Interspeech 2018
2-6 September 2018, Hyderabad

A Study of Lexical and Prosodic Cues to Segmentation in a Hindi-English Code-switched Discourse

Preeti Rao[1], *Mugdha Pandya*[2], *Kamini Sabu*[1], *Kanhaiya Kumar*[1], *Nandini Bondale*[3]

[1]Department of Electrical Engineering, Indian Institute of Technology Bombay, India
[2]Vellore Institute of Technology, India
[3]School of Technology and Comp Science, Tata Institute of Fundamental Research, India

prao@ee.iitb.ac.in, pandya.mugdha4@gmail.com, kaminisabu@ee.iitb.ac.in, kanhaiya00210100@gmail.com, drnandini.bondale@gmail.com

Abstract

Bilingualism, almost universal in India, routinely appears in communication in many forms. Code-switching with English is common among city dwellers with the matrix language typically being the speaker's native tongue. While a number of English words have made their way into the lexicon of Indian languages, also prevalent is insertional code-switching, i.e. switching at sentence or clause level. We consider an interesting and widely encountered variety of code-switched speech in the form of public discourses by a popular motivational speaker who uses English, probably for effect, in her Hindi language speeches. We effectively observe three categories of segments in the discourse: Hindi, Hindi with embedded English words and English. In this work, we present the characteristics of our data, and investigate the discrimination potential of lexical and prosodic cues on manually segmented fragments. Lexical cues are obtained via Google Speech API for Indian English recognition. Prosodic cues computed from pitch, intensity and syllable duration estimates are found to demonstrate significant differences between Hindi and English segments, indicating more careful articulation of the embedded language.

Index Terms: code-switching, Hindi-English, prosody

1. Introduction

To make communication more effective, bilingual communities use a phenomenon known as 'code-switching'. It is a natural process that often occurs between multilingual speakers who share two or more languages. Code-switching can be defined as the use of more than one language, variety, or style by a speaker within an utterance or discourse, or between different interlocutors [1]. In India, bilingualism is commonplace and code-switching between native language and English is often encountered.

Code-switching, sometimes referred to as code mixing, can be classified as Inter-Sentential and Intra-Sentential. In inter-sentential, code-switching occurs at sentence boundaries and is seen most often between fluent bilingual speakers. In intra-sentential code-switching, the shift can occur in mid-sentence, at clause or at word level. In this study, we consider code-switching between Hindi and English, both official languages in India. Hindi is the matrix language or dominant language and English is the embedded language [2]. Apart from the scientific merit of such a study, the observed acoustic cues can fruitfully impact speech technologies such as text-to-speech synthesis (TTS) in terms of introducing naturalness, and automatic speech recognition (ASR) in terms of more accurate recognition with language-specific models.

In the Indian context, Kachru [3] has studied the observed varieties of code mixing, such as grammatical unit (noun or verb phrase) insertion or hybridization, sentence insertion, idiom or collocation insertion and word level inflectional attachment. Ahire [4] provides a summary with examples of the types of code mixing in Marathi and English. Lyu and Lyu [5] considered Mandarin-Taiwanese intra-sentential code-switching utterances in the context of ASR. Mandarin is the matrix language and Taiwanese is the embedded one. They integrated phonetic (syllable identity) and prosodic cues (duration and fundamental frequency of tonal syllables) to develop a language identification (LID) system for code-switched speech. With acoustic, duration and language model trained on these cues, they achieved reduced error rates for LID systems [5].

For LID, Vu et al. [6] classified transcribed words into four categories in Mandarin-English code-switched speech; Mandarin and English words, silence, and others (discourse particle, other languages, and hesitations) to obtain the percentages as 44%, 26%, 21%, and 7% respectively. They found that the duration of monolingual segments was very short; more than 82% English and 73% Mandarin segments were less than 1 second long.

For intra-sentence Chinese-English code-switching, Zhang and Tao [7] found that the context of matrix language influences the embedded target language words in terms of altering the prosody. They observed that word duration is higher for English compared to Chinese words whereas mean F0 mean was similar. This matched observations on Hindi-English code-switched speech, further exploited in code-switched speech synthesis [8]. It was found that word durations increase during code-switching from Hindi to English. Spanish-Basque code-switched segments too indicate difference in maximum pitch value and also in pitch accent peak positions [9]. For Spanish-English, code-switched segments are observed to be produced with increased F0 of tonic syllables and stressed vowel duration indicating hyper-articulation of embedded language [10]. In another study of Spanish-English code-switched speech compared with the monolingual speech in each of the two languages, stressed syllable F0 in English-only segments was observed to be higher compared to that of code-switched segments which further is

10.21437/Interspeech.2018-1600

higher compared to that in Spanish segments [11].

