Automatic Detection of Code-switching Style from Acoustics

SaiKrishna Rallabandi*, Sunayana Sitaram, Alan W Black*
*Language Technologies Institute, Carnegie Mellon University, USA
Microsoft Research India
`srallaba@cs.cmu.edu, susitara@microsoft.com, awb@cs.cmu.edu`

Abstract

Multilingual speakers switch between languages displaying inter sentential, intra sentential, and congruent lexicalization based transitions. While monolingual ASR systems may be capable of recognizing a few words from a foreign language, they are usually not robust enough to handle these varied styles of code-switching. There is also a lack of large code-switched speech corpora capturing all these styles making it difficult to build code-switched speech recognition systems. We hypothesize that it may be useful for an ASR system to be able to first detect the switching style of a particular utterance from acoustics, and then use specialized language models or other adaptation techniques for decoding the speech. In this paper, we look at the first problem of detecting code-switching style from acoustics. We classify code-switched Spanish-English and Hindi-English corpora using two metrics and show that features extracted from acoustics alone can distinguish between different kinds of code-switching in these language pairs.

Index Terms: speech recognition, code-switching, language identification

1 Introduction

Code-switching refers to the phenomenon where bilingual speakers alternate between the languages while speaking. It occurs in multilingual societies around the world. As Automatic Speech Recognition (ASR) systems are now recognizing conversational speech, it becomes important that they handle code-switching. Furthermore, code-switching affects co-articulation and context dependent acoustic modeling (Elias et al., 2017). Therefore, developing systems for such speech requires careful handling of unexpected language switches that may occur in a single utterance. We hypothesize that in such scenarios it would be desirable to condition the recognition systems on the type (Muysken, 2000) or style of language mixing that might be expected in the signal. In this paper, we present approaches to detecting code-switching 'style' from acoustics. We first define style of an utterance based on two metrics that indicate the level of mixing in the utterance: CodeMixing Index(CMI) and CodeMixing Span Index. Based on these, we classify each mixed utterance into 5 style classes. We also obtain an utterance level acoustic representation for each of the utterances using a variant of SoundNet. Using this acoustic representation as features, we try to predict the style of utterance.

2 Related Work

Prior work on building Acoustic and Language Models for ASR systems for code-switched speech can be categorized into the following approaches: (1) Detecting code-switching points in an utterance, followed by the application of monolingual acoustic and language models to the individual segments (Chan et al., 2004; Lyu and Lyu, 2008; Shia et al., 2004). (2)Employing a shared phone set to build acoustic models for mixed speech with standard language models trained on code-switched text (Imseng et al., 2011; Li et al., 2011; Bhuvanagiri and Kopparapu, 2010; Yeh et al., 2010). (3) Training Acoustic or Language models on monolingual data in both languages with little or no code-switched data (Lyu et al., 2006; Vu et al., 2012; Bhuvanagirir and Kopparapu, 2012; Yeh and Lee, 2015). We attempt to approach this

Class	CMI	Hi-En Utts	En-Es Utts
C1	0	6771	41624
C2	0-0.15	13986	2284
C3	0.15-0.30	492	2453
C4	0.30-0.45	8865	1025
C5	0.45-1	2496	1562

Table 1: Distribution of CMI classes for Hinglish and Spanglish

problem by first identifying the style of code mixing from acoustics. This is similar to the problem of language identification from acoustics, which is typically done over the span of an entire utterance.

Deep Learning based methods have recently proven very effective in speaker and language recognition tasks. Prior work in Deep Neural Networks (DNN) based language recognition can be grouped into two categories: (1) Approaches that use DNNs as feature extractors followed by separate classifiers to predict the identity of the language (Jiang et al., 2014; Matejka et al., 2014; Song et al., 2013) and (2) Approaches that employ DNNs to directly predict the language ID (Richardson et al., 2015b,a; Lopez-Moreno et al., 2014). Although DNN based systems outperform the iVector based approaches, the output decision is dependent on the outcome from every frame. This limits the real time deployment capabilities for such systems. Moreover, such systems typically use a fixed contextual window which spans hundreds of milliseconds of speech while the language effects in a code-switched scenario are suprasegmental and typically span a longer range. In addition, the accuracies of such systems, especially ones that employ some variant of iVectors drop as the duration of the utterance is reduced. We follow the approach of using DNNs as utterance level feature extractors. Our interest is in adding long term information to influence the recognition model, particularly at the level of the complete utterance, representing stylistic aspects of the degree and style of code-switching throughout the utterances.

