Joint Language Identification of Code-Switching Speech using Attention-based E2E Network

Ganji Sreeram, Kunal Dhawan, Kumar Priyadarshi, and Rohit Sinha

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati - 781039, India
Email: {s.ganji, k.dhawan, k.priyadarshi, rsinha}@iitg.ac.in

***Abstract*—Language identification (LID) has relevance in many speech processing applications. For the automatic recognition of code-switching speech, the conventional approaches often employ an LID system for detecting the languages present within an utterance. In the existing works, the LID on code-switching speech involves modeling of the underlying languages separately. In this work, we propose a joint modeling based LID system for code-switching speech. To achieve the same, an attention-based end-to-end (E2E) network has been explored. For the development and evaluation of the proposed approach, a recently created Hindi-English code-switching corpus has been used. For the contrast purpose, an LID system employing the connectionist temporal classification-based E2E network is also developed. On comparing both the systems, the attention based approach is noted to result in better LID accuracy. The effective location of code-switching boundaries within the utterance by the proposed approach has been demonstrated by plotting the attention weights of E2E network.**

Index Terms: language identification, code-switching, end-to-end models, attention mechanism

I. Introduction

The phenomenon of switching between two or more languages while speaking in multilingual communities is referred to as the code-switching [1], [2], [3]. It occurs not only in verbal discourse but also in textual chats on social media platforms [4], [5]. In a typical code-switching sentence, the syntactical composition belongs to one language while the words from the other language are used for emphasis or ease of delivery [6], [7]. The syntax providing language is referred to as the *native* language while the other language is called as the *foreign* language. Code-switching has become a common practice in several multilingual communities across the world. The salient examples of those include: Arabic-English [8], French-Arabic [9], French-German [10], Frisian-Dutch [11], Hindi-English [12], [13], Malay-English [14], Mandarin-English [15], Mandarin-Taiwanese [16], and Spanish-English [17]. Code-switching poses some interesting research challenges to speech recognition [16], [18], language identification [19], language modeling [20], [21] and speech synthesis [22]. However, researchers working in code-switching domain are often constrained by the lack of domain specific resources . Towards addressing the constraint of resources in Indian context, a Hindi-English code-switching text and speech corpora referred to as the *HingCoS Corpus*[1] has been recently developed [23].

The task of detecting the languages present in spoken or written data using machines is referred to as language identification (LID). It finds applications in many areas including automatic recognition of code-switching speech. In [19], the authors developed an LID system for code-switching speech by employing separate large vocabulary continuous speech recognizers (LVCSRs). In this work, we aim to develop an LID system that can directly identify the code-switching instances instead of separately modeling the underlying languages. Recently, researchers have explored end-to-end (E2E) networks in many speech/text processing applications. The E2E networks can be trained by employing two techniques: (i) connectionist temporal classification (CTC) [24], and (ii) sequence to sequence modeling with attention mechanism [25]. Current literature amply demonstrates that the attention-based E2E systems outperform the CTC-based E2E systems. Recently, an utterance-level LID system employing attention-mechanism is explored [26]. In that, for producing attentional vectors, a set of pretrained language category embeddings are used as a look-up table. Motivated by those works, we develop a joint LID system for code-switching speech using an attention-based E2E network. Unlike [26], the attention provided for the LID system is intra-sentential [27] and is dynamic. The salient contributions of this work include: (i) a novel application of E2E networks in developing a joint LID system for code-switching speech, and (ii) demonstration of the effectiveness of attention mechanism in locating the code-switching instances.

The remainder of this paper is organized as follows: In Section II, the proposed joint LID system trained by employing E2E networks has been discussed in detail. The detailed description of the HingCoS corpus, the system tuning parameters and the evaluations metrics involved in this study are described in Section III. The evaluation results along with a brief discussion, followed by the demonstration of attention mechanism for the LID task has been presented in Section IV. Finally, the paper is concluded in Section V.

