AUTOMATIC IDENTIFICATION OF NATIVE LANGUAGE FROM SPOKEN ENGLISH

Siddika Imani1,2, Parismita Sarma1, and Samudravijaya K2
1Department of IT, Gauhati University, Guwahati
2Centre for Linguistic Science and Technology, IITG, Guwahati
e-mail: {siddika.115, parismita.sarma, samudravijaya}@gmail.com

Automatic Speech Recognition (ASR) systems that facilitate voice based search and information retrieval have gained importance recently. While the performance of ASR systems for Indian languages have improved in the recent past, they have yet to gain wide acceptability as much as the ASR systems for English spoken by Indians. Almost all Indians learn English as a second or third language. So, the phoneme set and the prosody of native language of Indians influences the acoustic characteristics of spoken English. Since Indians speak a wide variety of languages, the acoustic characteristics of English spoken by Indians vary a lot. Thus, the recognition accuracy of Indian English could be improved by employing native language dependent English ASR systems. This approach requires automatic identification of the native language of the speaker. Here, we report the performance of an automatic Native Language Identification (NLI) system that recognises the native language of the speaker as Assamese or Bengali or Bodo after analysis of an English sentence spoken by the speaker. Training and performance evaluation of a NLI system needs appropriate linguistic resources. These include (a) speech data, in each of the 3 languages from several speakers, (b) corresponding word level transcriptions and (c) a pronunciation dictionary. While pronunciation dictionaries for English language are freely available, spoken English by speakers of the above mentioned 3 languages and transcriptions are not publicly available. So, we created a relevant speech database. We recorded English spoken by native speakers, both male and female, of these 3 scheduled languages. Each speaker read 100 sentences out of a set of 700 English sentences; these were either proverbs or digit sequences. Each sentence contained 5 to 10 words. The digitised speech, recorded under ambient conditions using a laptop, had the following characteristics: 16000Hz, 16bit, mono. The database contains spoken English from 35 native Assamese speakers, 33 Bengali and 30 Bodo speakers. In order to carry out a 3-fold evaluation of the performance of the system, the speakers from each language were grouped into 3 subsets such that each subset contains nearly equal number of speakers. In each fold, one subset was designated as test data, and the remaining two subsets were used to train the system. We used Kaldi, an open source ASR toolkit, for implementation of the NLI system. As a first step in the development of NLI system, we implemented 3 English ASR systems, each trained using training data from one of the 3 languages: Assamese, Bengali and Bodo. A 3-state Hidden Markov Model (HMM) represented a phone. Each state of HMM was associated with a Gaussian mixture model. We used Mel frequency cepstral coefficients and their temporal derivatives as features, and bigram as the language model. In order to identify the native language of a speaker, the test speech file was fed to each of the 3 ASR systems. An ASR system not only generates the decoded word sequence, but also the corresponding log likelihood. The NLI system follows the maximum likelihood criterion. The language corresponding to the ASR system that yielded the highest likelihood for the test speech was declared as the native
language of the speaker. The overall accuracy of the NLI system was computed as the unweighted average recall, computed from the confusion matrix. The NLI accuracy of the system, averaged over 3-fold cross evaluations, was 59% for test speech of just 3 seconds. The confusion was largest among Assamese and Bengali languages as both are close members of Indo-Aryan language family. In contrast, Bodo belongs to the Sino-Tibetan language family. We discuss the performance of the NLI system using different models such as context-dependent and context independent HMMs, employing Gaussian mixture model or deep neural network to estimate the likelihood of a feature vector emitted from a state of HMM.

1. Introduction

Research on automatic speech recognition (ASR), speaker recognition or language identification by machine is going on for the past 7 decades. The initial attempts to devise automatic speech recognition by machine were made in 1950s based on acoustic phonetics approach [1]. Recently, research in speech technology has diversified into emotion recognition, dialect recognition, accent recognition etc. Mother tongue recognition or native language identification is a quickly growing subfield in natural language processing and speech analysis.

