Accepted Manuscript

SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP) and of the International Speech Communication Association (ISCA)

IITG-HingCoS Corpus: A Hinglish Code-Switching Database for Automatic Speech Recognition

Sreeram Ganji, Kunal Dhawan, Rohit Sinha

PII:	S0167-6393(18)30421-7
DOI:	https://doi.org/10.1016/j.specom.2019.04.007
Reference:	SPECOM 2643

To appear in: *Speech Communication*

Received date:	6 December 2018
Revised date:	22 March 2019
Accepted date:	16 April 2019

Please cite this article as: Sreeram Ganji, Kunal Dhawan, Rohit Sinha, IITG-HingCoS Corpus: A Hinglish Code-Switching Database for Automatic Speech Recognition, *Speech Communication* (2019), doi: https://doi.org/10.1016/j.specom.2019.04.007

IITG-HingCoS Corpus: A Hinglish Code-Switching Database for Automatic Speech Recognition

Sreeram Ganji, Kunal Dhawan, and Rohit Sinha

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati - 781039, India

Abstract

Code-switching is a phenomenon in linguistics which refers to the use of two or more languages, especially within the same discourse. This phenomenon has been observed in many multilingual communities across the globe. In the recent past, there have been increasing demand for automatic speech recognition (ASR) systems to deal with code-switching. However, for training such systems, very limited code-switching resources are available as yet. Thus, the development of code-switching resources is highly desirable. In this work, we describe the collection of a Hinglish (Hindi-English) code-switching database at the Indian Institute of Technology Guwahati (IITG) which is referred to as the *IITG-HingCoS* corpus. This corpus consists of code-switching text data having 25,988 sentences with a total of 0.58 million words. In addition to that, the corpus also contains 25 hours of matching speech data corresponding to 9,251 code-switching sentences covering a vocabulary of 6,542 words. This paper elaborates the sources and the protocol used for collecting the corpus. The baseline experimental results on the collected corpus for language modeling and ASR tasks are also presented.

Keywords: code-switching, speech and text corpora, automatic speech recognition, language modeling

In multilingual communities, the speakers often switch or mix between two or more languages or language varieties during communication. In linguistics, this phenomenon is referred to as code-switching [1, 2]. Over the decades, due to colonization and globalization, a lot of people have migrated from one linguistic region to another for better trade opportunities and livelihood. In such situations, communicating in two or more languages helps people to interact better. It has been observed that, over the passage of time, such mixed linguistic communities tend to code-switch in order to mingle well culturally. Towards emphasizing the importance of more research on code-switching, we highlight the salient examples of multilingual communities across the world. In the United States of America, the use of Spanglish (Spanish-English) grew with immigration from many Spanish speaking countries in South-America and even Spain [3, 4]. The code-switching societies in North-Africa include the Arabic-English in Egypt [5] and the French-Arabic in Algeria, Morocco, and Tunisia [6]. South-Africa being a multi-ethnic country provides ample examples of code-switching communities such as Sepedi-English, English-isiZulu, English-isiXhosa, English-Setswana, and English-Sesotho [7]. Whereas, the French-German in Switzerland [8] and the Frisian-Dutch in the Netherlands [9] are two dominant examples of code-switching communities in Europe. In Asia, we find code-switching examples in the form of Malay-English [10] and Mandarin-English [11] in Malaysia, the Cantonese-English in Hong Kong [12], and the Mandarin-Taiwanese [13, 14] in Taiwan. Though the phenomenon of code-switching is widespread across the world, the research activity in this domain is somewhat limited due to lack of availability of the domain-specific resources.

India is the second most populous country in the world and has 23 official languages including English. After gaining independence from the British rule, though the Indian constitution declared Hindi as the primary official language, the usage of English was continued as a secondary language for its dominance in administration, education, and law [15, 16]. Thereby, communicating with frequent use of English

Email address: {s.ganji, k.dhawan, rsinha}@iitg.ac.in (Sreeram Ganji, Kunal Dhawan, and Rohit Sinha)

Table 1: Example Hinglish sentences along with their English translated versions showing the inter-sentential code-switching and the variants of the intra-sentential code-switching. Type-1 and Type-2 variants of intra-sentential code-switching refer to high and low contextual information being carried by the non-native (English) words, respectively.

Inter-sentential code-switching	Example-1		she is the daughter of our ceo, वह यहाँ दो दिन के लिए आई है she is the daughter of our ceo, she has come here for two days
	Example-2		मुझे अमेरीका में रहते चार साल हो गए, but I still miss my country I have been living in america for four years, but I still miss my country
Intra-sentential code-switching	Type-1	Example-1	कृपया मुझे मेरा current account balance बताएं please tell me my current account balance
		Example-2	देश की currency every year change होनी चाहिए the country's currency should change every year
	Type-2	Example-1	class और object के बीच relationship क्या है what is the relationship between class and object
		Example-2	meeting का outcome क्या था what was the outcome of meeting

words/phrases was considered as a sign of not only being in power/educated but also more trendy. Over the years, substantial code-switching to English while speaking the native Indian languages has become a common trend across India, in particular, by the urban population [17, 18]. As per the 1991 census of India [1], 40.2% of the population are native Hindi speakers which is followed by Bengali (8.3%) and Telugu (7.9%). Therefore, one can find frequent code-switching between Indian languages and English, with Hindi-English (Hinglish) being the most dominant one. However, a large sized code-switching speech corpus is yet to be created for Indian languages. At present, a small sized Hinglish corpus which contains about 30 minutes of speech data recorded from 9 speakers is reported [19]. In this work, we attempt towards addressing that gap by creating a larger Hinglish corpus. During British colonial rule, many words from English were borrowed in Hindi. Majority of them were proper nouns including a few collective/abstract nouns which, over the period, got internalized in Hindi for the ease of reference or the lack of acceptable equivalent Hindi words. In this study, except for the proper nouns borrowed from English, all remaining embedded English words are assumed to be code-switched.

A number of causes have been attributed to the prevalence of code-switching phenomenon. The people belonging to bilingual communities opine that the primary reason for code-switching between languages is the lack of words in the vocabulary of the respective native languages [20]. According to [19, 21–25], some of the other possible reasons for code-switching are: (i) to qualify the message by emphasizing specific words, (ii) to maintain confidentiality during the verbal communication, (iii) to show expertise, authority, status, etc, and (iv) to enrich communication between speakers without any change in the message. The recent works [26, 27] have highlighted that the code-switching phenomenon is also observed in the textual chats, comments, and messages posted on the social media sites like Facebook, Twitter, WhatsApp, YouTube, etc. Current literature on code-switching can be grouped into three broad research areas: (i) linguistics, (ii) natural language processing (NLP), and (iii) speech processing. The researchers in linguistics have studied the impact of code-switching from the point of view of socio-linguistics [21, 28] and syntactics [25, 29]. In [21], the researchers highlighted the socio-psychological and linguistic factors behind the code-switching phenomenon. The authors in [28] examined the grammatical structure and specific syntactic boundaries of a language that occurs due to code-switching. In [25], based on the locations of the non-native words, code-switching was broadly classified into two modes. When the switching happens within the sentence it is referred to as the *intra-sentential* code-switching and the one predominantly happening at the sentence boundary is referred to as the *inter-sentential* code-switching [29]. Intra-sentential mode of switching is a common phenomenon and has become an identifying characteristic in bilingual communities [30]. Table 1

[1] http://www.ciil-lisindia.net/

shows a few example Hinglish and their translated English sentences with different modes of code-switching while highlighting the differences in the contextual information carried by the non-native words. In Type-1 intra-sentential code-switching, the non-native language words either occur in sequence or form a phrase, thus carry some contextual information. Whereas, in Type-2 case, the non-native language words are embedded into the native language sentences in such a manner that virtually no contextual information could be derived from those words alone. Also, in our collected database, it is observed that about 70% of the sentences belong to Type-2 intra-sentential mode. In NLP domain, the researchers explored the code-switching phenomenon for language modeling [5, 12, 31] and machine translation [32] tasks. In speech processing area, there have been interesting research challenges in code-switching domain which include acoustic modeling [10, 13, 33], language identification [14, 34] and speech synthesis [35].

However, the effective handling of code-switching in the above said applications is still quite challenging in contrast to that of the monolingual case. According to [34, 36, 37], the possible causes for the same are: (i) the duration of the embedded foreign words/phrases can be very short, (ii) unpredictability of the occurrence of the code-switching instances in an utterance, (iii) pronunciation variation from the native language to the foreign language within the same utterance, (iv) requirement of dedicated tools, and (v) lack of publicly available resources. Motivated by the efforts done elsewhere, we endeavored to create an open resource for research in automatic speech recognition (ASR) of code-switching speech in the Indian context. A monolingual ASR system may be capable of recognizing a few words from a foreign language but are unable to handle a significant amount of code-switching in the data. In the recent past, researchers have reported that the native language of the speaker influences the foreign (non-native) language acquisition [38]. In India, English is taught from the elementary level in the schools across the country. Despite that, the English pronunciations of the majority of the pupils carry significant native language influences. On account of the existence of variants of English pronunciations and code-switching effects, the development of an ASR system for Hinglish code-switching speech data is a challenging task. To the best of our knowledge, there is no large-sized Hinglish corpus publicly available. Towards addressing that constraint, we created a Hinglish code-switching text corpus. Along with that, a Hinglish speech corpus is also created that covers all typical sources of variations such as accent, session, channel, age, gender, the influence of the mother tongue. The sentences spoken in the speech corpus are a subset of the text corpus. This paper describes the details of the created Hinglish text and speech corpora and reports the baseline performances for the same.

The remainder of this paper is organized as follows. In Section 1, the details about the developed Hinglish text and speech corpora along with those of the lexical resources necessary for developing the Hinglish ASR system, are presented. The statistical analyses of both the corpora have been presented in Section 2. In Section 3, we review the code-switching text and speech corpora currently reported in the literature while contrasting with the one reported in this work. The experimental evaluations using the created Hinglish corpora have been presented in Section 4. The paper is concluded in Section 5.

