IE-CPS Lexicon: An Automatic Speech Recognition Oriented Indian-English Pronunciation Dictionary

Shelly Jain* **Aditya Yadavalli*** **Ganesh S Mirishkar***

Chiranjeevi Yarra **Anil Kumar Vuppala**

Speech Processing Laboratory
International Institute of Information Technology, Hyderabad
Gachibowli, Hyderabad, Telangana, 500032
{shelly.jain, aditya.yadavalli, mirishkar.ganesh}@research.iiit.ac.in
{chiranjeevi.yarra, anil.vuppala}@iiit.ac.in

Abstract

Indian English (IE), on the surface, seems quite similar to standard English. However, closer observation shows that it has actually been influenced by the surrounding vernacular languages at several levels from phonology to vocabulary and syntax. Due to this, automatic speech recognition (ASR) systems developed for American or British varieties of English result in poor performance on Indian English data. The most prominent feature of Indian English is the characteristic pronunciation of the speakers. The systems are unable to learn these acoustic variations while modelling and cannot parse the non-standard articulation of non-native speakers. For this purpose, we propose a new phone dictionary developed based on the Indian language Common Phone Set (CPS). The dictionary maps the phone set of American English to existing Indian phones based on perceptual similarity. This dictionary is named Indian English Common Phone Set (IE-CPS). Using this, we build an Indian English ASR system and compare its performance with an American English ASR system on speech data of both varieties of English. Our experiments on the IE-CPS show that it is quite effective at modelling the pronunciation of the average speaker of Indian English. ASR systems trained on Indian English data perform much better when modelled using IE-CPS, achieving a reduction in the word error rate (WER) of upto 3.95% when used in place of CMUdict. This shows the need for a different lexicon for Indian English.

1 Introduction

Today, speech processing technology is gaining an undeniable importance. Automatic speech recognition (ASR) systems in particular are being sought after, as smart assistants like Siri, Alexa and Google Assistant grow in popularity at remarkable rates. Effective and accurate speech recognition is also a major concern in India for two major reasons – firstly, India is home to thousands of spoken languages, not all of which possess written forms; secondly, India has a low literacy rate, so a considerable portion of the population can communicate only via speech. Unfortunately, there is a dearth of available speech data on many languages spoken in India, so most ASR systems tend to be ineffective at modelling these. In order to both overcome the drawbacks caused by lack of available speech data, and leverage the similarities between the various Indian languages, linguistically motivated approaches are gaining appreciation among researchers in the field.

Many research groups have started working towards having a good speech recognition system for all the major Indian languages. To achieve this, it is necessary to have a good pronunciation modelling block, consisting of a grapheme-to-phoneme (G2P) system. A G2P is needed by Text-to-Speech (TTS) systems as well. Nair et al. (2013) work on building a rule-based G2P on the low-resource Malayalam language. Parlikar et al. (2016) work on building a G2P for major Indian languages such as Hindi, Tamil and Telugu and test it on TTS systems. Mortensen et al. (2018) build a multilingual G2P that works for 61 languages out of the box, in which 7 Indian languages are supported. They also provide a guideline to integrate any language into their system. Arora et al. (2020) work on statistical G2P that predicts schwa deletion in Hindi more accurately and show improvements over their baseline models. Wasala et al. (2006) work on Sinhala

*These authors contributed equally to the paper

Proceedings of the 18th International Conference on Natural Language Processing, pages 195–204
Silchar, India. December 16 - 19, 2021.

G2P, a language that is closely related to other Indian languages, and further, propose rules to handle schwa epenthesis.

English is also a language with widespread use in India. Given India's rich language diversity, English has largely been adopted in the spheres of the educated public as a *lingua franca* (Ng and Hirose, 2012; Kim et al., 2011) like in other parts of the world. The English spoken in India is influenced by the other languages with which it exists in constant contact, affecting its structure and vocabulary as well as its pronunciation. This means that existing ASR systems for General American English (GAE) or British English (BrE) (Ferragne and Pellegrino, 2004) are markedly ineffective when dealing with Indian English (IE). New ASR systems need to be created which are capable of processing the variety of English spoken in India, catering to the accents and language patterns (Jain et al., 2018) of the people speaking it.