Native Language Identification (NLI) is the task of identifying a speaker's native language (L1) based only on their expression in a second language (L2). NLI can be achieved by inspecting the text in written in L2 language [2] or based on speech in L2 [3]. Here, we confine ourselves to NLI based on the speech in L2. The text of speech is not a matter of concern to a NLI System. NLI works by identifying (a) the influence of L1 on the acoustic properties of speech sounds of L2 and (b) the language use patterns that are common to groups of speakers of the same native language. A Report on the 2017 Native Language Identification Shared Task [3] concluded that acoustic features are highly informative for speech-based NLI.

Krishna et al. [4] presented a work of identifying the mother tongue of a south Indian speaker based on spoken English. Telugu, Tamil and Kannada languages are considered as L1 languages and English as the L2 language. Each training file was 30 minutes long, and duration of test speech varied from 0.5 to 1.5 minutes. They used Mel Frequency Cepstral Coefficients (MFCC) features to train Gaussian Mixture Model (GMM). The NLI accuracy was 80%. Greater confusion was observed between Kannada and Tamil speakers. This confusion is found to be less when acoustic prosodic features are introduced. In 2018, the authors compared the performance of two classifiers, GMM and GMM-Universal Background Model (GMM-UBM) [5]. The GMM-UBM based classifier resulted in 91% accuracy in comparison to 88% accuracy obtained by the acoustic feature based GMM. They observed that additional usage of prosodic features did not aid the GMM system. The authors used 300 seconds of speech from each speaker for training and 120 seconds for testing the NLI system.

Senoussaoui et al. [6] proposed the use of i-vector representation to detect the native language of an English speaker. They evaluated different ways to extract i-vectors in order to adapt them to the specifics of the native language detection task. The experimental results on the 2016 ComParE Native language sub-challenge test set [4a] showed that their proposed system based on a conventional i-vector extractor outperforms the baseline system with a 42% relative improvement.

Ahmed et al. [7] studied the use of MFCC and gammatone filter cepstral coefficients (GFCC) features along with i-vector approach to identify the nativeness of the speaker. The framework was tested on the 2016 ComParE Native language sub-challenge dataset which has English language speakers from 11 different native language backgrounds. The accuracies of the i-vector based NLIs based on MFCC, and GFCCs were 67% and 68% respectively.

In previous works of identifying native language of Indians speaking English, the duration of speech signal used to identify the native language was long, of the order of a minute or so. For example, the durations of test speech were 120 seconds in [5], and 45 seconds in [3], [6] and [7]. While longer test speech increases the accuracy of NLI system, the practical utility of such systems become weak. On the other hand, NLI system that identifies the native language of a person speaking shorter speech would have greater
practical importance. The work presented here aims to recognise the native language of a speaker based on a spoken sentence comprising of 5 to 8 words whose typical duration is 3 seconds.

The remaining part of the paper is organised as follows. Section 2 gives a brief description of the linguistic resources created and used in this work. Section 3 describes the details of the experimental setup. Section 4 presents the outcome of various experiments, and analyses the results. A summary of the work is presented in Section 5.

2. Linguistic Resources

This section presents a brief description of the written and spoken linguistic resources created for training and evaluating the NLI system.

2.1 Text Corpus

The preparation of text corpora started with listing of English sentences that can be easily read by any person. The sentences chosen were mainly from high school story books, quotations from famous people or writers, proverbs, digit sequence, sequences of the names of states, their capital cities and languages of India. The length of the sentence varies from 5 to 8 words. A set of 700 unique sentences, comprising of 1500 unique words, forms the text corpus. The set of 700 sentences was subdivided into 7 subsets of 100 sentences each where each subset consists of 50 proverbs, 35 sentences from story books, 10 digit sequences and 5 state sequences. A subject was asked to read one of the 7 subsets containing 100 sentences.

2.2 Speech Corpus

Speech data was collected from the native speakers of Assamese, Bengali and Bodo languages who are residents of Assam state. The reading material used in this corpus consists of 700 different English sentences as described in section 2.1. Each person was asked to read a subset of 100 sentences. The subsets of sentences were assigned to persons in such a manner that each subset of 100 sentences was read by nearly equal number of male and female speakers of each language. Speech data was recorded using a laptop and an earphone at a sampling frequency of 16 kHz and resolution of 16 bits/sample. A typical recording session was as follows: the recording script displayed a sentence. The subject pressed the "Enter" button, and read the sentence within 6 seconds. Then, the next sentence was displayed, and so on. Each spoken sentence was stored in a separate file whose name indicated the identity and gender of the speaker, as well as the serial number of the sentence spoken. The duration of each speech file was 6 seconds that contained about 3 seconds of speech corresponding to a sentence of 5 to 8 words. The silence at the beginning and end of each speech file was detected by a simple energy based silence detection program, and the end-utterance silences were removed. Table 1 shows the details of the speech corpus created and used in this work.