978-1-7281-8895-9/20/$31.00

[1] *www.iitg.ac.in/eee/emstlab/HingCoS_Database/HingCoS.html*

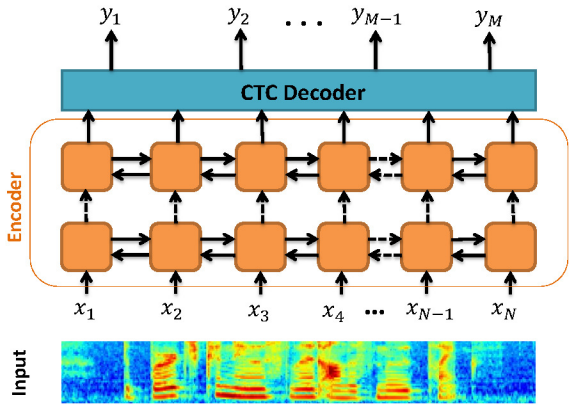

Fig. 1: Architecture of the CTC-based E2E network. The encoder is a deep network consisting of BiLSTMs.

II. E2E Network based Joint LID System

In this section, we describe the creation of E2E LID systems for identifying the languages in code-switching data. The developed systems jointly model the underlying languages and can handle intra-sentential code-switching types. So, in this work, we refer to them as *joint LID systems*. For developing those systems, we first explore the CTC-based E2E network and it is followed by experimentation on listen-attend-spell (LAS) [28] network which employs an attention mechanism. The details of those approaches are discussed below.

A. CTC-based E2E Network

The CTC-based E2E architecture consists of two modules: a deep bidirectional long-short-term-memory (BiLSTM) network as an encoder, and a CTC decoder. The deep BiLSTM network encodes input feature vector $\boldsymbol{x}$ into a higher level representation vector. CTC enables the training of E2E models without requiring a prior alignment between input and output sequences. It assumes the outputs at different time steps to be conditionally independent. The CTC decoder outputs a probability distribution over all possible output labels $\boldsymbol{y}$, conditioned on a given input sequence $\boldsymbol{x}$. A dynamic programming based forward-backward algorithm is employed to obtain the sum over all possible alignments and produces the probability of output sequence given a speech input. The typical architecture of the CTC-based E2E network is shown in Figure 1. Given a target transcription $\boldsymbol{y}$ and the input feature vector $\boldsymbol{x}$, the network is trained to minimize the CTC cost function as

$$\text{CTC}(\vec{x}) = -\log \text{P}(\boldsymbol{y}|\boldsymbol{x}) \tag{1}$$

where $\text{P}(\vec{y}|\vec{x}) = \sum_{\vec{a}\in\vec{\beta}(y,^{\rightarrow}x)} \text{P}(\vec{a}|\vec{x})$, a is an alignment, and $\beta(y,^{\rightarrow}x)$ is the set of all possible sequences between $\vec{y}$ and $\vec{x}$.

B. Attention-based E2E Network

The LAS architecture comprises 3 modules: listener, attender, and speller. The listener is a pyramidal architecture consisting of BiLSTM cells. It acts as an encoder and transforms an input feature vector $\boldsymbol{x}$ into a higher order vector representation $\boldsymbol{h}$. The encoded output vector $\boldsymbol{h}$ along with the decoder state s_i is passed to the attender. At every time instance, the attender takes $\boldsymbol{h}$ and decoder state s_i as the inputs and outputs the context c_i. It acts like an alignment generator determining which encoded features in $\boldsymbol{h}$ should be attended for accurate prediction of the current output symbol y_i. The output of this attention module c_i is then passed to the speller, which is an LSTM decoder. It takes the context information c_i as well as the previous prediction y_{i-1} in order to predict the current symbol y_i. The listener, attender and the speller are trained together to minimize the cross-entropy loss and thus making it a complete end-to-end system. The typical architecture of the LAS network is shown in Figure 2. The mathematical representations of each step in the LAS architecture are given as