Considering the limited work in this direction and the impact it may have, we study the major phonetic variation between Indian English and standard English, and propose an Indian English lexicon which can be used in an IE ASR system. To show its effectiveness, the performance of the two different ASR systems is compared on utterances of Indian English – an American English ASR system trained on Librispeech (Panayotov et al., 2015) and an Indian English ASR system trained on National Programme on Technology Enhanced Learning (NPTEL) speech corpus (described further in Section 4.1).

The organisation of the remainder of the paper is as follows. Section 2 describes prior work. Section 3 details the proposed lexicon and the rationale behind the chosen mappings. Sections 4 and 5 explain the experimental setup and results when using the new dictionary for ASR. Further discussion on the need for this lexicon is done in Section 6. Finally, possible future ventures are briefly mentioned in Section 7, and the paper is concluded in Section 8.

2 Related Work

ARPAbet is a phonetic transcription code used to represent the phonemes and allophones of GAE using distinct ASCII sequences. The CMU Pronouncing Dictionary (CMUdict) (Weide, 2005) is an open-source pronunciation dictionary which makes use of a modified form of ARPAbet. Currently, there has been very limited work done to replicate such efforts for other languages, or even other varieties of English. Our proposed lexicon attempts to fill this gap in the field of language phonology.

The work by Ganji et al. (2019) is also closely related to our line of research in this paper. They collected a Hindi-English code-switched speech corpus. In addition to releasing that corpus, they proposed a few measures that could help model such speech better. One of the measures they proposed was a way to handle the pronunciation model block for this code-switched corpus that is most often used in ASR and text-to-speech (TTS) systems. They proposed a CMUdict-based phone mapping with some of the English phones mapped to a common Indian language phone set (Ramani et al., 2013) based on perceptual similarity. The paper, however, did not present any experimental results for ASR with their proposed pronunciation model. We explain how we build upon this in greater detail, in Section 3.

Another study by Huang et al. (2020) compared the phonology of GAE and the phonology of (Hindi-based) IE and developed a phoneme set for IE in X-SAMPA (Wells, 1995) similar to that of GAE. However, the dataset on which they built the ASR system has its own limitations. It has speech data from just 10 native Hindi speakers with long-term exposure to only the English spoken in New Delhi. As such, the language modelled was a very specific variety of IE, which is not representative of a large-scale Indian population.

Yarra et al. (2019) collected a corpus of Indian English, consisting of 240 hours of recorded speech. This is called Indic TIMIT. They extensively analysed Indian English accent and compared it to standard English. They further came up with various kinds of rules to improve the pronunciation model for Indian English speech. However, such rules needed another complex module – a letter to phoneme aligner. The letter to phoneme aligner is used to align a word's letters to its individual sound units. Our work eliminates the need for this module, by applying only those rules which were exhibited by native speakers of a majority of Indian languages. In addition to that, Yarra et al. (2019) also did not use the now largely standardised phone codes for six major Indian languages – Hindi, Marathi, Bengali, Tamil, Malayalam and Telugu – proposed by Ramani et al. (2013), the Common Phone Set (CPS). This makes Yarra et al.

(2019)'s work difficult to use in a multilingual or a code-switched setting where one of the languages is an Indian language. They also did not compare the performance of their proposed-lexicon-based ASR system and CMUdict-based ASR system. Therefore, it is unclear what benefits their proposed lexicon might have to an ASR system.

In addition to the above works, the studies by Schilk (2010); Bansal (1969); Mohan (2021) were helpful in providing the general perspective on Indian English. They described some specific differences made in the pronunciation of English by Indians, at both the sound and word level. We incorporate some of the differences they point out at the sound level in our lexicon. In addition to the differences, they also provided insight into the question of whether Indian English is better treated as a variety of standard English or as a completely separate language. This is discussed further in Section 6.

3 Indian English-Common Phone Set (IE-CPS)

Given the limited availability of annotated IE speech data, there is need for a mechanism to provide the pronunciations for this kind of speech. Our lexicon, Indian English Common Phone Set (IE-CPS), is able to provide a simple yet effective solution to this by leveraging the existing CMUdict, avoiding use of large amounts of data and complex architectures to model IE speech.