1. Hinglish Code-Switching (HingCoS) Corpus

In this section, we describe a Hinglish code-switching corpus created primarily for ASR task at the Indian Institute of Technology Guwahati (IITG). The created corpus contains the code-switching text as well as the speech data and is formally referred to as the *IITG-HingCoS* corpus. But for brevity, we refer to that as the *HingCoS* (हिंगकोष) corpus in the rest of the paper. It has been pointed out by the researchers that the code-switching takes place both in spoken and written formats during spontaneous interactions [26, 27]. Like any monolingual speech corpus, the code-switching speech data can be collected in any of the three modes: (i) read speech, (ii) conversational/interview speech and (iii) radio/television broadcast speech. On the other hand, the collection of code-switching text data is tedious since its availability on traditional sources like newspaper/broadcaster websites is rather limited. For accessing the same, one has to access specific blogging websites and other social media portals. In the following, we describe the procedure followed for the collection of HingCoS text and speech data. Finally, the creation of lexical resources such as the common Hinglish phone set and the dictionary are presented.

1.1. HingCoS Text Corpus

At first, the text data has been collected by crawling a few Indian blogging websites that contain different contexts. The details of those websites and their context is discussed below.

- *ShoutMeHindi* [2]: It contains information about how to start a blogging website and how to earn money from it. It also explains about the social media websites (Facebook, Twitter, YouTube, etc.) and gives some tips and tricks to use them.
- *Computing Notes in Hinglis*h [3]: This blog explains about object oriented database management system in detail along with its design methodology and behavioral concepts.
- *Techyukti* [4] and *HindiMe* [5]: These websites discuss about more than 100 varieties of recent advancements in technology. The salient topics include mobiles, cameras, PUBG, WhatsApp, Paytm, online voter ID enrollment process, adhaar card application process, search engine optimization, etc.
- *LearnCpp* [6]: It is a free website having tutorials about C++ programming language that includes the steps to write, compile, and debug the C++ programs along with plenty of examples.

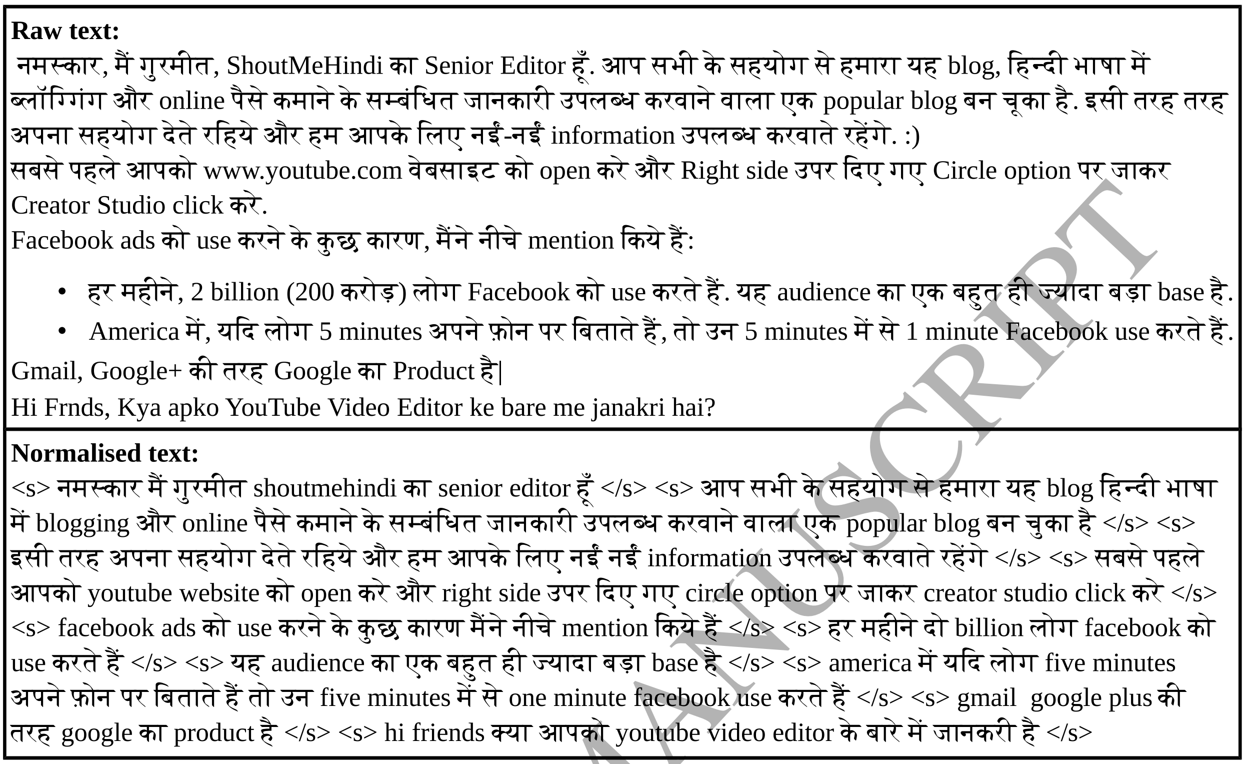
In all the above blogs, the bloggers have tried to explain the chosen context by writing in Hindi with frequent code-switching to English words/phrases. Also, the bloggers have followed their individualistic and rather casual writing styles. At times, even some Hindi and English words are written in cross scripts. As a result of that the collected data exhibits a lot of variabilities and its removal is not only essential but also a bit challenging. Therefore, the crawled data is first normalized before converting it into meaningful sentences and the steps followed for the text normalization is described below.

1.1.1. Steps involved in text normalization:

- Extra spaces, punctuation marks, special characters, bullet marks, emoticons etc., are filtered out.
- The braced explanations, website links, directory paths, etc., are filtered out while retaining the key information.
- All numerals, mathematical and currency symbols are replaced by their spellings either in Hindi or English based on the context.
- All the uppercase English characters in the data are converted to lowercase characters.
- Except for the proper nouns, all English and Hindi words written in cross scripts are fixed to have the correct scripts. Also, any error spotted in the spellings of Hindi and English words is fixed.
- All shorthand words are converted to their respective full forms, while all standard abbreviations are left as they are.
- Finally, the sentence begin and end markers are inserted to parse the data into meaningful sentences while removing any erroneous repetitions of words and phrases if spotted.

A few examples illustrating the above mentioned text normalizations are shown in Figure 1. Table 2 quantifies the impact of the text normalization process. Further, the validation of the text normalization in language modeling domain is reported in Table 3. A detailed analysis of the HingCoS text corpus is presented later in Section 2.

[2] https://shoutmehindi.com
[3] https://notesinhinglish.blogspot.in
[4] https://www.techyukti.com
[5] https://hindime.net
[6] http://www.learncpp.com



Raw text:

नमस्कार, मैं गुरमीत, ShoutMeHindi का Senior Editor हूँ. आप सभी के सहयोग से हमारा यह blog, हिन्दी भाषा में ब्लॉग्गिंग और online पैसे कमाने के सम्बंधित जानकारी उपलब्ध करवाने वाला एक popular blog बन चूका है. इसी तरह तरह अपना सहयोग देते रहिये और हम आपके लिए नईं-नईं information उपलब्ध करवाते रहेंगे. :)

सबसे पहले आपको www.youtube.com वेबसाइट को open करे और Right side उपर दिए गए Circle option पर जाकर Creator Studio click करे.

Facebook ads को use करने के कुछ कारण, मैंने नीचे mention किये हैं:

- हर महीने, 2 billion (200 करोड़) लोग Facebook को use करते हैं. यह audience का एक बहुत ही ज्यादा बड़ा base है.
- America में, यदि लोग 5 minutes अपने फ़ोन पर बिताते हैं, तो उन 5 minutes में से 1 minute Facebook use करते हैं.

Gmail, Google+ की तरह Google का Product है|

Hi Frnds, Kya apko YouTube Video Editor ke bare me janakri hai?

Normalised text:

<s> नमस्कार मैं गुरमीत shoutmehindi का senior editor हूँ </s> <s> आप सभी के सहयोग से हमारा यह blog हिन्दी भाषा में blogging और online पैसे कमाने के सम्बंधित जानकारी उपलब्ध करवाने वाला एक popular blog बन चुका है </s> <s> इसी तरह अपना सहयोग देते रहिये और हम आपके लिए नईं नईं information उपलब्ध करवाते रहेंगे </s> <s> सबसे पहले आपको youtube website को open करे और right side उपर दिए गए circle option पर जाकर creator studio click करे </s> <s> facebook ads को use करने के कुछ कारण मैंने नीचे mention किये हैं </s> <s> हर महीने दो billion लोग facebook को use करते हैं </s> <s> यह audience का एक बहुत ही ज्यादा बड़ा base है </s> <s> america में यदि लोग five minutes अपने फ़ोन पर बिताते हैं तो उन five minutes में से one minute facebook use करते हैं </s> <s> gmail google plus की तरह google का product है </s> <s> hi friends क्या आपको youtube video editor के बारे में जानकरी है </s>

Figure 1: Illustration of salient text normalizations that have been applied to raw Hinglish data collected from the web sources.

Table 2: Quantifying the impact of text normalization of the raw text data as described in Section 1.1.1. The non-characters include punctuation marks, braces, numerals, mathematical symbols, emoticons, etc.

Attribute	Raw	Normalized
Number of unique non-characters	1,657	Nil
Number of unique Hindi words	10,737	6,029
Number of unique English words	24,498	8,614
Vocabulary size	36,892	14,643
Number of sentences	12,284	25,988

Table 3: Validation of the text normalization process. In this experiment, two language models (LMs) are created using the raw and the normalized versions of the training set. For performance evaluation, a development and a test datasets which are non-overlapping to the training dataset, are also created. The corresponding perplexities of the LMs trained on raw- and normalized-text are reported. Note that, the out of vocabulary rate of the development and the test datasets with respect to the normalized training dataset turns out to be 16.9% and 18.2%, respectively.

Training dataset	Evaluation dataset	Perplexity
Raw	Development	351.02
	Test	359.61
Normalized	Development	174.93
	Test	176.43

1.2. HingCoS Speech Corpus

For creating the HingCoS speech corpus, about 30% of the sentences available in the HingCoS text corpus were randomly selected for recording the speech data from native Indian speakers. The selection of the sentences has been done in such a way that they cover the majority of the contexts described in Section 1.1. The speech data is collected over a toll-free telephone-based voice-server available in the Electro-Medical and Speech Technology (EMST) Laboratory at IITG. A group of students and residents of IITG participated in the speech data collection. Each of the volunteers was given 100 unique sentences for recording the speech data by calling to the voice-server. Those 100 sentences were further partitioned into 5 groups of 20 sentences each. The volunteers were asked to record each group of sentences in a separate session. We also recorded the meta-data of volunteers comprising the name, age, gender, mother tongue, and native state. The call flow of the voice-server used for recording the HingCoS speech corpus is shown in Figure 2.

