IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2020, accepted April 1, 2020, date of publication April 7, 2020, date of current version April 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986255

Exploration of End-to-End Framework for Code-Switching Speech Recognition Task: Challenges and Enhancements

GANJI SREERAM AND ROHIT SINHA, (Member, IEEE)
Department of Electronics and Electrical Engineering, IIT Guwahati, Guwahati 781039, India
Corresponding author: Ganji Sreeram (s.ganji@iitg.ac.in)

ABSTRACT The end-to-end (E2E) framework has emerged as a viable alternative to conventional hybrid systems in automatic speech recognition (ASR) domain. Unlike the monolingual case, the challenges faced by an E2E system in code-switching ASR task include (i) the expansion of target set to account for multiple languages involved, (ii) the requirement of a robust target-to-word (T2W) transduction, and (iii) the need for more effective context modeling. In this paper, we aim to address those challenges for reliable training of the E2E ASR system on a limited amount of code-switching data. The main contribution of this work lies in the E2E target set reduction by exploiting the acoustic similarity and the proposal of a novel context-dependent T2W transduction scheme. Additionally, a novel textual feature has been proposed to enhance the context modeling in the case of code-switching data. The experiments are performed on a recently created Hindi-English code-switching corpus. For contrast purposes, the existing combined target set based system is also evaluated. The proposed system outperforms the existing one and yields a target error rate of 18.1% along with a word error rate of 29.79%.

INDEX TERMS Code-switching, speech recognition, end-to-end system, factored language model, target-to-word transduction.

I. INTRODUCTION

Multilingual speakers often alternate between two or more languages (or dialects) during the conversation. In literature, this phenomenon is referred to as code-switching [1], [2]. The language to which the syntax of a code-switching sentence belongs is referred to as a native language, while that of the embedded foreign words is referred to as a non-native language [3]. The broad domains that carry out research on code-switching phenomenon are (i) linguistics [4], [5], (ii) language identification and diarization [6], [7], (iii) automatic speech recognition (ASR) [8]–[10], and (iv) language modeling [11], [12]. The scope of this work is limited to building an ASR system for the code-switching data.

Early works on code-switching ASR [9], [13], [14] happen to employ the hybrid framework typically developed for monolingual ASR task. The hybrid framework comprises three sub-modules, namely, a pronunciation model (PM), an acoustic model (AM), and a language model (LM). The PM takes into account the typical pronunciation variations of words by employing a dictionary. The AM involves the creation of the statistical models for the sub-word units such as phonemes or senones, given the acoustic features. The LM captures the conditional probabilities of the next words given the observed word sequences and helps in reducing the search space. All these sub-modules are trained and optimized separately, hence the resulting system can be sub-optimal. Towards addressing that, the end-to-end (E2E) framework was proposed and successfully explored in the monolingual ASR task [15]–[20]. Two variants of the E2E framework include (i) the connectionist temporal classification (CTC) [21], [22], and (ii) the sequence-to-sequence modeling with an attention mechanism [15], [16]. In both these variants, the network is trained with characters as the output targets and does not include any explicit PM or LM. Thus, the E2E ASR framework does not require the phonetically labeled training data. For multiple languages being involved, these attributes become more attractive in the case of code-switching ASR. Motivated by that, the recent works have explored the E2E framework in the code-switching ASR domain. In the very first work [23], Seki *et al.* explored an E2E ASR system for code-switching task on an artificially created dataset obtained by concatenating the monolingual utterances. In contrast, Shan *et al.* [24] employed

The associate editor coordinating the review of this manuscript and approving it for publication was Bohui Wang.

a real Mandarin-English code-switching dataset for developing the attention-based E2E ASR system. For improving the ASR performance, the multi-task learning (MTL) framework involving the language identification (LID) [25] was employed. In another work, Li *et al.* [26] explored a CTC-based E2E ASR system combined with frame-level LID for recognizing Chinese-English code-switching data. In a recent work on Mandarin-English code-switching ASR, Zeng *et al.* [27] experimented with data augmentation, MTL for LID, byte-pair encoding, and expansion of vocabulary in LM for N-best rescoring in the context of attention-based E2E framework.

In the existing code-switching E2E ASR works, the target set is derived by simply combining the character sets of the languages involved. It is argued that such systems would suffer from high confusability among the cross-language targets unless a sufficiently large amount of data is available for training. The possible cause of the confusability lies in the broad acoustic similarity among sound units involved in most of the code-switching language pairs. Also, for the enhanced target set, such systems would exhibit high computational complexity. In the context of low resourced modeling, one can avoid such a confusability if a common phone set covering the underlying languages in code-switching data is used as the output target. In an earlier work [28], we have proposed a common phone set towards building a hybrid ASR system for the Hindi-English code-switching task. Motivated by the above-cited reasons, we first explore the earlier defined common phone set as a reduced target set for developing an attention-based E2E Hindi-English code-switching ASR system using the *HingCoS* corpus [28]. Interestingly, the reduced target set based E2E ASR system outperformed the combined target set one in terms of the target error rate (TER). But, a reverse trend was noted when those target sequences were converted to word sequences, i.e., for computing the word error rate (WER). This degradation in WER is because of the enhanced confusability among the homophones (the words having identical pronunciation but different spellings) within or across the languages involved. For addressing the same, we have also proposed a context-dependent target-to-word (T2W) transduction scheme that employs an explicit *error model* (EM) along with an LM to provide context information. Thus, this work is focused on addressing the following two issues in the context of code-switching E2E ASR:

- High confusability among cross-lingual targets for combined target set modeling.
- Enhanced confusability among within- and across-language homophones.

Further, to enhance the context information of the non-native language words, we have also proposed a novel textual feature referred to as the *code-switching identification* (CSI) feature. With the incorporation of the CSI feature in the training of factored language model (FLM), in particular, the recurrent neural network-based FLM (RNN-FLM) [29], more effective modeling of code-switching is achieved.

The proposed context-dependent T2W transduction scheme is noted to achieve a relative improvement of 22% over the naive transduction scheme in the context of reduced target set based Hindi-English code-switching E2E ASR. Further improvements in the WER performances are achieved with the use of FLMs in T2W transduction. The proposed approaches are generic enough to be applied in any other code-switching context. In the context of code-switching E2E ASR, the notable contributions of this work are summarized as below.

- Development of a Hindi-English code-switching ASR system on the *HingCoS* corpus.
- Exploitation of acoustic similarity for the target set confusability reduction.
- Proposition of the context-dependent T2W transduction scheme for achieving enhanced WER performance.
- Proposition of a novel textual feature to deal with the intra-sentential code-switching.

