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

Language Identification of Assamese, Bengali and English Speech

Joyshree Chakraborty[1,2], *Shikhamoni Nath*[2], *S R Nirmala*[1], *Samudravijaya K*[2]

[1]Department of Electronics and Communication Engineering,
GUIST, Gauhati University-781014, India
[2]Centre for Linguistic Science and Technology,
Indian Institute of Technology Guwahati, Guwahati-781039, India
joyshreemlg@gmail.com, shikhanath2034@gmail.com, nirmalasr3@gmail.com, samudravijaya@gmail.com

Abstract

Machine identification of the language of input speech is of practical interest in regions where people are either bilingual or multi-lingual. Here, we present the development of automatic language identification system that identifies the language of input speech as one of Assamese or Bengali or English spoken by them. The speech databases comprise of sentences read by multiple speakers using their mobile phones. Kaldi toolkit was used to train acoustic models based on hidden Markov model in conjunction with Gaussian mixture models and deep neural networks. The accuracy of the implemented language identification system for test data is 99.3%.

Index Terms: automatic language identification, HMM, GMM, DNN, Assamese, Bengali.

1. Introduction

The objective of an automatic language identification (LID) system is to identify the language of the given speech signal. LID systems are technologically important as they facilitate spoken interaction of multilingual customers with online customer service/query systems. The query could be responded to by either human agents or speech recognition systems who could listen and respond in the language of the customer as identified by the LID system.

Spoken language technology for under-resourced languages is receiving more attention of the speech research community recently [1]. India is home to 121 languages and 270 mother tongues, each spoken by more than 10,000 speakers [2]. Most languages do not have sufficient linguistic resources for development of spoken language technologies. Majority of the endangered languages of India are spoken in north eastern part of India [3]. Assamese and Bengali are two major languages spoken in this region. Here, we describe the development of an LID system that can recognise the language of an utterance as one of Assamese, Bengali and English (spoken by native speakers of Assamese or Bengali).

The rest of the paper is organised as follows. Section 2 gives a short overview of LID approaches and a description of the linguistic resources collected for the LID task. An outline of the signal processing used to extract features from speech, and the statistical models used to represent speech units is provided in section 3. The results of experiments conducted are presented and discussed in section 4. Section 5 presents the conclusions of the work.

2. Linguistic resources

An overview of recent approaches to spoken language identification systems is given in [4]. A list of prior works following explicit and implicit LID approaches is given in [5]. The latter also lists previous LID works in the Indian context as well as usage of source features in addition to vocal tract features for LID.

The LID system described here follows the explicit LID approach, wherein the input speech is decoded in terms of linguistic units by Automatic Speech Recognition (ASR) systems of the relevant languages. The approach followed here is similar to the ‘parallel phone recognizer followed by language modeling’ approach. This approach uses a front end of language dependent phone recognizers that capture the acoustic characteristics of phonemes of the individual languages. The sequence of phonemes generated by the phone recognizer of a language reflect the phonotactic rules of the language. The latter can be modeled by a suitable language model. As a preliminary step, the LID system described here, however, uses traditional speech recognition systems to jointly exploit acoustic and linguistic knowledge.

2.1. Spoken language resources

The training of a speech recognition system not only requires a collection of speech data files, but also the corresponding transcriptions at word level as well as a pronunciation dictionary. These linguistic resources are to be created for each language of interest.

Assamese is the official language of the state of Assam, and is spoken by over 15 million native speakers. Bengali is the official language of the state of West Bengal. These two languages are major languages of North-East India. English is the medium of instruction of higher education in India. So, English is spoken by a significant number of Indians, and sometimes is used as a language for communication between persons whose mother tongues are different. So, we also recorded spoken English from native speakers of Assamese and Bengali. A detailed description of the database of these 3 languages is given in [6]. Salient features of the text and speech corpora are presented in the next subsections.