In the present work we carry out lexical and prosodic analyses for the chosen style of Hindi-English code-switching speech, a dialect that has not been researched much. We specifically investigate acoustic-prosodic features and present our results in view of the previous work on prosody for other code-switched language pairs. We describe next the dataset and annotation methods. Section 3 presents the implementation of lexical and prosodic features. Finally, our results are discussed in Section 4.

2. Dataset and Annotation

Public discourses by spiritual and motivational speakers have a wide following across India. Some of the well known speakers feature in daily or weekly television series with the archives further available on dedicated YouTube channels. The speakers use an easy and familiar speaking style to connect with their audience who come from a wide cross-section of the society. In the present study, we use the lectures of BK Shivani with a large number of video discourses available on YouTube [12]. We randomly selected 5 Hindi discourse videos (a total duration of 1 hour) for our study from the 700 videos available on the website. The videos are tagged by the dominant language used by the speaker (i.e. Hindi or English). Hindi discourses contain inter-sentential code-switching to English in extents of 10-15% of the total speech duration. English videos are monolingual, probably intended for audiences outside the country.

The speaker is a fluent bilingual and switches naturally and effortlessly between Hindi and English. The dominant language is Hindi with occasional spurts of English in a case of inter-sentential code-switching. When she switches to English, it is a continuation of the thought process, or at least a variation of the previous phrase, rather than a mere translation. In her Hindi speech (i.e. syntax is that of Hindi), she uses a number of English words, some of which belong to the present day extended Hindi lexicon. The recording of the particular speaker is selected because it is a good example of inter-sentential code-switched speech while also being fairly typical of a large class of speakers. Secondly, the quality of the recording is good with little background noise. Thirdly, a large amount of data is available in public domain.

The recordings were transcribed at word-level by manual correction of a transcript generated automatically using the Google Speech API recognizer [13] with Hindi setting (since this is the dominant language in the recording). In this setting, the output comprises Hindi words with common English words used by Hindi speakers, all in Devanagari script. The code-switched English sentences were manually transcribed. We next manually annotated code-switching boundaries, which separated the English spurts from Hindi spurts (the latter being defined as speech based on Hindi syntax even if English words are embedded within). In the inter-sentential code-switching cases encountered here, the code-switch boundaries coincide with phrase boundaries in the speech. Table 1 summarizes the duration statistics of code-switched segments. It can be seen that the average duration for Hindi (matrix language) segments is much longer than that for English (embedded language).

Table 1: *Dataset statistics for code-switched segments*

Segment Category	Hindi	English
Number of Segments	97	100
Total Duration (sec)	2854.64	744.82
Average Duration (sec)	29.43	7.45
Minimum Duration (sec)	1.73	0.91
Maximum Duration (sec)	137.31	35.2

In order to increase the number of speech segments for lexical and prosodic analyses, we further segmented the single-language spurts obtained above into 0.5–10 seconds chunks in a manner that the newly formed boundaries potentially qualify as candidate code-switch boundaries in the context of inter-sentential switching. The smaller segments (termed 'fragments') thus comprise one or a few phrases each, and these are each labeled with one of the following three tags: Hindi only (H), Hindi with embedded English words (HE) and English only (E). We eventually have 240 Hindi, 456 Hindi-English and 218 English fragments for the analyses. The preceding and succeeding silences are removed from the fragments.

3. Analysis methods

We describe the implementation of lexical and prosodic cues corresponding to each fragment.

3.1. Lexical Analysis

To determine lexical features that distinguish the two categories of fragments (namely H and E, where H includes HE), we first obtain an automatic transcription of each segment with the Google Speech API Indian English (en-IN) recognizer [13]. The recognizer provides an output in Roman script for the recognized English words as well as Hindi words (which appear to be part of the en-IN language model) together with a confidence score in 0-1 range. See Table 2 for sample pairs of manual and API transcripts. The Word Error Rate (WER), averaged across H and E segments, of the Speech API (en-IN) on our data was observed to be about 15 percent.

