Discovering Canonical Indian English Accents: A Crowdsourcing-based Approach

Sunayana Sitaram*, Varun Manjunath†, Varun Bharadwaj†, Monojit Choudhury*, Kalika Bali*, Michael Tjalve+

*Microsoft Research, Bangalore, India
susitara, monojitc, kalikab@microsoft.com

†PES University, Bangalore, India
varunmanjunath2012, varunbharadwaj1995@gmail.com

+Microsoft Corporation, Redmond, USA
University of Washington, Seattle, USA
mitjalve@microsoft.com

Abstract

Automatic Speech Recognition (ASR) systems typically degrade in performance when recognizing an accent different from the accents in the training data. One way to overcome this problem without training new models for every accent is adaptation. India has over a hundred major languages, which leads to many variants in Indian English accents. Making an ASR system work well for Indian English would involve collecting data for all representative accents in Indian English and then adapting Acoustic Models for each of those accents. However, given the number of languages that exist in India and the lack of a prior work in literature about how many Indian English accents exist, it is difficult to come up with a set of canonical accents that could sufficiently capture the variations observed in Indian English. In addition, there is a lack of labeled corpora of accents in Indian English. We approach the problem of determining a set of canonical Indian English accents by taking a crowdsourcing based approach. We conduct a mobile app based user study in which we play audio samples collected from all over India and ask users to identify the geographical origin of the speaker. We measure the consensus among users to come up with a set of candidate accents in Indian English and identify which accents are best recognized and which ones are confusable. We extend our preliminary user study to a web app-based study that can potentially generate more labeled data for Indian English accents. We describe results and challenges encountered in a pilot study conducted using the web-app and future work to scale up the study.

Keywords: non-native accents, crowdsourcing, Indian languages, speech processing

1. Introduction

Automatic Speech Recognition (ASR) systems have reached performance on par with humans on some tasks and languages. However, the performance of ASR systems is significantly worse in the presence of accents that are different from the accents in the data used for training the ASR system. Since commercial ASR systems require hundreds or thousands of hours to train, it is not always feasible to train a separate system for every accent the ASR system will have to recognize. One solution to this is to adapt Acoustic Models that are trained on one accent to a target accent by using a small amount of speech data in the target accent (Huang et al., 2014).

India is an extremely multilingual country. According to the 2001 Census of India (Banthia, 2001) there are 122 major and 1599 other languages in India; 23 languages have been granted the status of official languages. The existing literature on Indian English (IE) accents either focuses on its difference with British or American English, or on very specific phonetic features that mark IE (Sahgal and Agnihotri, 1988) (Kachru, 2005) but fails to define what a canonical IE accent is, or what a set of canonical IE accents could be. Recently, it has been shown that using native language (L1) data could help adapt Acoustic Models to an accent influenced by that L1 (Aditya Siddhant, 2017). However, in the case of IE, we face the issue of not knowing which native language(s) should be chosen for adaptation. Labeled data exists for some major accents in English, however, there does not exist any labeled data covering all the possible accents in IE.

In this work, we follow a crowdsourcing based approach to finding out what the set of canonical accents in IE could be. We build mobile and web-based interfaces using which users listen to audio samples of IE and annotate on a map of India where they think the speaker is originally from. Users have the option of choosing one of five geographical regions in India, or one of the 29 states. Then, we analyze responses from all users to find geographical regions where labels have high agreement and regions that are confusable. In this case, we use geographical region as a proxy for L1, which in turn is assumed to influence the IE accent of a speaker. This assumption is reasonable because many states in India have their own language, and there is a correspondence between state and a major language except in the case of some states in North India.

Because of the lack of annotated accent data, or information about the L1 of the speakers in our audio samples, we have no ground truth L1 labels. So, we use geographical location of an audio file initially as a proxy of what the accent of the speaker could be. This clearly does not hold true for people who have migrated from one part of the country to another. In addition, exposure to different languages, travel, the level of education and other socio-linguistic factors play an important role in determining one's accent. We describe how we propose to handle some of these challenges in the pilot and the large-scale web-based study.