2.1.1. Text corpora

The text corpora of the 3 languages consists of sentences. Each sentence comprises of 5 to 10 words. The Assamese text corpus contains 1000 unique sentences with a vocabulary of 2704 unique words. There are 400 distinct sentences each in Bengali and English languages text corpora. Each text corpora

10.21437/SLTU.2018-37

was organised in sets of 20 sentences. Each set contained sentences from various online sources, proverbs and digit sequences.

2.1.2. Recording of speech

Subjects were asked to read sentences printed on paper after calling a toll-free number. Narrow band speech data was stored in a voice server. All the speakers were residents of Guwahati city. Metadata of speakers were recorded in DTMF mode. Table 1 shows a subset of metadata.

Table 1: *Information about the speakers of the speech databases.*

Language	Mother Tongue	Number of Speakers		
		Male	Female	Total
Assamese	Assamese	14	10	24
	Bengali	0	3	3
Bengali	Bengali	8	13	21
English	Assamese	2	4	6
	Bengali	3	8	11

Speakers were asked to read as many sets of sentences as possible. So, the number of speech files per speaker varies from 20 to 400. The total number of speech files in Assamese, Bengali and English speech databases are 5658, 2500 and 2500 respectively.

2.1.3. Transcription and lexicon

The speakers were assumed to read the sentences faithfully. Thus, the word level transcriptions corresponding to speech files are expected/ideal word sequences. No markers were added corresponding to extraneous sounds such as cough, babble etc. as presence of such noise markers seemed to have no beneficial effect on recognition accuracy in a similar ASR system of Marathi language [7]. Pronunciation dictionaries were manually created following the ILSL12 convention for labels for phone level linguistic units [8]. The number of phone labels were 46, 63 and 49 in the dictionaries of Assamese, Bengali and English. The number of unique words in Assamese, Bengali and English lexicon were 2704, 1579 and 1145 respectively.

3. Experimental details

The details of signal processing, models and the toolkit used to implement the LID system are given in this section.

3.1. Features and models

The long silence segments at either end of the speech files were detected using a simple energy based silence detector; long silences were truncated to 0.3 sec. Feature vector is based on 13 Mel frequency cepstral coefficients (MFCC) [9] and their time derivatives. Phones were modeled by hidden Markov models (HMM) with 3 emitting states [10]. The distribution of features in each state was modelled by a Gaussian mixture model (GMM), The number of Gaussian mixtures in the probability density function of a state depended on the occupation count of the state. The upper limit for the total number of Gaussian mixtures in the model set of a language was set to 1000.

3.2. Kaldi toolkit

We used Kaldi as the toolkit for implementing the speech recognition system [11]. Bigram language model was used to capture the syntactic constraints of the language.

Speech data was divided into two parts: one for training models and another for testing. The test data sets of Assamese, Bengali and English languages comprised of 840, 500 and 500 speech files respectively. The rest of the files were used to train acoustic models. It was ensured that the sets of speakers in the test and train set were mutually exclusive. This was done to avoid speaker characteristics influencing the decision of the LID system.

The bigram language model was trained using the transcripts of train data alone. However, since a sentence was read by multiple speakers, the language model thus trained was adequate to recognise test speech.

4. Results and discussion

In this section, we report the performance of speech recognition systems of the 3 languages first. Then, we report the performance of language identification system based on these speech recognition systems.