All of the English phones in CMUdict have been mapped to the existing Common Phone Set (CPS) phones proposed by Ramani et al. (2013) (Black et al., 1998). CPS is a language-independent set of the phones of six major Indian languages (Hindi, Marathi, Bengali, Tamil, Malayalam and Telugu) which maps parallel phones to the same code. Reusing CPS was favoured over creating a new encoding because it already provides a phone dictionary for a large set of Indian languages, and is continuously extended to include a greater number. Since English, too, is becoming more widespread in India, extending CPS to include the proposed Indian English phone dictionary is the logical consequence. While some of the mappings in this paper are inspired by Ganji et al. (2019), we did make changes to further reflect the accent that native Indian language speakers would have when speaking English. These changes are elaborated on in Section 3.2.

3.1 Lexicon Construction

Indian English differs notably from standard American or British English. IE speech also exhibits further variation according to the other languages prevalent in each region IE-CPS models only those sound transformations which can be observed among a majority of the different linguistic regions. The first of these is the absence of dental fricatives ([ð], [θ]) and alveolar plosives ([d], [t]) in Indian languages, replaced by dental plosives ([d̪], [t̪ʰ]) and retroflex plosives ([ɖ], [ʈ]) respectively. The second notable distinction is the use of long vowels in place of diphthongs (barring two cases – [aʊ] and [aɪ]). The third is the lack of distinction between the sounds [v] and [w] in Indian languages. The final difference is the approximation of vowel sounds to the nearest vowel in the speakers' native language. Additional differences between American English and Indian English, such as the prosodic features of GAE and the phonemic aspiration in IE, do not exist as differences at the phonetic level; hence these are not reflected in the proposed lexicon.

As Yarra et al. (2019) have already pointed out, Indians speaking English do not add or delete phonemes as often as they substitute them. Our independent analysis also leads to similar numbers to Yarra et al. (2019)'s work. While deletions happen very rarely ($< 1\%$), insertions happen a little more often ($< 15\%$). However, substitutions occur frequently when Indians speak English ($> 30\%$)[1]. Later in this paper, we show that just substituting phones causes a drastic drop in WER. IE-CPS describes these substitutions, but has no rules for insertion or deletion. The advantage of such an approach is the ability to remove the letter to phoneme alignment module, which considerably simplifies the pronunciation model.

We also noticed that the behaviours like insertions differ significantly based on regional accent – for example, frequent vowel epenthesis by speakers of Hindi. Since such sound changes are not common among all accents, these rules are required only when modelling specific regional accents. Thus, we have not included rules of phone insertion in our lexicon. Additionally, this also simplifies the use of the IE-CPS in multilingual data containing English – IE-CPS can easily be applied to each variety of Indian English since accent specific behaviour is not replicated.

[1]For the remaining pronunciations, no errors were observed.

3.2 Indian English Common Phone Set

Based on previously mentioned observations, the changes made to the CMU pronunciation dictionary are broadly classified into four categories:

- **Mapping of English phones to existing CPS codes.** Ganji et al. (2019) introduce new codes for phones AA and AO, i.e., ao. The IE-CPS maps these phones to the existing CPS code, ou. We chose this specific CPS code because the phone that it denotes is perceptually similar, being the closest phone in the Indian phone inventory. This has the advantage of minimising the need for introducing new phonetic codes. Along with these, various other consonant and vowel sounds have been mapped to codes in CPS which are perceptually similar but not the same phone.

- **Phones with multiple mappings.** Some examples are Z and ZH. Z gets mapped to z and j. Similarly, ZH gets mapped to three different CPS codes – z, j and jhq. This is because while many Indian languages exhibit free variation between the sounds of j and z, the increasing exposure of Indians to both American and British varieties of English is resulting in a new phone (represented by jhq) being gradually realised as a separate phoneme. The use of multiple allophones is absent in the codes proposed by Ganji et al. (2019), and both Z and ZH are mapped to z.

- **Addition of new mappings.** Huang et al. (2020); Yarra et al. (2019) explain how the diphthongs in English (barring ai and au) are pronounced as a single long vowel when Indians speak English. To reflect this, we map EY and OW to existing CPS codes ee and oo respectively, which represent long utterances of the first vowels in their corresponding diphthongs. This is in contrast to the use of ei and o respectively by Ganji et al. (2019), which are codes for shorter vowel utterances.

