Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Automatic Speech Recognition of BODO Alpha digits Using Hidden Morkov Models

Laba Thakuria ,Akalpita Das, Purnendu Acharjee.Pran Hari Talukdar
Department of Instrumentation and USIC
Gauhati University, India

Abstract—Through this paper we are going to recognized spoken alphabets and digits in the process of Automatic speech recognition of of bodo language which is needed in many different field by taking alphadigit as input. Bodo language is a Semitic language which differs from other languages such as Assamese,English etc. the major differences is how to pronounce the all alphabets and ten digits. In our research, spoken Bodo digits and alphabets pronunciation are experienced from the speech recognition point of view. It is designed to recognize an isolated whole-word speech based on Hidden Morkov Models which is based on phoneme recognition. In the training and testing phase of the system, the GU_Bodo corpus is used. Ten different observations are performed on GU_Bodo corpus database in this experiment. The first three of them are trained and tested by using each individual digital subset. The fourth one is conducted on these three subsets i.e., trained by using all three training subsets and tested by using all three testing subsets. In these three experiments, the training subset is the same as the fourth experiment but the testing subsets are the same as the first three experiments. The eighth experiment is on the bodo alphabets, and the ninth is applied on the digits and the alphabets respectively. The complete experiment is done in three phases; first one is the designing of the system by using Bodo digits, the second one is working on Bodo alphabets to be evaluated, analyzed and recognized and the third is the combination of the Bodo digits and alphabets together. Through our experiment we achieved 90.60% correct digit recognition in the noisy environment using mixed training and testing subsets. But in case of the alphabet, the system performance is almost 70.17%. again in case of mixed alphabets and digits, the system accuracy is about 81.20% which is better than the alphabet experiments but much less than those conducted for the digits.

Key words--Alphabets, Digits, Bodo, GU_ Bodo , Noisy, HMM, speech recognition.

I. INTRODUCTION

Automatic Speech Recognition is a technique that allows a system to identify the words that a person speaks into a telephone or in front of a microphone. It is our ultimate goal to have a system that would recognize the spoken Bodo alphabets and digits regardless of the environmental noise, gender, and dialect. Automatic Speech Recognition has gained success due to increases in productivity by greatly assisting human operators or by replacing the human element altogether. The basic applications of ASR are: voice response ,command recognition, information inquiry, dictation, automated telephone services etc. It is also used in education for language learning. It has also used for handicapped people to help them to communication with the rest of the society and their daily life.

A. Bodo Language

Before 1953, the Bodo language had no standard form of writing. It had a history of using Deodhai, Roman and Assamese scripts. At present, Bodos adopted the Devanagari script. But, there is a huge difference in the usage of the letters in Bodo language from the Devanagari script. Bodo language shares some common salient features with other languages belonging to the Bodo group. These features are similar in terms of phonology, morphology, syntax, and vocabulary. Bodo language is closely associated with the Dimasa language of the state of Assam and with the Garo language of the state of Meghalaya, and also with Kokborok language of Tripura. It important to note that ,among the four districts of present Bodo land , namely, Kokrajhar, Chirang, Baksa and Udalguri, the language is heard in pure form only in the district of Udalguri. The language is affected by other communities , mostly, Assamese, Bengali and Hindi speaking communities. Like other language, the structure of Bodo language can be studied in the following three level of category [3]- Phonological Structure, Morphological Structure, Syntactic Structure.

The Bodo phonemes consists of 6(six) vowels and 16(sixteen) consonants. Out of these 16 consonants 2(two) are semi vowel. They are as shown below-

- Vowels : अ, आ, इ, उ, ए, औ
- Consonants : ख, ग, ङ, ज, थ, द, न, फ, ब, म, र, ल, स, ह
- Semi Vowels : य, व

Table 1
Bodo Phone-Set

Sl No	Label	IPA	Bodo
1	A	/a/	অ
2	Aa	/a: /	আ
3	I	/ɪ /,/i/	इ
4	U	/u/,/ʊ /	उ
5	E	/e/	এ
6	O	/o/	ओ
7	Kh	/kʰ /	খ
8	G	/g/	গ
9	Ng	/ŋ /	ङ
10	J	/dʒ/	জ
11	Th	/t̪ʰ /	থ
12	D	/d̪/	দ
13	N	/n/	ন
14	Ph	/pʰ /	ফ
15	B	/b/	ব
16	M	/m/	ম
17	Y	/j/	য
18	R	/r/	র
19	L	/l/	ল
20	S	/s /	স
21	H	/ɦ /	হ
22	W	/y/	ব

In addition, Bodo vowels cannot be word initial and must occur either between two consonants or at word-final position. Bodo syllables can be classified as short or long. Syllables can also be classified as open or closed. An open syllable ends with a vowel while a closed syllable ends with a consonant.