4.1. Speech recognition systems

Different types of acoustic models can be trained using Kaldi toolkit. When a HMM models a context independent phone, it is denoted by 'Mono' in this paper. The phonetic context dependent models are called triphones. Three major types of triphone models can be trained using Kaldi. The plain-vanilla triphone model based on MFCCs and their first and second order derivatives is denoted by 'Tri1'. Further, a revised feature vector is computed as follows. The static MFCC feature vectors are spliced in time with 4 frames to the right and 4 frames to the left of the central frame thereby making the feature vector dimension of 117 (13x9). The dimensionality of the resultant feature vector is reduced to 40 by carrying out linear discriminant analysis. The features are further decorrelated using maximum likelihood linear transformations of full-covariance matrices shared among phone classes. The triphone models trained with the resultant feature vectors are denoted as 'Tri2'. When the triphone models are trained with features that have undergone (speaker dependent) feature maximum likelihood linear regression in addition to linear discriminant analysis and maximum likelihood linear transformation, the resultant models are denoted as 'Tri3'. The kaldi can train HMM models whose emission probabilities can be estimated either by a subspace GMM or by employing a deep neural net. These models are denoted as 'sGMM' and 'DNN' respectively.

The performance of speech recognition system can be evaluated in terms of Word Error Rate (WER), defined as

$$\text{WER (\%)} = 100 * (D+S+I) / N \quad (1)$$

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].

During the testing phase, a trained ASR system was fed with test data. The output of the decoder was compared with the reference transcription in order to compute the errors at word level. The WERs of the 3 ASR systems employing various types of acoustic models are shown in Table 2. All ASR systems used bigram language models. HMM models the temporal properties of the speech signal in all cases. Different acoustic models differ (i) by whether HMM models a context independent unit (Mono) or otherwise, and (ii) by the method used to estimate the likelihoods of a test feature vector corresponding to various states of HMMs. GMM is used for estimating the likelihoods in all but the last case. The last model uses deep neural network to estimate the posterior probabilities of the states, from which the likelihoods are estimated. The 3 versions of triphone models (Tri1, Tri2 and Tri3) differ by the details of feature transformation and speaker adaptive training. As expected, the WER decreases with increasing sophistication of feature extraction/transformation, and of acoustic models. The word error rates of the 3 ASR systems with the best acoustic model (DNN) varies in the range 2% to 4%.

Table 2: *Word Error Rates of speech recognition systems employing different acoustic models*

Acoustic Models (HMM)	WER(in %)		
	Bengali	Assamese	English
Mono	4.7	5.2	11.9
Tri1	3.8	6.0	6.4
Tri2	4.5	6.1	8.1
Tri3	3.4	4.7	4.7
sGMM	3.1	3.8	4.3
DNN	2.1	2.7	4.0

4.2. Language Identification Systems

In this section, we report the performance of 2 LID systems, using two types of acoustic models: (i) HMM represents a context independent phone (Mono) and a GMM models the emission probability density function of a state, (ii) HMM represents a context dependent phone and a DNN is used to estimate the likelihood of a feature vector being emitted from a state.

4.2.1 LID using monophone GMM-HMM model

A baseline LID system was built based on the individual ASR systems, employing GMM-HMM acoustic models, of the 3 languages as described in section 4.1. The ASR system of a language utilizes not only properties of the language dependent phone set but also the lexical and syntactic constraints of the language. Hence, the likelihood of a given ASR system generating/matching a speech signal of its native language is likely to be higher than that of an alien language. Hence, the Maximum Likelihood criterion was used to identify the language of a test speech by comparing the log-likelihoods of the test speech generated by the models of the 3 ASR systems.

Table 3 shows the confusion matrix of the baseline LID system employing GMM-HMM. The rows show the correct language of test speech. Various columns denote the language of test data as identified by the LID system. For example, 99.4% of the Assamese test files were correctly identified as containing Assamese speech. The language of the remaining 0.6% of the Assamese test files were mis-identified as Bengali. Similarly, all (6.6%) mis-identified Bengali speech files were labeled as Assamese. None of the Assamese or Bengali speech files were classified as English. This result is expected, and is satisfying as Bengali and Assamese are close languages whereas English is a very distant cousin of theirs. The overall accuracy of the LID system is 95.1%.

