International Journal of Research in Signal Processing, Computing & Communication System Design
4 (2), 2018, 09-12
http://www.publishingindia.com/ijrsp

Assamese Connected Digit Recognition System

Barsha Deka[1], Abhishek Dey[2*] and S. R. Nirmala[3]
[1]Department of Electronics and Communication Engineering, Gauhati University Institute of Science and Technology, Gauhati University, Guwahati, Assam, India. barsha.deka4@gmail.com
[2]Department of Electronics and Communication Engineering, Gauhati University Institute of Science and Technology, Gauhati University, Guwahati, Assam, India. Email: abhishekdey.gu @gmail.com
[3]Department of Electronics and Communication Engineering, Gauhati University Institute of Science and Technology, Gauhati University, Guwahati, Assam, India. Email: nirmalasr3@gmail.com
*Corresponding Author

Abstract: In this work, we present the development of a connected digit recognition system in Assamese language. Assamese is an under-resourced language of North-East India that is widely spoken in the state of Assam. The text corpus used in this work, consists of a sequence 7 digits spoken in continuous manner. In order to capture the variations in phonetic context, the sequence of digits were arranged in such a way that, each digit occur in all the 7 positions. The speech corpus used in this work was collected from 11 native Assamese speakers out of which 5 were female while 6 were male. Mel Frequency Cepstral Coefficient (MFCC) features have been used as front-end features. We have explored the Subspace Gaussian Mixture Model (SGMM) based acoustic modeling approach in addition to the Gaussian Mixture Model (GMM) within the Hidden Markov Model (HMM) framework. Accuracies of 95.7% and 95.9% are achieved in GMM-HMM and SGMM-HMM systems respectively.

Keywords: Assamese language, Digit recognition, SGMM-HMM.

I. Introduction

Automatic Speech Recognition (ASR) in under-resourced languages is considered as an active area of research in the speech research community. During the past decade, with the progress of digital evolution, speech recognition applications have gained immense popularity in the commercial market. In most of these applications, digit recognition plays a vital role such as a PIN code recognizer. However, such systems are readily available in highly resourced languages like English. Considering these advancements, there is need to develop these systems in under-resourced languages as well. Motivated by this, we have made an attempt to develop a connected digit recognition system in Assamese, an under-resourced language of North-Eastern region of India. Assamese is an Eastern Indo-Aryan language spoken by more than 15 million people primarily in the state of Assam [12]. Apart from Assam, it is also spoken in neighboring states like Meghalaya and Nagaland.

Previous works in literature report a number of works in digit recognition task for under-resourced Indian languages. Babita *et al.* in [1] developed an isolated digit recognition system in Hindi using HTK toolkit. They have achieved recognition rates of 86.17% and 85% on phone model and word model respectively. CiniKurian *et al.* in [2] proposed a connected digit recognition system in Malayalam. They have employed Perceptual Linear Predictive (PLP) as the front-end features. Accuracy of 99.5% is reported with unseen test data. In [3], isolated digit recognition in Tamil is proposed. A comparative study on template based approach and HMM based approach is reported in this work. Devyani *et al*. in [4] reported an isolated digit recognition system in Marathi using HTK toolkit. Biswajit *et al.* in [5] reported a noise robust digit recognition system in Mizo using data augmentation and tonal features.

However, in the context of Assamese language, although a number of works on speech recognition area have been reported in the literature, limited work has been reported on digit recognition task. Himangshu *et al.* in [6] reported development of an automatic syllabification model with 11 rules for Assamese with accuracy of more than 95%. They have also reported a deep learning based neural network model for Assamese speech recognition task with an accuracy of 78.05%. Mousmita *et al.* reported a work on Assamese numeral recognition system using Linear Predictive Coding (LPC), Principal Component Analysis (PCA) features and other features to tackle mood and gender variations [7]. Biswajit *et al.* in [8] reported the development of Assamese phonetic engine in three different speaking modes, namely, reading, lecture and conversation. Accuracy of 47.31%, 45.30% and 36.13% are achieved in reading, lecture and conversation modes, respectively. A number of works have been reported related to the Assamese spoken query system for retrieving the price of agricultural commodities and weather information [9-11].

In the current work, we propose the development of a connected digit recognition system with limited data set. We have employed the Mel Frequency Cepstral Coefficients (MFCC) as the front-end feature set. In addition to the conventional GMM-HMM based acoustic approach, we have also explored the SGMM-HMM based approach in our present study.