B. Spoken Alpha Digits Recognition

On spoken alpha digits recognitions a extensive work has been done in the area of automatic speech recognition and different algorithms have been developed with different degrees of accuracy and efficiency to recognize. In Bodo language there are different accents depending on the geographical region. The researchers targeted spoken alphabets and digits for different languages for automatic speech recognition.

In spoken English alphabet recognition system developed by Cole *et al.* [9] by training on one token of each letter from 120 speakers got 95% performance but it increased to 96% for new set of 30 speakers. Loizou *et al.* [10] gets high performance spoken English recognizer which is implemented using Hidden Morkov Models (HMM). This system of recognizer achieved 97.3% accuracy in speaker-independent alphabet recognition, 55% accuracy in nasal discrimination, 95% accuracy in speaker-independent E-set recognition, and 91.7% accuracy in 300 last names recognition etc.

In Karnjanadecha *et al.* [11] system of Isolated English alphabet recognition the accuracy achieved by the system for speaker independent alphabet recognition is 97.9% and for regarding digits recognitions, Cosi *et al.* [12] designed and tested a telephone bandwidth speaker-independent continuous digit recognizer which is completely based on an artificial neural network and get 99.92% word accuracy rate and a 92.62% accuracy rate for sentence recognition. We have developed and implemented a speaker recognition system based on HMM through this paper.

C. Hidden Markov Models

The commonly used feature vectors are Mel Frequency Cepstrum Coefficient (MFCC) [2] and Linear Prediction derived Cepstrum Coefficient (LPCC) and On the other hand, the Gaussian Mixture Model (GMM) [17] and the Hidden Markov Model (HMM) [18] is the most popular algorithm to implement the recognition schemes. The Hidden Markov Model provides better performances compared to the GMM. HMM is a well-known and widely used statistical method for characterizing the spectral features of speech frame. The assumption of the HMM is that the speech signal can be well characterized as a parametric random access, and the parameters of the stochastic process can be predicted in a well-

defined manner. The HMM technique provides a natural and highly reliable way of recognizing speech for a wide range of applications [18], [19]. In this research, we use MFCC for feature vector and HMM for classification. The most widely used recognition engine based on HMM is Hidden Markov Tool Kit [20] which is mainly used for testing, designing, and implementing ASR systems.

D. Aim of this Research

The paper deals on analysis and evaluation of Bodo alpha digits from an Automatic Speech Recognition perspective. The number of alphadigits is limited to 40 characters; we have need to design a recognition system which is simple, fast and embeddable in other systems and applications. This can be accomplished by studying the similarity between each and every pair of Bodo alphadigits and analyzing the degree of similarity level. Through this system we may minimize ASR system errors. If all possible pairs of similar alphabets are consider and have a very low similarity, or if all pairs of utterances of the same alphabet are taken from various trails and speakers have a very high similarity, then the recognition system will unlikely produce errors in its output.

II. EXPERIMENTAL STUDY

A. System Overview

A ASR system based on HMM is developed to carry out the aim of our research. The system is divided into three modules according to their functionality, as shown in Figure 1. The first one is a training ,which is to create the knowledge about the speech and language to be used in the system, the second is the HMM models, which is to store and organize the system knowledge gained by the first module and the third module is the recognition, whose function is to figure out what is the meaning of the input speech given in the testing phase. As in Table 2, the parameters of the system were an 10KHz at sampling rate with a 16 bit sample resolution, a 30 millisecond Hamming window duration with a step size of 15 milliseconds, MFCC coefficients with 24 as the length of cepstral leftering, and 28 filter bank channels of which 14 were as the number of MFCC coefficients, and of which 0.87 are as are the pre-emphasis coefficients. The parameters are adopted from our experimental research experience, such as in [21], regarding the design of the ASR system . It is clear that our system is a limited-vocabulary and word dependent system; hence, there is no need for a complicated system which is normally needed in more advanced applications such as continuous and spontaneous speech recognition systems.