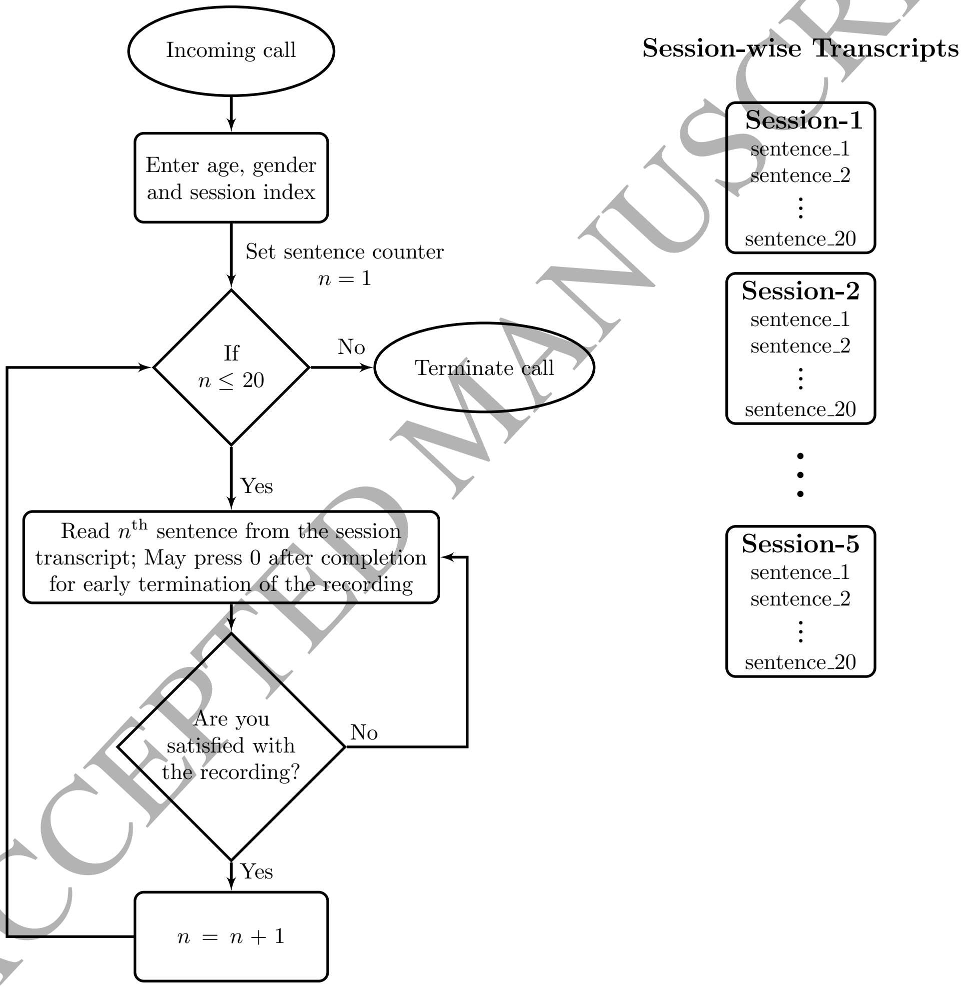

Incoming call
Enter age, gender and session index
Set sentence counter $n = 1$
If $n \leq 20$
No
Terminate call
Yes
Read n^{th} sentence from the session transcript; May press 0 after completion for early termination of the recording
Are you satisfied with the recording?
No
Yes
$n = n + 1$
Session-wise Transcripts
Session-1
sentence_1
sentence_2
sentence_20
Session-2
sentence_1
sentence_2
sentence_20
Session-5
sentence_1
sentence_2
sentence_20

Figure 2: The call flow of the voice-server used by the volunteers for the session-wise recording of Hinglish transcripts.

All the text data collected from the blogs is written in a style as if an expert is educating/explaining his/her audience about a topic. These sentences are read by the volunteers while recording the speech data, therefore we referred to it as "read" speech in this work. In order to ensure the quality and integrity of the collected speech data, each of the volunteers was given the following instructions:

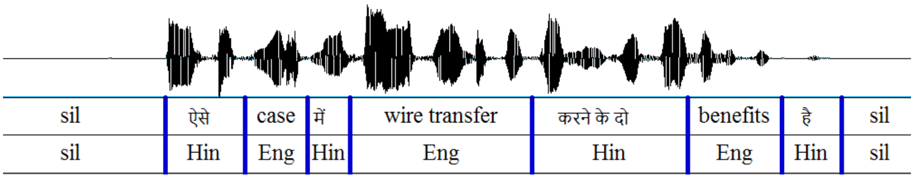


Figure 3: An example of the typical recorded Hinglish speech utterance marked with native and non-native word/phrase boundaries along with their corresponding language identity labels. The short-hand notations 'Hin' and 'Eng' are used to denote Hindi and English words/phrases, respectively.

- Read the given sentences at least two times before recording the speech data in order to sound them in a spontaneous manner.
- Prior to the actual recording, familiarize with the call-flow of the voice-server by conducting a few dummy recordings.
- Record each group of sentences in a separate session while keeping a gap of at least one day between the sessions and choosing different environments as far as possible.
- Enter the correct personal details as well as the session index. Carefully read the prompted sentence from the text transcript provided for that session.

Owing to the use of voice-server, the recorded speech data has the sampling rate of 8 kHz and the precision of 16-bits. Speech files corresponding to the read sentences are stored in *.wav* format and labeled as <speaker ID>_<gender>_<age>_<session index>_<sentence number> *.wav* (For example, if a 29 year old male speaker having unique speaker ID as Spk10, accessed the voice-server and read the 3^{rd} sentence from Session-1 transcript, then that particular speech file is labeled as Spk10_M_29_1_03*.wav*). Figure 3 shows a typical recorded Hinglish speech utterance marked with native and non-native word/phrase boundaries along with their corresponding language identity labels. Though the database collection protocol ensures a quality recording condition, a few challenges still exists due to technical issues, such as creation of empty/broken recordings due to call drop or power failure. Therefore, a lot of manual hours are invested to inspect and prune out any empty and broken recordings. We employed a voice activity detection (VAD) for removing the long silences and any non-overlapping background noises present in the recorded data. This VAD includes background noise suppression and was developed in an earlier work [39]. A detailed analysis of the HingCoS speech corpus is presented in Section 2.

1.3. Hinglish Lexical Resources

It is well known that lexicon plays an important role in the ASR task. It establishes the link between the acoustic representation of basic sound units and the symbols outputted by the ASR system. The design of a lexicon involves two steps: (i) fixing of the vocabulary covering the task, and (ii) the listing of all possible pronunciation variants of each word in the vocabulary. Unlike the monolingual task, the lexicon in code-switching task has to cover the words from two or more languages involved.

In this section, we discuss the development of a lexicon for Hinglish ASR system. Firstly, for compact acoustic modeling, we intended to define a phone set that covers all basic sounds present in both Hindi and Indian-English. To achieve that, a composite phone set is used which has been proposed recently in the context of computer processing of major Indian languages [40]. This composite phone set consists of 81 romanized labels that cover sounds in Hindi, Bengali, Marathi, Malayalam, Tamil and Telugu languages. Common romanized labels are assigned to the sounds across different languages based on their perceptual similarity. We extended that idea to define a common romanized phone set that covers the sounds in both Hindi and Indian-English languages without making any changes to the labels already defined for Hindi in [40]. For defining phone labels for Indian-English, we made use of 39 CMU ARPAbet labels along with their root words [7]. Each ARPAbet has been assigned to an existing Hindi label based on the perceptual similarity of the respective English root word being typically pronounced by Indian speakers. Whereas, we assigned our own romanized labels to a few ARPAbets that do not have perceptual similarity with any of the Hindi phones. The complete set consists of 62 romanized labels and is referred to as the Indian real pronunciation alphabets (IRPAbet) in this work. It is worth noting that owing to Indian accent, many ARPAbet labels get mapped to a single IRPAbet label.

Secondly, a unique word list is extracted from the developed HingCoS text corpus. The phoneme-level transcriptions for all those words have been done manually using the IRPAbet labels while covering all the pronunciation variations present in the speech data. Both the phonetic labels and the pronunciations are finally cross-checked by a linguist at our end. The lists of the IRPAbet labels that cover the sound units in Hindi and Indian-English languages are given in Table 4. The distribution of the vocabulary based on the pronunciation count is given in Figure 4.

Table 4: The IRPAbet consisting of 62 labels defined in this work for labeling sounds in both Indian-English and Hindi languages. For Indian-English, the CMU ARPAbet labels are assigned with Hindi phone set borrowed from [40] based on perceptual similarity, while a few new labels introduced to cover Indian-English are marked in gray colour. Owing to Indian accent, many ARPAbet labels get mapped to a single IRPAbet label.