The remainder of this paper is organized as follows: Section II presents a brief review of the code-switching phenomenon. Section III describes the development of attention-based E2E ASR systems on both the combined and the reduced target sets. It also includes a discussion about the challenges arising with the reduced target set. The proposed context-dependent T2W transduction scheme is presented in Section IV. The discussion about the proposed CSI textual feature for enhancing the context modeling of the code-switching data is presented in Section V. The details of the experimental setup, system description, and tuning of parameters for various systems developed are given in Section VI. The experimental results and their discussion are reported in Section VII. Finally, the paper is concluded in Section VIII.

II. A REVIEW OF CODE-SWITCHING PHENOMENON

Code-switching is a phenomenon in linguistics which refers to the use of two or more languages, especially within the same discourse. This phenomenon can be broadly classified into two modes based on the locations of the non-native language words in the sentence. The code-switching that occurs at the boundary of the sentence is referred to as *inter-sentential* code-switching. While the one occurring within the sentence is referred to as the *intra-sentential* code-switching [30]. In literature [4], [5], [31]–[34], the possible reasons for code-switching are attributed to the lack of appropriate words in the native language, emphasizing specific word/phrase, and showing expertise. As a result of the colonization and other historical factors, many multilingual communities have emerged across the globe. In turn, that led to the emergence of code-switching. The salient examples of the same are as follows. Spanish-English [35] in the United States of America, Arabic-English [12] in Egypt, French-German [36] in Switzerland, Frisian-Dutch [37] in Netherlands, Malay-English [9] and Mandarin-English [38]

TABLE 1. Sample Hindi-English code-switching sentences extracted from the HingCoS corpus [28] along with their corresponding English translations. The majority of sentences in the HingCoS corpus correspond to intra-sentential mode.

Hindi-English code-switching sentences	meeting का outcome क्या था rajadhani express के बारे में information कैसे प्राप्त करें class और object के बीच relationship क्या है
English translations	what is the outcome of the meeting how to get information about rajadhani express what is the relationship between class and object

in Malaysia, Mandarin-Taiwanese [6], [13] in Taiwan, Cantonese-English [39] in Hong Kong, English-isiXhosa, English-isiZulu, and English-Setswana [40] in South-Africa, and Hindi-English [41]–[43] in India.

In recent times, there is a greater need to handle the code-switching phenomenon by the spoken-input systems. According to [7], [20], [44], unlike the monolingual case, the salient challenges posed by the code-switching phenomenon are (i) influence of native language on the pronunciation of non-native language words within an utterance, (ii) requirement of expert linguistic knowledge and dedicated tools to handle the involved languages and (iii) lack of publicly available domain-specific resources. For augmenting resources in the Indian context, we recently created the *HingCoS* corpus [28] containing Hindi-English code-switching text and speech data. A few example sentences from the said corpus along with their respective English translations are given in Table 1.

III. EXPLORATION OF E2E ASR FRAMEWORK FOR HINDI-ENGLISH CODE-SWITCHING TASK

In this section, we report a maiden exploration of the E2E framework for the Hindi-English code-switching ASR task. The primary experimentation has been done in the context of attention-based E2E framework. Later, in Section VII-C, the proposed techniques are revalidated in the context of the CTC-based E2E framework for completeness.

The attention-based E2E framework, used in the primary experimentation, employs a popular architecture referred to as the *listen*, *attend*, and *spell* (LAS) [19]. The LAS architecture consists of three sub-modules: Listener, Attender, and Speller as shown in Figure 1. The Listener is a pyramidal bidirectional long short-term memory (BLSTM) network which acts as an encoder. Given a set of input features $\{x_1, \ldots, x_N\}$ corresponding to a predetermined context length N, the Listener produces an embedding $\boldsymbol{h}$ as

$$\boldsymbol{h} = \text{Listener}(x_1, \ldots, x_N).$$

For every time instance, the Attender produces a context c_i, given the embedding $\boldsymbol{h}$ and the decoder state s_i as

$$c_i = \text{Attender}(\boldsymbol{h}, s_i),$$

The decoder state s_i is computed by employing an LSTM network that takes the past decoded output label y_{i-1}, state s_{i-1}, and the context information c_{i-1} as

$$s_i = \text{LSTM}(y_{i-1}, s_{i-1}, c_{i-1}).$$

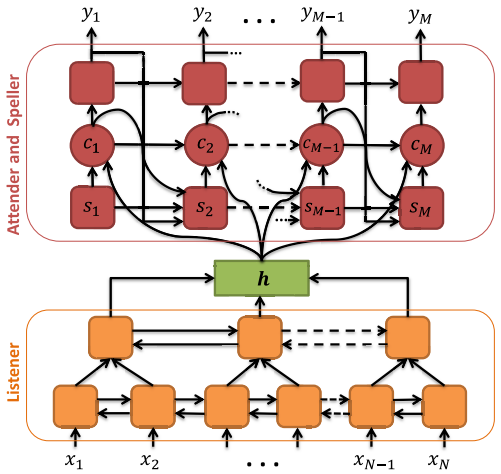

FIGURE 1. The LAS network used for E2E ASR system.

The Attender acts like an alignment generator that determines which portion of $\boldsymbol{h}$ need to be considered for accurate prediction of the current output label y_i. In order to predict y_i, the current context c_i and the previous predicted output label y_{i-1} are passed to the Speller which is an LSTM decoder as

$$p(y_i|\boldsymbol{x}) = \text{Speller}(c_i, y_{i-1}). \quad (1)$$

The entire network is trained to optimize the posterior probability defined as

$$\max_{\lambda} \sum_{i} \log \text{P}(y_i|\boldsymbol{x}, \boldsymbol{y}^*_{<i}; \lambda)$$

where λ represents the LAS model parameters and $\boldsymbol{y}^*_{<i}$ refers to the ground truth of the previously decoded targets.

A. COMBINED TARGET SET MODELING

Typically, the E2E ASR systems are trained for the character set of the spoken language as the output target, given the acoustic features. In the languages which involve both upper and lower case characters, the transcription is normalized to either of the cases. Thus, in the context of code-switching, such systems have to model the combined character set of the underlying languages. In this work, this approach is referred to as the combined target set modeling. The character sets of the Hindi and English languages are shown in Table 2. Using those, we built an E2E Hindi-English code-switching ASR system with 95 targets comprising of Hindi (68), English (26), and a special character '_' used for separating the words. For the details of the database and the LAS system used for the experimental evaluation, the readers are referred to Section VI. The TER of the developed system using the combined target set is reported in Table 3. For reference purposes, a few decoded character sequences are shown in Table 4. For converting the hypothesized target sequences to their corresponding word sequences, a scheme usually followed in

TABLE 2. The character sets of Hindi (68) and English (26) languages that are used as targets in building the conventional E2E Hindi-English code-switching ASR system.