The lexical feature we compute is based on the counts of Hindi and English words in the speech decoder hypotheses of every input fragment. We use Pyenchant [14] (a Python based spelling-check library for English text) to check whether each of the hypothesized words is a valid English word. The fraction of words in the fragment that are valid English words forms the feature value for the given fragment. Potential inaccuracies arise from words that are common to both languages (e.g. *hum*).

Another lexical feature we compute is the silence-to-speech ratio over code-switched segments. For each Hindi and English segment, we calculate the total duration of silence and speech to obtain the respective silence to speech ratios. Further, the length of pauses may not be same for Hindi-to-English switching and for English-to-Hindi switching. To investigate this, we compute the pause duration of each code-switched boundary.

3.2. Prosody Analysis

As per [15], code-switching is mainly used by bilinguals to indicate narrow focus or prominence. We therefore

Table 2: *Google Speech API (en-IN) output for sample segments from our dataset*

Category	Manual Transcript	Google Speech API Transcript (Indian-English Setting)	Confidence
E	So relationship is not on the label, relationship is based on the quality of energy we exchange with each other.	relationship is not on the label relationship is based on the quality of energy we Exchange with each other	0.94
H	यह भी बचपन से सुना है, जो करेगा सो पाएगा. (yeh bhi bachpan se suna hai, jo Karega so payega.)	EB Bachpan se suna hai jo Karega so payega	0.68
HE	कुछ लोग है जिनकी लास्ट डे पे टिकेट कॅन्सल हुई है और कुछ लोग है जिनकी लास्ट डे पे टिकेट बुक हुई है. (kuch log hai jinki last day pe ticket cancel hui hai aur kuch log hai jinki last day pe ticket book hui hai.)	kuch log Hain Jinke last Deputy cat cancel hui hai aur kuch log Hain Jinke last day patikot book	0.87

test prosodic features normally used for prominence detection to see whether embedded English phrases can be discriminated from the Hindi. We prefer syllable duration features over word duration since word duration can be affected by number of syllables rather than speaking style. Pitch and intensity measures were considered across fragments instead of syllables since we are interested in overall behavior across phrases.

To obtain syllable boundaries, we need the time alignment of the transcript with the recording. This is not readily available from the Google Speech API output. Therefore, we carry out instead forced alignment of the recording with the manual transcription using an available state-of-the-art Automatic Speech Recognition (ASR) system. The ASR system uses an Indian English lexicon of Hindi and English words from the manual transcript. The acoustic models were trained for a different purpose on bilingual children reading English and Hindi texts, amounting to 5 hours of speech, to obtain 47 acoustic models of the phones and silence fillers expected in the English speech of native Hindi or Marathi speakers. The system is developed in Kaldi framework and uses hybrid DNN-HMM acoustic models configured as described in [16]. We have made a word-phone-syllable mapping dictionary manually and alternate pronunciations, if any, are included.

We perform prosodic analysis on the fragments annotated as H, HE, and E. The silence regions/pauses between two fragments are discarded during fragment level prosodic analysis, thus making each fragment span the region from the onset of the first word to the end of the last word in the fragment. For every fragment, syllable duration based features viz., mean, maximum and minimum syllable duration, and the ratio of minimum-to-maximum duration are computed. Next, for each recording, we extract the following prosodic contours at 10 ms intervals:

1. Pitch: F0 contour is extracted with the Praat autocorrelation function (Settings: 30 ms Hanning window, silence threshold=0.03, voicing threshold=0.45, octave cost=0.01, octave-jump cost=0.35, voice/unvoiced cost=0.14 with pitch range restricted to 100Hz-500Hz). The resulting pitch contour is linearly interpolated in the unvoiced regions and converted to semitone scale.
2. Intensity: The signal is first scaled to get unit maximum amplitude across each discourse to remove the recording gain effect. Then we take the squared sum of samples across the window to get energy. Intensity is calculated as the energy in dB.
3. Spectral Tilt: It is calculated as the first MFCC coefficient. This gives us the measure of proportion of energy in the lower frequency range compared to higher frequencies.
4. Spectral Balance: It is the relative energy in specific frequency bands. We considered four frequency bands in this work - 0-0.5 kHz, 0.5-1 kHz, 1-2 kHz and 2-4 kHz, spectral bands commonly used in prosodic analyses [17] .

For each of these contours, six statistical features are computed across each utterance fragment - mean, median, standard deviation, span or min-max range, minimum, maximum. We compare the feature values for H, HE and E fragments through box plots. Welch's t-test [18] is used to quantify the distinction between H (including HE) and E fragments.