In addition to finding accents that are well identified, we are also interested in knowing which geographical regions may produce accents that are not easily distinguishable as a particular accent, as such accents can be thought of as neutral or mild IE accents. A neutral IE accent may also be useful for a personal digital assistant that has to have a common accent for users from all over the country.

The rest of this paper is organized as follows. First, we describe datasets available in other English accents and prior work on Indian English accents. Next, we describe our mobile app-based pilot study and data analysis, followed by the design of our larger scale web-based study. We describe preliminary findings from a pilot study conducted using the web app, and some of the challenges faced. We conclude with ongoing and future work.

2. Relation to Prior Work

Labeled data for accent modeling and adaptation exists for some English accents. The CLSU Foreign Accent Corpus (Lander, 2007) consists of 1-2 hours of spontaneous speech by native speakers of 22 languages including Hindi and Tamil. In addition, the CSLU corpus also has judgments on the heaviness or mildness of each accent on a four point scale. The ABI (D'Arcy et al., 2004) corpus consists of 95 hours of recordings from 300 speakers, representing 15 accents of the British Isles. There does not exist a comprehensive labeled corpus that covers all the major L1s of speakers of Indian English.

Most studies on IE accents have focused on vowel analysis. (Phull and Kumar, 2016) describe a study on vowel analysis for four IE accents - North, South, East and West Indian. They found that there was a significant difference in the first four formants in these accents. (Maxwell and Fletcher, 2009) study the acoustic and durational properties of vowels of speakers whose L1 is Hindi or Punjabi, and find contrasts between the two. (Kalashnik and Fletcher, 2007) suggest that North Indian English shows distinct vowel patterns making it a separate sub-variety of IE.

(Sirsa and Redford, 2013) carried out a study to compare the sound structures of IE produced by native Hindi and Telugu speakers. They found that the L1 influenced the production of some segments in IE, but L1 temporal patterns were not found in IE. They also asked experienced and naive listeners to distinguish the speech based on L1. Experienced listeners could do so better than naive listeners. (Maxwell, 2014) studied the intonational phonology of Kannada and Bengali Indian English and found intonational differences within IE and between IE and other Englishes.

Although there have been studies on specific varieties of Indian English accents, to the best of our knowledge, there does not exist prior work or data to identify a set of canonical Indian English accents.

3. User Study

3.1. Data

We used an in-house data set of spoken queries to a speech recognition system as data for the pilot study. We divided India into 5 geographical regions and selected 25 representative cities and towns in total spread out over these regions. We tried selecting cities without a very high immigrant population to circumvent the problem of not knowing the true L1 of a speaker. We tried to avoid very large cities, because speakers from large cities may have milder, more urban accents that are harder to identify. However, in some cases our choice of city was dependent on the availability of data. In preliminary user studies, we found that users could not identify accents when the utterances were shorter than a few seconds. So, we manually listened to and collected data that was sufficiently long and that did not reveal any location based information, thus avoiding sentences like 'What is the weather like in Mysore?'.

Initially, we created an Android app which would play an audio file and show the user either a map of the country, or a drop-down menu of states and regions of India. In preliminary studies, we found that users preferred the map-based interface and decided to use that as the interface for the study. In the map-based interface, users were shown a map of the country with state boundaries. If they clicked on the map, one of 5 geographical regions would be highlighted. They had the option of zooming into a region once to see all the states in the region, if they wished to make a finer-grained decision at the state-level.

We selected 10 sentences from each of the 25 cities and towns, leading to 250 audio examples in all. Each user listened to 15 examples, with 3 examples from each of the 5 geographical regions. Even though this did not guarantee that they were listening to accents from all the regions, overall this gave users a reasonable distribution to listen to. We wanted each audio file to be labeled by at least three users, so we conducted the user study with 60 users, with each user listening to 15 utterances and having the option of skipping an audio file if they did not want to provide a judgment for it.