Hindi characters	अ आ इ ई उ ऊ ऋ ए ऐ ओ औ क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द ध न प फ ब भ म य र ल व श ष स ह क़ ख़ ग़ ज़ झ़ ड़ ढ़ फ़ ॉ ॉ ॊ ॊ ो ो ौ ौ ् ् ॎ ॎ ॏ ॏ ॐ ॑ ॒ ॓
English characters	a b c d e f g h i j k l m n o p q r s t u v w x y z

TABLE 3. Evaluation of the E2E ASR system trained on the combined target set for the Hindi-English code-switching task. The WER is computed after the transduction of the hypothesized targets. The percentage of invalid words generated indicate the naiveness of the transduction scheme employed.

%TER	%WER	%⟨*unk*⟩
21.85	33.92	8.2

the literature is employed, and the same has been referred to as the *naive transduction* scheme in this work. In that scheme, first, all-white spaces between the targets are removed, and then each of the '_' labels is replaced by a single space to derive the hypothesized word sequence.

From Table 4, it can be noted that the output character sequences often get corrupted with cross-lingual character substitutions. Thus, even for a single character in the output sequence being misclassified, the naive transduction scheme would not find any valid word match in the task wordlist. In those cases, it outputs a *unknown* (⟨*unk*⟩) label, which in turn degrades the WER. Table 3 also shows the WER for the developed E2E system along with the percentage of ⟨*unk*⟩ labels in the output. The high percentage of ⟨*unk*⟩ labels shows the naiveness of the transduction scheme. The issue of cross-lingual confusability can be addressed either by considering a sufficiently large amount of speech data for acoustic modeling or by deriving the target labels by exploiting the acoustic similarity. In the next subsection, we attempt to reduce the confusability among the cross-lingual targets.

B. REDUCED TARGET SET MODELING

The reduced target set modeling refers to employing a lesser number of target labels than those involved in the combined target set modeling based E2E code-switching ASR system. Recently, in the context of the multilingual ASR task [45], the authors successfully used the union of phone sets of the underlying languages as targets to the E2E ASR system instead of the combined character set. Motivated by that, in an earlier work [28], we had defined a common phone set having 62 labels that cover both Hindi and English languages. In the same work, that phone set was also explored for developing a hybrid Hindi-English code-switching ASR system. For the ease of reference, the said phone set creation is briefly outlined next. We borrowed the phone set for the Hindi language from a composite phone set covering the majority of the Indian languages already defined for computer processing [46]. As the Hindi phone set is bigger, the English phones were heuristically mapped to corresponding Hindi phones having a broad acoustic similarity. And those which could not be mapped to Hindi phones were given unique labels. For more details about that mapping, the readers are referred to Table 4 in [28]. In this work, we employ that common phone set along with the special character '_' as the reduced target set in training the attention-based E2E ASR system for the Hindi-English code-switching task.

The reduced target set based E2E ASR system was trained following the identical setup as used for the combined target set based system discussed in the previous subsection. In Table 5, we show the decoded outputs produced by the reduced target set based E2E ASR system for the same set of example sentences as considered in Table 4. On comparing those tables, it can be noted that the reduced target set based E2E system exhibits a reduction in the cross-lingual target confusability and thus resulting in improved TER performance measure as given in Table 6.

For converting the reduced target set sequence to corresponding word sequence, a pronunciation dictionary for all Hindi and English words in the HingCoS corpus is created. During T2W transduction, each target segment separated by '_' labels is searched in the created pronunciation dictionary and is replaced with the word corresponding to it. In the case of homophone words, the one having the highest unigram count is chosen. If there is no match, then that target segment is replaced with the ⟨*unk*⟩ label. Following that, each of the '_' labels is replaced by a single space to produce the hypothesized word sequence. The WER, along with the percentage of ⟨*unk*⟩ labels, are also reported in Table 6. On comparing Tables 3 and 6, the proposed reduced target set modeling scheme is noted to provide a substantial reduction in the TER as well as the ⟨*unk*⟩ labels in the output. On the flip side, the WER gets significantly degraded. The possible causes of WER degradation are discussed in the following.

1) NAIVETY IN T2W TRANSDUCTION

We first highlight the weakness of the T2W transduction approach typically employed to produce WERs for E2E ASR systems. Let $\{T_{h_i}\}_{i=1}^{n}$ be the segmented hypothesized target sequence, $\{T_{c_j}\}_{j=1}^{m}$ be the segmented correct target sequence and $\{W_j\}_{j=1}^{m}$ be the desired word sequence associated to $\{T_{c_j}\}_{j=1}^{m}$. The objective of the decoder is to determine the desired word, given the corresponding segment of the hypothesized target sequence. In this transduction approach, the desired word W_{h_i} is produced by the decoder only when an exact match for a segment T_{h_i} is found in the pronunciation

TABLE 4. Two sample decoded output sequences of the attention-based E2E ASR system developed using a combined target set for the Hindi-English code-switching task. The English translations of the sentences are given in the braces. The errors obtained in the hypothesized character sequences have been highlighted. Note that the symbol '_' is used to mark the word boundaries. The invalid words produced by the transduction process are labeled as ⟨*unk*⟩.

Example 1	**Ref. sentence:** company के about us page में जानकारी है (*information is in the company's about us page*) **Ref. target sequence:** c o m p a n y _क ◌े _ a b o u t _ u s _ p a g e _ म ◌े ◌ं _ ज ◌ा न क ◌ा र ◌ी _ ह ◌ै **Hyp. target sequence:** क o m p a n y _क ◌े _ए b o u t _ u s _ p a g e _ म ◌े ◌ं _ ज ◌ा न क ◌ा र ◌ी _ ह ◌ै **Hyp. sentence:** <*unk*> के <*unk*> us page में जानकारी है
Example 2	**Ref. sentence:** आपको hindi में blogging शुरू करनी चाहिए (*you should start blogging in hindi*) **Ref. target sequence:** आ प क ◌ो _ h i n d i _ म ◌े ◌ं _ b l o g g i n g _ श ◌ु र ◌ू _ क र न ◌ी _ च ◌ा ह ◌ि ए **Hyp. target sequence:** आ प क ◌ो _ h i n d i _ म ◌े ◌ं _ b l o g g i n g _ श ◌ु र ◌ू _ क र न ◌ी _ च ◌ा ह ◌ि ए **Hyp. sentence:** आपको hindi में blogging शुरू करनी चाहिए

TABLE 5. Two sample decoded output sequences of the attention-based E2E ASR system trained on the reduced target set for the Hindi-English code-switching task. For contrast purposes, the sentences are kept the same as considered in Table 4.