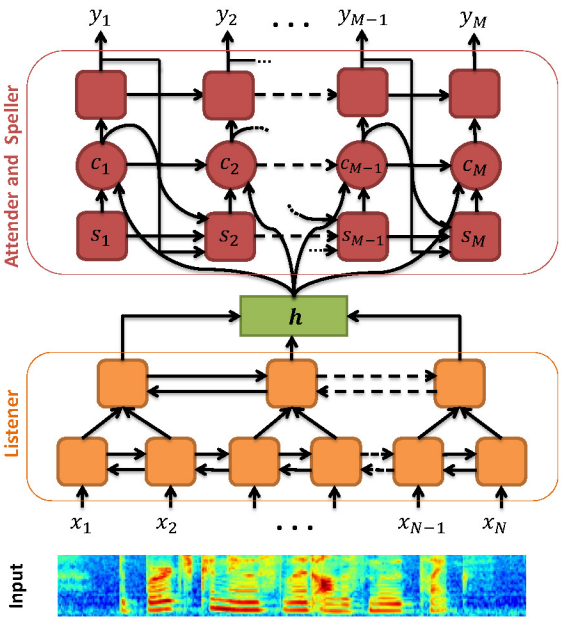

Fig. 2: Architecture of LAS network. It consists of three modules namely: listener (encoder), attender (alignment generator), and speller (decoder).

$$\begin{aligned} \boldsymbol{h} &= \text{Listener}(\boldsymbol{x}) && (2)\\ c_i &= \text{Attender}(\boldsymbol{h}, s_i) && (3)\\ s_i &= \text{LSTM}(y_{i-1}, s_{i-1}, c_{i-1}) && (4)\\ p(y_i|\boldsymbol{x}) &= \text{Speller}(c_i, y_{i-1}). && (5) \end{aligned}$$

C. Creation of Target Labels for LID

In intra-sentential code-switching utterances, the duration of the embedded foreign words/phrases could be very short. Thus, LID is required to be performed at the word level rather than the utterance level. Further, the explored E2E networks are required to be conditioned to perform the LID task on code-switching speech data. For achieving that, for each of the training utterance, first the given orthographic transcription is transformed into character level transcription. Later, each character in the transcription is mapped to the corresponding LID tags. This process is illustrated in Figure 3. This is to highlight that, in the orthographic transcription of the training data, the Hindi and English words are written

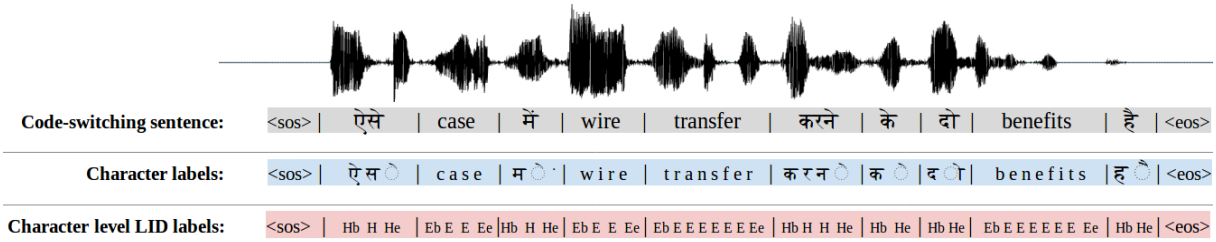

Fig. 3: Creation of character-level LID tags for the training data towards conditioning the E2E networks to perform LID task on code-switching speech. The 'H/E' denotes Hindi/English LID tag. The 'b/e' label is appended to the 'H/E' LID tag to mark the begin/end characters.

in their respective scripts. So, the character-level LID tags as 'H/E' are produced in a straight forward manner, except that additional labels 'b' and 'e' are appended to the tags of *begin* and *end* characters of each word, respectively. Also, a blank symbol '|' has been inserted between words to ease the marking of the word boundary. For training the E2E models, a total of 8 labels which include 6 LID tags (Hb, H, He, Eb, E, Ee), one blank label (|), and a silence label (sil) are given as targets to generate the output posterior probabilities. With the proposed target labeling scheme, the attention-based E2E system is hypothesized to predict the language boundaries more accurately. The experimental results discussed later in Section IV support the same.

III. Experimental Setup

This section describes the code-switching corpus used for experimentation purposes. The description of the tuning of parameters of different LID systems developed and the metric employed for evaluating their performances are also presented.