3 Style of Mixing and Motivation

Multiple metrics have been proposed to quantify codemixing (Guzmán et al., 2017; Gambäck and Das, 2014) such as span of the participating languages, burstiness and complexity. For our current study, we categorize the utterances into different styles based on two metrics: (1) Code Mixing index (Gamback and Das, 2014) which attempts to quantify the codemixing based on the word counts and (2) CodeMixed Span information which attempts to quantify codemixing of an utterance based on the span of participating languages.

Class	Description	Hi-En Utts	En-Es Utts
S1	Mono En	5413	27960
S2	Mono Hi/Es	0	12749
S3	En Matrix	626	2883
S4	Hi/Es Matrix	36454	1986
S5	Others	8307	3345

Table 2: Distribution of span based classes for Hinglish and Spanglish. Note that the term 'Matrix' is used just here notionally to indicate larger word span of the language.

3.1 Categorization based on Code Mixing Index

Code Mixing Index (Gamback and Das, 2014) was introduced to quantify the level of mixing between the participating languages in a codemixed utterance. CMI can be calculated at the corpus and utterance level. We use utterance CMI, which is defined as:

$$C_u(x) = 100 \frac{w_m(N(x) - \max_{L_i \in L}\{t_{L_i}\}(x)) + w_p P(x)}{N(x)} \tag{1}$$

where N is the number of languages, t_{L_i} are the tokens in language L_i, P is the number of code alternation points in utterance x and w_m and w_p are weights. In our current study, we quantize the range of codemixed index (0 to 1) into 5 styles and categorize each utterance as shown in Table 1. A CMI of 0 indicates that the utterance is monolingual. We experimented with various CMI ranges and found that the chosen ranges led to a reasonable distribution within the corpus. For example, the C2 CMI class in Hindi-English code switched data has utterances such as "पंधरा पे start किये थे ग्यारा पंधरा पे यार अभी तो कुछ नही हुआ" ('started at fifteen, eleven or fifteen but buddy nothing has happened so far'). The C4 class on the other hand, has utterances such as ”actual में आज यह rainy season का मौसम था ना" ('actually the weather today was like rainy season, right?'). An example of a C5 utterance is ”ohh English अच्छा English कौनसा favourite singer मतलब English में?" ('Ohh English, ok who is your favorite English singer?')

3.2 Categorization based on Span of codemixing

While CMI captures the level of mixing, it does not take to account the span information (regularity) of mixing. Therefore, we use language span information (Guzmán et al., 2017) to categorize the utterances into 5 different styles as shown in Table 2. We divide each utterance based on the span of the participating languages into five classes - monolingual English, monolingual Hindi or Spanish, classes where the two languages are dominant (70% or more) and all other utterances. The classes S3 and S4 indicate that the primary language in the utterance has a span of at least 70% with respect to the length of utterance. This criterion makes these classes notionally similar to the construct of ‘matrix’ language. However, we do not consider any information related to the word identity in this approach. As we can see from both the CMI and span-based classes, the distributions of the two language pairs are very different. The Spanglish data contains much more monolingual data, while the Hinglish data is predominantly Hindi matrix with English embeddings. The Hinglish data set does not have monolingual Hindi utterances which is due to the way the data was selected, as explained in Section 4.1.

3.3 Style Modeling using Modified SoundNet

SoundNet (Aytar et al., 2016) is a deep convolutional network that takes raw waveforms as input and is trained to predict objects and scenes in video streams. Once the network is trained, the activations of intermediate layers can be considered as a high level representation which can be used for other tasks. However, SoundNet is a fully convolutional network, in which the frame rate decreases with each layer. Each convolutional layer doubles the number of feature maps and halves the frame rate. The network is trained to minimize the KL divergence from the ground truth distributions to the predicted distributions. The higher layers in SoundNet are subsampled too much to be used directly for feature extraction. To alleviate this, we train a fully connected variant of Soundnet (Wang and Metze, 2017): Instead of using convolutional layers all the way up, we switch to fully connected layers after the 5th layer. We have also change the input sampling rate to 16 KHz to match the rate of provided data.