4. Results and Discussion

4.1. Lexical

We successfully decoded, with the Google Speech API, 910 fragments out of a total of 914. An API error or null transcript occurred for the remaining cases where typically the speaker spoke too fast or too softly. Figure 1 shows the box-plot for fraction of words of English for each category of fragments using the obtained automatic transcription. Clearly, for English fragments, the fraction is one for the most of the time and it is near zero for pure Hindi segments. We observe a good separation indicating the potential value of this feature in discriminating the two types of segments even in the presence of recognition errors. An alternate to word-based cues could be morpheme based; we have not explored this here.

It can be seen from the boxplot of silence-to-speech ratio in Figure 1 that there is a distinct difference in the silence-to-speech ratio between Hindi and English segments. The median value for E segments is 0.16 while it is 0.32 for H segments. This makes it a potential feature in discriminating the two types of segments. The length of the pauses between code-switched segments were found to be similar for both directions of change.

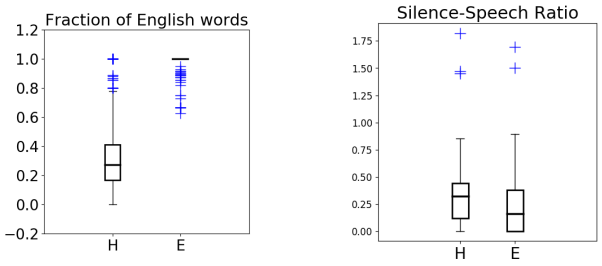

Figure 1: *Boxplots for (a) fraction of detected words in fragment that are English and (b) silence-speech ratio for code-switched segments.*

4.2. Prosody

As can be seen from Figure 2, H segments tend to be more intense compared to E segments. Although no significant differences were found between H and E segments for any of the pitch features, we observe in Figure 2 (middle), that maximum pitch value is higher for HE segments, i.e. the mixed language segments, compared to either the single-language fragments. Further, the lower spectral tilt in E segments indicates higher vocal effort indicative of some hyper-articulation.

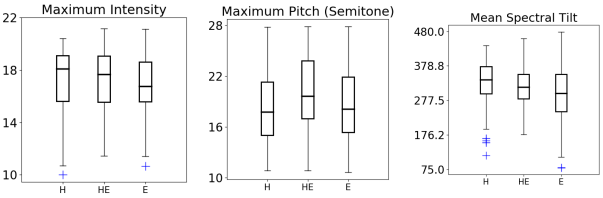

Figure 2: *Boxplots for (a) maximum intensity, (b) maximum pitch (semitone) and (c) mean spectral tilt.*

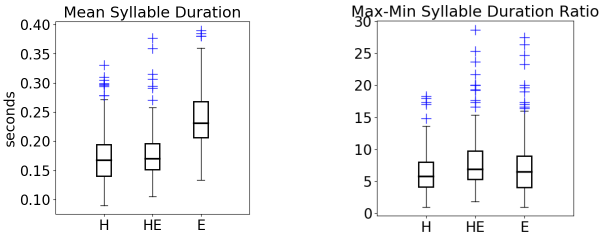

Figure 3: *Boxplots for (a) average syllable duration and (b) max-min syllable duration ratio.*

Figure 3 shows that average syllable duration is higher for E compared to H. The max-min syllable duration ratio is, however, almost equal for E and H cases and higher for HE. This suggests that the speaker speaks Hindi words fast and English words slowly, maintaining this trend throughout the speech. Table 3 shows the features analyzed using Welch's t-test between H (including HE) and E segments. Spectral balance was calculated for all the 4 bands, but other three bands did not show low p values and are not shown in the table.

Table 3: *List of prosodic features analyzed. Features in bold have $p < 0.01$ during Welch's t-test.*

Feature Class	Statistical Features
Syllable Duration	**mean**, **maximum**, **minimum**, max-min ratio
Pitch (Semitone)	mean, median, maximum, minimum, span, standard deviation
Intensity (dB)	mean, median, **maximum**, minimum, span, standard deviation
Spectral Tilt	**mean**, median, maximum, minimum, span, **standard deviation**
Spectral Balance Band2 (1-2kHz)	**mean**, median, **maximum**, minimum, **span**, standard deviation