Phoneme- based models are good at capturing phonetic details. Also, context-dependent phoneme models can be used to characterize formant transition information, which is very important to discriminate between digits that can be confused. The Hidden Markov Model Toolkit (HTK) is used for designing and testing the speech recognition systems throughout our all experiments. The system is initially designed as a phoneme level recognizer with three active states, one Gaussian mixture per state, continuous, left-to-right, and no skip HMM models. The system is designed by considering all Modern Standard Bodo monophones as given by the GU_Bodo. Since most of the digits consisted of more than two phonemes, context-dependent tri-phone models were created from the monophone models . A decision tree method is used to align and tie the model before the last step of the training phase. The last step in the training phase is to re-estimate HMM parameters. Also, we considered phoneme context by using tri-state models instead of mono-phoneme models. Moreover, the selection of other parameters, such as number of states in each HMM model, type and configuration of transition between states, number mixture per state, and way of initialization of the parameters, are considered in the past outcome of the research and way of initialization of the parameters are considered in the more advanced HMM- based systems.

B. Database

The GU_Bodo corpus [24] is created and contains a database of speech waves and their transcriptions of 2000 speakers covering all the regions in Kokrajhar District with statistical distribution of region, age, gender, and telephones. The Gu_Bodo corpus is designed to be rich in terms of its speech sound content and speaker diversity within Kokrajhar. It is designed to train and test automatic speech recognition and to be used in speaker, gender, accent, and language identification systems. The database has more than 500,000 electronic files.

III. RESULTS

A. Digits Subset

We have evaluated different experiments and obtained the results shown in the tables below. The experiments included the systems that are trained and tested by GU_Bodo1, GU_Bodo2,GU_Bodo3, and GU_Bodo4, in addition to the combination of these three parts . . The confusion matrixes of the results of these experiments are shown in Tables 3, 4, 5, and 6 respectively. In each experiment, the system is trained and tested by sets within the same parts of Gu_Bodo corpus. In the first Experiment , the system is and tested by 30% and trained by 70% of Gu_Bodo1.

TABLE 2
Confusion Matrix of the System when Trained and Tested With GU_Bodo1 (Del=Deleted, Corr=Correct)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**del**	**Corr%**	**Total tokens**	**Missed**
O	8	0	4	5	0	2	2	0	0	0	0	42.16	21	12
1	0	23	0	0	0	0	0	0	0	0	0	99	23	0
2	0	0	22	0	0	0	0	0	0	0	0	91.5	22	2

3	0	0	0	21	0	1	1	0	0	0	0	91	23	2
4	1	0	0	0	23	0	0	1	0	0	0	93	25	2
5	0	0	0	1		24	0	0	0	0	0	96	24	1
6	3	1		1		1	19				0	77	25	6
7					1			23			0	94.5	24	1
8	1			1					22		0	88.17	24	3
9		1	1	1		1		1		18	0	80.17	23	5
Ins												84.75	234	

TABLE 3
Confusion Matrix of the System when Trained and Tested With GU_Bodo2 Corpus (Del=Deleted, Corr=Correct)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**del**	**Corr%**	**Total Tokens**	**Missed**
O	176	0	18	13	1	5	7	8	1	11	2	72.92	242	67
1	0	232	0	2	1	1	1	2	0	0	3	97.07	242	10
2	3	1	259	3	0	0	0	0	0	11	0	97	267	8
3	3	0	0	236	0	1	1	0	1	0	1	96.5	243	7
4	1	1	0	0	255	0	0	5	0	0	4	96.3	266	11
5	1	1	0	2	0	220	4	0	0	0	2	97.4	230	10
6	2	3	1	9	0	0	223	1	0	2	3	92.5	244	21
7	2	0	1	11	1	0	0	240	1	I	0	93.3	257	17
8	1	0	1	2	0	0	1	0	236	0	1	96.4	242	6
9	3	0	3	3	0	1	2	5	0	220	0	93.8	237	17
Ins											16	92.5	2470	174