S. No.	Hindi char.	IRPAbet	S. No.	Hindi char.	IRPAbet
1	अ	a	29	द	d
2	आ	aa	30	ध	dh
3	इ	i	31	न,ऩ	n
4	ई	ii	32	प	p
5	उ	u	33	फ	ph
6	ऊ	uu	34	ब	b
7	ऋ,ॠ	rq	35	भ	bh
8	ए	ee	36	म	m
9	ऐ	ei	37	य,य़	y
10	ओ	o	38	र,ऱ	r
11	औ	ou	39	ल	l
12	क	k	40	व	w
13	ख	kh	41	श	sh
14	ग	g	42	ष	sx
15	घ	gh	43	स	s
16	ङ	ng	44	ह	h
17	च	c	45	क़	kq
18	छ	ch	46	ख़	khq
19	ज	j	47	ग़	gq
20	झ	jh	48	ज़	z
21	ञ	nj	49	झ़	jhq
22	ट	tx	50	ड़	dxq
23	ठ	txh	51	ढ़	dxhq
24	ड	dx	52	फ़	f
25	ढ	dxh	53	◌ं	q
26	ण	nx	54	◌ः	hq
27	त	t	55	◌ँ	mq
28	थ	th			

S. No.	ARPAbet (example)	IRPAbet	S. No.	ARPAbet (example)	IRPAbet
1	AA (odd)	ao	29	S (sea)	s
2	AE (at)	ae	30	SH (she)	sh
3	AH (hut)	a	31	T (tea)	tx
4	AO (ought)	ao	32	TH (theta)	th
5	AW (cow)	au	33	UH (hood)	u
6	AY (hide)	ai	34	UW (two)	uu
7	B (be)	b	35	V (vee)	w
8	CH (cheese)	c	36	W (we)	w
9	D (dee)	dx	37	Y (yield)	y
10	DH (thee)	d	38	Z (zee)	z
11	EH (Ed)	e	39	ZH (seizure)	z
12	ER (hurt)	er			
13	EY (ate)	ei			
14	F (fee)	f			
15	G (green)	g			
16	HH (he)	h			
17	IH (it)	i			
18	IY (eat)	ii			
19	JH (gee)	j			
20	K (key)	k			
21	L (lee)	l			
22	M (me)	m			
23	N (knee)	n			
24	NG (ping)	ng			
25	OW (oat)	o			
26	OY (toy)	oy			
27	P (pee)	p			
28	R (read)	r			

[7] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Figure 4: The distribution of the vocabulary based on the pronunciation count. Note that, the lexicon contains a total of 8,911 entries out of which 6,616 entries are unique.

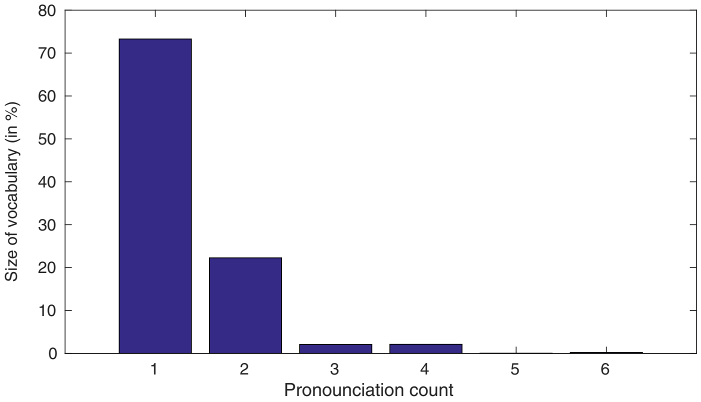


Table 5: Key statistics of the HingCoS text corpus.

# sentences	# words		# unique words		# code-switching instances
	Hindi	English	Hindi	English	
25,988	381,603	196,556	6,029	8,614	104,912

2. Statistical Analysis of the HingCoS Corpus

In this section, we present the salient attributes of the HingCoS text and speech corpora. First, we describe the key statistics of the HingCoS text corpus and it is followed by the details of the HingCoS speech corpus. Later, different distributions of the attributes of the text and speech corpora are plotted.

2.1. Analysis of the HingCoS Text Corpus

The HingCoS text corpus consists of $25,988$ sentences that are covered by a vocabulary of $14,643$ words ($6,029$ Hindi and $8,614$ English). The lengths of the sentences vary from 3 to 57 words. It is worth highlighting that the collected text corpus contains $104,912$ code-switching instances, i.e., where the bloggers have switched to English words/phrases while writing Hindi sentences. The key statistics of the created Hinglish code-switching text corpus are summarized in Table 5. The most frequent code-switching word pairs (Hindi-to-English and vice versa) that occur in HingCoS text corpus are given in Table 6. The distributions of code-switching instances with respect to the varying length of the sentences in the HingCoS text corpus are given in Figure 5. Further, the plot of average code-switching instances for varying length of the sentences in the HingCoS text corpus is also computed and is shown in Figure 6.

The existing approach of finding POS tags for the code-switching data employs separate POS taggers corresponding to the involved languages. But it is found to yield incorrect POS tags in particular to the non-native words due to limited context information. To address the same, we recently proposed a more efficient POS tagging scheme [41] for Hinglish text data. It consists of two steps: (i) the POS tags for the native (Hindi) words are derived conventionally, i.e., by passing the Hinglish text to a Hindi POS tagger, and (ii) the given Hinglish text is converted to pure English text through a machine translator and then the POS tags are derived using an English POS tagger. The final POS tags are derived through the distillation of the POS tags for the native and non-native words in the given Hinglish sentence derived in the above two steps. Following the above mentioned scheme, the POS tags for the HingCoS text corpus are derived. The distributions of the English and Hindi words present in the HingCoS text corpus with respect to their POS labels are shown in Figure 7. For validation purpose, a small set consisting of 100 sentences from the HingCoS text corpus are randomly selected and manually labeled for POS tags. On evaluating, the validation accuracy of the proposed POS scheme turns out to be 89.1%.

Table 6: Top 10 most frequent code-switching word pairs (Hindi-to-English and vice versa) that occur in HingCoS text corpus

Hindi to English		English to Hindi	
Word pair	**Log-likelihood**	**Word pair**	**Log-likelihood**
helpful रहा	-0.026	पाचों website	-0.080
proceed पर	-0.045	आप debit	-0.131
bitcoins के	-0.080	मैरा blog	-0.131
complexity को	-0.080	आपको ios	-0.156
infected है	-0.080	की audio	-0.156
modification के	-0.080	लिए unity	-0.156
depend करता	-0.101	निर्माता company	-0.228
cases में	-0.109	आपका memory	-0.249
click करें	-0.131	आपके account	-0.249
sms नही	-0.131	आपभी social	-0.249

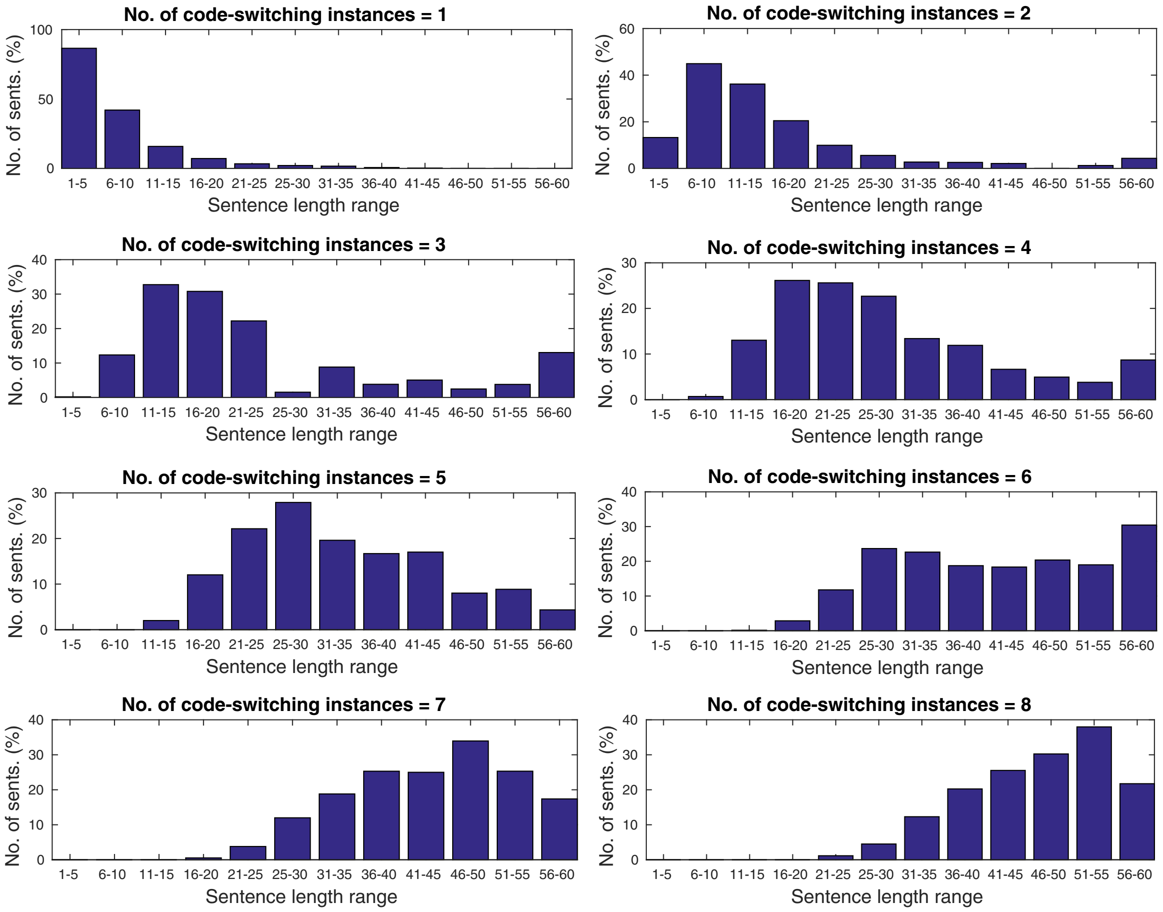


Figure 5: Distribution of code-switching instances for varying length of the sentences in the HingCoS text corpus. Note that, for the ease in display, the counts in each sentence length group are normalized separately, i.e., the sentence counts in a length group across 8 code-switching instances considered sum up to 100%.

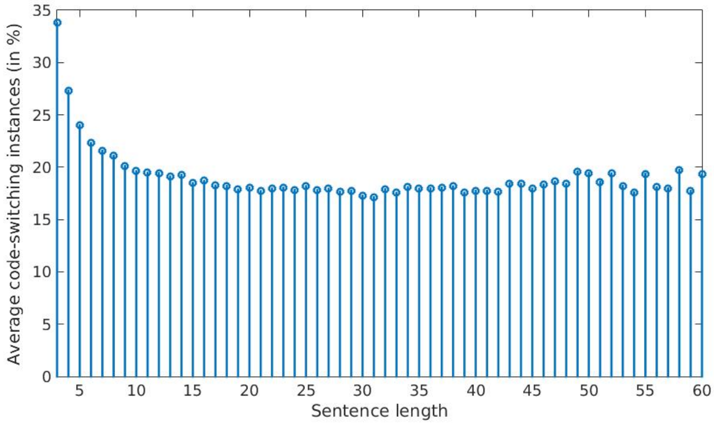


Figure 6: Plot of the average code-switching instances with respect to the length of sentences in HingCoS text corpus.