We also collected optional demographic information including the users' L1, other languages they knew, their educational qualifications and a list of places they had lived in for more than a year. Participants in this pilot consisted of students and researchers from our research lab or visitors to the lab. Most participants were between 20-35 years of age and were undergraduates, graduate students or post graduates. The participants spoke the following L1s: Hindi, Tamil, Telugu, Konkani, Malayalam, Punjabi, Bengali, Kannada, Gujarati, and Marathi. Among these L1s, Hindi, Tamil, and Telugu were the most common.

At the end of the study, participants were given the option of recording a short paragraph taken from the Accent Project at GMU (Weinberger, 2014). This paragraph was designed to capture most sounds in English and consisted of familiar words, but some difficult sound sequences. All participants volunteered to record this paragraph. The recording was conducted using the mobile app, so there was some background noise present in the recordings, leading to realistic training data for a speech recognizer. We obtained around 60 audio recordings of the paragraph by participants with corresponding demographic data.

3.2. Analysis

Next, we present an analysis of the data we collected in the pilot study. We compared participant responses to the ge-

ographical location of the audio sample from our in-house database of queries. Although we used geographical location as a proxy for accent, which has some limitations as described earlier, we saw some general trends in the data.

Participants had the option of choosing two levels of granularity while making their selection: region or state. 32% of the annotations were state level, while the rest were region-level. Figure 1 and Figure 2 show the confusion matrix of the region and inferred region-based judgments for all the audio samples. Figure 1 shows the actual regions picked by the users, while Figure 2 also shows the region that we inferred based on the state that the users chose. The color of the boxes and the numbers inside them indicate the absolute number of judgments, with the Y axis containing the true label (actual geographical origin) of the audio sample and the X axis indicating the judgments. From both the figures, we see that South India is the region that has maximum agreement between the geographical origin and judgment, while North India is second. We also see that there is a large difference in the counts between the correctly predicted region of South India in the two figures, which indicates that users were confident enough to also pick a South Indian state while making the judgment - roughly one-third of the time, users picked a South Indian state, and picked the entire region of South India two-thirds of the time. We see similar trends for North India as well. Both figures also show that Central India was confused with North India.

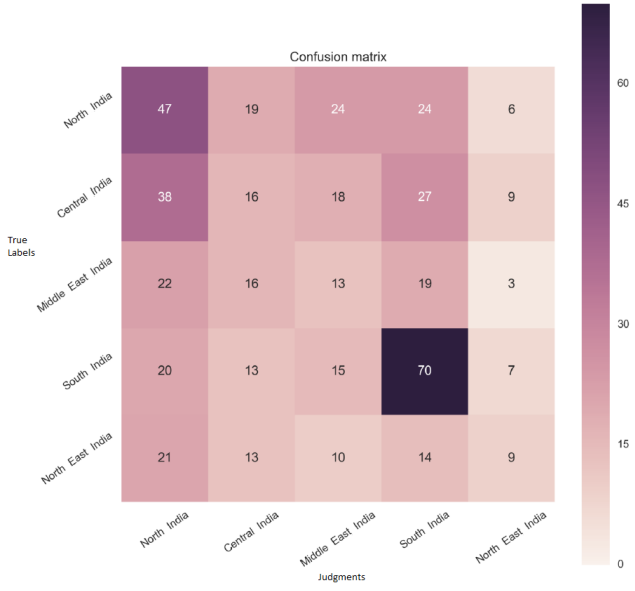

Figure 1: Region confusion matrix

It is important to note that our definition of regions was based on a particular grouping of Indian states into zones. A different grouping of regions based on the similarity of Indian languages could lead to more interpretable results. Similarly, we could choose to replace state boundaries with regions where a major Indian language is spoken.

