Assamese Spoken Query System to Access the Price of Agricultural Commodities

S Shahnawazuddin, Deepak Thotappa, B D Sarma, A Deka, S R M Prasanna and R Sinha
Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati -781039, India
Email: {s.syed, deepakkt, s.biswajit, ani.deka, prasanna, rsinha}@iitg.ernet.in

***Abstract*—In this work, a spoken query system developed for accessing the price of agricultural commodities in Assamese language is described. The developed system enables the user to access the latest price of the commodity by calling the system using a landline/mobile phone. The spoken query system consists of interactive voice response (IVR) and automatic speech recognition (ASR) modules which are developed using open source resources. For the development of ASR models, the task specific speech data is collected from different dialect regions of Assam. The issues in test data adaptation are highlighted and a constrained data-unseen speaker adaptation approach is implemented which is found to give a relative improvement by 8% in baseline performance.**

Index Terms: Automatic speech recognition, spoken query system, Assamese phone set, speaker adaptation.

I. Introduction

There are over 3.8 million farmers spread across different regions of Assam and they deal with different agricultural commodities [1]. The Ministry of Agriculture, Government of India, maintains a website by the name AGMARKNET (Agricultural Marketing Information Network) [2] to facilitate the farmers about the current prices of different agricultural commodities. This website lists the maximum, minimum and modal prices of commonly grown/traded agricultural commodities for all the major states of India and are updated frequently. In India the majority of farmers are semi-literate or not computer savvy; it is hence difficult for them to browse the AGMARKNET website for the price information. With increasing teledensity in rural India [3], majority of the farmers are getting conversant with handling the phones. It is easy and cost effective to integrate the telephone network with an automatic speech recognition (ASR) system. As a result developing a spoken query (SQ) system for accessing the prices of agricultural commodities over the telephone appears more appropriate solution for this problem [4]. The SQ systems are prevalent across the world since early 1990s and are reported to be cost effective as well as easy to maintain [5]. In literature, we do find some spoken query systems reported for other Indian languages [6], [7] but not in Assamese. Although there are a few works reporting the recognition of limited set of vowels and the digits in Assamese language, to the best of our knowledge no moderate vocabulary Assamese ASR system is reported in literature. This work is the first of its kind in Assamese language targeting the practically deployable SQ system. Also there is no publicly available Assamese speech corpus large enough for training the state-of-the-art ASR systems which are based on hidden Markov models (HMM).

978-1-4673-5952-8/13/$31.00

The developed spoken query system consists of two obvious modules : an interactive voice response (IVR) and an ASR system which are developed using open source resources. For the development of ASR system, the task specific speech data was collected from different dialect regions of Assam using the same IVR module and is described in detail later.

In name recognition systems usually the user response is of a very small duration (less than 1 second) and also marked by a significant amount of silence at the beginning and end of the user input. In our case we have designed the IVR system to illicit the commodity names only from the user which are typically no longer than 1 second. Consequently adapting spoken query system to the user becomes a challenging task. Conventional MAP [8] and MLLR [9] adaptation methods fail to be effective as these require a significant amount of adaptation data. Fast adaptation techniques, like eigenvoices [10] and cluster adaptive training [11], are reported to be effective in case of small adaptation data. However these methods are found to be non-convergent when the adaptation data is extremely small, i.e., less than 1 second. It has been reported that with dynamically chosen clusters/eigenvoices [12]–[15] adaptation performance is found to be further improved. Though this adds to the latency of the system but may be feasible for low adaptation data case like name recognition systems. In this work we have also implemented a simple scheme of adaptation which is found to give significant improvement in the performance of the developed SQ system.

This paper is organized as follows: In Section II the spoken query system is described. The details of the task specific speech data collection are given in Section III. The phone set used for developing the Assamese ASR system is described in Section IV. The details of the ASR system are mentioned in Section V.The adaptation issues in such a scenario are discussed in Section VI. Finally the paper is concluded in Section VII.

