The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Development of Assamese Continuous Speech Recognition System

Barsha Deka[1,2], *S.R. Nirmala*[1], *Samudravijaya K.*[2]

[1]Department of Electronics and Communication Engineering,
Gauhati University Institute of Science and Technology,
Gauhati University, Guwahati-781014, India
[2]Centre for Linguistic Science and Technology,
Indian Institute of Technology Guwahati, Guwahati-781039, India
{barsha.deka4, nirmalasr3, samudravijaya}@gmail.com

Abstract

This paper describes the development of a continuous speech recognition system for Assamese, an under-resourced language of North-East India. The Speech corpus used in this work consists of 5658 spoken utterances collected from 27 speakers over telephone channel. The baseline speech recognition system was implemented using conventional hidden Markov model in conjunction with Gaussian mixture model, employing Mel-frequency cepstral coefficients as features. ASR systems using subspace Gaussian mixture model and deep neural networks together with hidden Markov model were implemented. The systems were evaluated with 3-fold cross validation method. The average word error rate of the best ASR system is 4.3%.

Index Terms: ASR, Assamese language, GMM-HMM, SGMM-HMM, DNN-HMM

1. Introduction

Over the past two decades, development of speech recognition systems for under-resourced languages has been an active area of research in the speech community. Languages spoken in India belong to several language families. 75% Indians speak Indo-Aryan languages, 20% speak Dravidian languages while the other languages belong to the Austro-Asiatic, Sino-Tibetan, Tai-Kadai, and a few other minor language families and isolates. Many Indian languages have little written or spoken language resources on the digital platform. Assamese is one such under-resourced language spoken by over 15 million people in North-East India. Assamese belongs to the Indo-Aryan language family, and is the official language of the state of Assam.

As per the 2011 census, there are 121 languages and 270 mother tongues in India [1]. But only a few major Indian languages have been explored for the development of speech recognition systems. Melvin *et al.* reported a large vocabulary Continuous Speech Recognition (CSR) system for Tamil language [2]. The authors reported syllable error rate of 17.44% on read mode newspaper speech. In [3], CSR systems in two Indian languages, Tamil and Telugu are reported with syllable recognition accuracies of 43.3% and 32.9% respectively on Indian television news bulletins. A Bengali language CSR system that used monophone level acoustic models was reported with an accuracy of 71.6% [4].

In case of Assamese language, a few speech recognition systems have been reported. Mousmita *et al.* [5] reported a work on Assamese numeral recognition system using isolated digits. A digit recognition accuracy of 98% was achieved by an ASR system that used recurrent neural networks. Biswajit *et al.* [6] reported the development of an Assamese Phonetic Engine using speech collected in reading, lecture and conversation modes. The phone recognition accuracies of the system were 47.3%, 45.3% and 36.1% respectively. Shahnawazuddin *et al.* [7, 8] reported work on a spoken query system developed for accessing the price of agricultural commodities in Assamese language. This isolated word/phrase recognition system was trained using speech data recorded over landline and mobile phone channel. Recently, Sarma *et al.* [9] developed an ASR system for Assamese language using HTK, a hidden Markov model toolkit [10]. The authors used deep neural net for modeling context dependent acoustic units. The accuracy of word level transcription was 78.0%.

Here, we report the development of an ASR system that can recognise Assamese sentences spoken in a continuous manner. The system uses Hidden Markov Model (HMM) to model temporal variations speech signal. The best version of the system uses Deep Neural Network (DNN) to estimate the posterior probabilities of context dependent acoustic units generating feature vectors derived from short-time signal processing of input speech. The system was implemented using an open source toolkit, kaldi [11].

The rest of the paper is organized as follows. Information about the phone level units modeled by the ASR system, and the linguistic resources used to train and test the ASR system is given in Section 2. The details of the experimental setup is given in Section 3. The results of experiments are discussed and compared with those of other systems in Section 4. A few concluding remarks are given in Section 5.

2. Linguistic Resources

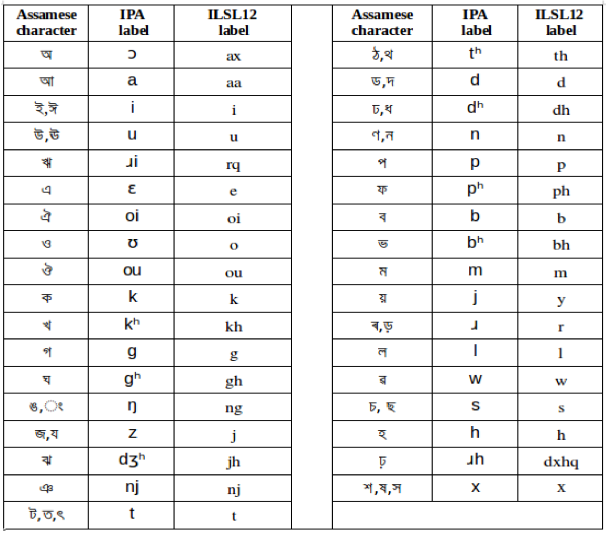
This section presents information about the acoustic units of Assamese language for which statistical models were trained. In addition, a brief description of the written and spoken linguistic resources created for training and evaluating the ASR system is also given.