We created a similar confusion matrix for the state-level judgments, shown in Figure 3. Some observations are as follows.

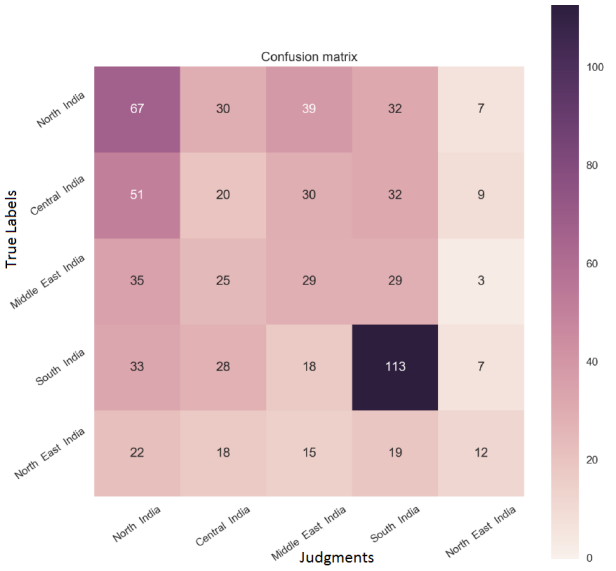

Figure 2: Inferred region confusion matrix

1. Punjab was most likely to be identified as Uttar Pradesh, but Uttar Pradesh was most likely to be identified as itself.

3. Tamil Nadu was well identified in general but sometimes confused with Maharashtra and UP

4. Karnataka was identified most often as itself, but was sometimes confused with Tamil Nadu, Kerala, Andhra Pradesh

5. Andhra Pradesh was most often identified as Maharashtra

6. None of the other states had strong diagonal values in the confusion matrix, which meant that they were not easily distinguishable

7. Some states, such as Telangana, Uttarakhand, Jharkhand, Himachal Pradesh and Sikkim were not present in the data and were not picked as candidate states by users. The states of Telangana, Uttarakhand and Jharkhand are newly-formed states in India.

Each file received at least three annotations and the geographical origin was used as the reference. We calculated normalized scores for each state that had been annotated by the users as follows. For each label, if it was an exact match with the geographical origin state of the file, we gave it a score of 1. If it was a neighboring state, we gave it a score of 0.5. If the state was not a neighboring state but in the same geographical region, we gave it a score of 0.25. We aggregated the scores for all the files for each state, which is shown in Figure 5.

Next, we wanted to calculate the agreement among participants in choosing labels. If there was high agreement for a particular file but a mismatch with the geographical origin, this could indicate that the speaker could have been an immigrant. If there was low agreement, it could indicate that the accent was difficult to guess. For state-level judgments, we found that there was very poor agreement due to the lack of more than one state-level label for most files. The states Karnataka, Kerala, Maharashtra and Tamil Nadu had only one instance of agreement between users, while Uttar Pradesh, Punjab and Bihar had two. So, we extended this