<table>
<thead>
<tr>
<th></th>
<th>Native language</th>
<th>Assamese</th>
<th>Bengali</th>
<th>Bodo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech data (Hrs)</td>
<td></td>
<td>5.8</td>
<td>5.5</td>
<td>5.0</td>
</tr>
<tr>
<td>No. of utterances</td>
<td></td>
<td>3500</td>
<td>3300</td>
<td>3000</td>
</tr>
<tr>
<td>Total no. of speakers</td>
<td></td>
<td>35</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td>No. of male speakers</td>
<td></td>
<td>19</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>No. of female speakers</td>
<td></td>
<td>16</td>
<td>21</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the corpus of English spoken by native speakers of 3 Indian languages.
3. Experimental Details

This section discusses the details of the implementation of NLI system that recognises the native language of a speaker as one of Assamese, Bengali or Bodo based on an English sentence spoken by her/him. The NLI system utilises the evidence provided 3 ASR systems, each trained with English speech from speakers of one of these 3 native languages. Kaldi toolkit [8] was used for implementing the ASR and NLI systems.

Each ASR system uses a 3-state Hidden Markov Model (HMM) to represent a phone. The probability density function associated with a state of a HMM was modeled as a GMM. We used MFCCs and their temporal derivatives as features; bigram language model was used. In order to identify the native language of a speaker, English spoken by her/him was fed to each of the 3 ASR systems. An ASR system not only generates the decoded word sequence, but also the corresponding likelihood. The language corresponding to the ASR system that yielded the highest likelihood for the test speech would be declared as the native language of the speaker by the NLI system, following the maximum likelihood criterion.

The performance of the system was evaluated using a 3-fold cross validation method. The speech corpus was divided into three subsets such that each subset contains almost equal number of speakers and speech files; such a division was carried out independently for each native language. 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. In Fold1 experiment, the first two subsets (train data) were used for training the system, and the remaining subset (test data) was used to evaluate the performance of the system. The confusion matrix of NLI system was generated for the test data of Fold1 experiment. The overall accuracy of the NLI system was computed as the average of the diagonal elements of the confusion matrix. This procedure was repeated for the other two folds. Finally, the identification accuracy of the NLI system was reported as the average of the 3 accuracy values corresponding to the 3 fold experiments.

4. Results and Discussion

In this section, we present the performance of the speech recognition system first. The accuracy of native language identification based on the 3 speech recognition systems is presented and discussed next.

4.1 Speech Recognition System

During training phase different types of acoustic models are trained using Kaldi toolkit [8]. The system is initialized with a context independent monophone acoustic model, called 'mono'. In addition, 3 types of context dependent (triphone) models were trained; we will call these as 'tri1', 'tri2' and 'tri3'. The kaldi can also train HMM models whose emission probabilities can be estimated either by a subspace GMM or by employing a deep neural net. In this work, these models are denoted as ‘sGMM’ and ‘DNN’ respectively.

The performance of an ASR system is evaluated in terms of word error rate (WER) as:

\[
\text{WER} (\%) = 100 * \frac{D+S+I}{N}
\]

Here, S, D and I denote the number of substitution, deletion and insertion errors respectively, and N denotes the total number of words present in the reference transcription. Lower WER, the better the system is.