A. Database and Front-end Features

The HingCoS speech corpus, used for the experimentation, is collected over the telephone by speakers uttering predefined Hindi-English code-switching sentences in varying acoustic environments. These sentences were collected by crawling a few web-blogs having different contexts [23]. In this corpus, the native language of the sentences is Hindi and the foreign language is English. The speech data is contributed by 101 speakers (64 male and 37 female) and recorded at a sampling frequency of 8 kHz and a resolution of 8 bits/sample. This database contains 9251 Hindi-English code-switching utterances (about 25 hours) with utterance duration varying between 2 to 30 seconds. For experimental purpose, the corpus is partitioned into train, development, and test sets containing 7015, 100 and 2136 sentences, respectively.

The acoustic features comprise 26-dimensional log filter bank energies computed using the Hamming window having the length as 25 ms, window shift of 10 ms, and pre-emphasis factor of 0.97. These features are then used to develop both CTC- and attention-based E2E LID systems. The development of these systems has been done on the Nabu toolkit [29].

B. Parameter Tuning

In this section, we describe the tuning of parameters for both the developed LID systems done on the development set defined earlier. In this work, the attention-based E2E LID system is trained by employing the LAS network in which the encoder (listener) has 2 hidden layers, each with 128 BiLSTM nodes. The dropout rate of the encoder is set as 0.5. The number of hidden layers and nodes of the decoder (speller) are kept same as that of the encoder, except that the nodes are simple LSTMs. The LAS network is trained by setting the number of epochs as 300, batch size as 32, and the learning rate decay as 0.1. During decoding, the beam width is set as 8. For contrast purpose, a CTC-based E2E LID model is trained with 3 hidden layers, each having 128 BiLSTM nodes. The dropout rate of the encoder is set to be similar to that of the LAS model. The network is trained with number of epochs as 300. Also, the parameters corresponding to the batch size, and the learning rate decay are set as 8 and 0.1, respectively. During decoding, the CTC cost function is employed to produce 1-best output sequence.

C. Evaluation Measures

The developed E2E systems are evaluated in terms of the LID error rate computed as

$$\text{LID error rate} = \frac{N_S + N_I + N_D}{N} \times 100$$

where, the numerator terms N_S, N_I, and N_D refer to the number of substitutions, insertions, and deletions, respectively. The denominator N refers to the total number of labels in the reference. For this evaluation, the reference transcriptions for all test utterances labeled in terms of the proposed LID tags are aligned with the corresponding outputs produced by the E2E network. In addition to this character-level LID error rate, a corresponding word-level LID error rate is also computed in a similar fashion by applying majority voting scheme [30] on the character-level LID labels.

IV. Results and Discussion

In this work, two different kinds of E2E joint LID systems are developed and evaluated on the HingCoS corpus. The LID error rates computed both at character and word levels for these systems are reported in Table I. Note that, there are 6 target labels (Hb, He, H, Eb, Ee, E) for character-level