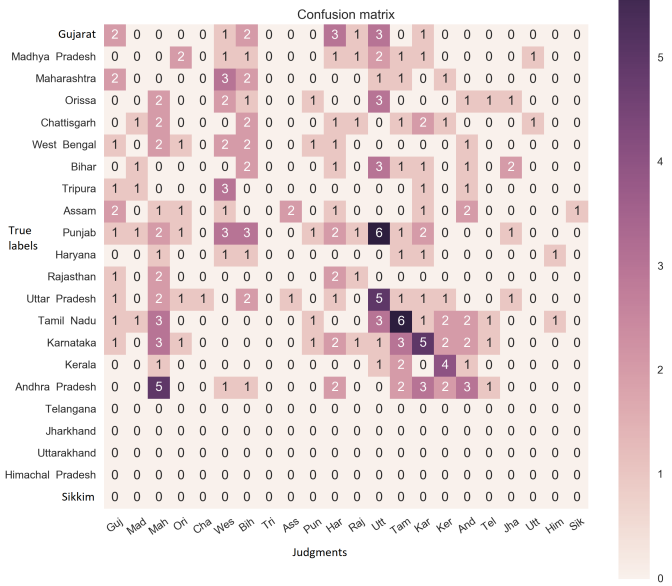

Figure 3: State-level confusion matrix

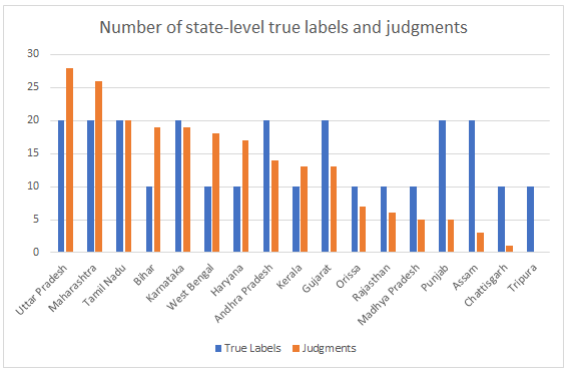

Figure 4: Number of state-level true labels and judgments

score to include region-level agreement as well. If there was any agreement, either at the state or region level for a file, it received a score of one. We added a point to the agreement score for a file for each pair of annotators who agreed, and aggregated this score over each geographical origin state, as shown in Figure 5.

From the figure we can see that some states had high agreement among participants, but a low match with the geographical origin of the audio sample. This could be due to the presence of immigrants in those states.

Although we found some interesting trends, to achieve our original goal of discovering canonical accents in Indian English, we needed to have higher confidence of what the L1 of each speaker was. Going forward, we wanted to scale up the study by using user-reported L1 as the true label instead of using geographical location as a proxy.

In addition, some of the findings of the pilot study may have been influenced by the fact that all participants were either living in or visiting Bangalore, Karnataka, which is why they may have been able to distinguish Southern Indian accents better. Participants also had similar educational and socio-economic backgrounds. So, we wanted to create a web-based application going forward that could potentially be accessed by anyone in the country.

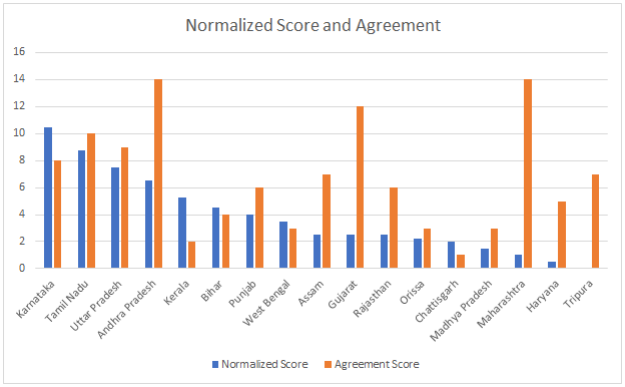

Figure 5: Normalized Scores and Agreement between users

3.3. Feedback from users

As part of the user study, we collected feedback from users about the task and interface. Users felt that the task was difficult and reported that selecting a state was hard compared to selecting a region. However, from our analysis we see that a third of the judgments were state-level.

Users wanted to see their score at the end of the study to know how well they did at identifying where the speakers were from. They also wanted to know which accents (or regions) they did well on, and which ones they got wrong. This was not possible in the pilot study because we did not have ground truth labels for the files. We incorporated this feedback while designing our web-based study.

4. Web-based Study

We designed a web-based study in the form of a game to scale up our pilot study, in which we used the audio recorded by participants of the pilot as the audio examples to play to users. Since we had demographic information about the users who recorded audio including their L1, we could use it as a ground truth label for the accent.

Figure 6 shows the interface of the web-based study. Users were shown a map of the country and a button to toggle a region view or a state view. They could play the audio sample and annotate a region or state on the map that they felt corresponded with the place of origin of the speaker.