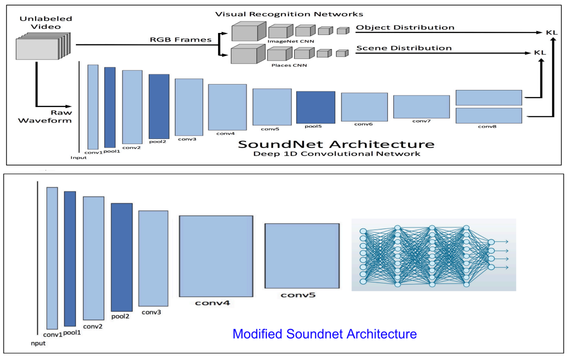

Figure 1: Architecture for style modeling using modified Soundnet

4 Experimental Setup

4.1 Data

We use code-switched Spanish English (referred to as Spanglish hereafter) released as a part of Miami Corpus (Deuchar et al., 2014) for training and testing. The corpus consists of 56 audio recordings and their corresponding transcripts of informal conversation between two or more speakers, involving a total of 84 speakers. We segment the files based on the transcriptions provided and obtain a total of 51993 utterances. For Hinglish, we use an in-house speech corpus of conversational speech. Participants were given a topic and asked to have a conversation in Hindi with another speaker. 40% of the data had at least one English word in it, which was transcribed in English, while the Hindi portion of the data was transcribed in Devanagari script. We split the data into Hindi and Hinglish by filtering for English words, hence the Hinglish data does not contain monolingual Hindi utterances. Note that this data did contain a few monolingual English sentences, but they were typically single word sentences. Such English utterances were considered to be part of the Hinglish class. The number of Hinglish utterances is 54279.

4.2 Style Identification

For style identification we perform the following procedure: We first categorize the utterances into 5 styles based on the criteria described in section 3. We pass each utterance through pretrained modified SoundNet and obtain the representations at all the layers. We use the representation from 7th (penultimate) layer as embedding for the utterance. We experimented with combining the repre-

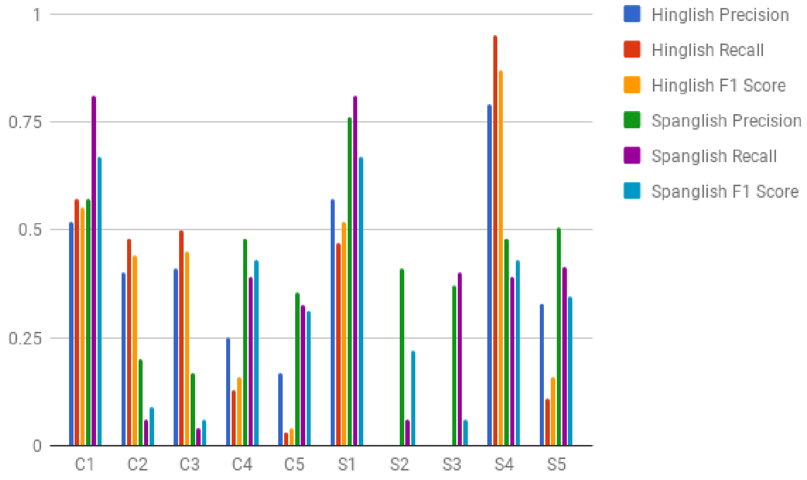

Figure 2: Precision, Recall and F1 scores for 5 way style classification of Hinglish and Spanglish

sentations at multiple layers but found that they do not outperform the representation at the 7th layer alone. Therefore for the purposes of this paper, we restrict ourselves to the representation at the penultimate layer. The embedding is obtained by performing mean pooling on the representation. Finally, we train a Random Forest classifier using the obtained embedding to predict the style of mixing.

4.3 Results and Discussion

Figure 2 shows the results for 5 class classification for Hinglish and Spanglish based on CMI (classes C1-C5) and span (classes S1-S5). Some classes (C1, C2, C3, S1, S4 for Hi-En and C1, C4, C5, S1, S4, S5 for En-Es) are easier to predict and are not always the majority classes. In our current implementation, we use a two stage approach for feature extraction and classification. We hypothesize that there might be better approaches to perform each of the components independently. It might also be possible to incorporate a style discovery module in an end to end fashion (Wang et al., 2018). As we plan to include the predicted style information in our recognition system, we also evaluate our approach using language models. For this, we build style specific language models tested on style specific test sets and include the average perplexity values for all of them in table 3. Ground Truth indicates that the model was built on the classes segregated based on approaches described in section 3. Predicted indicates that the language model was built based on the classes predicted by the model described in section 4.2. We also build a language model on utterances from the majority class for CMI and Span, as well as all the Spanglish data with no style information. As can be observed, the perplexity has a considerable reduction when using style specific information, while the majority style does not lead to the same reduction over the model with no style information. This further validates our hypothesis that style specific models may help decrease LM perplexities and ASR error rates.