2.1. Acoustic units

The acoustic units employed in this work are listed in Table 1 in 3 different representations. The 3 columns show the Assamese character, the corresponding IPA symbol, and its ASCII representation following the Indian Language Sound Label (ILSL12) convention [12]. ILSL12 is a common label set, for representing phone level units of many Indian languages, being used by many ASR researchers in India. The pronunciation dictionary, created in this work, follows the ILSL12 convention. The As-

10.21437/SLTU.2018-46

Table 1: *The list of acoustic units in 3 notations: Assamese script, IPA, ILSL12*



Assamese character	IPA label	ILSL12 label	Assamese character	IPA label	ILSL12 label
অ	ɔ	ax	ঠ,থ	t^h	th
আ	a	aa	ড,দ	d	d
ই,ঈ	i	i	ঢ,ধ	d^h	dh
উ,ঊ	u	u	ণ,ন	n	n
ঋ	ɹi	rq	প	p	p
এ	ɛ	e	ফ	p^h	ph
ঐ	oi	oi	ব	b	b
ও	ʊ	o	ভ	b^h	bh
ঔ	ou	ou	ম	m	m
ক	k	k	য়	j	y
খ	k^h	kh	ৰ,ড়	ɹ	r
গ	g	g	ল	l	l
ঘ	g^h	gh	ৱ	w	w
ঙ,◌ং	ŋ	ng	চ, ছ	s	s
জ,য	z	j	হ	h	h
ঝ	$dʒ^h$	jh	ঢ়	ɹh	dxhq
ঞ	nj	nj	শ,ষ,স	x	x
ট,ত,ৎ	t	t			

samese label set, used in this work, represents 46 phonemes: 10 vowels and 36 consonants.

2.2. Text corpus

The sentences for the text corpus were drawn from different sources such as story books, articles, online newspapers. The length of the selected sentences ranged from 5-10 words. In addition, proverbs and digit sequences, designed to contain all digit pairs [13], were also included. The text corpus used in this work contains 1000 unique sentences comprising of 2777 unique words. The set of 1000 sentences were grouped in 50 different subsets, each containing 20 sentences. Each subset contains 1 digit sequence, 4 proverbs and 15 sentences randomly selected from various sources in proportions as shown in Figure 1.

2.3. Speech Corpus

A short description of the speech corpus used in this work is given here; a detailed description is given in [14]. The speech corpus contains speech data collected from 27 speakers belonging to different age groups. Out of 27 speakers, 24 were native speakers of Assamese. The mother-tongue of 3 speakers is Bengali; however, these 3 persons could read and speak Assamese fluently. The data was collected over mobile telephone channel using an interactive voice response system configured with Asterisk software. The speakers were asked to call the voice-server using their own mobile phones and to respond to predefined prompts. Each speaker was provided with a printed sheet that contained a set of 20 sentences to be read, one after another after a beep. Speakers contributed speech data on a voluntary basis without any remuneration. They were asked to read as many sentence sets as possible. So, the number of sentences (and speech files) per speaker varied from 20 to 500. The speech corpus used in this work contains a total of 5658 files, each containing one spoken sentence. The numbers of occurrence of vowels and consonants in the Assamese speech corpus are 67899 and 107783 respectively.

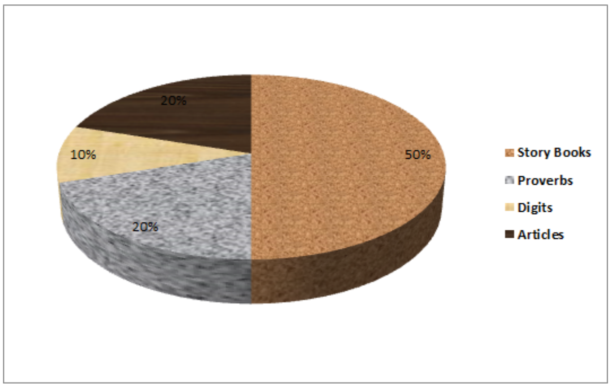


Figure 1: *Distribution of the sentences from different sources*

Table 2: *Information about training and test data of 3-fold experiments.*

No. of	Fold 1		Fold 2		Fold 3	
	Train	Test	Train	Test	Train	Test
Utterances	3759	1899	3759	1899	3798	1860
Speakers	18	9	19	8	17	10
Male	9	4	9	4	8	5
Female	9	5	10	4	9	5