Table 2 shows WER (%), averaged over 3-fold experiments, of an exemplar English ASR system. This ASR system was trained with English speech from native speakers of Assamese language. The WERs of test English speech from people with different native languages are shown in Table 2 as a function of the acoustic models employed. For example, the WERs of the English ASR system, using monophone HMM, are 9.0%, 9.9% and 13.8% for test speech by people whose native language is Assamese, Bengali and Bodo respectively. As expected, WER of the NLI system corresponding to speech by native Assamese speakers is the lowest since the system had been trained with speech from native Assamese speakers. Similar outcome is observed in case of the other 2 ASR systems as well. Consequently, corresponding to English spoken by an Assamese speaker, the likelihood of an ASR system trained with English from native Assamese speakers is likely to higher than the likelihood of any other ASR system. This logic is used by the NLI system to identify the native language of any speaker speaking an English sentence.
Table 2: Word error rate (%) of English ASR system as a function of different types of acoustic models. The ASR systems were trained with English spoken by Assamese speakers. The test data consists of English spoken by persons whose native language could be either Assamese, Bengali or Bodo.

<table>
<thead>
<tr>
<th>Native language of test speakers</th>
<th>Word Error Rate (%) of various systems</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mono</td>
</tr>
<tr>
<td>Assamese</td>
<td>9.0</td>
</tr>
<tr>
<td>Bengali</td>
<td>9.9</td>
</tr>
<tr>
<td>Bodo</td>
<td>13.8</td>
</tr>
</tbody>
</table>

4.2 Native Language Identification System

In this section, we report the performance of NLI systems. Table 3 shows the accuracy (%), averaged over 3-fold experiments, of NLI systems using different types of acoustic models, for test data. The subspace GMM based NLI system achieves 59% native language identification accuracy for test speech of just 3 seconds.

Table 3: Identification accuracy of the NLI systems for different kinds of acoustic models, for test data.

<table>
<thead>
<tr>
<th>Acoustic Models</th>
<th>mono</th>
<th>tri1</th>
<th>tri2</th>
<th>tri3</th>
<th>sGMM</th>
<th>DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification accuracy (%)</td>
<td>55.2</td>
<td>58.0</td>
<td>51.4</td>
<td>59.0</td>
<td>52.1</td>
<td>58.2</td>
</tr>
</tbody>
</table>

4.3 Discussion

In order to analyse the results of NLI system, we show confusion matrix of the NLI system, using DNN-HMM acoustic model, for all the 3-fold experiments in Table 4. For example, the numbers in the top-left most row show that, in Fold1 experiment, the native language of Assamese speakers were identified as 57.2% times as Assamese and 29.9% times as Bengali. In general, the confusion was largest among Assamese and Bengali languages as both are close members of the Indo-Aryan language family. In contrast, Bodo belongs to the Sino-Tibetan language family even though all the languages are spoken in geographically adjacent areas.

<table>
<thead>
<tr>
<th>Fold1</th>
<th>Fold2</th>
<th>Fold3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nativ e Language (%)</td>
<td>Accu racy</td>
<td>Recognized Language (%)</td>
</tr>
<tr>
<td>AS</td>
<td>BN</td>
<td>BO</td>
</tr>
<tr>
<td>57.5</td>
<td>58.3</td>
<td>58.8</td>
</tr>
</tbody>
</table>

FRSM 2019, Kanpur, India, July 6-7, 2019
A drawback of the current NLI system is that it is tested on sentences to which the system is exposed during training. The performance of the system for unrestricted vocabulary is not evaluated. For such an evaluation, a (text independent) phone recognition system should be preferred over a word recognition system. Then, a phone n-gram model should be used instead of word bigram language model used here.

In all the experiments that we conducted, we used the default settings of the Kaldi toolkit in the implementation of this preliminary NLI system. The accuracy of the NLI system can be improved by tuning the parameters of the ASR systems. Bodo is a tonal language whereas Assamese and Bengali are not. So, using the fundamental frequency of vocal folds as an additional feature is likely to improve the performance. Also, the ranges of likelihood of the English ASR systems trained with speech of native speakers Assamese, Bengali and Bodo are different. Suitable normalization of the likelihood values is likely to lead to better results.

5. Conclusion

A preliminary native language identification system was implemented that identifies the native language of a person as one of Assamese, Bengali or Bodo, based on analysis of English spoken by her/him. The accuracy of the system is 59% when the duration of English speech is as short as 3 seconds. As far as we know, this is the first such work carried out on spoken languages of north-east India. Further tuning of this baseline system is likely to improve the performance of the system, leading to its use in practical scenarios since the system needs short duration spoken English.

REFERENCES