- **Exclusion of some CPS mappings.** English does not have phones analogous to all the CPS codes in its phoneme inventory. Those absent in Indian English (like the phones corresponding to q, txh, lx, etc) have thus not been included in the proposed lexicon.

Example	CMUdict		IE-CPS	
	ARPAbet	IPA	Code	IPA
odd	AA	ɑ	ou[2]	ɔ
at	AE	æ	ae	æ
h**u**t	AH	ʌ	a	a
ought	AO	ɔ	ou	ɔ
c**ow**	AW	aʊ	au	aʊ
h**i**de	AY	aɪ	ai	aɪ
be	B	b	b	b
cheese	CH	ʧ	c	ʧ
dark	D	d	dx	ɖ
this	DH	ð	d	d̪
egg	EH	ɛ	e[3]	ɛ
h**ur**t	ER	ɝ	**er**	əː
w**ai**t	EY	eɪ	ee	eː
fin	F	f	f	f
game	G	g	g	g
hill	HH	h	h	ɦ
s**i**t	IH	ɪ	i	ɪ
eat	IY	i	ii	iː
joke	JH	ʤ	j	ʤ
key	K	k	k	k
let	L	l	l	l
map	M	m	m	m
note	N	n	n	n
si**ng**	NG	ŋ	ng	ŋ
b**oa**t	OW	oʊ	oo	oː
t**oy**	OY	ɔi	**oy**	ɔi
pen	P	p	p	p
read	R	ɹ	r	ɾ
sea	S	s	s	s
show	SH	ʃ	sh	ʃ
tea	T	t	tx	ʈ
thin	TH	θ	th	t̪ʰ
p**u**t	UH	ʊ	u	u
f**oo**d	UW	u	uu	uː
villa	V	v	w	ʋ
we	W	w	w	ʋ
yes	Y	j	y	j
zip	Z	z	z	z
zip	Z	z	j	ʤ
sei**z**ure	ZH	ʒ	jhq[4]	ʒ
sei**z**ure	ZH	ʒ	z	z
sei**z**ure	ZH	ʒ	j	ʤ

Table 1: The proposed Indian English Common Phone Set (IE-CPS). The IE-CPS codes marked in **bold** in column 4 indicate the phones unique to Indian English.

The complete set of Indian English phones is given in Table 1. Each of the phones of GAE, as provided by CMUdict, has been mapped to existing phones from CPS. In cases where this was not possible, additional phones unique to the Indian English phone set have been added to CPS. This lexicon was verified by a trained linguist specialised in the study of Indian languages.

Word	CMUdict	IE-CPS
thought	TH AO T	th ou tx
waited	W EY T IH D	w ee tx i dx

Table 2: Comparison of CMUdict and IE-CPS with example words.

In Table 2, we show the phone break down of two examples in the case of CMUdict and proposed dictionary. In the examples taken, i.e, the words 'waited' and 'thought', one can notice the following differences:

- **Shift in place of articulation.** This change can be seen in several places in the examples taken. In both the example words, the code T in CMUdict is mapped to tx in IE-CPS. tx denotes a retroflex plosive whereas in the case of CMUdict the place of articulation is alveolar. As the latter is absent from Indian languages, the closest phoneme (here, tx) is substituted. Similar substitution is seen for the CMUdict codes TH, W and D (replaced by th, w and dx respectively).
- **Diphthong to long vowel.** As already mentioned, the diphthong in "waited" is pronounced as a long vowel when Indians speak English. The same is encoded when the code EY in CMUdict is mapped to ee in IE-CPS

Note here that not all sounds are changed in the Indian pronunciation of the words. The codes AO (CMUdict) and ou (IE-CPS) both represent the same phone [ɔ]. Hence, the vowel sound in 'thought' is the same in both American and Indian varieties of English.

[2]In CPS, this code is mapped to [oʊ] but the character this code represents is also allophonic to [ɔ]. Hence it is used in the latter sense here.