II. Spoken Query System

The Spoken query (SQ) system is developed to provide a value addition to the AGMARKNET website. This website under the Assamese section lists the max, min and modal

prices of 109 different agricultural commodities spread across main *mandis* (agriculture marts) in various districts of Assam. The SQ system consists of a server running Asterisk, an ASR system and a price information database. Asterisk is an open source software on Linux/Unix platform that enables in connecting the server to the telephone network [16]. The Asterisk server in turn consists of an IVR and a computer telephone interface (CTI) card. The CTI card is connected to the integrated services digital network (ISDN) primary rate interface (PRI) digital line and is capable of supporting 30 simultaneous time division multiplexed (TDM) telephone channels. Devices such as IP phone, and mobile phones or landline can access the Asterisk server through the ISDN-PRI line. The developed system enables the user to make a query about the commodity price, records and processes that and disseminates the current price of that commodity in his/her district through a pre-recorded voice response. The block diagram of the SQ system is given in Fig. 1.

A. *Callflow*

The SQ system interacts with the user following a call-flow which consists of two major branches. In the main branch, the user is prompted to utter commodity name whose price is required to be accessed followed by the district name in which that commodity price is sought. On successful recognition of both the queries, the user is provided with the price of the asked commodity in the required district. If there is a failure in recognizing the district name in two attempts, the modal price of the asked commodity across Assam is provided. Whereas if there is a failure in recognizing the commodity in two attempts itself then the flow switches to the second branch. In this second branch, the user is prompted first to provide the district name followed by the commodity name. Two trials are provided for the recognition of the district name, whether success or failure, it moves to the query about the commodity name. Like the district name recognition, the commodity name recognition is also provided with two trials. In case of failing to recognise the commodity name again, the call-flow terminates with a sorry message. In case the district name recognition has failed but the commodity name is successfully recognised, the modal price of the asked commodity is disseminated. The detailed call-flow is shown in Figure 2. This calflow structure has been adapted a number of times based on the user feedbacks.

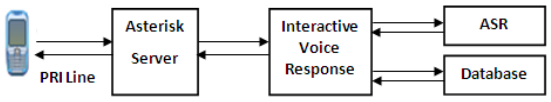

Figure 1: Spoken Query System

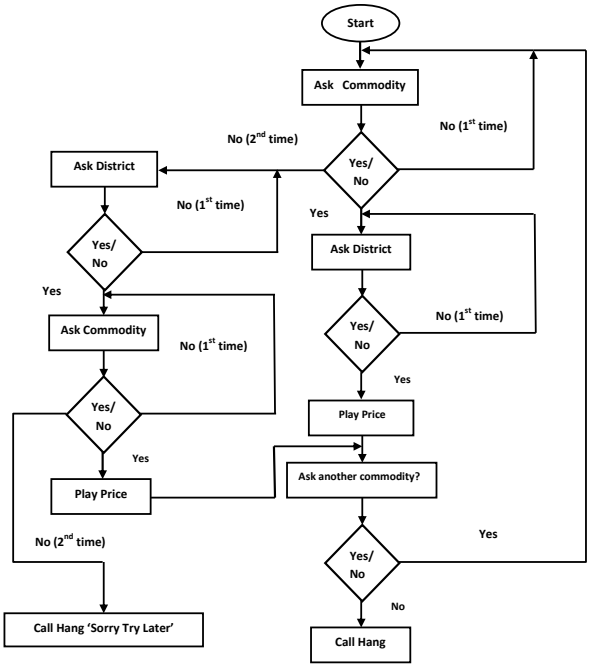

Figure 2: Callflow Structure of the SQ system

B. *Error Recovery Mechanisms*

Error recovery mechanisms are integral part of all SQ systems. In the developed SQ system the following error recovery mechanisms are adopted.