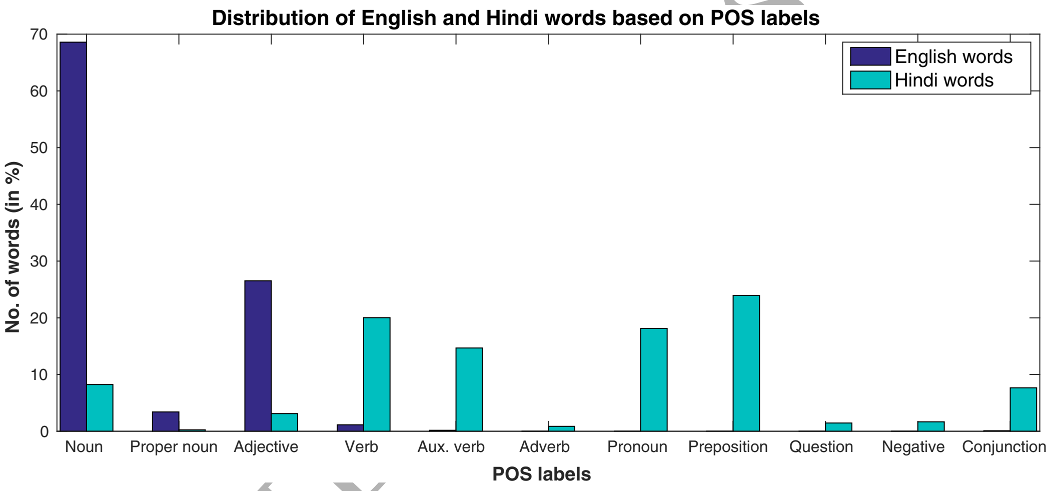


Figure 7: Distributions of the English and Hindi Words in the HingCoS text corpus based on their parts of speech (POS) labels. Note that, for the ease in display, the POS tags for Hindi and English words have been normalized separately, i.e., all POS labels corresponding to English/Hindi words sum up to 100%.

Table 7: The details of the HingCoS speech corpus developed in this study.

# utterances	**# words**		**# unique words**		**# code-switching instances**
	Hindi	**English**	**Hindi**	**English**	
9,251	125,653	50,719	2,644	3,901	30,035

2.2. Analysis of the HingCoS Speech Corpus

The HingCoS speech corpus contains about 25 hours of Hinglish speech data contributed by 101 speakers (64 male and 37 female). Table 7 summarizes the details about the total number of sentences, the total and the unique number of words spoken in Hindi and English portion of the data and the total number of code-switching instances in the HingCoS speech corpus. As per the meta-data collected, the speakers belong to 13 native language backgrounds in India. The distribution of the speakers' age along with the gender, and the distribution of the speakers based on their mother tongue (native language) are shown as a bar-plot

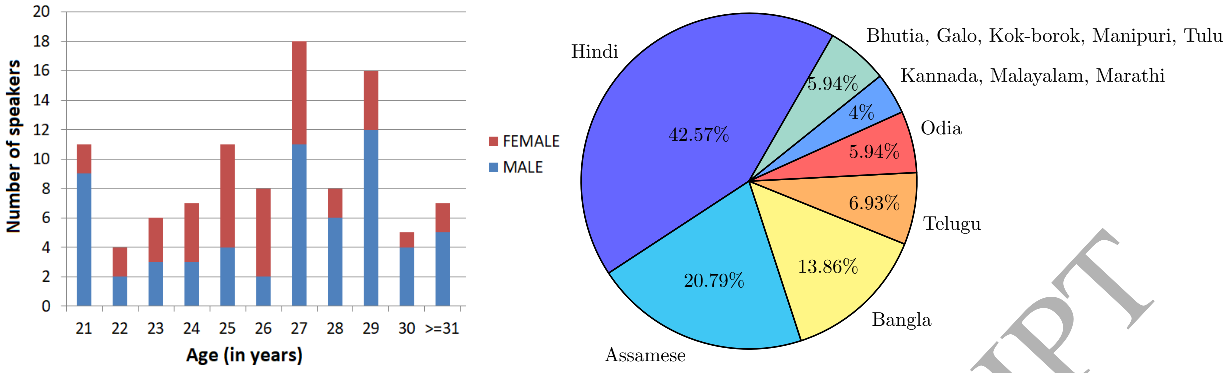
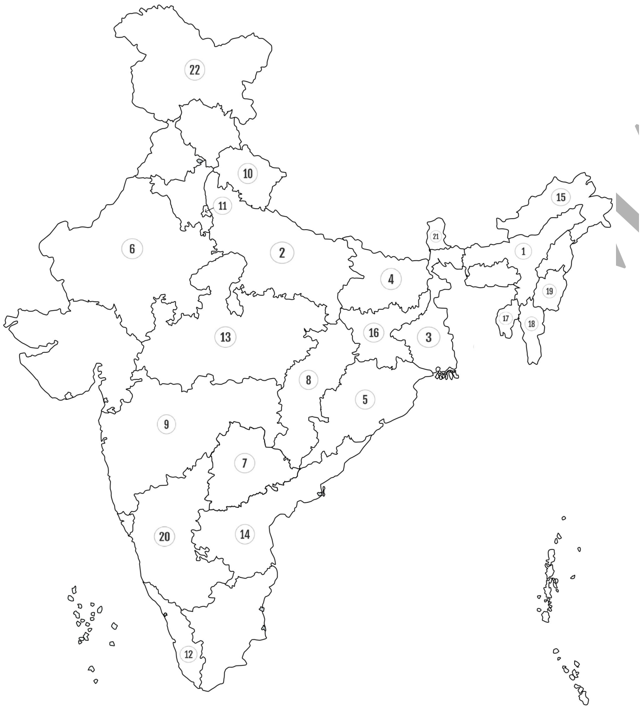


Figure 8: Distribution of the speakers' age along with the gender is shown as bar-plot. Whereas, distribution of mother tongues (native language) of the speakers in the collected Hinglish speech data is shown as pie-chart. The Hindi forms the mother tongue of the majority of speakers and is followed by Assamese and Bangla native speakers.

Figure 9: Distribution of the native states of the speakers in the collected speech data. The majority of speakers are from Assam and is followed by Uttar Pradesh. A total of 22 states are covered in the HingCoS speech corpus. The states of India represented in the HingCoS speech corpus are marked in the associated map.



State ID	State Name	# Speakers (M/F)
1	Assam	27 (M:10/F:17)
2	Uttar Pradesh	15 (M:13/F:2)
3	West Bengal	8 (M:6/F:2)
4	Bihar	8 (M:7/F:1)
5	Odisha	7 (M:3/F:4)
6	Rajasthan	6 (M:3/F:3)
7	Telangana	4 (M:3/F:1)
8	Chhattisgarh	3 (M:3/F:0)
9	Maharashtra	3 (M:3/F:0)
10	Uttrakhand	3 (M:2/F:1)
11	Delhi	2 (M:2/F:0)
12	Kerala	2 (M:1/F:1)
13	Madhya Pradesh	2 (M:1/F:1)
14	Andhra Pradesh	2 (M:2/F:0)
15	Arunachal Pradesh	2 (M:1/F:1)
16	Jharkhand	1 (M:1/F:0)
17	Tripura	1 (M:1/F:0)
18	Mizoram	1 (M:0/F:1)
19	Manipur	1 (M:1/F:0)
20	Karnataka	1 (M:1/F:0)
21	Sikkim	1 (M:0/F:1)
22	Jammu & Kashmir	1 (M:0/F:1)
	Total	**101 (M:64/F:37)**

and pie-chart, respectively in Figure 8. These speakers are native to 22 different states of India. Most of the states of India are associated with at least one distinct native language or regional dialect. Thus, the collected Hinglish speech data happen to carry wide variations in the accent of the speakers. The native state-wise distribution of speakers is shown on the political map of India in Figure 9.

3. Review of Existing Code-switching Corpora

In literature, a few code-switching speech and text corpora are already reported and they happen to cover different native and non-native language combinations. In this section, we briefly review those code-switching corpora while summarizing their salient attributes in Table 8 and Table 9.

The CUMIX [12], a Cantonese-English code-switching speech corpus was developed at the Chinese University of Hong Kong. This database contains 17 hours of speech data read by 40 speakers. The purpose of this corpus is to develop Cantonese-English code-switching ASR system. Lyu et al. [42] created a training dataset consisting of monolingual Taiwanese and Mandarin speech data from 100 speakers, with each speaker uttering 700 utterances in both the languages. For evaluation purpose, a small Mandarin-Taiwanese code-switching test set containing 4000 utterances recorded from 16 speakers was also developed. The English-Spanish code-switching speech corpus was compiled by Franco and Solorio at the University of Texas [3]. This corpus contains 40 minutes of transcribed spontaneous conversations of 3 speakers. The SEAME corpus [11, 43], a Mandarin-English code-switching conversational speech corpus is developed at Nanyang Technological University, Singapore, and Universiti Sains Malaysia. This database contains 63 hours of spontaneous Mandarin-English code-switching interview and conversational speech uttered by 157 Singaporean and Malaysian speakers. The CECOS [44], a Chinese-English code-switching speech corpus containing 12.1 hours of speech data collected from 77 speakers uttering prompted code-switching sentences is developed at the National Cheng Kung University in Taiwan. A corpus of Sepedi-English code-switching speech corpus was created by the South African CSIR [45]. This database consists of 10 hours of prompted speech, sourced from radio broadcasts and read by 20 Sepedi speakers. FAME! [9], a Frisian-Dutch code-switching speech corpus of radio broadcast speech is developed at Radboud University, Nijmegen. The recordings are collected from the archives of Omrop Fryslan, the regional public broadcaster of the province Fryslan. The database covers almost a 50 years time span. The Malay-English corpus consists of 100 hours of Malaysian Malay-English code-switching speech data from 120 Chinese, 72 Malay and 16 Indian speakers. [10]. MediaParl is a Swiss accented bilingual database containing recordings in both French and German as they are spoken in Switzerland. The data was recorded at the Valais Parliament. Valais is a bi-lingual Swiss canton with many local accents and dialects [8]. The FACST, a French-Arabic speech corpus consists of records of code-switching read and conversational utterances by 20 bilingual speakers who tend to code-switch in their daily lives [6]. It contains about 7.30 hours of data. Westhuizen et al. created a South African speech corpus containing English-isiZulu, English-isiXhosa, English-Setswana, and English-Sesotho code-switching speech utterances from South African soap operas. The soap opera speech is typically fast, spontaneous and may express emotion, with a speech rate higher than prompted speech in the same languages [7]. Injy Hamed et al. recently developed an Arabic-English code-switching corpus by conducting the interviews with 12 participants [46]. A small Hindi-English code-switching speech corpus was collected at Hong Kong University of Science and Technology [19]. This corpus is primarily made up of student interview speech which is about 30 minutes of data collected from 9 speakers.