Example 1	**Ref. sentence:** company के about us page में जानकारी है (*information is in the company's about us page*) **Ref. target sequence:** k a m p a n ii _ k ee _ a b au tx _ a z _ p ei j _ m ee _ j aa n k aa r ii _ h ei **Hyp. target sequence:** k a m p a n ii _ k ee _ a b au tx _ a s _ p ei j _ m ee q _ j aa n k aa r ii _ h ei **Hyp. sentence:** company के about <*unk*> page में जानकारी है
Example 2	**Ref. sentence:** आपको hindi में blogging शुरू करनी चाहिए (*you should start blogging in hindi*) **Ref. target sequence:** aa p k o _ h i n dx ii _ m ee _ b l ao g i ng _ sh u r uu _ k a r n ii _ c aa h i ee **Hyp. target sequence:** aa p k o _ h i n dx ii _ m ee _ b l ao g i ng _ sh u r uu _ k a r n ii _ c aa h i ee **Hyp. sentence:** आपको हिंदी में blogging शुरू करनी चाहिए

TABLE 6. Evaluation of the E2E ASR system trained on reduced target set for Hindi-English code-switching task.

%TER	%WER	%⟨*unk*⟩
18.1	40.19	6.3

dictionary D. For a single entry of T_{h_i} being in error, no valid word corresponding to it would be found in D. In such cases, the ⟨*unk*⟩ label is emitted. More formally, the same can be expressed by a decoder function $\mathcal{F}$ as

$$\mathcal{F}(T_{h_i}) = \begin{cases} W_{h_i} & \text{if } T_{h_i} \in D \\ \langle unk \rangle & \text{otherwise.} \end{cases} \quad (2)$$

From Eqn. 2, it can be noted that, unless the TER of the E2E ASR system is very low, the derived word-level hypothesis will lead to a degraded WER due to the presence of ⟨*unk*⟩ labels.

2) HOMOPHONE CONFUSABILITY

Almost every language has some homophones, with the English language having a large number of them. In the context of code-switching, the homophone issue gets further enhanced as they may occur across the languages too. For the ease of illustration, we have grouped them into five broad categories, namely proper nouns, internalized collective/abstract nouns, abbreviations, intra-language, and inter-language homophones. This categorization has been done based on parts-of-speech and language information. Table 7 lists a few examples of those homophone categories for the Hindi-English code-switching case. There are 96 homo-

TABLE 7. A few examples of different types of homophones present in the Hindi-English code-switching data.

Proper nouns	Collective/ abstract nouns	Abbreviations	Intra-language	Inter-language
amit - अमित hindi - हिंदी japan - जापान	internet - इंटरनेट ticket - टिकट station - स्टेशन	ATM - एटीएम USA - यूएसए CEO - सीईओ	sea-see due - dew dye - die	fool - फूल bus - बस say - से

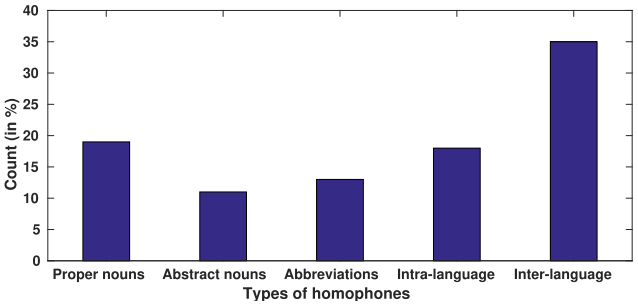

FIGURE 2. Histogram of different types of homophones present in the HingCoS corpus.

phones in the HingCoS corpus and their distribution in terms of earlier defined broad categories is shown in Figure 2. It is worth noting that, a large number of homophones belong to the inter-language category. Despite the reduction of target set confusability with the proposed target set, it was observed that the confusability during T2W transduction gets enhanced for homophones. Referring to Example 2 in Table 5, it can be observed that, for a proper noun having the hypothesized target sequence "h i n dx ii", the T2W transduction process can yield either "hindi" or "हिंदी" as the word output. A high frequency of such errors ends up degrading the WER.

In the context of hybrid ASR task involving Hindi-English code-switching, the authors in [47] faced a similar challenge and explored merging of some identical sounding words based on the unigram counts in their database. It is argued that, by following such an approach, we can handle only a sub-set of homophones (proper nouns, abstract nouns, and abbreviations) but not the set comprising intra- and inter-language homophones. For effective handling of the latter set of homophones, we would require a more in-depth context information rather than unigram counts.

IV. CONTEXT-DEPENDENT T2W TRANSDUCTION

In this section, we propose a context-dependent T2W transduction process developed by exploiting modularized decoding for addressing the earlier highlighted issues.

In hybrid ASR literature, a few works have already explored the modularized decoding for T2W transduction. Demuynck *et al.* [48], [49] proposed a two-step decoding process that employed morpho-syntactic and morpho-phonologic constraints for T2W transduction. Following that work, Zweig and Nedel [50] presented an empirical study on the error-robustness of T2W transduction across a variety of languages. For that study, the decoding objective for the transduction of i^{th} hypothesized segment is formulated as shown below.

$$
\begin{aligned}
&\arg\max_{W_i} \mathrm{P}(W_i|T_{h_i}) \\
&\quad = \arg\max_{W_i} \mathrm{P}(W_i)\,\mathrm{P}(T_{h_i}|W_i) \\
&\quad = \arg\max_{W_i} \mathrm{P}(W_i) \sum_{T_{c_i}} \mathrm{P}(T_{c_i}, T_{h_i}|W_i) \\
&\quad = \arg\max_{W_i} \mathrm{P}(W_i) \sum_{T_{c_i}} \mathrm{P}(T_{h_i}|T_{c_i}, W_i)\,\mathrm{P}(T_{c_i}|W_i) \\
&\quad \approx \arg\max_{W_i,\, T_{c_i}} \mathrm{P}(W_i)\ \mathrm{P}(T_{c_i}|W_i)\ \mathrm{P}(T_{h_i}|T_{c_i})
\end{aligned}
$$

In the above formulation, the first and second factors respectively denote the LM and the PM, while the third factor accounts for the EM. With the maximization performed over all possible correct target sequences T_{c_i} and their corresponding words W_i, the earlier discussed $\langle unk \rangle$ and homophone issues can be resolved. Exploiting these observations, a novel T2W transduction scheme for E2E ASR systems has been evolved and is explained below.

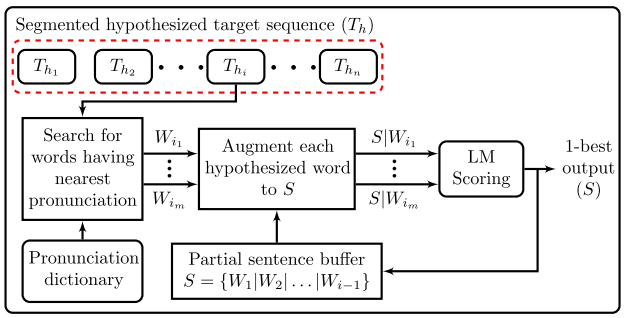

FIGURE 3. Flow chart of the proposed context-dependent T2W transduction scheme.