Table 3: *The confusion matrix of the LID system based on monophone GMM-HMMs. The diagonal entries are measures of the performance of the LID system.*

	Recognised Language (%)		
True Language	Assamese	Bengali	English
Assamese	**99.4**	0.6	0
Bengali	6.6	**93.4**	0
English	7.2	0.4	**92.4**

The accuracy with which the Assamese files are correctly recognised is significantly higher than that for Bengali or English. This can be attributed to higher amount of labelled data that was available to train the Assamese ASR system. The English and Bengali ASR models were trained with 2000 speech files each whereas the Assamese acoustic model was trained with 4818 files, larger by a factor of nearly 2.5. ASR experiments with other databases showed similar trend of higher accuracy with larger amount of data.

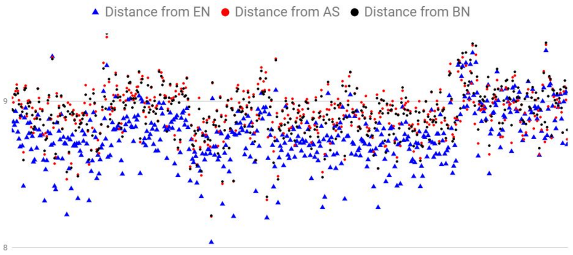

Figure 1: *Distances of English 500 test speech data from the 3 language models. The distances are smaller from the English (blue triangles) model than the others.*

In Figure 1, the distances (negative log likelihoods) of test English files from the 3 ASR models are plotted for all 500 test English speech files. The blue triangle, the red circle and the black circle denote the distances of a test English file from the English, Assamese and Bengali ASR models respectively. Most blue triangles are located lower in the graph indicating that the test files whose actual language is English are closer to English ASR model than to other language ASR models.

Figure 2 shows this trend more clearly. On the y-axis are the differences between the distances of test English data from English ASR model and non-English ASR models. In particular, a blue triangle shows such a difference between distance from the English model and the Assamese model for one test English speech file. It is computed as d_EN - d_AS, where d_EN and d_AS are the distances of the test file from the ENglish and ASsamese models respectively. A red square indicates such a difference with respect to English and Bengali models. It is clear that most points have negative values indicating that the distance to English model is smaller than that to either Assamese or Bengali model.

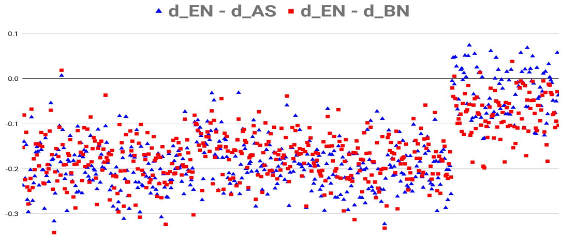

Figure 2: *Difference in distances from English and Assamese (blue triangles) or Bengali (red squares) ASR models are shown for 500 test English speech data. The negative values indicate that English test files are closer to English than non-English ASR models.*

In Fig. 2, the rightmost 100 speech files are of particular interest. Both blue triangles and red squares are shifted upwards for these files. These speech files belong to a male speaker whereas the initial 400 speech files belong to a female speaker. The speaker dependent performance of ASR/LID systems is evident here. Such a speaker dependence is reduced when DNNs are used instead of GMMs as described below.

4.2.2 LID using triphone DNN-HMM model

Another version of LID was implemented wherein HMM models a triphone (instead of monophone), and a DNN (instead of GMM) is used to estimate the likelihoods. The confusion matrix is shown in Table 4. The overall accuracy of the LID system employing DNN-HMM models is 99.3%.

Table 4: *The confusion matrix of the LID system using triphone DNN-HMM acoustic models.*

	Recognised Language (%)		
True Language	Assamese	Bengali	English
Assamese	**99.9**	0	0.1
Bengali	0.4	**99.6**	0
English	1.6	0	**98.4**

The differences between distances of the 500 English test speech files from English DNN-HMM based ASR model and the corresponding non-English ASR models are plotted in Fig. 3. Two observations can be made by comparing Fig. 3 (corresponding to triphone DNN-HMM models) with Fig. 2 (corresponding to mono GMM-HMM models): (i) the differences are more negative when DNN-HMM models are used. This is expected as the WER of triphone DNN-HMM system is less than half of WER of monophone GMM-HMM system, as shown in Table 2. (ii) the magnitude of the gender dependence, apparent in Fig. 2, has reduced due to speaker adaptive training and usage of sophisticated acoustic model.