[3]Similar to the previous case, the character associated with this code is mapped to the allophones [e] and [ɛ], the latter of which is used here.

[4]Officially this is included in CPS, but in practical use the character representing this phone is never seen in text, or is pronounced the same as jh ([dʒʰ]).

4 Experimental Setup

In this subsection, we elaborate on the GAE and IE speech corpora upon which we ran our experiments. We also briefly describe two components of the ASR system – the acoustic model and the language model. These components have not been changed in the study, and are only used to highlight the effectiveness of the IE-CPS as a lexicon for Indian English ASR.

4.1 Corpora

The experiments are conducted on the following datasets. The details about the datasets are explained below:

1. **NPTEL Data**: National Programme on Technology Enhanced Learning (NPTEL)[5] is an open-source e-learning platform which is mainly maintained and organised by top-tier academic institutes from India (like Indian Institute of Technology (IIT) and Indian Institute of Science (IISc)). It covers a wide range of courses, including engineering, basic sciences, management, law, and personality development. As part of the challenge conducted by IIT Madras, the organisers released 80 hours of read speech and 200 hours of recorded lectures (on Computer Science and Electrical Engineering) from NPTEL[6]. The test set contains 12 hours of read speech and 10 hours of lectures (spontaneous speech). We report the results on both kinds of speech separately in Section 5. All of the audio is sampled at 16Khz. There is no explicit speaker information given because the dataset was shared by the organisers of the ASR challenge. However, we do confirm that all of the recorded speech is in Indian English. The speech is well distributed among several regions and diverse accents, thus preventing an issue of overfitting on a single variety of Indian English.
2. **Librispeech**: This is a corpus of approximately 1000 hours of 16Khz American English[7] read speech. The dataset is derived from the audiobooks that are part of the Librivox project. We take 960 hours of the corpus

[5]https://nptel.ac.in/

[6]https://sites.google.com/view/englishasrchallenge/home

[7]Strictly speaking this is not fully American English but a mix. However, most of the audiobooks part of the dataset use American English.

for training the model and 5 hours for testing it. There are no overlapping speakers in the train and test sets. There are a total of 1166 speakers in this dataset. The dataset is gender balanced to ensure that there is no bias towards one gender.

4.2 Acoustic Model

The acoustic model (AM) includes a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), a hybrid Deep Neural Network (DNN-HMM), and a Time Delay Neural Network (TDNN) (Peddinti et al., 2015).

In this work, Mel-frequency cepstral coefficients (MFCC) features, computed for a 20 ms window with 10 ms overlap, are fed into the system for initial speaker-independent GMM-HMM training. The speaker-dependent GMM-HMM model is built using Feature space Maximum Likelihood Linear Regression (fMLLR) features (Gales, 1998). For DNN-based acoustic model training, we use three hidden layers. Each hidden layer contains 2000 dimensional hidden units with p-norm activation. As the input p-norm dimension is 2000, the resultant output p-norm dimension is of 400 dimensions ($p = 2$ and group size=5). The DNN has an input layer that takes 360-dimensional input, and the DNN's output is 2365 context-dependent phonemes states.

The outputs are obtained by GMM-HMM alignment. The cross-entropy loss is minimised by using back-propagation with an initial learning rate of 0.01 and a final learning rate of 0.001. In addition to these 39-dimensional MFCCs, 100-dimensional iVectors (Dehak et al., 2010) are appended at each time step. It has been noted that iVectors capture both speaker and environment-specific information and are useful for rapid and discriminative adaptation of the neural network. The training data was increased 3-fold artificially through time-warping of raw audio (Ko et al., 2015). The training procedure for chain models is a Lattice-Free (LF) version of the Maximum Mutual Information (MMI) (Povey et al., 2016) criterion without the need for frame-level cross-entropy pre-training.

We use the chain TDNN model comprising 6 layers and 725 Rectified Linear Units (ReLU) at the input layer. The input features are at the original frame rate of 100 per second, and the output frame is reduced by 3 times. The first splicing is removed before the Linear Discriminant Analysis (LDA) transformation layer. The spliced indices in the consecutive layers were $[-1, 0, 1; -3, 0, 3; -3, 0, 3; -3, 0, 3; -6, -3, 0]$ with LDA applied to the input features.