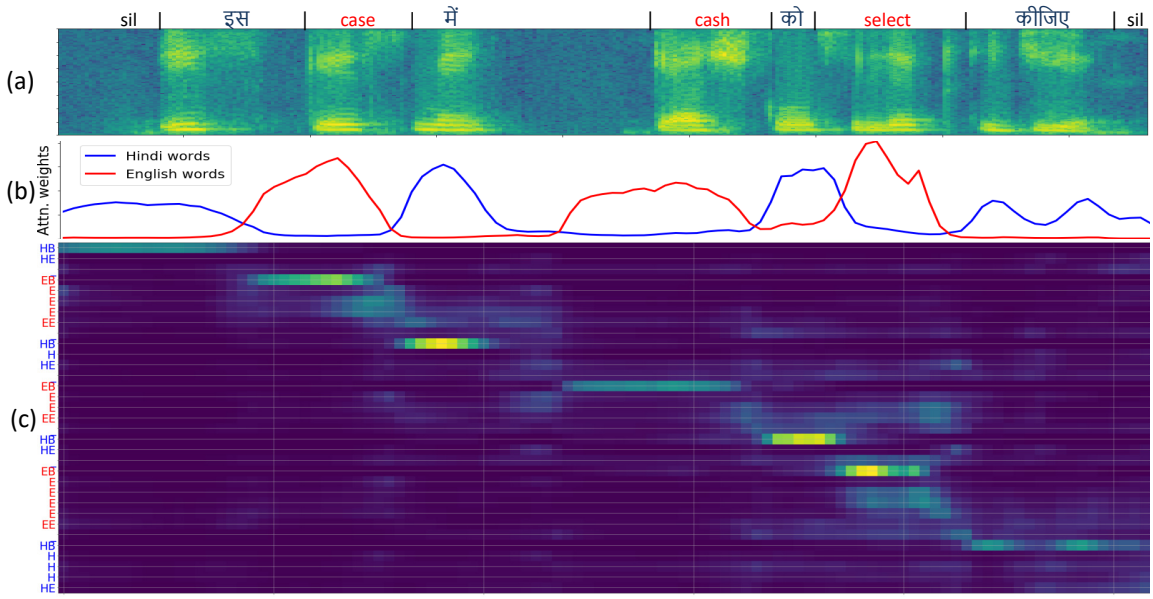

Fig. 4: Visualization of attention mechanism for LID task. For a given Hindi-English code-switching utterance: a) spectrogram labeled with Hindi and English word boundaries for reference purpose. (b) variation of attention weights with respect to time for Hindi and English languages, and (c) alignment produced by the attention network for the input speech and the decoded output LID labels.

TABLE I: Evaluation of the developed E2E LID systems in Hindi-English code-switching task. The LID error rates have been computed both for character and word levels. The total number of characters/words (N) in the reference transcription is $198,855/41,025$.

LID system	Target label	N_D	N_I	N_S	LID error rate (%)
CTC	Character	73,502	3,384	20,957	49.20
	Word	3,576	3,655	5,136	30.14
Attention	Character	20,299	13,587	12,789	**23.47**
	Word	2,713	1,616	2,484	**16.60**

LID, whereas the word-level LID has 2 target labels (H, E). Therefore, better word-level LID error rates are achieved when compared to character-level LID error rate as the former is computed by applying majority voting scheme [30] on the character-level LID labels. In contrast to CTC, the use of LAS architecture in E2E LID system is noted to yield substantial reduction in the error rates. This is attributed to the ability of attention mechanism in LAS network to accurately predict the languages switching in the data. To highlight that, we have computed the language-specific averaged attention weights with respect to the decoded LID label sequence and the plot for the same is shown in Figure 4. The description of each of the subplots in Figure 4 is presented next.

Figure 4(a) shows the spectrogram of a typical Hindi-English speech utterance in the test set. Note that, the spectrogram is manually labeled with spoken words and their boundaries for the reference purposes. The variations of the averaged attention weights for Hindi and English language targets present in the input speech data with respect to time, are shown in Figure 4(b). The sequence alignment produced by the attention network for the input speech data (on the x-axis) and the decoded output LID labels (on the y-axis) is plotted in Figure 4(c). From Figures 4(b) and 4(c), we observe that the attention weights for Hindi and English languages mostly peak around the corresponding word locations.

It is worth highlighting here that both CTC-based and attention-based E2E systems are provided with identical target-level supervision while training. Unlike the attention-based system, the CTC-based system could not exploit that supervision. This is attributed to the fact that CTC assumes the outputs at different time steps to be conditionally independent, hence making it less capable of learning the sequence. To support this argument, for the very utterance in Figure 4, the character level decoded outputs of the CTC- and attention-based E2E LID systems are listed Table II. The word-level LID labels for both the considered systems are also shown in that table. On comparing the hypothesized sequence of output labels, it can be noted that the inclusion of attention mechanism in E2E LID system leads to more effective language identification within code-switching speech data.

V. Conclusions

In this work, we propose joint E2E LID systems employing CTC and attention mechanism for identifying the languages present in code-switching speech. The development and evaluation of the proposed systems are done on a recently created Hindi-English code-switching speech corpus. Towards developing the LID systems, a novel target labeling scheme has been introduced which is found to be very effective for the attention-based system. On comparing the attention and CTC mechanisms, the former is noted to achieve a two-fold reduction in both character- and word-level LID error rates.

TABLE II: The character and word level decoded outputs for CTC- and attention-based E2E LID systems for the utterance considered in Figure 4. A majority voting scheme is employed for mapping the character-level LID label sequences to word-level LID label sequences. The attention-based system is able to decode the LID label sequences more accurately when compared to the CTC-based system.

