},\end{aligned}$$

where P(W) is the prior probability of the word sequence W, P(O|W) is the likelihood of the observation sequence O given the word sequence W, and P(O) is the probability of observing O. Since P(O) is not a variable of W, Equation can be written as:

$$\hat{W} = \arg\max_{W} P(W)P(O|W).$$

Although the true distribution of P(O|W) and P(W), those probabilities can be estimated from the predefined acoustic model and language model.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

Acoustic model: Most ASR frameworks embrace the Hidden Markov models (HMMs) [Baum and Petrie, 1966; Baum and Egon, 1967] to catch the acoustic qualities of discourse sounds. Figure 2 demonstrates the run of the typical topology of HMMs utilized in discourse acknowledgment. The model has three concealed states linked from left to right. In the wake of going into an express, an example can either stay in that state for some time or travel to the following state. The perception of succession O is created by each state and just relies upon that state. To train a HMM, we need to estimate the initial state distribution $\pi = \{\pi_i = P(q_1 = Si)\}$, the transition probability matrix $A = \{a_{ij} = P(q_{t+1} = S_j \,|qt = S_i)\}$ and the observation distribution $B = \{b_i(O_t) = P(O_t \,|q_t = S_i)\}$, where $O = \{O_1, O_2, \cdots, O_T\}$ is a T-frame

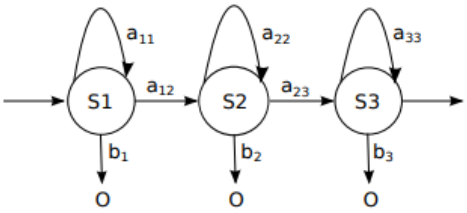

Figure 2: Acoustic Model Framework for Discourse Sounds using Hidden Markov Model vide Speech State Distribution.

observation sequence and $Q = \{q_1, q_2, \ldots, q_T\}$ is the underlying state sequence. The Gaussian mixture model (GMM) is usually used to approximate the observation distribution B, hence the likelihood P(O|W) of the observation sequence O given W can be calculated as:

$$\begin{aligned} P(O|W) &= \sum_{allQ} P(O|Q,\lambda)P(Q|\lambda) \\ &= \sum_{q_1,q_2,\cdots,q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T). \end{aligned}$$

However, HMM parameters $\lambda = (\pi, A, B)$ can be estimated using the well known Baum-Welch (BW) algorithm [Baum et al., 1970], a special case of the classical Expectation-Maximization (EM) algorithm [Dempster et al., 1977]. HMM can be prepared on various units, for example, telephones, syllables, words, and so on. As there are less one of a kind telephones than words in a language, preparing telephone HMMs requires considerably less preparing information than preparing word HMMs. Then again, on the grounds that co-verbalizations frequently show up in consistent discourse, the discourse flag of a telephone can be vigorously impacted by encompassing telephones. Just preparing a HMM for each telephone isn't adequate to show the acoustic properties of discourse sounds in various settings. Thus, in discourse acknowledgment, HMMs are typically prepared on triphone, which is a telephone unit gone before and pursued by explicit telephones. In any case, indeed, even simply preparing triphone HMMs, there are still such a large number of triphone units to work with. For model, in our English ASR framework, there are just 39 telephones, however up to 393 = 59319 special triphones. Accordingly, to diminish the measure of preparing information, we bunch triphone HMM states or on the other hand Gaussian blends into gatherings and utilize the information from each gathering for preparing [Hwang, 1993; Huang, 1989].

Language model: The language model is utilized to ascertain the earlier likelihood P(W) of watching the word grouping W in a language. In speech acknowledgment, the language demonstrates is exceptionally useful to separate acoustic questionable speech sounds and lessen the pursuit space amid interpreting. For instance, it is hard to segregate the accompanying two expressions, "SANTA BANTA" and "CENTA BENTA", utilizing acoustic properties. In any case, from our earlier information of English, we realize that the main articulation is bound to hear that the second expression, all things considered. Scientifically, P(W) can be decayed as

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

$$\begin{aligned} P(W) &= P(w_1, w_2, \cdots, w_n) \\ &= P(w_1)P(w_2|w_1)\cdots P(w_n|w_1, w_2, \cdots, w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i|w_1, w_2, \cdots, w_{i-1}), \end{aligned}$$

which is a product of the probabilities of observing word w_1 given is defined under the ordinary speech models.