TABLE 8. Demonstration of T2W transduction achieved by the naive and the proposed schemes for the reduced target set case. The hypothesized target segments and words in error are shown in 'red' colour. For the proposed scheme, the highlighted LM score corresponds to the 1-best output.

Ref. sentence	क्या आपने google web light से अपने stats में traffic को notice किया
Ref. target sequence	k y aa _ aa p n ee _ g uu g a l _ w ae b _ l ii tx _ s ee _ a p n ee _ s tx ae tx s _ m ee _ tx r ae f i k _ k o _ n o tx i s _ k i y aa
Hyp. target sequence	k y aa _ aa p n ee _ g uu g a l _ w ae b _ l i tx _ s ee _ a p n ee _ s tx ei tx s _ m ee _ tx r ae f i k _ k o _ n o tx i s _ k i y aa
Naive T2W transduction	क्या आपने google web <unk> से अपने <unk> में traffic co notice किया
Proposed T2W transduction (Candidate sentences with LM scores)	-32.91 क्या आपने google web light से अपने stats में traffic co notice किया **-27.91** क्या आपने google web light से अपने stats में traffic को notice किया -32.33 क्या आपने google web light से अपने status में traffic co notice किया -28.58 क्या आपने google web light से अपने status में traffic को notice किया -35.79 क्या आपने google web lite से अपने stats में traffic co notice किया -32.51 क्या आपने google web lite से अपने stats में traffic को notice किया -37.10 क्या आपने google web lite से अपने status में traffic co notice किया -33.22 क्या आपने google web lite से अपने status में traffic को notice किया

The hypothesized target sequences produced by an E2E ASR system may contain one or more errors. For the error modeling purpose, we have employed the Levenshtein (edit) distance-based search. Let $\{T_{h_i}\}_{i=1}^{n}$ denote a hypothesized target sequence having n segments. For each segment T_{h_i}, we have determined all possible (say p) pronunciation sequences $\{T_{c_j}\}_{j=1}^{p}$ in PM having edit distances up to a predetermined threshold.[1] For the reduced target set case, those sequences may further map to homophones within or across the languages. Let a set $\{W_{i_k}\}_{k=1}^{m}$ denote all possible (say m) words returned by that search. On appending each of those words to the current partial sentence S, a corresponding new partial candidate sentence is constructed. All those constructed sentences are now pruned based on the context information derived from an appropriate LM and the 1-best sentence is generated. When all segments in T_h get processed, the 1-best output yields the final transducted output. The overall flow diagram of the proposed T2W transduction scheme is shown in Figure 3.

[1] In this work, the threshold is set as zero when the minimum edit distance value of the matches is zero; otherwise, it is set as one more than the minimum edit distance value.

The innovation in the proposed scheme is demonstrated with the help of an example shown in Table 8. The top-two rows of that table show the word- and target-level reference transcriptions for an example utterance. The output generated by the reduced target set-based E2E ASR system for that utterance is given in the third row. Whereas, the last-two rows correspond to the outputs produced by the naive and the proposed transduction schemes. On comparison, it can be noted that the proposed scheme not only avoids $\langle unk \rangle$ labels but also can handle intra- and inter-language homophone pairs such as "light–lite" and "co–को" , respectively. The first attribute refers to effective error modeling, while the second one is the result of context modeling. The LM scores of the candidate sentences clearly show how effectively the homophone issue gets resolved. Thus, LM plays a vital role in the proposed T2W transduction scheme.

The effective language modeling of code-switching text data is a research problem in itself. Towards that end, a novel textual feature is discussed in the following section.

V. CODE-SWITCHING TEXTUAL FEATURE FOR ENHANCED CONTEXT MODELING

In the context of code-switching, a few works have already shown that the FLMs trained with parts-of-speech (POS) information are more effective than the traditional LMs [35], [51]. In intra-sentential code-switching, the non-native words are embedded into the native sentences, mostly without affecting their structure. In an earlier work [52], we had exploited those structures to adapt the monolingual LM to deal with code-switching data. In this section, we demonstrate how those structures could be exploited in the direct modeling of code-switching data. For that purpose, we propose a textual feature that can be used in FLM training similar to POS tags. In the following, extraction of the proposed textual feature and the RNN-FLM paradigm in which it is included, are described.

A. PROPOSED TEXTUAL FEATURE

The proposed textual feature is referred to as *code-switching identification* (CSI) feature in this work. It not only marks the location of code-switching but also provides information about the equivalent native (Hindi) word. The procedure for extracting the CSI feature for Hindi-English code-switching data is described next. In the HingCoS corpus, the Hindi and English words appear in Devanagari and Latin scripts, respectively. First, we pass every English-scripted word (w_E) through a machine translator to get its equivalent Hindi-scripted word (w_H). Later, a string {W-w_E:S-w_H:C-Yes} is emitted, which comprises the word w_E along with the CSI feature. Where the identifiers W-, S-, and C- denote input-word, switched-word, and code-switching status, respectively. Similarly, for Hindi-scripted word (w_H), the string including the CSI feature, is emitted as {W-w_H:S-w_H:C-No}. Note that, as there is no code-switching, W- and S- are tagged with the same word w_H. The structure of the above strings follows the syntax of the FLM toolkits [53], [54].

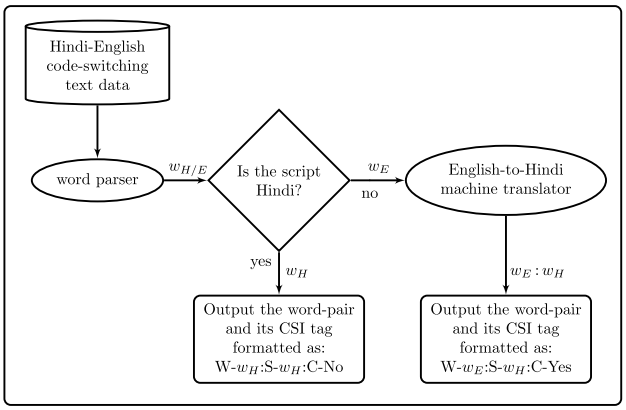

FIGURE 4. Flow chart of the scheme employed to tag the code-switching text data with the proposed CSI feature. In this work, English-to-Hindi translation is done by employing the *Google Translate*, an online machine translation tool.

CSI tagged output: W-meeting:S-मुलाकात:C-Yes W-का:S-का:C-No W-outcome:S-परिणाम:C-Yes W-क्या:S-क्या:C-No W-था:S-था:C-No

Hindi-English sentence: meeting का outcome क्या था
(*what is the outcome of the meeting*)

FIGURE 5. The proposed CSI tagged output for an example Hindi-English code-switching sentence. The English translation of example sentence is given in bracket for reference.