Since we had the ground truth labels in this study, we could calculate a score that we could show users at the end of the game. The score was calculated by matching the state or region that the user's L1 and place of birth corresponded to the state or region that was selected by the user. An exact match at the state-level was given a score of 10, while a region-level match got a score of 5. Each user listened to 10 files in the web-based study. We divided scores into 5 equal buckets, each of 20 points and assigned a humorous label to each bucket to show to the users at the end of the game.

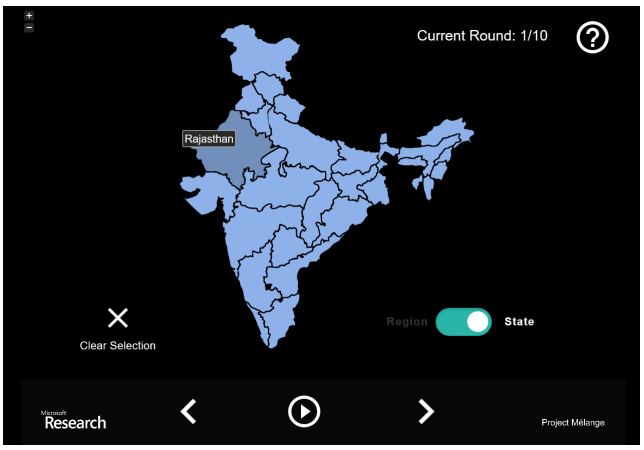

Figure 6: Screenshot of the web-based study interface

Figure 7: Screenshot of the final judgments with correct answers

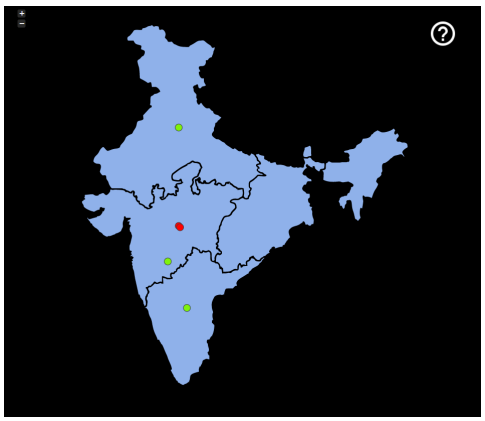

Figure 8: Screenshot of a map with markers indicating correct and wrong judgments

4.1. Feedback from users

10 users from the research lab tested the web-based application. All users said that they enjoyed playing the game, however, their scores fell into the lowermost or second lowest bucket, which meant that they got a score of 40 or less out of 100. This could be attributed to the following reasons. Users said that some of the files had background noise, which was due to the fact that these had been recorded during the pilot on mobile phones. More importantly, users pointed out that the accents were difficult to guess because the speakers were very urban with mild accents. This was due to the fact that we had collected the data used for the web app-based study from participants visiting our research lab. The data used for the earlier mobile app-based study was from users of a speech recognition system which was a more diverse, but presumably quite urban population.

We also received feedback that users wanted to know which accents they got correct and which ones they made mistakes on. We modified the interface to show this to users, by giving them the choice of listening to every file that they got incorrect and by showing them a map with dots in different colors indicating which ones they got correct and wrong. This would allow participants to see if they performed differently on accents from different regions. Figures 7 and 8 show screenshots of what the participants see once they finish the game. Figure 7 shows participants the state or region they selected along with the true state or region, and also gives them the option of listening to the corresponding audio files again. Figure 8 shows a map with green dots indicating correct judgments by state or region, and red dots indicating incorrect judgments.

Going forward, we would like to use the web-based study to scale up both the annotation and collection of accented data. Our first challenge is to make the task easier for users to do by providing them with a diverse set of accents from urban, semi-urban and if possible rural speakers. The purpose of the web based-study is to collect such data automatically by participants uploading their own audio files. To bootstrap this process, we plan to collect some data which includes some strong accents, so that participants are able to identify some accents in the game more easily.