The remainder of the paper is discussed as follows: Section II details the speech corpus used in this work. Section III discusses the phonetic structure along with their distribution. In Section IV, the models of speech recognition are discussed. Section V discusses the experimental setup and results. Finally the paper is concluded in Section VI.

II. Speech Corpus

The speech corpus is collected from 11 native Assamese speakers (5 female and 6 male) consisting a vocabulary size of 10 digits (0-9). The speech data is recorded in microphone channel using laptop. The speech files are recorded at a sampling frequency of 16 kHz and bit resolution of 16 bits / sample. Each recorded speech files are of 5 seconds duration, including short pauses and silences. The text corpus used in this study, comprises of 80 different digit sequences. Each digit sequence consists of 7 digits. Each speaker is asked to read all the 80 digit sequence which resulted in 880 speech samples (11x80). The overall speech corpus is divided in such a way that 80% of the data is used for training the acoustic models while remaining 20% is used for testing the performance of the trained models. Hence, out of 880 speech files, 560 files are used in the training set and 320 files are used in the testing set. While disparting the speech corpus into two parts, it is ensured that the speakers in the training set are different from that of the testing set.

III. Phonemics Inventory

The phonetic composition of the Assamese digit database is discussed in this section. The Assamese digit database consists of 14 unique phones out of which 5 are vowels and 9 are consonants. Table I shows the pronunciation dictionary of all the digits. The labeling of the phonetic units are created following the ILSL12 [13] naming convention. The frequency distribution of occurrences of the 14 phonetic units in the training data is shown in Fig. 1.

Table I: Pronunciation Dictionary of Assamese Digits

Digits	Digits in Assamese	Phone Level Break Up
0	suunya	s u n y ax
1	ek	e k
2	dui	d u i
3	tini	t i n i
4	saari	s aa r i
5	paas	p aa s
6	soy	s ax y
7	saat	s aa t
8	aath	aa th
9	na	n ax

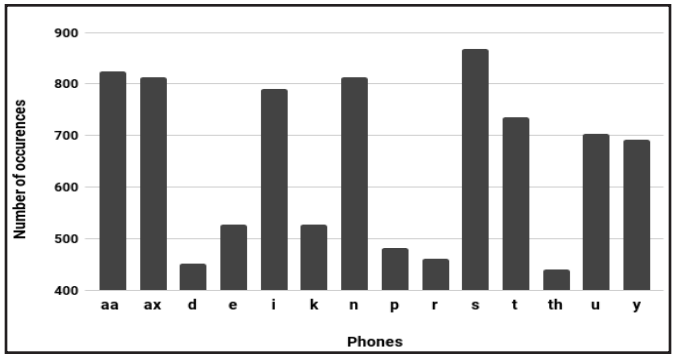

Fig. 1: Frequency of Occurrence of Phones in Assamese Digit Database

IV. Models of Speech Recognition

The goal of an ASR system is to hypothesize the best word sequence W from a given acoustic feature sequence A extracted from the test data.

$$W=\text{argmax}\ (P(W|A))$$

The probability distribution P(W|A) can further be determined using Bayes' theorem shown below:

$$P(W|A) = (P(A|W)\ .\ P(W))\ /\ P(A)$$

where, P(A|W) is the likelihood of the feature sequence, P(W) is the prior probability for the word sequence, P(A) is the observation probability. P(A) is constant for each W, therefore, the previous equation can be written as:

$$P(W|A) = P(A|W)\ .\ P(W)$$

Therefore, the final equation is,

$$W=\text{argmax}\ (P(A|W)\ .\ P(W))$$

The likelihood P(A|W) is determined by the acoustic model and the prior probability P(W) is estimated by the language model.

A. Acoustic Modeling

The current study explores two different acoustic modeling approaches which are discussed as follows:

i. GMM-HMM System

GMM-HMM is the standard approach used in ASR systems. Gaussian Mixture Model (GMM) based Hidden Markov Models (HMMs) is used to represent the sequential structure of speech signals. Gaussian Mixture Model (GMM) is used to model the HMM states by distributing the feature vectors of all the phonemes for the given state [14]. The parameters of multi-variate GMM are mean, covariance and weights of the components. The GMM is given by the equation:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \; g(x|\mu_i, \Sigma_i)$$

where, x is the multi-dimensional vector, $g(x|\mu_i,\Sigma_i)$ are the Gaussian densities for M components and μ_i and Σ_i are the mean vector and covariance matrix respectively.