1) Voice activity detection (VAD) is tuned to detect not only no-response but also the feeble response from the user. In those cases the user is asked to give the give query or repeat the query more loudly.
2) The two branched call-flow discussed earlier which provides four trials for the commodity name recognition interspersed with district name recognition has been introduced with purpose. The interspersing of commodity name and district name recognition imparts user conditioning while reducing the monotonicity of the task.
3) Speaker is prompted to utter *hoy/nohoy* (yes/no) at each stage of the decision making based on the ASR output. The accuracy of yes/no recognition is more than 97 %.
4) To increase the confidence level in the ASR output, the user response is recognized using three parallel decoders and their outputs are polled to generate the final response.
    a) If all the three decoders are in consensus, then no confirmation is required from the user and the price of the recognized commodity is directly disseminated. It has been found that in case of consensus among the decoders the hypothesis is correct 95% of times, thus allowing us to reduce the latency of the calflow.
    b) If at least 2 decoders are having the same hypothe-

sis then the system prompts the matched hypothesis and seeks the user confirmation.

c) If none of the decoder outputs match then the system simply prompts the user to repeat the query.

5) Even though having parallel decoders reduces the system performance slightly, it does help in building confidence in the system output over the single decoder system.

C. Price Database

On the AGMARKNET website, the all India commodity prices are available in number of Indian languages but the state wise commodity prices are given in English language only. This may be big hindrance for the semi-literate farmers. So for disseminating the commodity names and their prices in Assamese, a mapping has been performed. To serve this purpose, the MySQL database management system is used. The price information is updated using a webcrawler every 24 hours. This is realized using Crontab scheduler. The webcrawler is interfaced with MySQL using python-MySQLdb connector. The query from the farmer is used to retrieve the price information that is stored in the database. The prices are disseminated through the pre-recorded messages.

III. Speech Data Collection

The collection of task specific data is done following a call-flow which is slightly different than the final call-flow shown earlier in Figure 2. In the call-flow used for the data collection, all the decision making processes were taken care by a wizard-of-Oz. The wizard-of-Oz first sets up a conference call using his/her mobile among the farmer's phone, the Asterisk server and an additional mobile phone whose purpose is discussed later. The appropriate codes in DTMF are keyed in by the wizard-of-Oz for controlling the call-flow based on the user's response to call-flow. As the mobile which sets up the conference call cannot send any DTMF input, for that purpose an additional mobile is used for keying-in the appropriate DTMF codes. The speech corpus is collected from 1159 farmers (875 males and 284 females) across different regions of Assam. It has a vocabulary of 138 words that covers the commodity enlisted on AGMARKNET website, *Hoy/Nohoy* (yes/no) for confirmation purposes and 27 district names of Assam. In addition to that four broad category names: *Phal* (fruits), *Xaisya* (cereals and pulses), *Pasoli* (vegetables), *Ainya* (all others) are also recorded into which the commodities can be classified to reduce the search space. Though the use of this classification was later dropped in final system based on user feedbacks. The following points were taken into account while collecting the data:

1) Assamese language has four major dialect groups:
   - *Eastern group* is spoken in Sibsagar, Dibrugarh, Jorhat, Golaghat, Dhemaji, Lakhimpur, Tinsukia districts.
   - *Central group* which is spoken in Nagaon, Sonitpur, Marigaon districts.
   - *Kamrupi group* spoken in Nalbari, Barpeta, Bongaigaon, Darrang, Kokrajhar districts.
   - *Goalporia group* which is spoken in Goalpara, Dhubri districts.

   The collected data is well spread across the different dialect regions of Assam to capture the pronunciation variations.
2) The data is collected keeping the sensor (different make mobile handsets) and the channel (different service providers) variabilities into consideration.
3) A male to female gender ratio of 3:1 is maintained while collecting the data.
4) With the limited vocabulary from the agricultural commodities not all context dependent phones are expected to get properly trained. To overcome this problem, an additional data of 3 hours from 25 speakers (17 males and 8 females) uttering a set of 28 phonetically balanced sentences is also collected. On adding this data to earlier collected corpus, the triphone modeling is found to be significantly improved.