4.3. Discussion

We find that the Google Speech API (en-IN) does a better job of decoding the variety of Hindi-English code-switched speech considered here compared with the Google Speech API (Hindi). This indicates the robustness of the language model trained on real-world speech corpora of Indian English. Our observations on lexical features of code-switched segments provide useful insights that can be exploited in dialog generation. Given that code-switched speech is expected to be a natural mode of communication between humans and automatic agents, we can see that the embedding language (syntax) in a phrase must typically be the dominant one in terms of the word count for natural sounding speech. The prosodic analysis has shown that the embedded language (English) segments tend to be spoken more slowly and with higher vocal effort indicating that it is more clearly articulated. This is consistent with the observations of Olson [10] who found that duration and pitch range are higher for code-switched segments compared to non code-switched segments. We also observed that pitch variation is more pronounced in the HE (i.e. bilingual) segments compared to either of the monolingual fragments (which are similar to each other in this aspect); this is unlike previous reports on Spanish-English code-switching [11]. The outcomes of the prosodic study can contribute to better prosody modeling in code-switched speech TTS on the lines of previous attempts [8].

An extension of this study to new public figures using Hindi-English code-switched speech is currently in progress. The preliminary findings for actor Alia Bhatt [19] are similar to those from BK Shivani in terms of the higher salience of code-switched (i.e. English) fragments. The phonetic correlates appear to be slightly different however, e.g. intensity increase in the code-switched speech accompanied by mean syllable duration increase, and silence-to-speech ratio decrease at the segment level.

5. Acknowledgement

Supported by National Programme on Perception Engineering, sponsored by the Department of Electronics and Information Technology, Government of India.

6. References

[1] S. Romaine, *Bilingualism*. Wiley-Blackwell, 1995.

[2] D. Winford, *An introduction to contact linguistics*. Wiley-Blackwell, 2003.

[3] B. Kachru, "Toward structuring code-mixing: An indian perspective," *International Journal of the Sociology of Language*, vol. 16, no. 34, pp. 27–46, 1978.

[4] M. Ahire, "Form and function of code mixing in marathi." *Language in India*, vol. 15, no. 7, 2015.

[5] G. Lyu and M. Lyu, "Language identification on code-switching utterances using multiple cues," in *Proceedings of INTERSPEECH*, Australia, 2008.

[6] N. Vu, D. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," in *Proceedings of ICASSP*, Japan, 2012.

[7] Y. Zhang and J. Tao, "Prosody modification on mixed-language speech synthesis," in *Proceedings of International Symposium on CSLP 2008*, Kunming, China.

[8] S. Rallabandi and A. Black, "On building mixed lingual speech synthesis systems," in *Proceedings of INTERSPEECH 2017*, Stockholm, Sweden.

[9] A. Aly, "Prosodic effects of code-switching in spanish-basque bilinguals," Ph.D. dissertation, University of California, Los Angeles, 2017.

[10] D. Olson, "The phonetics of insertional code-switching:suprasegmental analysis and a case of hyper-articulation," vol. 2, no. 4, pp. 439–457, 2012.

[11] P. Piccinini and M. Garellek, "Prosodic cues to monolingual versus code-switching sentences in english and spanish," in *Proceedings of 7th Speech Prosody Conference*, Dublin, Ireland, 2014.

[12] "BK Shivani, Awakening with Brahmakumaris," https://www.youtube.com/user/bkshivani/videos, Youtube, last accessed 20/3/2018.

[13] "Google Cloud Speech API, Language Support," https://cloud.google.com/speech/docs/languages, last accessed 16/3/2018.

[14] "pyenchant 2.0.0:Python Package Index," https://pypi.python.org/pypi/pyenchant/, last accessed 20/3/2018.

[15] D. Olson and M. Ortega-Llebaria, "The perceptual relevance of code switching and intonation in creating narrow focus," in *Proceedings of fourth Conference on Laboratory Approaches to Spanish Phonology*, 2010.

[16] K. Sabu, P. Swarup, H. Tulsiani, and P. Rao, "Automatic assessment of children's L2 reading for accuracy and fluency," in *Proceedings of SLaTE*, Stockholm, Sweden, 2017.

[17] G. Christodoulides and M. Avanzi, "An evaluation of machine learning methods for prominence detection in french," in *Proceedings of INTERSPEECH*, Singapore, 2014.

[18] B. Welch, "The generalization of "student's" problem when several different population variances are involved," *Biometrika*, vol. 1-2, no. 34, pp. 28–35, 1947.

[19] "Alia Bhatt Full Interview," https://www.youtube.com/watch?v=VtzV2Ad6n54&t=121s, Youtube, last accessed 14/6/2018.