The algorithm for the proposed CSI feature extraction is given as a flowchart in Figure 4. Also, an example Hindi-English code-switching sentence tagged with the proposed CSI feature is shown in Figure 5.

B. RNN-FLM ARCHITECTURE

The FLMs incorporate morphological features or linguistic features of the word w_t while training the LM [55]. For training the FLMs, the appropriate set of features are derived either using linguistic knowledge or using the data-driven techniques. In this technique, each word w_t in the vocabulary is represented as a group of k features denoted as

$$w_t \equiv \{f_t^1, f_t^2, \ldots, f_t^k\} = f_t^{1:k} = F_t.$$

In recent works, the RNNs are also used in training the FLMs and have shown significant improvement in recognition performances [29], [56]. Motivated by that, in this work, we have employed the RNN-FLM architecture to model the code-switching data by including the CSI textual features. The RNN-FLM predicts the posterior probability of the current word w_t as

$$\begin{aligned} &\mathrm{P}(w_t|\mathrm{F}_{t-1}, s_{t-1}) \\ &\quad = \sum_{c_t} \mathrm{P}(w_t|F_{t-1}, s_{t-1}, c_t)\,\mathrm{P}(c_t|F_{t-1}, s_{t-1}) \\ &\quad \approx \arg\max_{c_t} \mathrm{P}(w_t|F_{t-1}, s_{t-1}, c_t)\,\mathrm{P}(c_t|F_{t-1}, s_{t-1}) \end{aligned}$$

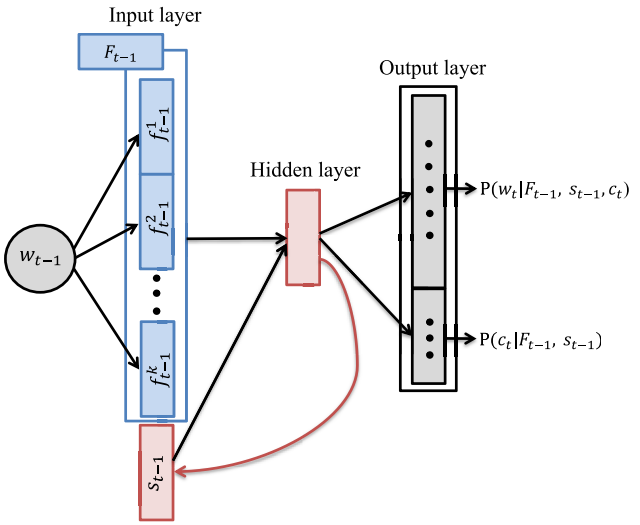

FIGURE 6. Architecture of the RNN-FLM.

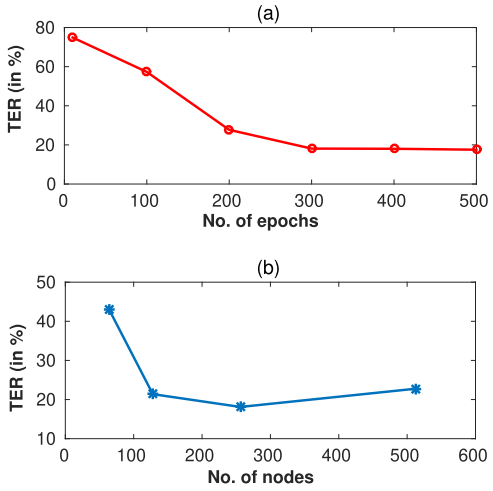

FIGURE 7. The tuning of parameters for attention-based E2E ASR system. (a) Selection of number of epochs, and (b) selection of number of nodes in the encoder.

TABLE 9. Salient statistics of text and speech components of the HingCoS corpus. The CS count refers to the number of code-switching instances in the data.

Type	# sent.	# words		# unique words		CS count
		Hindi	English	Hindi	English	
Text	25,988	381,603	196,556	6,029	8,614	104,912
Speech	9,251	125,653	50,719	2,644	3,901	30,035

where, F_{t-1} denotes the feature vector corresponding to w_{t-1}, i.e., the previous word, s_{t-1} refers to the RNN state [57], and c_t represents the class to which the word w_t belongs to [58]. Those classes are derived by partitioning the vocabulary of training data into groups based on the word counts which helps in reducing the search complexity. The network architecture of RNN-FLM while highlighting the component variables is shown in Figure 6.

VI. EXPERIMENTAL SETUP

A. DATABASE

In this work, a recently created *HingCoS Corpus*[2] has been used for the experimentation purpose. The text data in the HingCoS corpus consists of 25988 Hindi-English code-switching sentences and has a vocabulary of 14643 words (6029 Hindi and 8614 English). The lengths of sentences vary from 3–57 words and on an average, there are 3–4 code-switching instances per sentence. For a total of 9251 Hindi-English text sentences in the HingCoS corpus, the corresponding speech data spoken by 101 speakers (61 males and 40 females) is also available. The speech data, being collected over telephones, is sampled at 8 kHz with a resolution of 16 bits/sample. The total size of the speech data is about 25 hours. The salient statistics of the HingCoS corpus are summarized in Table 9.

The available speech data is divided into three non-overlapping sets having 7115, 160, and 1976 utterances for training, development, and testing of the E2E ASR systems, respectively. These sets are also non-overlapping in terms of the speakers involved. For language modeling, excluding the earlier defined acoustic test set, the remaining text data is partitioned into training and development sets having 22700 and 1312 sentences, respectively. In this way, both acoustic and language models are evaluated on the same test set.

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

B. SYSTEM DESCRIPTION AND PARAMETER TUNING

1) ATTENTION-BASED E2E ASR SYSTEM

The Nabu toolkit [59] is used for developing the LAS architecture-based E2E ASR system. The parameter settings used for analyzing the speech data include window length of 25 ms, window shift of 10 ms, and pre-emphasis factor of 0.97. The 40-dimensional log Mel-filterbank energies per speech frame are used as features for acoustic modeling. The details of the LAS architecture are as follows. The listener has 3 pyramidal BLSTM layers, with 256 units in each layer. The pyramidal step size is kept as 2, and the dropout rate in training is set to 0.5. The speller has 2 LSTM layers, with 256 units in the input layer. The dropout rate for the speller is also set to 0.5. The average cross-entropy loss is used as a loss function. The model is trained for 300 epochs with a batch size of 32 and learning rate set to 0.1 with decay 0.01. For decoding, a beam-search decoder with beam width set to 10 is employed.