Further, the gaussian density is given by:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi\,|\Sigma|)}} \; \exp\{-1/2(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\}$$

ii. SGMM-HMM System

The parameters of GMM that are mean, covariance and weights, are estimated independently whereas for SGMM, this are estimated globally. GMM models are trained from a lower dimensional subspace instead, we train a globally shared low dimensional subspace from which the GMM models are trained. This results in reduction of the total number of parameter estimation and makes it possible to learn the model parameters with a limited amount of training data [14].

B. Language Model

Language model essentially models the transition between words by estimating the prior probabilities. Such models are usually estimated by counting on a large collection of text. In our present work, the prior probability of the word sequence P(W) is learned using a statistical bi-gram language model. The prior probabilities are learned from transcription of the training data.

V. Experimental Setup and Results

This section discusses the experimental setup and results observed across different acoustic modeling techniques. All the experimental evaluations are performed using KALDI [15] speech recognition toolkit.

In this work, we have employed Mel Frequency Cepstral Coefficients (MFCC) [16] as front-end features. The MFCC feature vectors are extracted for speech frames of 25 ms duration with a frame-shift of 10 ms duration. A pre-emphasis factor of 0.97 is used and frames are windowed using Hamming window. In addition to the 13 dimensional base MFCC feature vectors, the delta and delta-delta variants are appended in order to capture the dynamic characteristics. The resulting 39 dimensional MFCC features are used for initializing the context-independent GMM-HMM monophone system. For each of the 14 phonetic units along with silence, a 3-state left to right HMM model is trained. Using monophone model, the boundaries are forced aligned and fed as initial alignment to the context-dependent GMM-HMM [15] system. This system is labeled as Tri1. In order to get the optimum performance, the number of Gaussians per state (G) is varied from 2, 4, 8 and 16 while the number of senons (S) is varied from 100 to 1000 in steps of 100. It is observed that for G=2 and S=100, WER of 11.9% is achieved in the Tri1 system. Using the Tri1 acoustic model, the boundaries are again forced aligned and Tri2 GMM-HMM system is developed. In Tri2 system, the dynamic characteristics are captured from the neighboring frames. The 13 dimensional static MFCC features are spliced in four frames to the left and four features to the right thereby resulting in 117 (13x9) dimensional feature vector which is reduced to 40 using Linear Discriminant Analysis (LDA). Using Maximum Likelihood Linear Transform (MLLT), the resulting feature vectors are further decorrelated. Similar to the Tri1 system, G and S are tuned and 11.1% WER is achieved for G=4 and S=100 in Tri2 system. The phone boundaries are again realigned using Tri2 model and fed as input to the Tri3 model. In this variant of GMM-HMM system, the feature vectors are further normalized using Feature space Maximum Likelihood Linear Regression (fMLLR) and using these transformations Speaker Adaptive Training (SAT) is incorporated. The WER is reduced from 11.1 % in Tri2 system to 4.3% in Tri3 system for G=8 and S=100. Using the Tri3 model, the phone boundaries are further realigned and fed as input alignment to the SGMM [15] system. For G=2 and S=400, WER of 4.1% is noted in SGMM-HMM system. The results obtained in the experimental studies are tabulated in Table II. The first column of Table II lists the different ASR systems, the second column shows the %WER and the third column gives the corresponding accuracies.

Table II: Results for Different Training Models

ASR System	WER (%)	Accuracy (%)
GMM-HMM (Monophone)	18.4	81.6
GMM-HMM (Tri1)	11.9	88.1
GMM-HMM (Tri2)	11.1	88.9
GMM-HMM (Tri3)	4.3	95.7
SGMM-HMM	4.1	95.9

VI. Conclusion

This paper illustrates the work on development of a connected digit recognition system in Assamese language. The ASR systems developed in this study, comprises of limited resources. However, while recording the digit sequences, the arrangement of the digit sequences helped in learning the contextual information well. Further, tuning of senones and Gaussians per state helped in getting the optimum performance in each of the ASR systems. Thus, it can be concluded that, even with small size speech databases, limited vocabulary speech recognition tasks like digit recognition systems can be developed with high recognition rates that can serve as the backbone for different ASR applications.