The speech data corpus collected is transcribed at word level manually and cross-checked by two independent supervisors.

IV. Assamese Phone Set

The Assamese ASR system is developed by modeling the 36 phonetic symbols as shown in Figure 3. We have used Bangla symbols as defined in [17] with a few exceptions. In Assamese phone inventory, there are 8 vowels (/ɔ/, /o/, /ɑ/, /i/, /u/, /ʊ/, /ɛ/, /e/) and 23 consonants. The agricultural commodity names also contain 5 diphthongs (/oi/, /ou/, /ɑi/, /ɑu/, /ɔi/). To represent sounds /w/ and /x/ that are not present in Bangla, additional symbols w and x are used. As the vowel sound /ʊ/ is only slightly lower than /u/, they are merged to the same symbol u. On the other hand, front vowels /e/ and /ɛ/ are more spread out allowing significant height differences. Hence they are treated as two different symbols. Symbols for Assamese alveolar plosive are taken from Bangla dental plosive (t, th, d, dh, n) as Assamese does not have any dental sound.

Another phonological feature of Assamese is that it does not have retroflex sounds. Instead the language has a whole series of alveolar sounds, which includes oral/nasal stops, fricatives, laterals, approximants, flaps and trills, unlike other Indo-Aryan and Dravidian languages [18]. Affricates are also not present as distinctive sound units, but appear as allophones for some speakers. Extensive use of velar nasal /ŋ/ is one of the important features of Assamese language. Unlike other Indo-Aryan languages, this /ŋ/ can appear independent of homorganic consonants like /k/ or /g/. The voiceless velar fricative /x/ is unique in Assamese which is not present in any other Indian languages. The alveolar approximant /ɹ/ is subject to optional deletion in non-initial position, giving way to phonetic lengthening in the place of the deleted segment. e.g. [*j u r h aa t*] becomes [*j u u h aa t*] [18].

V. Automatic Speech Recognition

The speech corpus discussed in Section III, is split into orthogonal test and training sets with training set comprising of 885 speaker (669 males and 216 females) and testing set

LETTER	IPA	USED SYMBOL
অ	/ɔ/	a
	/o/	o
আ	/ɑ/	aa
ই/ ঈ	/i/	i
উ/ ঊ	/u/	u
এ	/ɛ/	e
	/e/	ee
ও	/ʊ/	u
ঐ	/oi/	oi
ঔ	/ou/	ou
ক	/k/	k
খ	/k^h/	kh
গ	/g/	g
ঘ	/g^h/	gh
ঙ	/ŋ/	ng
চ/ ছ	/s/	s
জ/ য	/z/	j
ট/ ত	/t/	t
ঠ/ থ	/t^h/	th
ড/ দ	/d/	d
ঢ/ ধ	/d^h/	dh
ণ/ ন	/n/	n
প	/p/	p
ফ	/p^h/	ph
ব	/b/	b
ভ	/b^h/	bh
ম	/m/	m
ৰ	/ɹ/	r
ল	/l/	l
ৱ	/w/	w
শ/ ষ/ স	/x/	x
হ	/h/	h
য়	/j/	y
	/ɔi/	ai
	/ɑi/	aai
	/ɑu/	aau

Figure 3: Table representing the Assamese script symbol, corresponding IPA symbol and grapheme used in ASR system. It is to note that some entries under the Assamese script are blank as there is no symbol defined for those sounds.

comprising of 275 speakers (206 males and 69 females). The splitting is done keeping the fact into consideration that we maintain roughly similar male to female ratios in both the sets. At the same time care has been taken to have enough utterances of each unique word in training as well as testing sets. The testing set is then split into two segregating the commodity and district names. The commodity set comprises of 2552 utterances spanning a set of 109 unique commodity names while the district set has 1125 utterances consisting of 27 unique district names. On an average each speaker in the training set contributes 40-50 utterances of commodity and district names while in the test set the number of utterances per speaker ranges from 1 to 30.