For training the LAS network, the number of epochs and the number of hidden units in the input layer of the encoder are selected performing the tuning experiments on the acoustic development set. The tuning of the LAS network is done for the combined target set case, and the same parameters are fixed even for the reduced target set case. The plots showing the trends of those experiments are given in Figure 7. Note that tuning for the number of epochs is done by keeping the number of nodes as fixed and vice-versa. The remaining parameters are set to their default values as defined in the

TABLE 10. Quality assessment of T2W transduction in terms of %WERs obtained for attention-based E2E ASR systems developed using reduced and combined target sets. In Naive method, neither an error model (EM) nor a language model (LM) is involved.

Transduction method	EM / LM	%WER	
		Reduced	Combined
Naive	No / No	40.19	33.92
Proposed	Yes / RNN	**31.09**	32.29
Contrast	Yes / Unigram	33.02	33.16
	Yes / No	36.77	33.56

toolkit. The TER is found to saturate after 300 epochs, while it degrades beyond 256 nodes.

2) RNN-BASED LMs

For the experimentation purpose, both simple and factor modeling-based RNN-LMs are developed using the RNN-LM toolkit [54]. Both kinds of RNN-LMs are developed employing identical network architecture with *sigmoid* as the non-linearity function. After performing the tuning of simple RNN-LM on the linguistic development set, the salient parameters of the architecture are a hidden layer with 200 nodes, the value of back-propagation through time variable set to 5, and the number of classes set to 100.

VII. EXPERIMENTAL RESULTS

In this section, we present the evaluation of both the proposed context-dependent T2W transduction scheme and the CSL textual feature in the context of the Hindi-English code-switching ASR task.

A. EVALUATION OF THE T2W TRANSDUCTION

For the primary proposal in this work, i.e., the reduced target set for the E2E ASR system, the results are already discussed in Section III. From the experiments done on the HingCoS corpus, it can be deduced that the proposed reduced target set modeling yields about 17% relative improvement in TER in contrast to the combined target set case. Towards addressing the challenges in T2W transduction with the reduced target set modeling, we have also proposed a context-dependent T2W transduction scheme as the secondary contribution. Table 10 presents the detailed evaluation of the same while studying the impact of both the error model and the inclusion of context information. The proposed transduction scheme yields a WER of 31.09%, which happens to be 22.6% relative improvement over that of the naive transduction scheme. Further, to study the impact of context information on the transduction performance, we also evaluated the proposed scheme with unigram LM and with no LM. The latter case refers to randomly choosing one among the word possibilities available during the error modeling. From those results, we can conclude that most of the improvement in the T2W transduction performance has been achieved on account of better context modeling. For comparison purposes, the results for the combined target set modeling case are also given in Table 10.

TABLE 11. The evaluation of CSI and POS features in the context of Hindi-English code-switching data. For contrast purpose, the perplexity (PPL) score for the default RNN-LM is also reported.

Model	Features	PPL
RNN-FLM	Word + CSI	88.59
	Word + POS	102.45
	Word + CSI + POS	**70.11**
RNN-LM	Word	115.23

Unlike the reduced target set modeling, the homophone issue does not crop up in the combined target set case, despite that the reduced target set modeling results in the best WER.

B. EVALUATION OF THE CSI TEXTUAL FEATURE

As argued earlier and also obvious from Table 10, for the reduced target set case, the T2W transduction performance is highly dependent on the quality of the context information provided by the LM. Therefore, as the third contribution of this work, we have proposed the CSI textual feature for improving the code-switching LM. The same has been evaluated separately in language modeling and speech recognition tasks. Table 11 shows that with the inclusion of CSI feature, about 23% relative reduction has been achieved in the perplexity (PPL) score in comparison to default RNN-LM. This improvement is attributed to (i) the binary categorization of code-switching, and (ii) tagging of code-switching (English) words in the training data to their corresponding native (Hindi) words. For the code-switched words having a little or no evidence in the training data, the FLM falls back to their equivalent Hindi words, if those exist in the vocabulary. For better contrast, we have also experimented with the POS textual features extracted by following a procedure similar to that used in [35]. The PPL scores for standalone inclusion of the POS features and their combination with the CSI feature are produced and given in Tables 11. Also, these textual features are evaluated for the proposed T2W transduction scheme, and the performances in terms of WER are reported in Table 12. From that table, it can be noted that a similar trend has been noted in WER scores as that of the PPL scores given in Tables 11. Note that, unlike the CSI feature, the grouping induced by the POS features are targeted towards sentence structure rather than code-switching. This could be the reason why the CSI feature not only outperformed the POS features but also provided an additive improvement when combined with the POS features.

C. REVALIDATION IN ALTERNATE E2E FRAMEWORK

For a thorough evaluation, the proposed approaches have also been evaluated in the CTC-based E2E ASR framework. In contrast to the attention-based system, the CTC-based E2E system consists of two modules: a deep BLSTM encoder, and a CTC decoder. The deep BLSTM network encodes input feature vector $\boldsymbol{x}$ into a higher-level representation. The decoding is performed using the CTC, a loss function that

TABLE 12. Assessment of textual feature augmented LMs on the proposed T2W transduction scheme for attention-based E2E ASR systems trained on reduced/combined target set.

Features for FLM	%WER	
	Reduced	Combined
Word + CSI	30.43	31.12
Word + POS	30.84	31.68
Word + CSI + POS	**29.79**	30.80

TABLE 13. Quality assessment of T2W transduction in terms of %WERs obtained for CTC-based E2E ASR systems developed using reduced and combined target sets.

Transduction method	EM / LM	%WER	
		Reduced	Combined
Naive	No / No	42.21	41.12
Proposed	Yes / RNN	**35.07**	37.97
Contrast	Yes / Unigram	37.92	39.23
	Yes / No	39.38	40.74

TABLE 14. Assessment of textual feature augmented language modeling on the proposed T2W transduction scheme for CTC-based E2E ASR systems trained on reduced/combined target set.

Features for FLM	%WER	
	Reduced	Combined
Word + CSI	34.09	35.23
Word + POS	34.83	35.78
Word + CSI + POS	**33.17**	34.64

assumes the outputs generated at different time steps to be conditionally independent. The CTC allows training of the network without requiring a prior alignment between input and output sequences. In the CTC decoder network, the output softmax layer has one unit each for the targets in addition to a blank symbol ϵ denoting the null emission. For a given training utterance, there are many possible alignments. At every time-step, the network decides whether to emit a symbol or not. As a result, the distribution over all possible alignments between the input and target sequences is obtained. To produce the probability of output sequence given the input, a dynamic programming-based algorithm (*forward-backward*) is employed to obtain the sum over all possible alignments. Given a target transcription $\boldsymbol{y}$ and the input feature vector $\boldsymbol{x}$, the network is trained to minimize the CTC loss function as

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

where $\text{P}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{a}\in\beta(\boldsymbol{y},\ \boldsymbol{x})} \text{P}(\boldsymbol{a}|\boldsymbol{x})$, $\boldsymbol{a}$ is an alignment, and $\beta(\boldsymbol{y},\ \boldsymbol{x})$ is set of all possible sequences between $\boldsymbol{y}$ and $\boldsymbol{x}$. In our implementation of the CTC-based system, the DBLSTM encoder network consists of 4 layers and 256 units in each layer. The remaining network training parameters are kept the same as mentioned in Section VI-B1. The CTC-based system is trained and evaluated on identical data partitions, as already described in Section VI-A.