Dictionary: We had talked about the acoustic model and language model in previous sectionss. As appeared in Figure 1, there is another module in the decoder, the word reference. Acoustic model estimates the acoustic properties of speech sounds. Language show assesses the earlier likelihood of word groupings in a language. While the word reference conquers any hindrance between acoustic model and language show with the lexical information. Lexicon gives elocutions of words, so decoder knows which HMMs to use for a specific word. Lexicon additionally gives a rundown of words to restrict the language show intricacy and the decoder's inquiry space. Thus, an ASR framework can just perceive a predetermined number of words displayed in the lexicon, which is regularly known as shut vocabulary search acknowledgment. Table 1. shows some portion of the lexicon utilized in ASR framework. We can find that for certain words, for example, "A", numerous elocutions are given in the lexicon, as there are typically a couple of various approaches to articulate those words.

An example of the dictionary used in our ASR system.

A	AH
A(2)	EY
ABANDON	AH B AE N D AH N
⋮	⋮
INK	IH NG K
⋮	⋮
ZURICH	Z UH R IH K

Table 1. Lexicon Utilization under ASR Framework

It is difficult to create a lexicon without any preparation. To acquire a lexicon explicitly for speech acknowledgment, it generally includes with different etymologists physically compose principles and check singular articulations. This procedure can be in all respects exorbitant and tedious. Not specify that numerous language specialists may not concur with one another and an etymologist may not be steady over a significant lot of time. Analysts had explored to anticipate elocutions of new words with models prepared from existing lexicons [Chen, 2003; Bisani and Ney, 2008]. There are additionally some works on refining a current word reference with expressed models [Bahl et al., 1991; Maison, 2003]. Be that as it may, most word references utilized in ASR frameworks still require human mediation. The extent of a lexicon, i.e., the quantity of one of kind words it contains, is an imperative parameter for an ASR framework. For some area explicit applications, a 5k-word lexicon might be sufficient. For a huge vocabulary persistent speech acknowledgment framework, a 64k-word or bigger lexicons are typically connected. While for voice look frameworks, it is exceptionally basic to apply a lexicon with more than 100k words. An exceptionally substantial word reference may make a few issues an ASR framework. To start with, it requires more information for preparing the acoustic model and language demonstrate, which will create bigger models with more parameters. Accordingly, the decoder will devour more memory to stack those models amid unravelling. Second, a bigger word reference will in general increment the perplexity of the language model to the testing information, which

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

will influence the speed and exactness of the recognizer, since it expands the measure of the pursuit space amid deciphering. Along these lines, we can't generally utilize a vast lexicon for all speech acknowledgment applications

II. RELATED WORK

Raymond Low and Roberto Togneri [1] depicted that, Most commonly used score normalization methods can improve the performance of speaker verification systems, but need extra speech data or cohort models, more memory and computation MIPS. In this paper we present a low-cost adaptive online score normalization (LAOSN) method to improve the performance of speaker verification without any extra data. The computation and memory cost of LAOSN is very small. The procedure begins with initialization of the normalization parameters with existing scores of enrolment utterances from a given enrolment speaker model, and the normalization parameters will be online updated with the scores of subsequent test utterances. By this means, an accurate estimation of the unknown score distribution is archived to normalize current test score. Experiments on the Polycost corpus suggest that the LAOSN can achieve much better performance comparing to the well-known Z-norm method without any extra memory and computation cost.

M. M. El Choubassi, H. E. El Khoury, C. E. Jabra Alagha, J. A. Skaf & M. A. AlAlaoui [2] depicted that, language processing make easier path to various speech recognition system to enhance the precision of various Language interaction systems. In this paper we project the work done on updating and improvement of the language model component of a continuous speech recognition system. This experiment proposes a recurrent neural network based language model for Tamil speech recognition system, which aims at reduction of perplexity, word error rate and generating more meaningful text. The RNN language model operates as a predictive model for the next word given in the previous schema. In our experiments, we show that the recurrent neural network language model outperforms the MFCC model and feed forward language model on various Tamil language datasets.