4.3 Language Model

The Language Model (LM) predicts the probability of a hypothesised word sequence by learning the correlation between words for given text corpora. In this study, the text data corresponding to the respective audio is normalised. The SRI Language modelling toolkit (Stolcke, 2002) is used to train Kneser-Ney (Chen and Goodman, 1999) smoothed tri-gram LM on the text corpora. No external text is used.

All the experiments (both acoustic and language modelling) have been conducted using the Kaldi speech recognition toolkit (Povey et al., 2011). The experiments on Librispeech were conducted using the standard Kaldi s5 recipe[8]. The same recipe was adapted to run experiments on the NPTEL data. The training was carried out using GeForceGTX 2080 Ti. The metric used for evaluating the ASR system's performance is word error rate (WER).

5 Results

This section presents our results and analysis on IE-CPS for acoustic model training. Further, we compare and analyse the scenarios in which it outperforms the existing CMUdict. All the experiments were conducted on Librispeech and NPTEL data as mentioned in Section 4.1. For evaluating IE-CPS on a speech recognition task, two sets of experiments were performed with varying conditions, i.e., read and spontaneous condition.

5.1 Experiments with CMUdict

In the first study, the acoustic model is trained using CMUdict on both the speech corpora. The obtained results are shown in Table 3. As expected, the performance of the model trained on the Librispeech corpus was much better than the performance of the model trained on the NPTEL database. This held true for all three architectures of the acoustic model. The primary reason for this is that Librispeech is oriented to the American accent, which was the reference for the pronunciation dictionary.

[8] https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/run.sh

Corpus	Training	Testing	Acoustic Model	WER (%)	
				CMUdict	IE-CPS
Librispeech	Read	Read	GMM	7.02	16.72
			DNN	4.90	15.85
			TDNN	4.43	13.08
NPTEL	Read + Lecture (Spontaneous)	Read	GMM	10.64	**6.52**
			DNN	8.25	**4.89**
			TDNN	6.67	**3.76**
		Lecture	GMM	12.45	**8.12**
			DNN	10.97	**6.48**
			TDNN	8.85	**4.90**

Table 3: Comparison of results using CMUdict and IE-CPS.

5.2 Experiments with IE-CPS

In this study, the ASR pipeline uses the proposed IE-CPS lexicon instead of the CMUdict. The WER (%) for the same is shown in Table 3. The use of the IE-CPS improves the performance of the ASR trained on NPTEL (Indian-English) data. On the other hand, it performs worse than CMUdict when the ASR is trained on Librispeech. This further demonstrates the significant difference between the Indian and American varieties of English. Table 4 shows an example of how the ASR trained on IE data has poorer performance when modelled using CMUdict, a lexicon not specifically designed for it. The highlighted mistake made with CMUdict is corrected when using IE-CPS, as a direct result of applying substitution rules with proper linguistic foundation.

In the example provided in Table 4, [ð] (from 'th' in 'this') is realised as [d̪] in IE speech due to their phonetic similarity. However, [d̪] is also phonetically similar to [d] (from 'd' in 'disappointing'). This poorly defined contrast between [ð] and [d] adversely affects the LM when the ASR is modelled using CMUdict, but since the distinction is clear in the IE-CPS mappings, the ASR using IE-CPS is successful at disambiguating such cases.

6 Discussion & Analysis

In this section we investigate the behaviour of the IE-CPS in more detail by asking the questions detailed below.

If Indian English varies so much with region, then how can this common lexicon be useful? It has been mentioned earlier that the regional languages have a significant effect on spoken English. This results in several different accents and thus phoneme inventories. However, IE-CPS is based on the general speech characteristics observed in speakers of a range of languages. The rules applied to map the phones from American to Indian English are generated from the most prominent and widespread approximations made by Indians speaking English, such as the transformation from alveolar to retroflex plosives. The notable improvement demonstrated by experimental results performed on the NPTEL speech data (consisting of a broad range of regional Indian accents) also proves the overall effectiveness of the lexicon despite the variation in Indian English.