Table I: MFCC Feature extraction parameters

Parameter	Value	Parameter	Value
Pre-emphasis coef.	0.97	Sampling rate	8000 Hz
No. of filters in filter	22	Base features	C1-C13
Lower cut-off freq.	130 Hz	Window length	25 ms
Upper cut-off freq.	3500 Hz	Frame rate	100 Hz

Table II: Word Error Rate (WER) for ASR baseline system

Decoder	WER in %	No. of test files
Commodity	15.99	2552
District	6.13	1125

The ASR system is developed using the open-source speech recognition toolkit HTK. MFCC features of 39 dimensions, comprising of 13 base features with its first and second derivatives, are used for speech parametrization. The various parameters that are used for feature extraction are summarized in Table I. The system is trained using three state left-to-right HMMs. A cross-word triphone based modeling is performed with 16 GMM/state while 32 GMM/state is used for modeling silence. Equally likely wordnet is used as language model for all testing with separate wordnets for commodity name, district name, and confirmation (yes/no) recognition. The performance of the baseline system over the commodity and district test sets are given in Table II.

VI. Constrained Data Unseen Speaker Adaptation

In context of the system discussed in this work, both conventional and fast adaptation approaches would not be effective since we neither have sufficient amount of adaptation data and supervision about the speaker as well as hypotheses of the adaptation data available. In other words, the system has no idea of who the caller may be and at the same time what the word uttered is. We may refer to these cases of adaptation as *constrained data unseen speaker adaptation*. In such cases we can hope to adapt the ASR system to the test data based on broad acoustic similarity only and not the speaker specific similarity. As we have a large number of speakers, it would not be feasible to search for the closest speaker space without seriously affecting the latency of the system. The available training data can be clustered into acoustically similar regions using a number of techniques such vector quantization (VQ), eigen-analysis and nearest neighbor search. In this work we have used VQ of the adapted mean supervectors, commonly used in speaker recognition domain [19], for acoustic clustering.

For each of the 885 training speakers, the speaker dependent models are estimated by performing either mean-only MAP or global MLLR transform on varying mixture monophone-HMM as well as 128 mixture UBM models. Mean supervectors are then extracted from the speaker dependent models and VQ-based clustering is performed. The model parameters for the derived clusters are then obtained using MAP adaptation or 32 regression class based MLLR adaptation of triphone based

SI HMM model. For MAP adaptation based cluster models we have tried adapting a number of combination among mean, variance and mixture-weight parameters.

For determining the optimal cluster the test utterance is aligned against each of the cluster models against the hypothesis generated in the first recognition pass using the baseline SI model. The one which results in highest likelihood for the test utterance is used for decoding that test utterance to get the final hypothesis.The results of the different experiments performed exploring the type of adaptation for speaker and cluster model generation as well as the different parameters adapted for creating cluster models are given in Table III. The best performance is noted for the case when the speaker models for acoustic clustering are generated using MLLR adaptation of 8-mixture monophone HMM means while the cluster models are obtained using MAP adaption of mean and mixture-weight parameters. Compared to baseline performance a relative improvement of 8% is obtained using the constrained data unseen speaker adaptation approach. To check the performance for more refined clustering case, we have also tried 16 cluster with speaker models created using mean-only MAP adaptation of 8-mix monophone HMM and the results for the varying parameter adaptation at cluster model level are tabulated in Table IV. It is to note that performance for 16-clusters case turns out to be inferior to that of 8-clusters. At the same time cluster scoring in the former is more time consuming than the latter.

Table III: Performances for the proposed constrained data unseen speaker adaptation approach for 8 clusters case with varying adaptation parameter space as well as adaptation types.