From the literature review, it can be noted that a very small sized code-switching acoustic and linguistic resources are available so far covering the Indian context. This motivated us to create moderate sized Hinglish resources so that current technological advances in acoustic and language modeling can be explored. Among the publicly accessible code-switching corpora created in different contexts, the SEAME corpus happens to be the largest one and is followed by the HingCoS corpus reported in this work. In all the code-switching speech corpora reported so far, the speech data is recorded in clean conditions using good quality microphones. Unlike those, HingCoS speech corpus contains speech recorded in a realistic environment using a landline and mobile phones. This choice facilitates the development of telephone-based speech applications. The speech data in HingCoS corpus is collected in read-style, while the transcripts correspond to web blogs are written in a conversational style with atleast one code-switching instance in each utterance. Note that, the data recording protocol allowed the volunteers to re-record in case of any hesitation or disfluency. Hence, the speaking style of the collected data has been referred to as the *read* speech instead of spontaneous (conversational) speech. Another unique feature of HingCoS corpus is the diversity in the linguistic background of the speakers who contributed the speech data. The native language of only 42.57% of the speakers is Hindi while the remaining ones come from other Indian language backgrounds.

Table 8: Contrastive comparison of existing code-switching speech databases reported in the literature including the HingCoS speech corpus described in this work.

Reference	Name	Language pair(s)	Speech style	Dur. (hrs)	# spkr. (M/F)	Age grp.	# uttr.	Vocab.	Data recording	Access
Cao, et al. [12]	CUMIX	Cantonese-English	read	17	40 (20/20)	19-26	8,000	–	microphone; 48 kHz	Public
Lyu, et al. [14, 42]	–	Mandarin-Taiwanese	read	4.8	24	–	4,600	–	–	–
Franco, et al. [3]	–	English-Spanish	conversational	0.7	3	–	–	1516	–	Public
Lyu, et al. [11]	SEAME	Mandarin-English	conversational	51.7	157	18-34	42,759	15,338	microphone; 16 kHz	Public
Shen, e al. [44]	CECOS	Chinese-English	read	12.1	77 (62/15)	20-35	6,700	–	microphone; 16 kHz	–
Modipa, et al. [45]	SPCS	Sepedi-English	read	10	20 (12/8)	17-27	450	–	–	–
Ylmaz, et al. [9]	FAME!	Frisian-Dutch	conversational	18.5	309	–	–	–	microphone; 48 kHz	Public
Ahmed, et al. [10]	–	Malay-English	read	100	208	–	–	–	microphone; 16 kHz	–
Imseng, et al. [8]	MediaParl	French-German	conversational (parliamentary debates)	∼ 6	7	–	2,617	–	microphone; 44.1 kHz	Public
Amazouz, et al. [6]	FACST	French-Arabic	read and stimulated spontaneous	7.3	20 (10/10)	23-39	–	–	–	–
Westhuizen, et al. [7]	–	English-isiZulu, English-isiXhosa, English-Setswana, and English-Sesotho	conversational (soap opera episodes)	14.3	–	–	10,343	–	microphone; 32 kHz	Public
Hamed, et al. [46]	–	Arabic-English	conversational	5.3	12 (6/6)	–	1,234	–	microphone; 16 kHz	Public
This work	HingCoS	Hindi-English	read	25	101 (61/40)	19-40	9,251	6,542	telephone; 8 kHz	Public

Table 9: Contrastive comparison of existing code-switching text databases reported in the literature including the HingCoS text corpus described in this work.

Reference	**Language pair**	**Corpus name**	**# utterances**	**# words**	**Vocubulary**	**# code-switching instances**
Hamed, et al. [5]	Arabic-English	–	240,874	2,590,954 81.3% Arabic and 16.5% English	–	–
Lyu, et al. [11]	Mandarin-English	SEAME	42,759	–	15,338	–
This work	Hindi-English	HingCoS	25,988	578,159 66% Hindi and 34% English	14,643	104,912

4. Experimental Setups and Evaluations

The main motivation behind the creation of HingCoS corpus is to facilitate more research in code-switching ASR task in the Indian context. In this section, the created corpus has been evaluated for Hinglish speech recognition task to benchmark its quality. For this purpose, both acoustic and language models are created using the appropriate data from the HingCoS corpus. The details of the acoustic and linguistic datasets, the front-end features, different acoustic and language modeling approaches employed and the tuning of model parameters are discussed in the following subsections.

4.1. Acoustic and Linguistic Datasets

For language modeling, there are 25,988 sentences available in the HingCoS text corpus. We have divided them into three non-overlapping groups containing 22737, 2136 and 1115 sentences for training, testing and development purposes, respectively. For acoustic modeling, a total of 9251 sentences out of the HingCoS text corpus spoken by native speakers are recorded. The created HingCoS speech corpus includes 2136 utterances corresponding to the above defined linguistic test set and about 26% of the remaining text corpus (totaling 7115 utterances) for acoustic modeling purpose.

4.2. Front-End Features

For a thorough experimental evaluation, the acoustic models created employing a number of front-end features are explored. The front-end signal processing has been primarily done using the standard Mel frequency cepstral coefficient (MFCC) features and the more contemporary i-vector [47] based features. The parametric details of all these features are described in the next paragraphs.

The computation of MFCC features has been done considering 25 ms of hamming windowed speech frames along with 10 ms frame shift. Feature vector corresponding to each frame consists of a log energy coefficient (C0) and 12-dimensional MFCC features (C1-C12). For incorporating the dynamic characteristics of the vocal tract system, the 13-dimensional features obtained above have been appended with corresponding velocity and acceleration components. Hence, we finally have a 39-dimensional feature vector which is used for acoustic modeling.

Motivated by a recent work [48], the i-vector based acoustic features are also employed for training the ASR systems. For deriving the i-vector representations of the speech data, the use of Gaussian mixture model based universal background model (GMM-UBM) has been done. First, the 13-dimensional static MFCC features are time-spliced to capture the dynamic information across the frames. In time-splicing, four frames on either sides of the central frame are concatenated to form a 117-dimensional (13 × 9) feature vector which is then projected to a 40-dimensional space using the linear discriminant analysis (LDA) [49]. On these low-dimensional features, a 1024 component gender-independent GMM-UBM is learned. The *total variability* matrix (T-matrix) for estimating the i-vectors is randomly initialized and trained using the expectation maximization (EM) algorithm [47]. The 150-dimensional i-vectors are used in acoustic modeling.

4.3. Acoustic Model Training

In this study, different kinds of hidden Markov model (HMM) based hybrid modeling paradigms have been employed to develop the ASR systems. We have explored the time delay deep neural network (TDNN), feed-forward neural network (FDNN) and subspace Gaussian mixture model (SGMM) based acoustic modeling approaches in addition to the traditional Gaussian mixture model (GMM)based approach. All the experimental evaluations are performed using the Kaldi speech recognition toolkit [50].

- **GMM-HMM system**:
  This system is initialized with a context-independent monophone model using 39-dimensional MFCC features. Each of the phonemes is modeled by a three state left-to-right HMM model. Later, the cross-word triphone models are trained with a decision tree-based state tying approach to capture the contextual information. The 40-dimensional LDA-based feature vectors derived earlier are further decorrelated by employing the maximum likelihood linear transform (MLLT) [51]. The resultant features are further normalized by using feature-space maximum likelihood linear regression (fMLLR) [52]

and the speaker adaptive training (SAT) [53] is performed. Later, the cross-word triphone models are trained on these normalized features.

- **SGMM-HMM system**:
  In the conventional GMM-HMM systems, a large number of model parameters are required to be estimated. The SGMM based acoustic modeling framework addresses this issue by representing the complex distribution of parameters in a compact way. Here, the HMM states share a common structure globally, and only the state-dependent model parameters are required to be estimated. Instead of estimating GMM parameters directly from the training data, the model parameters are derived from the low-dimensional model and speaker subspaces that can capture phonetic and speaker variations. As a result of that, the total number of parameter estimation is greatly reduced, which makes the learning of the model parameters possible on a limited amount of training data. In SGMM [54] based modeling techniques, the unit distributions are derived from a GMM-UBM learned on a part of training data.

- **FDNN-HMM system**:
  The FDNN-HMM [55] based acoustic modeling approach is also explored in this study. A multi-layered FDNN is trained using time-spliced features normalized with LDA+MLLT+fMLLR as the input and computes the posterior probabilities over HMM states as the output. The specifications of the parameters used in training of the DNN-HMM system are given in Subsection 4.5.

- **TDNN-HMM system**:
  In the typical DNN architectures, the initial layers try to learn an affine transformation for the entire temporal context while training the models. But, it requires the availability of a large amount of training data to learn good transformations. This issue can be addressed by using the TDNN [56] architecture where the initial transformations learn narrower context and deeper layers try to learn longer temporal relationships. The specifications of the parameters used in training the TDNN-HMM system are given in the Subsection 4.5.

4.4. Language Model Training

In ASR, the language model (LM) reduces the search space while decoding the sequence of words in a sentence. Also, it helps in computing the joint probability of the word sequence $P(W)$. In this section, we discuss different LM paradigms employed for evaluation. The n-gram LMs trained by employing the IRSTLM language modeling toolkit [57] and the recurrent neural network (RNN) based LM trained using the RNNLM language modeling toolkit [58] are explored in this study.