In addition, we also plan to make the game more engaging and fun by showing a leaderboard so that participants within a group (such as a university) can compare scores with each other more easily, if they wish to. Currently, all participants who wish to upload an audio file in their own voice read the same paragraph mentioned in Section 3.1. We plan to allow users to record any sentence they wish to in future iterations of the game, so that we can collect a variety of accented data that can potentially be useful for training ASR models.

5. Discussion

In this paper, we have described a first attempt to identify canonical accents in Indian English. Due to the lack of labeled accented data or a large corpus of speech that covers all regions of India, we used a crowdsourcing-based approach by asking users to identify the geographical origin of a speaker at a state or region level, which we used to represent their L1. In the mobile app-based pilot study, we found that users were able to distinguish North and South Indian accents better than other accents and were able to give state-level judgments a third of the time. The geographical location of speakers residing in states such as Uttar Pradesh, Tamil Nadu and Karnataka were identified with higher accuracies than other states, however, the agreement between participants was low for state-level judgments in general.

To overcome the limitations due to the lack of ground truth about the L1 of the speaker, we used speech data collected during the pilot with self-reported L1 to bootstrap a web-based accent recognition game. We conducted preliminary

user studies with the web-based app, and incorporated user feedback to show file-by-file results, and a map with markers to indicate how well the users performed on accents from different regions of the country. We plan to release the web app to a much larger set of people from all over the country to obtain more judgments and more accented speech data. To overcome the difficulty of providing judgments in the current game due to the presence of mild urban accents, we plan to collect some heavily accented data and use it to bootstrap the game till we can collect data from a diverse set of participants.

In our analysis of the data collected from this study, we plan to focus on the correlation between the reported L1 of the speaker and the judgment provided by participants. In addition, we plan to study the effects of L1 and language exposure of participants on their judgments.

6. References

Aditya Siddhant, Preethi Jyothi, S. G. (2017). Leveraging native language speech for accent identification using deep siamese networks. In *Proceedings of ASRU 2017*.

Banthia, J. K. (2001). *Census of India, 2001*, volume 1. Controller of Publications.

D'Arcy, S. M., Russell, M. J., Browning, S. R., and Tomlinson, M. J. (2004). The accents of the british isles (abi) corpus. *Proceedings Modélisations pour l'Identification des Langues*, pages 115–119.

Huang, Y., Yu, D., Liu, C., and Gong, Y. (2014). Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Kachru, B. B. (2005). *Asian Englishes: beyond the canon*, volume 1. Hong Kong University Press.

Kalashnik, O. and Fletcher, J. (2007). An acoustic study of vowel contrasts in north indian english. In *Proceedings of the 16th international congress of phonetic sciences*, pages 953–956.

Lander, T. (2007). Cslu: Foreign accented english release 1.2. *Linguistic Data Consortium, Philadelphia*.

Maxwell, O. and Fletcher, J. (2009). Acoustic and durational properties of indian english vowels. *World Englishes*, 28(1):52–69.

Maxwell, O. (2014). *The Intonational Phonology of Indian English: An Autosegmental-Metrical Analysis Based on Bengali and Kannada English*. Ph.D. thesis, University of Melbourne, School of Languages and Linguistics.

Phull, D. K. and Kumar, G. B. (2016). Vowel analysis for indian english. *Procedia Computer Science*, 93:533–538.

Sahgal, A. and Agnihotri, R. K. (1988). Indian english phonology: A sociolinguistic perspective. *English World-Wide*, 9(1):51–64.

Sirsa, H. and Redford, M. A. (2013). The effects of native language on indian english sounds and timing patterns. *Journal of phonetics*, 41(6):393–406.

Weinberger, S. H. (2014). Speech accent archive. george mason university. *http://accent.gmu.edu*.