Dongsuk YUK[3] depicted that, The performance of well-trained speech recognizers using high quality full bandwidth speech data is usually degraded when used in real world environments. In particular, telephone speech recognition is extremely difficult due to the limited bandwidth of transmission channels. In this paper, neural network based adaptation methods are applied to telephone speech recognition and a new unsupervised model adaptation method is proposed. The advantage of the neural network based approach is that the retr aining of speech recognizers for telephone speech is avoided. Furthermore, because the multi-layer neural network is able to compute nonlinear functions, it can accommodate for the non-linear mapping between full bandwidth speech and telephone speech. The new unsupervised model adaptation method does not require transcriptions and can be used with the neural networks. Experimental results on TIMIT/NTIMIT corpora show that the performance of the proposed methods is comparable to that of recognizers retrained on telephone speech.

Bülent Bolat, Ünal Küçük [4] depicted that, In this work an active learning PNN was used to recognize instrumental sounds. LPC and MFCC coefficients with different orders were used as features. The best analysis orders were found by using passive PNNs and these sets were used with active learning PNNs. By realizing some experiments, it was shown that the entire performance was improved by using the active learning algorithm.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

III. PROPOSED ALGORITHM & PSEUDO CODE

A. *Work Flow Model:*

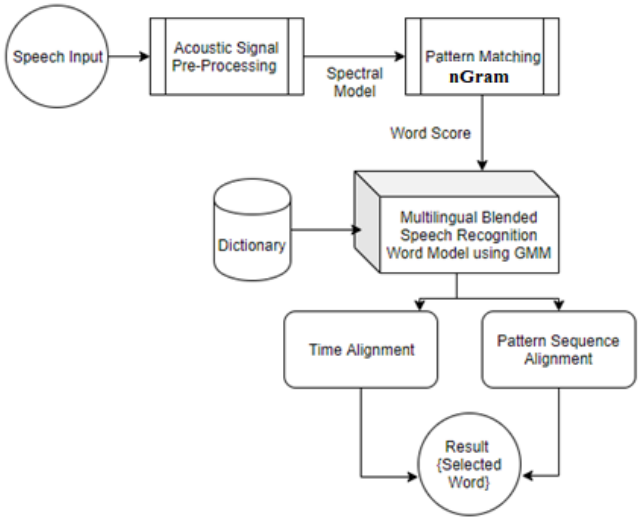

Figure 3: Work Flow of Proposed Scheme comprising of Speech Input, Acoustic Signal Pre-processing using Spectral Model vide N-Gram and Pattern Matching based on Multilingual Dictionary, GMM Model

B. *N-Gram Model:*

So considering these factors a model was designed in which when data is passed three types of probabilities are calculated.

1. **Unigram**: These are the probability of occurrence of single phoneme in language. These basically try to capture the distinct phonemes which are special to particular language like phonemes ending with \h\ are more probable in Hindi than in English.

$$Uni\_Prob(ph|X) = \frac{No.of\ time\ phoneme\ 'ph'\ occur\ in\ the\ training\ data}{Count\ of\ total\ no\ of\ phonemes\ in\ the\ training\ data\ for\ 'X'}$$

2. **Bigram & Trigram**: This is the probability of a phoneme being followed by a given phoneme or pair of phonemes in a given language. This basically tries to capture the sequential information related to phonemes which is also specific to a language. This method is called as n-Gram approach, and we can go till any value of n but as you increase value of n complexity increases exponentially. Therefore we have calculated till n=3.

$$Bi\_Prob(ph2 \mid X, ph1) = \frac{No.of\ time\ phoneme\ ph1\ is\ followed\ by\ phoneme\ ph2\ in\ the\ data}{No\ of\ times\ phonemes\ ph1\ occurs\ in\ the\ training\ data\ of\ language\ X'}$$

$$Tri\_Prob(ph3 \mid X, ph1, ph2) = \frac{No.of\ time\ phoneme\ ph1\ is\ followed\ by\ ph2\ and\ then\ ph3\ in\ the\ data}{No.of\ time\ phoneme\ ph1\ is\ followed\ by\ ph2\ in\ the\ training\ data\ of\ language\ X}$$

So in a simple language in case of trigrams what you do is that instead of considering phoneme as single unit you consider sequence of three phonemes as a single unit to calculate the probability. Following diagram gives a pictorial view of the whole process:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

C. *Gaussian Mixture Model And Maximizing The Likelihood Of Bilingual Language.:*

In this approach different parameters were generated using Gaussian Mixture Model approach in which optimal values of parameters are found by maximizing any given value, which is calculated by combination different values. In this approach value which was maximized was sum of combined probability obtained by using current value of parameters, which is represented by the following value:

SUM(alpha_eng*P(Seqj|Uni)+beta_eng*P(Seqj|Bi)+gamma_eng* P(Seqj|Tri))

where P(Seqj|Uni) represents the unigram probability for sequence j.
P(Seqj|Bi) represents the bigram probability for sequence j.
P(Seqj|Tri) represents the trigram probability for sequence j.