Character level LID lables	Reference sequence	Hb He | Eb E Ee | Hb H He | Eb E Ee | Hb He | Eb E E E E Ee | Hb H H H He
	CTC-based hypothesis	Hb E Ee | Eb E Ee | Eb He | Hb He | Eb Ee | Eb Ee | Hb He
	Attention-based hypothesis	Hb He | Eb E E E Ee | Hb H He | Eb E E Ee | Hb He | Eb E E E E Ee | Hb H H H He
Word level LID lables	Reference sequence	H | E | H | E | H | E | H
	CTC-based hypothesis	E | E | E | H | E | E | H
	Attention-based hypothesis	H | E | H | E | H | E | H

The work also demonstrates the ability of the attention mechanism in detecting the language boundaries in code-switching speech data. Despite the experiments being performed on Hindi-English code-switching data, the proposed approach can easily be extended to other code-switching contexts.

In a recent work [31], the authors reported improvement in Mandarin-English code-switching ASR by employing multi-task learning with the LID labels. Motivated by that work, in the future, we aim to explore the proposed LID labeling scheme as a supervision in the multi-task learning framework for code-switching ASR.

REFERENCES

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

[2] C. M. Eastman, "Codeswitching as an urban language-contact phenomenon," *Journal of Multilingual & Multicultural Development*, vol. 13, no. 1-2, pp. 1–17, 1992.

[3] C. M. Scotton, "Comparing codeswitching and borrowing," *Journal of Multilingual & Multicultural Development*, vol. 13, no. 1-2, pp. 19–39, 1992.

[4] K. Bali, J. Sharma, M. Choudhury, and 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.

[5] A. Das and B. Gambäck, "Code-mixing in social media text: The last language identification frontier?" in *Proc.of Traitement Automatique des Langues (ATALA)*, 2015.

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

[7] 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.

[8] I. Hamed, M. Elmahdy, and S. Abdennadher, "Building a First Language Model for Code-switch Arabic-English," *Procedia Computer Science*, vol. 117, pp. 208–216, 2017.

[9] D. Amazouz, M. Adda-Decker, and L. Lamel, "The French-Algerian code-switching triggered audio corpus (FACST)." in *Proc. of Language Resources and Evaluation Conference (LREC)*, 2018.

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

[11] E. Yilmaz, M. Andringa, S. Kingma, J. Dijkstra, F. Van der Kuip, H. Van de Velde, F. Kampstra, J. Algra, H. Heuvel, and 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.

[12] S. Malhotra, "Hindi-English, Code Switching and Language Choice in Urban, Uppermiddle-class Indian Families," *University of Kansas. Linguistics Graduate Student Association*, 1980.

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

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

[15] D.-C. Lyu, T.-P. Tan, E. S. Chng, and 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.

[16] D. C. Lyu, R. Y. Lyu, Y. C. Chiang, and 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.

[17] T. Solorio and 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.

[18] K. Bhuvanagirir and S. K. Kopparapu, "Mixed language speech recognition without explicit identification of language," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 92–97, 2012.

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

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

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

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

[23] S. Ganji, K. Dhawan, and R. Sinha, "IITG-HingCoS corpus: A Hinglish code-switching database for automatic speech recognition," *Speech Communication*, vol. 110, pp. 76–89, 2019.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the 23rd International Conference on Machine learning*, 2006, pp. 369–376.

[25] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *Proc. of Deep Learning and Representation Learning Workshop*, 2014.

[26] W. Geng, W. Wang, Y. Zhao, X. Cai, B. Xu, C. Xinyuan *et al.*, "End-to-end language identification using attention-based recurrent neural networks," in *Proc. of Interspeech, an Annual Conference of International Speech Communication Association*, 2016.

[27] K. A. H. Zirker, "Intrasentential vs. intersentential code switching in early and late bilinguals," 2007.

[28] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[29] Vincent, "Nabu: An end-to-end speech recognition toolkit," [Online] https://vrenkens.github.io/nabu/, accessed: 2019-03-24.

[30] B. Parhami, "Voting algorithms," *IEEE Transactions on Reliability*, vol. 43, no. 4, pp. 617–629, 1994.

[31] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, "Towards end-to-end code-switching speech recognition," *arXiv preprint arXiv:1810.13091*, 2018.