In this work, we evaluated the performance of the system using the k-fold validation methodology. Due to small size of the speech database, we set k=3. The speech corpus was divided into three subsets, each set containing approximately equal number of speakers, and equal number of speech files. While the set of speakers in the 3 subsets are mutually exclusive, there is a large degree of overlap in the sentences read by the speakers in these 3 subsets. The 3 subsets correspond to the test sets in 3 folds as shown in Table 2. Specifically, subsets 1, 2, and 3 consists of data from 9, 8 and 10 speakers respectively such that each subset contains nearly equal number of speech files. In 'Fold 3' experiment, the first two subsets are used as train data, and the remaining subset is used as test data. The Word Error Rates (WER) of the trained system when fed with train data set and test data set were computed. This procedure is repeated for the other two folds. In this work, we evaluate the WER of the system when fed with test data for the 3 folds, and report the average of the 3 WER values.

3. Experimental Setup

This section discusses the experimental setup of this work. All the experiments were conducted using Kaldi speech recognition toolkit [11]. Default values of the parameters were used in most cases.

Mel Frequency Cepstral Coefficients (MFCC) [15] were used for capturing the vocal tract information. The MFCC feature vectors were extracted from speech segments of 25ms duration, and successive frames were shifted by 10ms. Hamming window and pre-emphasis factor of 0.97 were used. In order to capture the dynamic characteristics of the speech signal, the 13 dimensional base MFCC feature vectors were appended with the velocity (Δ) and the acceleration ($\Delta\Delta$) components. The resultant 39 dimensional feature vectors were used for training the context-independent (monophone) GMM-HMM model, called **Mono** henceforth. For each of the 46 acoustic phonetic units, a 3-state left-to-right HMM model

Table 3: *Specifications of the DNN-HMM system*

Parameter	Specification
No. of hidden layers	4
No. of epochs	20
Dimension of hidden layer	1024
Mini batch size	128
Initial learning rate	0.0015
Final learning rate	0.0002

was trained. Using the trained monophone models, the phone boundaries were re-aligned. The information is used to initialize the context-dependent (triphone) GMM-HMM models. The resultant trained model is labeled as **Tri1**. In order to get optimum performance, the number of senones (shared states) was varied from 100 to 1000 in steps of 100, fixing the number of Gaussians per state as 16. The performance was optimal for 200 senones. Using the optimal Tri1 model, the boundaries of triphones were adjusted using Viterbi force-alignment. The 13 dimensional static MFCC features were spliced in four frames to the left and four frames to the right, thereby resulting in 117 (13 X 9) dimensional feature vector whose dimensionality was reduced to 40 using linear discriminant analysis (LDA) [16, 17]. The resulting trained model is referred to as **Tri2**. Using Maximum Likelihood Linear Transform (MLLT) [18, 19], the feature vectors are further decorrelated. The number of senones was varied from 100 to 1000 in steps of 100, keeping the number of Gaussians per state constant as 16. The triphone boundaries were again realigned using the optimal model. In this variant of GMM-HMM system, speaker adaptive training [20] was performed by normalizing the feature vectors using Feature space Maximum Likelihood Linear Regression (fMLLR) [21]. This model is referred to as **Tri3** model. Using the Tri3 model, the phone boundaries are further realigned and fed as input to the SGMM-HMM [22] system. In SGMM-HMM system, a globally shared model subspace is used to estimate the parameters of the statistical model of each and every state. This reduces the total number of parameters to be estimated, and makes it possible to learn the model parameters with a limited amount of training data. We also implemented the system following the DNN-HMM [23] acoustic modeling approach. A feed-forward deep neural network was trained using multiple hidden layers that takes time-spliced feature vectors with LDA+MLLT+fMLLR as input, and computes the posterior probabilities over HMM states as output. The specification of the parameters used in training the DNN-HMM system is shown in Table 3.

4. Results and Discussion

The baseline ASR systems were developed employing 3 different acoustic modelling techniques as discussed in section 3. The linguistic knowledge was captured by bi-gram language models. For each fold of k-fold evaluation, separate bi-gram language model was learned using the respective training transcript. Here, the performance of an ASR system is measured in terms of Word Error Rate (WER), defined as

$$WER(\%) = 100 * (D + S + I)/N$$

where D is the number of word deletions, S is the number of word substitutions, I is the number of words inserted by the decoder and N is the number of the words in the reference transcription [11]. The performances of the baseline systems, measured in terms of WER (in %), are shown in Table 4. For each k-fold dataset, WER was computed for different ASR systems.