Here all training and testing subsets were disjointed. The number of tokens in the testing subsets of GU_Bodo1, GU_Bodo2,GU_Bodo3, and GU_Bodo4 is 300, 400, 5000, and 1,162, respectively, and the number of tokens in the training subsets was almost three times that of their respective testing subsets. From the given in Tables 3, 4, and 5, we have the obtained the correct rates of the system are 88.21%, 93.16%, and 90.57% when using GU_Bodo1, GU_Bodo2, and GU_Bodo3, respectively. The least correct rate is obtained in case of GU_Bodo1 and the best is obtained in the case of GU_Bodo2. In these three experiments, digit zero got a very low recognition rate compared to the other digits. Depending on testing database subset, the system had to recognize 4000 samples for all 10 digits. The overall system performance is 96.39%.

B. Alphabets Subset

For Bodo alphabets subset, we evaluated different experiments and get four confusion matrices. The system is tested and trained by GU_Bodo12, GU_Bodo21, GU_Bodo31 separately and, then, together these three into one. We get the correct rates of the system are 68.14%, 70%, and 75.26% respectively. The least correct rate is evaluated in the case of GU_Bodo21 while the best correct rate is encountered in the case of GU_Bodo12. These reflect the normal distribution of letters in Bodo where some letters are more frequent while the others are less frequent.

C. Alphadigits Subset

For Bodo alphadigits, we evaluated an experiment by using the same system but with all data used for both digits and alphabets. In the same way, the six individual testing subsets were combined in one testing subset. From this dataset, we are getting using a total of 40 vocabularies. We get the overall system accuracy rate is 75.15%, which is better than the accuracies of the alphabet experiments but much less than those conducted for the digits experiments. Although the overall accuracy rate of the alphabets improved in this experiment, the accuracy rate of the digits was lower than that of the previous experiments.

IV. CONCLUSION

Bodo speech data recognizer is designed to investigate the accuracy and the process of automatic speeech recognition. It is completely HMM based and by using a speech in a noisy environment. Our system is developed by using Hidden Markov Model Tool Kit . It consists of HMM modules storage, recognition module and training module. The speech data are partitioned to training (80%) and testing (20%). At the beginning by using the speech data (3 subsets) various experiments are done in our study. The result is varied in all the experiments but we find 84.17% as its minimum correct rate and 93.67% as its maximum correct rate. By considering the digit recognition analysis, the greatest accuracy was encountered in the case of digit 1 while the least accuracy is encountered in the case of digit 0. The main cause for such a variation may be attributed to the tokens themselves; 0 is the only monosyllabic Bodo digit wherewith only one vowel is there and it is short with a lower amplitude than the other Bodo vowels. The case of speech perception for acoustic characteristics vowels are very important which include comparatively wide range and high amplitude of spectrum coverage. On the other hand, digit 1 is disyllabic where the vowel in the first syllable is the low long /a/. Since vowel possesses the highest amplitude and wide range of spectrum coverage in Bodo language system which helps in recognizing words using such a vowel. In the next part , we consider again (3 subset) of different experiments by using Bodo alphabets. Our experiments are conducted. We get the correct rate as 70.14%. The rate of

alphabet is lower than the rate of the digits. The alphabets are parts of Bodo words the alphabets differ greatly in terms of their frequencies and digits are considered as independent characters which can be maintained in terms of frequency. This difference of frequency affects in the testing phase as well as in the training phase.

At the very last ,digits and alphabet are together in one experiment for the use of maximum dataset for all digits and alphabets and we get 78.17%. as the correct rate. We can conclude that the performance of our system is far better for the alphabets than the digits.

REFERENCES

[1] Top Internet Languages - Internet World Stats, http://www.internetworldstats.com/stats7.htm, 2010.

[2] Y. A. Alotaibi, "Investigating Spoken Arabic Digits in Speech Recognition Setting", *Journal of Information Sciences*, **173(1–3)**(2005), Elsevier, pp. 115–139.

[3] Muhammad Alkhouli, "*Alaswaat Alaghawaiyah*", Daar Alfalah: Jordan, 1990 (in Arabic).