Generated using Gaussian Mixture Model and maximizing the difference of likelihoods from different languages.

This approach is same as the last one just the value which is maximized is now the difference between the combined probabilities for different languages which is represented by the following value:

SUM(alpha_eng *(P(Seqj | UniE)/ P(Seqj | UniH)) +
beta_eng *(P(Seqj | BiE) / P(Seqj | BiH)) +
gamma_eng * (P(Seqj | TriE)/P(Seqj | TriH)))

IV. SIMULATION RESULTS

So an attempt was made to design a language independent front end processor. For this a GMM with N-Gram based algorithm was designed, in which concept of classes is used instead of phonemes. The feature vectors which were generated after first phase of language ID were segmented into different 'broad phonetic category' or classes, based on some properties and then the whole speech data was converted into sequence of classes, and then plays the same role, which phoneme sequences played in previous approach. Now because of its probabilistic approach, it makes it language independent and thus better than Hidden Markov Models based phoneme recognizers although it has not been properly tested. But the system designed on this principle has a lower performance (as shown in the table) in comparison to phoneme based models.

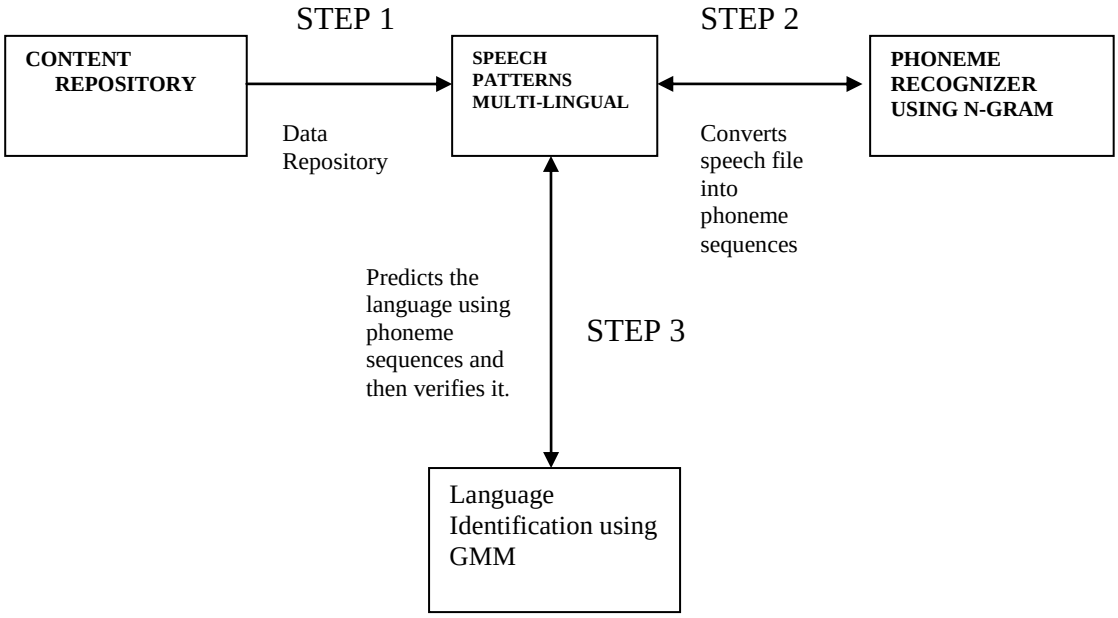

Figure 4 Language Identification using GMM using Pattern Match vide N-Gram

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

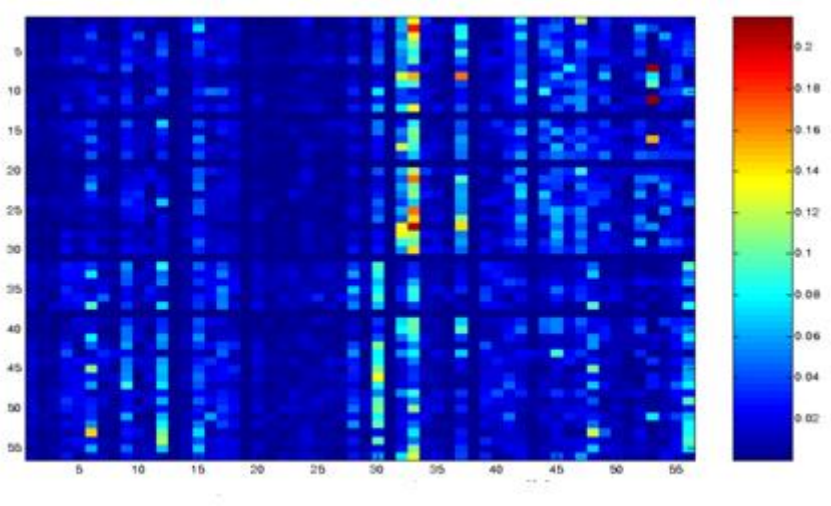

Figure 5 Matrix Scale GMM using Pattern Match vide N-Gram

V. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed algorithm delivers or produces better results as amalgamation of two best techniques i.e. N-Gram and GMM which produces better and effective results in the context of non-dictionary words based on native language vocabulary rules and grammar. Consequently, the facilitation achieves very powerful systems that provide sensible speech recognition results based on complexity lies in non-dictionary words. Hence, for the future the same can be incorporated with non-dictionary datasets either using machine learning models for text classification based on non-dictionary words to form a new model or framework in the dimension of Automatic speech recognition systems.

REFERENCES