Table 4: *WER (in %) of the Assamese ASR systems for three different kinds of acoustic models, for test data. The WERs for the 3 folds and the average WER are shown.*

Acoustic Model	% WER			Avg . WER
	Fold 1	Fold 2	Fold 3	
GMM-HMM (Tri 3)	7.99	8.52	7.60	8.03
SGMM-HMM	7.04	7.24	5.28	6.58
DNN-HMM	5.93	6.41	4.62	5.65

The average WER of GMM-HMM, SGMM-HMM and DNN-HMM systems are 8.03%, 6.58% and 5.65% respectively.

During data collection, the duration of speech recording was set to a fixed duration that was long enough for any speaker to read every sentence completely. Consequently, the speech files contained varying lengths of silences at either end of the files. A second set of experiments were performed after removing the end silences from the speech files using energy based end-point detection method. The WERs of the systems trained and tested with speech data after removing the end silences, for 3-folds, are shown in Table 5. The average WER of GMM-HMM, SGMM-HMM and DNN-HMM systems are 5.74%, 5.30% snf 4.29% respectively.

Table 5: *% WER of the Assamese ASR systems for three different kinds of acoustic models for test data after the removal of long end silences*

Acoustic Model	% WER			Avg . %WER
	Fold 1	Fold 2	Fold 3	
GMM-HMM (Tri 3)	6.66	5.57	5.00	5.74
SGMM-HMM	6.38	5.04	4.49	5.30
DNN-HMM	4.94	4.62	3.32	4.29

By comparing the average WER figures in Table 4 with those in Table 5, one can note that reductions of 2.29%, 1.28% and 1.36% in WER were achieved by GMM-HMM, SGMM-HMM and DNN-HMM systems respectively thanks to the removal of long silences from the speech files. This observation is graphically shown in Figure 2.

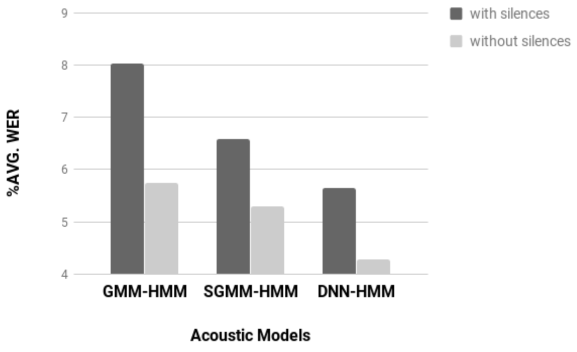


Figure 2: *Word Error Rates of Assamese ASR systems before and after removal of long end silences*

4.1. Comparison with other Assamese ASR systems

A summary of several systems for recognizing Assamese speech as a sequence of linguistic units such as phone or word is given in Section 1. Probably, the Assamese ASR system closest to the current ASR system in terms of system features is the one reported by Sarma *et al.* [9]. They developed an ASR system for Assamese language using HTK toolkit [10]. The authors used deep neural net for modeling context dependent acoustic units, similar to the current work. The number of phone level units employed by Sarma et al. was 38; the corresponding figure is 48 in the current work. The accuracy of word level transcription obtained by them was 78.0%. This corresponds to a WER of 22% in contrast to 6% WER in the current work. Sarma *et al.* used 3 hours of speech data (527 files) for training and about 0.5 hours of speech data (127 files) for testing the ASR system. This is about an order of magnitude smaller than the resources used in this work: about 3800 files for training and about 1900 files for testing. DNN is known to be data hungry; this could be one of the reasons for the better performance of the ASR system reported here.

Another DNN based ASR system for Assamese language is based on speech data released through IARPA Babel project [24]. Using the language packs distributed by IARPA, ASR systems in several languages, including Assamese, were developed by the Cambridge University speech group as part of the Lorelei team co-ordinated by IBM [25]. The ASR systems were based on bottle neck multi layer perceptrons. The performance of ASR system was measured in terms of Token Error Rate (TER), a measure similar to WER. Bigram language model was used, as in the current work. The lowest TER was achieved for Confusion Network Combination (CNC) system that combines the confusion networks generated by the Tandem HMM-GMM and Stacked Hybrid systems. The lowest TERs for the CNC system were 64.3% and 52.8% corresponding to limited (approximately 10 hours) and Full (approximately 80 hours) pack of speech data respectively. However, it is not fair to compare these TER figures with the WERs of the present system because the data distributed by IARPA is very challenging and close to real-life situation. In contrast, the speech corpus used here is scripted/read speech.