[4] J. Deller, J. Proakis, and J. H. Hansen, "*Discrete-Time Processing of Speech Signal*", Macmillan, 1993.

[5] M. Elshafei, "Toward an Arabic Text-to -Speech System", *The Arabian Journal for Science and Engineering*, **16**, **(4B)**(1991), pp. 565–583.

[6] Yousif A. El-Imam, "An Unrestricted Vocabulary Arabic Speech Synthesis System", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, **37(12)**(1989), pp. 1829–1845.

[7] A. Youssef and O. Emam, "An Arabic TTS System on the IBM Trainable Synthesizer", Le Traitement Automatique de l'Arabe, *JEP-TALN 2004*, Fes, 19–21 avril 2004.

[8] Y. A. Alotaibi, "High Performance Arabic Digits Recognizer Using Neural Networks", *Proceedings of the International Joint Conference on Neural Networks*, **1**(2003), pp. 670–674.

[9] R. Cole, M. Fanty, Y. Muthusamy, and M. Gopalakrishnan, "Speaker-Independent Recognition of Spoken English Letters", *International Joint Conference on Neural Networks (IJCNN)*, **2**(1990), pp. 45–51.

[10] P. C. Loizou and A. S. Spanias, "High-Performance Alphabet Recognition", *IEEE Trans. on Speech and Audio Processing*, **4**(**6**), pp. 430–445.

[11] M. Karnjanadecha and Z. Zahorian, "Signal Modeling for High-Performance Robust Isolated Word Recognition", *IEEE Trans. on Speech and Audio Processing*, **9**(**6)**(2001), pp. 647–654.

[12] P. Cosi, J. Hosom, and A. Valente, "High Performance Telephone Bandwidth Speaker Independent Continuous Digit Recognition", *Automatic Speech Recognition and Understanding Workshop (ASRU)*, Trento, Italy, 2001.

[13] Elias Hagos, "Implementation of an Isolated Word Recognition System", *UMI Dissertation Service*, 1985.

[14] W. Abdulah and M. Abdul-Karim, "Real-Time Spoken Arabic Recognizer", *Int. J. Electronics*, **59**(**5)**(1984), pp. 645–648.

[15] A. Al-Otaibi, "Speech Processing", The British Library in Association with UMI, 1988.

[16] Rafik Djemili, Mouldi Bedda, and Hocine Bourouba, "Recognition of Spoken Arabic Digits Using Neural Predictive Hidden Markov Models", *The International Arab Journal of Information Technology*, **1**(**2**)(2004), pp. 226–233.

[17] John Morgan, "Making a Speech Recognizer Tolerate Non-Native Speech Through Gaussian Mixture Merging", In *STIL/ICALL – NLP and Speech Technology in Advanced Language Learning Systems*, Venice, Italy, 2004.

[18] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, **77(2)**(1989), pp. 257–286.

[19] B. Juang and L. Rabiner, "Hidden Markov Models for Speech Recognition", *Technometrics*, **33(3)**(1991), pp. 251–272.

[20] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version. 3.4)", Cambridge University Engineering Department, 2006. http:///htk.eng.cam.ac.uk/prot-doc/ktkbook.pdf.

[21] Yousef Ajami Alotaibi, "Comparative Study of ANN and HMM to Arabic Digits Recognition Systems", *Journal of King Abdulaziz University: Engineering Sciences*, **19(1)**(2008), pp. 43–59.

[22] Mansour Alghamdi, Yahia El Hadj, and Mohamed Alkanhal, (2007) "A Manual System to Segment and Transcribe Arabic Speech", *IEEE International Conference on Signal Processing and Communication (ICSPC07)*. Dubai, UAE: 24–27 November 2007.

[23] Yousef Ajami Alotaibi, "Is Phoneme Level Better than Word Level for HMM Models in Limited Vocabulary ASR Systems?", *The International Conference on Information Technology - New Generations (ITNG2010)*, pp. 332–337, Las Vegas, USA, April 12-14, 2010.

[24] M. Alghamdi, F. Alhargan, M. Alkanhal, A. Alkhairi, and M. Aldusuqi, "Saudi Accented Arabic Voice Bank (SAAVB)", Final Report, Computer and Electronics Research Institute, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia, 2003**.**