Consequently, it may be extrapolated that such a lexicon can be applied to the English in regions where academic effort to model the speech is still deficient – such as North-East India, with languages Bodo, Khasi, Assamiya, Meitei, Mizo – without any extra linguistic expertise on the matter. Any required region specific tuning also becomes easier as there are fewer changes to make, which can be done using the already mostly standardised encoding of CPS. Furthermore, if there is the need for a lexicon for multilingual speech (also containing English), such a pronunciation dictionary with the phones common to most varieties of Indian English would serve to better represent the multiple possible accents.

End-to-End models do not need a lexicon. What is the need for such a lexicon? With sufficiently large amounts of data, it is possible to discard pronunciation models entirely and build end-to-end (E2E) ASR systems. However, as it

Lexicon	Sentence
IE-CPS	**Its disappointing** when you are convinced that there is nothing
CMUdict	**This appointing** when you are convinced that there is nothing
Ground Truth	Its disappointing when you are convinced that there is nothing

Table 4: Comparison of Indian English ASR system when using IE-CPS and CMUdict. The words in **bold** were the ones confused in the latter case.

is with many Indian languages, if the language (or the accented language in this case) is low resource then E2E models would not be a good solution. In such cases, conventional hybrid systems perform better. Hybrid systems require us to explicitly model the pronunciation of the language, i.e., a mapping from grapheme to phoneme (G2P) is required. While one can model this with sequence-to-sequence (seq2seq) architectures (Gorman et al., 2020; Peters et al., 2017), even that would require relatively large amounts of G2P data. In cases where even that is not available, rule-based parsers are the only choice for researchers wanting to model a language or an accent. Additionally, even with large training data, seq2seq models the *long tail* in the distribution of the data causes a deterioration in the overall performance. Rule-based parsers avoid this issue as they do not rely on learning the data. In such scenarios, the rule-based IE-CPS can be used to obtain considerable improvements as shown in Table 3. In fact, the lexicon can directly be substituted in place of another lexicon in an existing ASR pipeline, making its integration simpler. The other advantages of this lexicon are that, being completely rule-based, it would require no training data or additional computational resources as opposed to seq2seq-based G2P systems.

7 Future Work

The current IE-CPS generalises several characteristics of spoken Indian English. However, our preliminary analysis reveals that fine-tuning the proposed lexicon to a particular regional accent of Indian English, such as Bengali or Malayalam, results in a better ASR performance when testing on the same regional accent. For example, a Bengali accent would cause [a] to be pronounced as [ɔ], and aspiration of consonants would be largely absent for a Malayalam speaker. Such specific changes can further be adapted in IE-CPS to reflect regional varieties of Indian English. Therefore, we plan to propose different lexicons for different regional accents of Indian English in the future.

There is a growing amount of interest in modelling code-switched speech, especially in the Indian community (Manghat et al., 2020). This can be seen by how successful the code-switching subtask was in the 2021 *Multilingual and code-switching (MUCS)* ASR challenges for low resource Indian languages[9] (Diwan et al., 2021). We also noticed that there is an increased interest in having a specialised pronunciation block for such settings. In this direction, we plan to use IE-CPS to model the Indian English part of the code-switched speech. This is especially useful because the other Indian language can be modelled with the same set of phone codes – CPS. Therefore, such a lexicon would be easy to integrate into existing frameworks in such settings.

8 Conclusions

In this paper, we propose a lexicon designed for use in ASR – the Indian English Common Phone Set (IE-CPS). The IE-CPS is a lexicon that can be easily integrated into the existing, (largely) standardised phonetic code for Indian languages – the Common Phone Set (CPS). We show the improvements in an Indian English ASR system when IE-CPS is used as the pronunciation model. This proves that fine-tuning the pronunciation model to Indian English when the ASR system is deployed to work on Indian English speech is the correct way going forward. This is especially true when the user is limited by either the available data or the computational resources to utilise the data.

Acknowledgements

We would like to thank Dipti Misra Sharma for her assistance in verifying and validating this IE-CPS lexicon.

[9] https://navana-tech.github.io/IS21SS-indicASRchallenge/

References

Aryaman Arora, Luke Gessler, and Nathan Schneider. 2020. Supervised grapheme-to-phoneme conversion of orthographic schwas in hindi and punjabi. *ArXiv*, abs/2004.10353.