Type of Adapt.	Param. adapted in cluster HMM	WER in %			
		Type of model used in acoustic clustering			
		UBM	Monophone HMM Model		
			1mix	8mix	16mix
MAP	Mean-Only	15.32	15.09	14.97	15.63
	Mean+Var	16.07	15.71	15.63	15.73
	Mean+Mixwt	15.36	14.81	15.00	14.77
	Mean+Var+Mixwt	16.02	15.39	15.39	15.16
MLLR	Mean-Only	-	15.01	14.85	15.08
MLLR	Mean+Mixwt MAP	-	14.81	14.73	14.93

VII. Conclusion

The development of a spoken query system for accessing the price of agricultural commodities in Assamese language is described in detail in this work. Although the system is developed using open source tools available following the standard procedures, there are some contributions made in terms of Assamese phone set definition for ASR and the collection of the task specific Assamese speech data. We have also explored a constrained data unseen speaker adaptation approach which is found to give 8% relative improvement in the baseline performance.

Acknowledgment

This work is supported by the ongoing grant no. 11(12)/2009-HCC(TDIL) from the Department of Information Technology, Govt. of India.

Table IV: Performances for proposed adaptation approach for 16 clusters case for contrast purpose.

Type of Adaptation	WER in %
Mean-Only	15.67
Mean+Var	16.18
Mean+Mixwt	15.36
Mean+Var+Mixwt	15.71

References

[1] Assam Small Farmers Agri-Business Consortium, http:// assamagribusiness.nic.in.

[2] Agricultural Marketing Information Network - AGMARKNET, http://agmarknet.nic.in.

[3] India Telecom Online, http://www.indiatelecomonline.com.

[4] P. Kotkar, W. Thies, and S. Amarsinghe, "An audio wiki for publishing user-generated content in the developing world," in *HCI for Community and International Development*, Florence, Italy, April 2008.

[5] L. R. Rabiner, "Applications of Speech Recognition in the Area of Telecommunications," *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 501–510, 1997.

[6] S. Goel and M. Bhattacharya, "Speech based dialog query system over asterisk pbx server," in *2nd International Conference on Signal Processing Signal Processing Systems (ICSPS)*, Dalian, Jul 2010.

[7] M. Prabhaker, "Tamil market: A spoken dialog system for rural india," in *ACM CHI Conference*, April 2006.

[8] J. L. Gauvain and C. H. Lee, "Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[9] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[10] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[11] M. Gales, "Cluster adaptive training of hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, Julu 1999.

[12] B. Mak, T.-C. Lai, and R. Hsiao, "Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers," in *Proc. ICASSP*, 2006.

[13] T. Cai and J. Zhu, "A novel method for rapid speaker adaptation based on support speaker weighting," in *Proc. ICASSP*, 2005, pp. 993–996.

[14] J. Duchateau, T. Leroy, K. Demuynck, and H. V. Hamme, "Fast speaker adaptation using non-negative matrix factorization," in *Proc. ICASSP*, 2008, pp. 4269–4272.

[15] Y. Gomez, T. Toda, H. Saruwatari, and K. Shikano, "Improving rapid unsupervised speaker adaptation based on hmm-sufficient statistics," *ICASSP*, pp. 1001–1004, 2006.

[16] Asterisk open source communications, http://www.asterisk.org/home.

[17] M. M. Hasan, F. Hassan, G. M. M. Islam, M. Banik, M. R. A. Kotwal, S. M. M. Rahman, G. Muhammad, and N. H. Mohammad, "Bangla triphone hmm based word recognition," in *Circuits and Systems (APCCAS)*, Kuala Lumpur, 2010.

[18] K. D. Mousmita Sarma and K. K. Sarma, "Assamese Numeral Corpus for Speech Recognition using Cooperative ANN Architecture," *International Journal of Electrical and Electronics Engineering*, vol. 3, p. 456, 2009.

[19] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.