- **N-gram language model**:
  The n-gram language model predicts the next word in a sentence by using the previous $(n-1)$ words [59]. In this technique, the probability of observing the word sequence $w_1, \ldots, w_N$ is approximated as

$$\begin{aligned} P(W) &= \prod_{i=1}^{N} P(w_i|w_1, \ldots, w_{i-1}) \\ &\approx \prod_{i=1}^{N} P(w_i|w_{i-(n-1)}, \ldots, w_{i-1}) \end{aligned} \tag{1}$$

For this joint probability distribution, there is not enough data for modeling all the sequence lengths in a given language. So, the conditional probability is approximated to the history of previous L words, where the value of L is very much less than n. Hence, the joint probability $P(W)$ is approximated as

$$P(W) = \prod_{i=1}^{N} P(w_i|w_{i-(L-1)}, \ldots, w_{i-1}) \tag{2}$$

The frequency count in the n-gram LM is given by the following equation,

$$P(w_i|w_{i-(L-1)},\ldots,w_{i-1}) = \frac{\text{count}(w_{i-(L-1)},\ldots,w_{i-1},w_i)}{\text{count}(w_{i-(L-1)},\ldots,w_{i-1})} \tag{3}$$

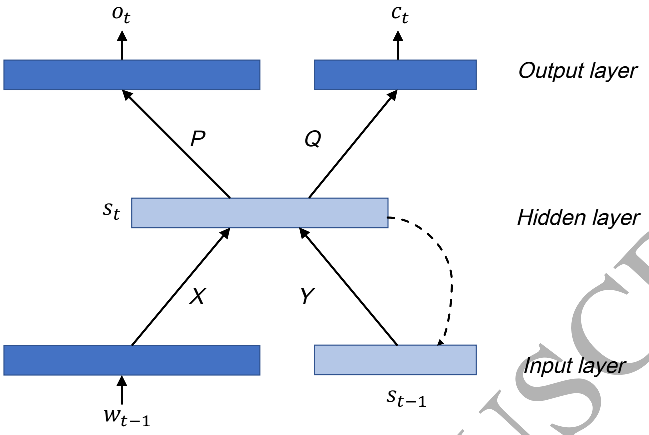


Figure 10: Network architecture for the class based RNN LM

- **Recurrent neural network (RNN) language model**:
The RNNs possesses the ability to model the long-term dependencies with the presence of the feed-back connections. This ability of RNN has been exploited for effective language modeling in contrast to the traditional n-gram LM. The architecture for the single layered RNN employed for language modeling task is shown in Figure 10. Theoretically, the RNN LM compute the probability of the next word w_t, by utilizing the full history of word sequence $(w_{t-1},\ldots,w_1)$ by the recurrent connections. From Figure 10, we note that, the RNN LM architecture has an input layer, a hidden layer, and an output layer. At time t, the input to the RNN is denoted by w_{t-1}, the state of the hidden layer is denoted by s_t. Whereas, the output layer is factorized to two parts: c_t for the word classes, and o_t for words conditioned on the classes [60]. The training of RNN LM having all the words in the vocabulary in the output layer is computationally complex. To overcome this issue, the words are clustered into classes s_t based on the word counts, and then the RNN LM is trained using this class information [61]. The previous context information s_{t-1} and the word w_{t-1} are fed as inputs for modeling the present context information s_t. The output layer uses this information s_t to compute the probability of the next word w_t in the sequence. Given the context information s_t, the probability of a word w_t is approximated as a product of the probability of the class to which w_t belongs and the class conditional probability of w_t. The computations of the RNN LM are defined by the following equations.

$$\begin{aligned} s_t &= f(w_{t-1}.X + s_{t-1}.Y) && (4)\\ c_t &= g(s_t.Q) && (5)\\ o_t &= g(s_t.P) && (6)\\ P(w_t|s_{t-1}) &= P(c_t|s_{t-1}) * P(w_t|s_{t-1},c_t) && (7) \end{aligned}$$

where, X, Y, P, Q are the weights computed for the corresponding layers and f, g are sigmoid and softmax functions[8], respectively. The RNN LM is trained by using back-propagation through time (BPTT) algorithm which helps the network to store the contextual information for several time steps in the hidden/context layer.

[8] Sigmoid function is defined as $f(x) = \frac{1}{1+e^{-x}}$ and Softmax function is defined as $g(x_i) = \frac{e^{x_i}}{\sum_i e^{x_i}}$

4.5. Parameter Tuning

This section describes the specifications of the parameters that are tuned to train the acoustic and language models. Unlike the LMs which are tuned on the development set, the tuning of the acoustic model parameters has been done on a very small subset extracted from the training set.

- **Language model**:
  The tuning of context length has been done on the traditional N-gram LM and the results are shown in Figure 11. It can be seen that the context length of 5 yields the best perplexity score on the development set.

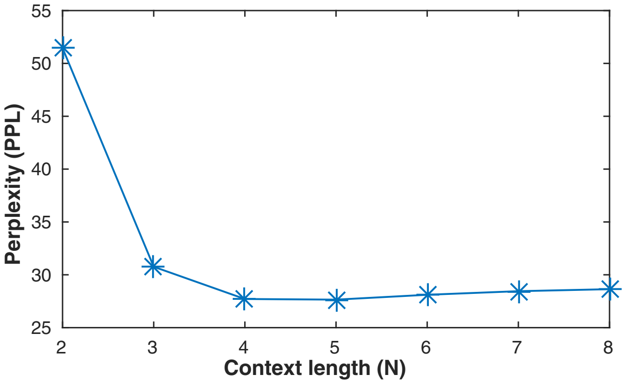


Figure 11: Tuning of the context length (N) on the development data. The optimal perplexity score is obtained for N = 5.

  The RNN LMs used in the experiments are trained with a single hidden layer having 150 nodes and *sigmoid* as the non-linearity function. By conducting tuning experiments on the Hinglish development data, the number of classes is set to be 100 and the variable corresponding to BPTT is set as 5.

Table 10: The parameters used for training the DNN and TDNN based hybrid models that are employed in this study.

Parameter	**Specification**	
	DNN	**TDNN**
No. of hidden layers	5	5
No. of hidden nodes	1,024	300
No. of epochs	20	5
Size of mini batch	128	512
Initial learning rate	0.015	0.015
Final learning rate	0.002	0.002

- **Acoustic model**:
  The context-dependent GMM acoustic models are trained by tuning the number of senones. After tuning, the number of senones is set to be 2500 and the number of Gaussian mixtures per senone is set to be 8 in all the cases. In this work, for training the SGMM, 400 Gaussians are selected for training the UBM. The parameters used for training the DNN and TDNN based AMs are given in Table 10.

Table 11: Recognition performances in terms of perplexity for N-gram and RNN-based language models (LMs) trained on the Hinglish training dataset and evaluated on Hinglish development and test sets.

Dataset	**Perplexity**		**%OOV**
	5-gram LM	**RNN LM**	
Development	27.65	18.77	0.04
Test	62.29	40.13	0.52

Table 12: Evaluation of Hinglish code-switching speech corpus in context od ASR task. The performance results in terms of percentage word error rate (%WER) are reported.

AM	**Phone model**	**Front-end features**	**LM**	**% WER**
GMM	Monophone	MFCC	5-gram	44.47
GMM	Triphone	MFCC	5-gram	26.44
GMM	Triphone	MFCC + LDA	5-gram	25.24
GMM	Triphone	MFCC + LDA + SAT	5-gram	21.89
SGMM	Triphone	MFCC + LDA + SAT	5-gram	19.14
DNN	Triphone	MFCC + LDA + SAT	5-gram	18.00
TDNN	Triphone	MFCC + i-vector	5-gram	17.67
TDNN	Triphone	MFCC + i-vector	RNN	17.09

4.6. Evaluation Results

The evaluation on HingCoS text and speech corpora has been done in the language modeling and speech recognition domains separately and the results are reported in the following.

- **Language modeling**:
  Both n-gram and RNN based LMs are developed using $22,737$ training sentences in the text data having a wordlist of 14k. Table 11 shows the LM performances in terms of the perplexity as well as the percentage out-of-vocabulary (OOV) words for both development and test datasets. In case of n-gram LM, the value of n is fixed to be 5 after tuning done on the development data as shown in Figure 11. From Table 11, we can note that the RNN-LM has resulted in consistently better recognition performances in contrast to the n-gram LM on both the development and test sets.

- **Speech recognition**:
  The experimental studies have been conducted for 4 different modeling paradigms: GMM-HMM, SGMM-HMM, DNN-HMM, and TDNN-HMM. In decoding, the 5-gram and the RNN LMs discussed in Table 11 are employed. The evaluation results of HingCoS speech corpus in terms of word error rate (WER) are given in Table 12. On considering the 5-gram LM, among all the systems developed, the TDNN-HMM based system trained using MFCC plus i-vector front-end features yields the best WER. On the use of RNN-LM for rescoring, further improvement in the performance of the TDNN-HMM system is noted. These trends are consistent with those reported in the literature.

5. Conclusion

In this work, we present the development of a moderate size Hinglish code-switching corpus and present the baseline evaluations of the same for both the language modeling and the speech recognition domains. This corpus is intended to be made public and the details of sharing the same would be available on the corpus webpage [9]. In the future, we plan to extend this corpus by adding more text and speech data to support more data-intensive modeling approaches.

[9] `https://www.iitg.ac.in/eee/emstlab/HingCoS_Database/HingCoS.html`

6. Acknowledgment

The authors wish to acknowledge with gratitude for the partial financial assistance received towards data collection from an ongoing project grant No. 11(18)/2012-HCC(TDIL) from the Ministry of Electronics and Information Technology, Govt. of India. The authors also sincerely thank the IITG fraternity members who gave their valuable time and support in the creation of the HingCoS corpus.

References

References

[1] J. J. Gumperz, Discourse Strategies, Cambridge University Press, 1982.

[2] C. Nilep, ‘Code switching’ in sociocultural linguistics, Colorado Research in Linguistics 19(1) (2006) 1–22.

[3] J. C. Franco, T. Solorio, Baby-steps towards building a Spanglish language model, in: Proc. of International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2007, pp. 75–84.

[4] T. Solorio, Y. Liu, Part-of-Speech tagging for English-Spanish code-switched text, in: Proc. of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 1051–1060.

[5] I. Hamed, M. Elmahdy, S. Abdennadher, Building a First Language Model for Code-switch Arabic-English, Procedia Computer Science 117 (2017) 208–216.