R. K. Bansal. 1969. *The intelligibility of Indian English: measurements of the intelligibility of connected speech, and sentence and word material, presented to listeners of different nationalities*. Central Institute of English.

Alan W Black, Kevin Lenzo, and Vincent Pagel. 1998. Issues in building general letter to sound rules. In *The third ESCA/COCOSDA workshop (ETRW) on speech synthesis*.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan K. M., Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish R. Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, Karthik Sankaranarayanan, Tejaswi Seeram, and Basil Abraham. 2021. Multilingual and code-switching ASR challenges for low resource indian languages. *CoRR*, abs/2104.00235.

Emmanuel Ferragne and François Pellegrino. 2004. A comparative account of the suprasegmental and rhythmic features of british english dialects. *Modelisations pour l'Identification des Langues*.

Mark JF Gales. 1998. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98.

Sreeram Ganji, Kunal Dhawan, and Rohit Sinha. 2019. Iitg-hingcos corpus: A hinglish code-switching database for automatic speech recognition. *Speech Communication*, 110:76–89.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50, Online. Association for Computational Linguistics.

Xian Huang, Xin Jin, Qike Li, and Keliang Zhang. 2020. On construction of the asr-oriented indian english pronunciation dictionary. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6593–6598.

Abhinav Jain, Minali Upreti, and Preethi Jyothi. 2018. Improved accented speech recognition using accent embeddings and multi-task learning. In *Interspeech*, pages 2454–2458.

Sunhee Kim, Kyuwhan Lee, and Minhwa Chung. 2011. A corpus-based study of english pronunciation variations. In *Twelfth Annual Conference of the International Speech Communication Association*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech*.

Sreeja Manghat, Sreeram Manghat, and Tanja Schultz. 2020. Malayalam-English Code-Switched: Grapheme to Phoneme System. In *Proc. Interspeech 2020*, pages 4133–4137.

Peggy Mohan. 2021. *WANDERERS, KINGS, MERCHANTS: the story of india through its languages*. PENGUIN.

David R Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision g2p for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

S. Nair, Cr Rechitha, and C. S. Kumar. 2013. Rule-based grapheme to phoneme converter for malayalam.

Raymond WM Ng and Keikichi Hirose. 2012. Syllable: A self-contained unit to model pronunciation variation. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4457–4460. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Alok Parlikar, Sunayana Sitaram, Andrew Wilkinson, and Alan W Black. 2016. The festvox indic frontend for grapheme to phoneme conversion. In *WILDRE: Workshop on Indian Language Data-Resources and Evaluation*.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively multilingual neural grapheme-to-phoneme conversion. *arXiv preprint arXiv:1708.01464*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Proc. Interspeech*, pages 2751–2755.

B Ramani, S Lilly Christina, G Anushiya Rachel, V Sherlin Solomi, Mahesh Kumar Nandwana, Anusha Prakash, S Aswin Shanmugam, Raghava Krishnan, S Kishore Prahalad, K Samudravijaya, et al. 2013. A common attribute based unified hts framework for speech synthesis in indian languages. In *Eighth ISCA Workshop on Speech Synthesis*.

Marco Schilk. 2010. Pingali sailaja, indian english (dialects of english). edinburgh: Edinburgh university press, 2009. pp x 172. hardback £60.00, isbn 978-0-7486-2594-9, paperback £19.99, isbn 978-0-7486-2595-6. *English Language and Linguistics*, 14(1):135–139.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proc. Seventh International Conference on Spoken language Processing*.

A. Wasala, R. Weerasinghe, and K. Gamage. 2006. Sinhala grapheme-to-phoneme conversion and rules for schwa epenthesis. In *ACL*.

Robert Weide. 2005. The carnegie mellon pronouncing dictionary [cmudict. 0.6]. *Pittsburgh, PA: Carnegie Mellon University*.

J. Wells. 1995. Computer-coding the ipa: a proposed extension of sampa.

Chiranjeevi Yarra, Ritu Aggarwal, Avni Rajpal, and Prasanta Kumar Ghosh. 2019. Indic timit and indic english lexicon: A speech database of indian speakers using timit stimuli and a lexicon from their mispronunciations. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.