Acknowledgement

The authors would like to thank IITG for making available the infrastructure in carrying out this experiments, and Prof. Samudravijaya K. for his support.

References

[1] D. C. Wyld, B. Saxena, and C. Wahi, "Hindi digits recognition system on speech data collected in different natural noise environments," *Computer Science and Information Technology*, vol. 5, pp. 23-30, 2015. DOI: 10.5121/csit.2015.50303.

[2] C. Kurian, and K. Balakrishnan, "Connected digit speech recognition system for Malayalam language," *Sadhana,* vol. 38, no. 6, December 2013. DOI: 10.1007/s12046-013-0160-2.

[3] S. Karpagavalli, R. Deepika, P. Kokila, K. U. Rani, and E. Chandra, "Isolated Tamil digit speech recognition using template-based and HMM-based approaches," in P. V. Krishna, M. R. Babu, and E. Ariwa, (eds.), *Global Trends in Information Systems and Software Applications, ObCom 2011. Communications in Computer and Information Science*, vol. 270, Springer, Berlin, Heidelberg, 2012.

[4] D. S. Kulkarni, R. R. Deshmukh, V. J. L. Patil, P. P. Shrishrimal, S. D. Waghmare, and A. M. Oirere, "Marathi isolated digit recognition system using HTK," *IJCA Proceedings on International Conference on Cognitive Knowledge Engineering, ICKE,* vol. 2016, no. 2, pp. 42-45, January 2018.

[5] B. D. Sarma, A. Dey, W. Lalhminghlui, P. Sarma, and S. R. M. Prasanna, "Robust Mizo digit recognition using data augmentation and tonal information," *9th International Conference on Speech Prosody*, Poland, 13-16 June 2018.

[6] H. Sarma, N. Saharia, and U. Sharma, "Development and analysis of speech recognition systems for Assamese language using HTK," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 1, pp. 7:1-7:14, November 2017. DOI: 10.1145/3137055.

[7] M. Sarma, K. Dutta, and K. K. Sarma, "Assamese numeral corpus for speech recognition using cooperative ANN architecture," *International Journal of Electrical and Electronics Engineering*, vol. 3, no. 8, pp. 456-465, 2009.

[8] B. D. Sarma, M. Sarma, M. Sarma and S. R. M. Prasanna, "Development of Assamese phonetic engine: Some issues," *2013 Annual IEEE India Conference (INDICON)*, pp. 1-6, Mumbai, India, 2013. DOI: 10.1109/INDCON.2013.6725966.

[9] S. Shahnawazuddin, D. Thotappa, B. D. Sarma, A. Deka, S. R. M. Prasanna and R. Sinha, "Assamese spoken query system to access the price of agricultural commodities," *2013 National Conference on Communications (NCC),* pp. 1-5, New Delhi, India, 2013. DOI: 10.1109/NCC.2013.6488011.

[10] A. Dey, S. Shahnawazuddin, Deepak K. T., S. Imani, S. R. M. Prasanna, and R. Sinha, "Enhancements in Assamese spoken query system: Enabling background noise suppression and flexible queries," *2016 Twenty Second National Conference on Communication (NCC)*, pp. 1-6, Guwahati, India, 2016. DOI: 10.1109/NCC.2016.7561193.

[11] S. Shahnawazuddin, D. Thotappa, A. Dey, S. Imani, S. R. M. Prasanna, and R. Sinha, "Improvements in IITG Assamese spoken query system: Background noise suppression and alternate acoustic modeling," *Journal of Signal Processing Systems*, vol. 88, no. 1, pp. 91-102, 2017.

[12] https://en.wikipedia.org/wiki/Assamese_language

[13] Indian Language Speech sound Label set (ILSL12) (Version 2.1.6). Available: https://www.iitm.ac.in/donlab/tts/downloads/cls/cls v2.1.6.pdf

[14] A. Y. Mon, W. Pa, and Y. K. Thu, "Building HMM-SGMM continuous automatic speech recognition on Myanmar web news," *Proc. of 15th International Conference on Computer Applications (ICCA 2017),* pp. 446-453, 2017.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, ….. K. Vesely, "The Kaldi speech recognition toolkit," *in IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2011),* p. 4, IEEE Signal Processing Society, Hawaïï, USA, 11-15 December 2011.

[16] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mf ccs/