5. Conclusion

We reported the development of an ASR system for Assamese language using limited linguistic resources. Several versions of ASR system were implemented that used GMM-HMM or SGMM-HMM or DNN-HMM models to represent the acoustic-phonetic units. The best performance was obtained by DNN-HMM model with word recognition accuracy of about 95.7%, a figure better than those of the systems reported in the literature. Using larger amount of speech data collected under real-life conditions, practical ASR systems can be developed with high recognition rates that can serve as the backbone of different ASR applications.

6. References

[1] *Office of the Registrar General & Census Commissioner India*, http://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf.

[2] M. J. J., N. T. Vu, and T. Schultz, “Initial experiments with tamil lvcsr,” in *2012 International Conference on Asian Language Processing*, Nov 2012, pp. 81–84.

[3] T. Nagarajan, H. A. Murthy, and N. Hemalatha, “Automatic segmentation and labeling of continuous speech without bootstrapping,” in *2004 12th European Signal Processing Conference*, Sept 2004, pp. 561–564.

[4] P. Banerjee, G. Garg, P. Mitra, and A. Basu, “Application of triphone clustering in acoustic modeling for continuous speech recognition in bengali,” in *2008 19th International Conference on Pattern Recognition*, Dec 2008, pp. 1–4.

[5] M. P. Sarma and K. K. Sarma, “Assamese numeral speech recognition using multiple features and cooperative lvq-architectures,” *International Journal of Electrical Electronics Engineering*, vol. 5, p. 27, February 2011.

[6] B. D. Sarma, M. Sarma, M. Sarma, and S. R. M. Prasanna, “Development of assamese phonetic engine: Some issues,” in *2013 Annual IEEE India Conference (INDICON)*, Dec 2013, pp. 1–6.

[7] 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,” in *2013 National Conference on Communications (NCC)*, Feb 2013, pp. 1–5.

[8] A. Dey, S. Shahnawazuddin, D. K. T., S. Imani, S. R. M. Prasanna, and R. Sinha, “Enhancements in assamese spoken query system: Enabling background noise suppression and flexible queries,” in *2016 Twenty Second National Conference on Communication (NCC)*, March 2016, pp. 1–6.

[9] H. Sarma, N. Saharia, and U. Sharma, “Development and analysis of speech recognition systems for assamese language using HTK,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 17, pp. 7:1–7:14, October 2017.

[10] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi Speech Recognition Toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[12] *Indian Language Speech sound Label set (ILSL12)*, https://www.iitm.ac.in/donlab/tts/downloads/cls/cls_v2.1.6.pdf.

[13] V. Chourasia, K. Samudravijaya, and M. Chandwani, “Phonetically rich hindi sentence corpus for creation of speech database,” *Int. Symp. on Speech Technology and Processing Systems and Oriental COCOSDA-2005*, pp. 132–137, 2005.

[14] B. Deka, J. Chakraborty, A. Dey, S. Nath, P. Sarmah, S.R.Nirmala, and SamudraVijaya, “Speech corpora of under resourced languages of north-east india,” in *Oriental COCOSDA, 2018, Miyazaki, Japan*, May 2018.

[15] S. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[16] R. Haeb-Umbach and H. Ney, “Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition,” in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP’92. Washington, DC, USA: IEEE Computer Society, 1992, pp. 13–16.

[17] H. Abbasian, B. Nasersharif, A. Akbari, M. Rahmani, and M. Moin, “Optimized linear discriminant analysis for extracting robust speech features,” in *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*. IEEE, 2008, pp. 819–824.

[18] M. Gales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition,” *Computer Speech & Language*, vol. 12, no. 2, pp. 75 – 98, 1998.

[19] R. A. Gopinath, “Maximum likelihood modeling with gaussian distributions for classification,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 661–664.

[20] T. Anastasakos, J. McDonough, and J. Makhoul, “Speaker adaptive training: A maximum likelihood approach to speaker normalization,” in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’97)-Volume 2 - Volume 2*, ser. ICASSP ’97. Washington, DC, USA: IEEE Computer Society, 1997, pp. 1043–.

[21] C. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[22] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafit, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, “Subspace Gaussian Mixture Models for Speech Recognition,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 4330–4333.

[23] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[24] *M. Harper, IARPA Babel Program*, http://www.iarpa.gov/Programs/ia/Babel/babel.html.

[25] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, “Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED,” *SLTU 2014*, pp. 16–23, 2014.