Table 3: Language Model Experiments

Language			Avg Ppl
Spanglish	CMI	GroundTruth	54.8
		Predicted	56.2
		Majority Class	81.2
	Span	GroundTruth	59.1
		Predicted	62.8
		Majority Class	80.2
No Style Info			82.1

5 Conclusion

In this paper, we present a preliminary attempt at categorizing code-switching style from acoustics, that can be used as a first pass by a speech recognition system. Language Model experiments indicate promising results with considerable reduction in perplexity for style-specific models. In future work, we plan to improve our feature extraction and classification models and test our language models on code-switched speech recognition.

References

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*. pages 892–900.

K Bhuvanagiri and Sunil Kopparapu. 2010. An approach to mixed language automatic speech recognition. *Oriental COCOSDA, Kathmandu, Nepal* .

Kiran Bhuvanagirir and Sunil Kumar Kopparapu. 2012. Mixed language speech recognition without explicit identification of language. *American Journal of Signal Processing* 2(5):92–97.

Joyce YC Chan, PC Ching, Tan Lee, and Helen M Meng. 2004. Detection of language boundary in code-switching utterances by bi-phone probabilities. In *Chinese Spoken Language Processing, 2004 International Symposium on*. IEEE, pages 293–296.

Margaret Deuchar, Peredur Davies, Jon Herring, M Carmen Parafita Couto, and Diana Carter. 2014. Building bilingual corpora. *Advances in the Study of Bilingualism* pages 93–111.

Vanessa Elias, Sean McKinnon, and Ángel Milla-Muñoz. 2017. The effects of code-switching and lexical stress on vowel quality and duration of heritage speakers of spanish. *Languages* 2(4):29.

B. Gamback and A Das. 2014. On measuring the complexity of code-mixing. In *Proc. of the 1st Workshop on Language Technologies for Indian Social Media (Social-India)*.

Björn Gambäck and Amitava Das. 2014. On measuring the complexity of code-mixing. In *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*. pages 1–7.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. *Proc. Interspeech 2017* pages 67–71.

David Imseng, Hervé Bourlard, Mathew Magimai Doss, and John Dines. 2011. Language dependent universal phoneme posterior estimation for mixed language speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, pages 5012–5015.

Bing Jiang, Yan Song, Si Wei, Jun-Hua Liu, Ian Vince McLoughlin, and Li-Rong Dai. 2014. Deep bottleneck features for spoken language identification. *PloS one* 9(7):e100795.

Ying Li, Pascale Fung, Ping Xu, and Yi Liu. 2011. Asymmetric acoustic modeling of mixed language speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, pages 5004–5007.

Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno. 2014. Automatic language identification using deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pages 5337–5341.

Dau-Cheng Lyu and Ren-Yuan Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *Ninth Annual Conference of the International Speech Communication Association*.

Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang, and Chun-Nan Hsu. 2006. Speech recognition on code-switching among the chinese dialects. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, volume 1, pages I–I.

Pavel Matejka, Le Zhang, Tim Ng, Harish Sri Mallidi, Ondrej Glembek, Jeff Ma, and Bing Zhang. 2014. Neural network bottleneck features for language identification. *Proceedings of IEEE Odyssey* pages 299–304.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Fred Richardson, Douglas Reynolds, and Najim Dehak. 2015a. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters* 22(10):1671–1675.

Fred Richardson, Douglas Reynolds, and Najim Dehak. 2015b. A unified deep neural network for speaker and language recognition. *arXiv preprint arXiv:1504.00923* .

Chi-Jiun Shia, Yu-Hsien Chiu, Jia-Hsin Hsieh, and Chung-Hsien Wu. 2004. Language boundary detection and identification of mixed-language speech based on map estimation. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, volume 1, pages I–381.

Yan Song, Bing Jiang, YeBo Bao, Si Wei, and Li-Rong Dai. 2013. I-vector representation based on bottleneck features for language identification. *Electronics Letters* 49(24):1569–1570.

Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li. 2012. A first speech recognition system for mandarin-english code-switch conversational speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 4889–4892.

Yun Wang and Florian Metze. 2017. A transfer learning based feature extractor for polyphonic sound event detection using connectionist temporal classification. *Proceedings of Interspeech, ISCA* pages 3097–3101.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017* .

Ching Feng Yeh, Chao Yu Huang, Liang Che Sun, and Lin Shan Lee. 2010. An integrated framework for transcribing mandarin-english code-mixed lectures with improved acoustic and language modeling. In *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. IEEE, pages 214–219.

Ching-Feng Yeh and Lin-Shan Lee. 2015. An improved framework for recognizing highly imbalanced bilingual code-switched lectures with cross-language acoustic modeling and frame-level language identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(7):1144–1159.