1. Raymond Low and Roberto Togneri, Speech Recognition Using the Probabilistic Neural Network, The University of Western Australia,Department of Electrical and Electronic Engineering
2. M. M. El Choubassi, H. E. El Khoury, C. E. Jabra Alagha, J. A. Skaf and M. A. AlAlaoui, Arabic Speech Recognition Using Recurrent Neural Networks, Faculty of Engineering and Architecture – American University of Beirut
3. Dongsuk YUK, Robust Speech Recognition Using Neural Networks and Hidden Markov Models, The State University of New Jersey
4. Bülent Bolat, Ünal Küçük, Speech Music Classification By Using Statistical Neural Networks, Yıldız Technical University, Department of Electric and Electronic Engineering
5. Paolo Marroe, A Complate Guide All You Need to Know Aboat Joone, 17.1.2007
6. Harshavardhana , Varun Ramesh , Sanjana Sundaresh, Vyshak B N, Speaker Recognition System Using MFCC, A Project Work Of 6th Semester Electronics &Communication Engineering, Visvesvaraya Technological University
7. http://cmusphinx.sourceforge.net/sphinx4/#what_is_sphinx4
8. Audio: A Feature Extraction Library, Daniel McEnnis, Cory McKay, Ichiro Fujinaga, Philippe Depalle, Faculty of Music, McGill University
9. B.H. Juang and S. Katigiri "Discriminative Learning for Minimum Error Classification", IEEE Trans. On Signal Processing, Vol. 40, n°12, pp. 3043-3054, 1992.
10. A. Biem, S. Katigiri, E. McDermott & B.-H. Juang An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition, IEEE Trans. SAP, Vol. 9, pp. 96-110, 2001.
11. H. Hermansky "Perceptual Linear Predictive (PLP) analysis of speech", Journal of the Acoustical Society of America, Vol. 4, pp. 1738-1752, 1990.
12. H. Hermansky, N. Morgan "RASTA Processing of Speech", IEEE Trans. On Speech and Audio Processing, Vol. 2, pp. 587-589, October 1994.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

13. B. Townshend "Nonlinear Prediction of Speech", ICASSP'91, Vol.1, pp. 425-428, 1991.
14. M. Faundez, E. Monte, F. Vallverdu "A Comparative Study Between Linear And Nonlinear Speech Prediction", Biological & artificial computation: from neuroscience to technology. IWANN'97, pp. 1154-1163, 1997.
15. P. Chevalier, P. Duvaut, B. Pincinbone "Le Filtrage de Volterra Transverse Réel et Complexe en Traitement du Signal", Traitement du Signal, Vol. 7, n°5, pp. 451-476, 1990.