[6] D. Amazouz, M. Adda-Decker, L. Lamel, The French-Algerian Code-Switching Triggered audio corpus (FACST)., in: Proc. of Language Resources and Evaluation Conference (LREC), 2018.

[7] E. van der Westhuizen, T. Niesler, A first South African corpus of multilingual code-switched Soap Opera speech., in: Proc. of Language Resources and Evaluation Conference (LREC), 2018.

[8] D. Imseng, H. Bourlard, H. Caesar, P. N. Garner, G. Lecorvé, A. Nanchen, MediaParl: Bilingual mixed language accented speech database, in: Proc. of Spoken Language Technology Workshop (SLT), IEEE, 2012, pp. 263–268.

[9] E. Yilmaz, M. Andringa, S. Kingma, J. Dijkstra, F. Van der Kuip, H. Van de Velde, F. Kampstra, J. Algra, H. Heuvel, D. Van Leeuwen, A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC), 2016.

[10] B. H. Ahmed, T.-P. Tan, Automatic speech recognition of code switching speech using 1-best rescoring, in: Proc. of the International Conference on Asian Language Processing (IALP), IEEE, 2012, pp. 137–140.

[11] D.-C. Lyu, T.-P. Tan, E. S. Chng, H. Li, SEAME: a Mandarin-English code-switching speech corpus in south-east asia, in: Proc. of Interspeech, an Annual Conference of International Speech Communication Association, 2010.

[12] H. Cao, P. Ching, T. Lee, Y. T. Yeung, Semantics-based language modeling for Cantonese-English code-mixing speech recognition, in: Proc. of the 7th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2010, pp. 246–250.

[13] D. C. Lyu, R. Y. Lyu, Y. C. Chiang, C. N. Hsu, Speech recognition on code-switching among the Chinese dialects, in: Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, IEEE, 2006.

[14] D. C. Lyu, R. Y. Lyu, Language identification on code-switching utterances using multiple cues, in: Proc. of the Interspeech, an Annual Conference of International Speech Communication Association, 2008.

[15] A. C. Chandola, Some linguistic influences of English on Hindi, Anthropological Linguistics (1963) 9–13.

[16] S. Malhotra, Hindi-English, Code Switching and Language Choice in Urban, Uppermiddle-class Indian Families, Kansas Working Papers in Linguistics 5 (2) (1980) 39–46.

[17] A. Kumar, Certain aspects of the form and functions of Hindi-English code-switching, Anthropological Linguistics (1986) 195–205.

[18] S. Sinha, Code Switching and Code Mixing Among Oriya Trilingual Children - A Study, Academic Journal on Language in India 9(4) (2009) 274.

[19] A. Dey, P. Fung, A Hindi-English Code-Switching Corpus., in: Proc. of the Language Resources and Evaluation Conference (LREC), 2014, pp. 2410–2413.

[20] F. Grosjean, Life with Two Languages: An Introduction to Bilingualism, Harvard University Press, 1982.

[21] L. Malik, Socio-linguistics: A study of code-switching, Anmol Publications PVT. LTD., 1994.

[22] L. Milroy, P. Muysken, One speaker, two languages: Cross-disciplinary perspectives on code-switching, Cambridge University Press, 1995.

[23] H.-Y. Su, Code-switching between Mandarin and Taiwanese in three telephone conversation: The negotiation of interpersonal relationships among bilingual speakers in Taiwan, in: Proc. of the Symposium about Language and Society, 2001.

[24] W. Craig, Y. Harel-Fisch, H. Fogel-Grinvald, S. Dostaler, J. Hetland, B. Simons-Morton, M. Molcho, M. G. de Mato, M. Overpeck, P. Due, et al., A cross-national profile of bullying and victimization among adolescents in 40 countries, International journal of public health 54 (2) (2009) 216–224.

[25] C. Myers-Scotton, Social motivations for code-switching: evidence from Africa. Clarendon (1993).

[26] K. Bali, J. Sharma, M. Choudhury, Y. Vyas, I am borrowing ya mixing? An Analysis of English-Hindi Code Mixing in Facebook, in: Proc. of the First Workshop on Computational Approaches to Code Switching, 2014, pp. 116–126.

[27] A. Das, B. Gambäck, Code-mixing in social media text: the last language identification frontier?, in Proc. of the Traitement Automatique des Langues (TAL), Special Issue on Social Networks and NLP 54(3).

[28] J. MacSwan, Code-Switching and Grammatical Theory, The handbook of bilingualism and multilingualism (2012) 321–350.

[29] C. Myers-Scotton, Codeswitching with English: types of switching, types of communities, World Englishes 8 (3) (1989) 333–346.

[30] K. A. H. Zirker, Intrasentential vs. intersentential code switching in early and late bilinguals.

[31] C. F. Yeh, C. Y. Huang, L. C. Sun, C. Liang, L. S. Lee, An integrated framework for transcribing Mandarin-English code-mixed lectures with improved acoustic and language modeling, in: Proc. of the 7th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2010, pp. 214–219.

[32] M. Dhar, V. Kumar, M. Shrivastava, Enabling Code-Mixed Translation: Parallel Corpus Creation and MT Augmentation Approach, in: Proc. of the First Workshop on Linguistic Resources for Natural Language Processing, Association for Computational Linguistics, 2018, pp. 131–140.

[33] K. Bhuvanagirir, S. K. Kopparapu, Mixed language speech recognition without explicit identification of language, American Journal of Signal Processing 2 (5) (2012) 92–97.

[34] D.-C. Lyu, E.-S. Chng, H. Li, Language diarization for code-switch conversational speech, in: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 7314–7318.

[35] S. Sitaram, A. W. Black, Speech Synthesis of Code-Mixed Text., in: Proc. of Language Resources and Evaluation Conference LREC, 2016.

[36] G. I. Winata, A. Madotto, C.-S. Wu, P. Fung, Towards end-to-end automatic code-switching speech recognition, arXiv preprint arXiv:1810.12620.

[37] Ö. Çetinoğlu, S. Schulz, N. T. Vu, Challenges of computational processing of code-switching, Proc. of the Second Workshop on Computational Approaches to Code Switching.

[38] J. E. Flege, Second-language speech learning: Theory, findings, and problems, Speech perception and linguistic experience (1995) 233–272.

[39] S. Shahnawazuddin, D. Thotappa, A. Dey, S. Imani, S. Prasanna, R. Sinha, Improvements in iitg assamese spoken query system: Background noise suppression and alternate acoustic modeling, Journal of Signal Processing Systems 88 (1) (2017) 91–102.

[40] B. Ramani, S. L. Christina, G. A. Rachel, V. S. Solomi, M. K. Nandwana, A. Prakash, S. A. Shanmugam, R. Krishnan, S. K. Prahalad, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, H. A. Murthy, A common attribute based unified HTS framework for speech synthesis in Indian languages, in: Eighth ISCA Workshop on Speech Synthesis, 2013.

[41] G. Sreeram, K. Dhawan, R. Sinha, Novel textual features for language modelling of intra-sentential code-switching data, Under review for Journal of Computer Speech and Language, since March 2019.

[42] D.-C. Lyu, R.-Y. Lyu, Y.-c. Chiang, C.-N. Hsu, Speech recognition on code-switching among the Chinese dialects, in: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, IEEE, 2006.

[43] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, H. Li, A first speech recognition system for Mandarin-English code-switch conversational speech, in: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 4889–4892.

[44] H.-P. Shen, C.-H. Wu, Y.-T. Yang, C.-S. Hsu, Cecos: A Chinese-English code-switching speech database, in: Proc. of International Conference on Speech Database and Assessments (Oriental COCOSDA), IEEE, 2011, pp. 120–123.

[45] T. I. Modipa, M. H. Davel, F. De Wet, Implications of Sepedi/English code switching for ASR systems, 2013.

[46] I. Hamed, M. Elmahdy, S. Abdennadher, Collection and Analysis of Code-switch Egyptian Arabic-English Speech Corpus., in: Proc. of Language Resources and Evaluation Conference (LREC), 2018.

[47] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, Transactions on Audio, Speech, and Language Processing 19 (4) (2011) 788–798.

[48] S. Garimella, A. Mandal, N. Strom, B. Hoffmeister, S. Matsoukas, S. H. K. Parthasarathi, Robust i-vector based adaptation of dnn acoustic model for speech recognition, in: Proc of Interspeech, an Annual Conference of International Speech Communication Association, 2015.

[49] R. Haeb-Umbach, H. Ney, Linear discriminant analysis for improved large vocabulary continuous speech recognition, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, IEEE, 1992, pp. 13–16.

[50] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The Kaldi speech recognition toolkit, in: Workshop on automatic speech recognition and understanding, no. EPFL-CONF-192584, IEEE Signal Processing Society, 2011.

[51] M. J. Gales, et al., Maximum likelihood linear transformations for HMM-based speech recognition, Computer speech & language 12 (2) (1998) 75–98.

[52] C. J. Leggetter, P. C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models, Computer speech & language 9 (2) (1995) 171–185.

[53] T. Anastasakos, J. McDonough, J. Makhoul, Speaker adaptive training: A maximum likelihood approach to speaker normalization, in: International Conference Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, IEEE, 1997, pp. 1043–1046.

[54] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, et al., Subspace Gaussian mixture models for speech recognition, in: International Conference on Acoustics Speech and Signal Processing (ICASSP), IEEE, 2010, pp. 4330–4333.

[55] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Signal processing magazine 29 (6) (2012) 82–97.

[56] V. Peddinti, D. Povey, S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts, in: Proc of Interspeech, an Annual Conference of International Speech Communication Association, 2015.

[57] M. Federico, N. Bertoldi, M. Cettolo, IRSTLM: an open source toolkit for handling large scale language models, in: Proc. of Interspeech, an Annual Conference of International Speech Communication Association, 2008.

[58] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, J. Cernocky, RNNLM: Recurrent neural network language modeling toolkit, in: Proc. of the ASRU Workshop, 2011, pp. 196–201.

[59] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, J. C. Lai, Class-based n-gram models of natural language, Proc. of the Computational Linguistics 18 (4) (1992) 467–479.

[60] M. Song, Y. Zhao, S. Wang, Exploiting different word clusterings for class-based RNN language modeling in speech recognition, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 5735–5739.

[61] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, S. Khudanpur, Extensions of recurrent neural network language model, in: Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 5528–5531.