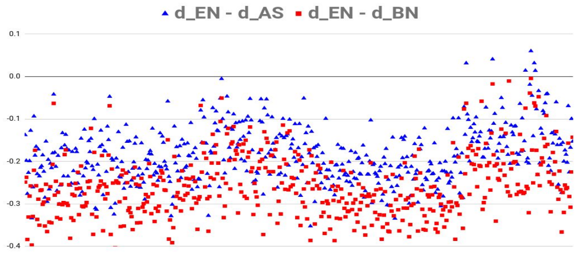

Figure 3: *Difference in distances from English and Assamese (blue triangles) or Bengali (red squares) ASR systems employing DNN-HMM to model triphones are shown for 500 test English speech data.*

The experiments conducted here are preliminary; they were done to illustrate the possibility of developing spoken language technologies for under resourced languages. There are a few limitations that can be corrected in future. All speech data were collected from residents of just one city. Also, the amount of speech data is uneven across languages and across speakers. While English is not an under resourced language, English spoken by Indians with a variety of mother tongues has a lot of variations. Since the number of speakers is small, the observations in the current experiment may be influenced by the particular set of speakers grouped into train and test data sets. Even though care has been taken to ensure mutual exclusivity of speakers in train and test data, a k-fold cross validation would be a better method of evaluation. Default values of parameters as set by Kaldi toolkit were used in our experiments. These can be tuned to obtain a better performance for larger datasets. Usage of source features in addition to that of vocal tract features is known to result in better accuracy. These are works that will be carried out in future.

5. Conclusions

The recent developments in speech and language technology have enabled development of spoken language technologies for languages with limited resources. A Language Identification system was implemented for two major languages of north-east India. This work can be extended to other languages of the region. Such systems will aid in implementation of better human machine interaction systems.

6. Acknowledgement

Our heartfelt thanks to Barsha Deka for collecting Assamese speech data used in this work, and also to all the speakers for participating voluntarily in the speech collection procedure.

7. References

[1] 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," in 4th Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14-16, 2014, 2014, pp. 16–23.

[2] Census of India, "Abstract of speakers' strength of languages and mother tongues – 2011", http://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf

[3] C. Moseley, "UNESCO Atlas of the World's Languages in Danger", Online version: http:// www.unesco.org/languages-atlas/index.php.

[4] E. Ambikarajah, H. Li, L. Wang, B. Yin, V. Sethu, "Language Identification: A Tutorial", IEEE Circuits and Systems Magazine, vol. 11, no. 2, pp. 82-108, 2011.

[5] K. S. Rao and D. Nandi, "Language Identification—A Brief Review", in "Language Identification Using Excitation Source Features", SpringerBriefs in Springer Technology, 2015, pp.11-30.

[6] B. Deka, J. Chakraborty, A. Dey, S. Nath, P. Sarmah, S.R. Nirmala and S. Vijaya, "SPEECH CORPORA OF UNDER RESOURCED LANGUAGES OF NORTH-EAST INDIA", Proc. of Oriental COCOSDA, Miyazaki, Japan, 2018.

[7] S. Paulose, S. Nath and Samudravijaya K, "Marathi Speech Recognition", to appear in Proc. of SLTU-2018, August 29-31, New Delhi, India.

[8] Indian Language Speech Label (ILSL12), https://www.iitm.ac.in/donlab/tts/downloads/cls/cls v2.1.6.pdf

[9] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[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, K. Vesely, "The Kaldi Speech Recognition Toolkit", In Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011, Hawaii, US