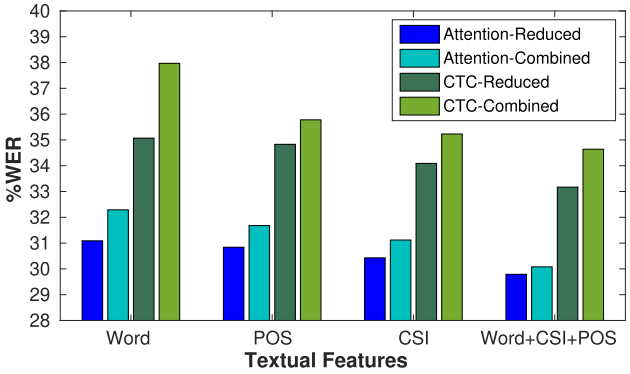

FIGURE 8. Assessment of the impact of proposed context-dependent T2W transduction with/without textual features on attention- and CTC-based E2E ASR systems.

TABLE 15. The details of the memory usage and the computational time for training the E2E ASR systems using both the reduced and combined target sets.

E2E system	Target set	GPU usage	RAM usage	Avg. minibatch time
Attention	Reduced	1412 MB	4.66 GB	2.36 sec
	Combined	1832 MB	12.60 GB	3.41 sec
CTC	Reduced	1223 MB	4.63 GB	1.57 sec
	Combined	1791 MB	12.57 GB	2.92 sec

†CPU: Intel® Xeon®, 64-bit, @3.60GHz × 12; RAM: 128 GB, DDR4; GPU: GeForce GTX 1060, 6 GB.

The evaluations of the proposed context-dependent T2W transduction scheme and the CSI textual feature for CTC-based Hindi-English code-switching E2E ASR have been done, and the results are reported in Tables 13 and 14, respectively. On comparing with the corresponding performances of the attention-based system, it can be noted that the CTC-based E2E framework has exhibited similar performance trends. The proposed transduction scheme yields a WER of 35.07%, which happens to be 16.91% relative improvement over that of the naive transduction scheme. Also, when the LM is trained with combined textual features, an improved WER of 33.17% has been achieved.

For the ease of assessment of the relative impact of the proposed context-dependent T2W transduction with/without textual features, the performances for the attention- and CTC-based E2E ASR systems are summarized in Figure 8. It can be noted that a similar trend in the performances is observed for both E2E frameworks. At the same time, we wish to point out that the developed CTC-based E2E system does not incorporate any character-level LM while decoding the combined/reduced targets. Therefore, the performances of the attention- and CTC-based E2E ASR systems cannot be directly compared.

D. COMPUTATIONAL COMPLEXITY

All systems are developed on a HP-Z440 workstation. The memory requirement and the computational complexity for different systems along with the key specifications of the said

workstation, are given in Table 15. From that table, it can be noted that, the reduced target set based E2E ASR system training takes much lesser memory and computational time when compared to the combined target set case.

VIII. CONCLUSION

This paper explores the development of code-switching E2E ASR system on limited resources. For efficient modeling of the code-switching E2E ASR system, the acoustic similarity-based target reduction scheme has been proposed. Towards converting the hypothesized target sequence to the desired word sequence, a context-dependent transduction scheme has been developed. Further, a novel textual feature has also been proposed, which enables more effective context modeling of code-switching data. The proposed approaches are noted to consistently outperform the combined target set based E2E ASR modeling in terms of target/word error rate. The work also presents a detailed description of the Hindi-English code-switching E2E ASR system. To the best of authors' knowledge, for the Hindi-English code-switching task, such a system is yet to be reported.

Despite the evaluations being performed in the context of Hindi-English code-switching ASR, the proposed techniques are generic enough to be applied for any other code-switching context. Also, the reduced target set based E2E ASR systems training take much lesser memory and computational time when compared to the combined target set case. A few shortcomings of the proposed methods include (i) the requirement of a pronunciation model in the context-dependent T2W transduction scheme for reduced target set case, and (ii) the dependency of the CSI and POS features on the quality of the machine translator and the POS tagger employed, respectively.

In the future, we would be interested in incorporating the T2W transduction into the E2E system to optimize the developed system directly at word-level instead of characters.

REFERENCES

[1] J. J. Gumperz, *Discourse Strategies*. Cambridge, U.K.: Cambridge Univ. Press, 1982.
[2] C. M. Eastman, "Codeswitching as an urban language-contact phenomenon," *J. Multilingual Multicultural Develop.*, vol. 13, nos. 1–2, pp. 1–17, Jan. 1992.
[3] C. Myers-Scotton, "Comparing codeswitching and borrowing," *J. Multilingual Multicultural Develop.*, vol. 13, nos. 1–2, pp. 19–39, Jan. 1992.
[4] C. Myers-Scotton, *Social Motivations for Code-Switching: Evidence From Africa*. Oxford, U.K.: Clarendon, 1993.
[5] L. Malik, *Socio-Linguistics: A Study of Code-Switching*. Bengaluru, Karnataka: Anmol Publications, 1994.
[6] D. C. Lyu and R. Y. Lyu, "Language identification on code-switching utterances using multiple cues," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2008, pp. 711–714.
[7] D.-C. Lyu, E.-S. Chng, and H. Li, "Language diarization for code-switch conversational speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7314–7318.
[8] K. Bhuvanagirir and S. Kumar Kopparapu, "Mixed language speech recognition without explicit identification of language," *Amer. J. Signal Process.*, vol. 2, no. 5, pp. 92–97, Dec. 2012.
[9] B. H. A. Ahmed and T.-P. Tan, "Automatic speech recognition of code switching speech using 1-Best rescoring," in *Proc. Int. Conf. Asian Lang. Process.*, Nov. 2012, pp. 137–140.
[10] S. Sitaram and A. W. Black, "Speech synthesis of code-mixed text," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2016, pp. 3422–3428.
[11] C. F. Yeh, C. Y. Huang, L. C. Sun, and L. S. Lee, "An integrated framework for transcribing mandarin-english code-mixed lectures with improved acoustic and language modeling," in *Proc. 7th Int. Symp. Chin. Spoken Lang. Process.*, Nov. 2010, pp. 214–219.
[12] I. Hamed, M. Elmahdy, and S. Abdennadher, "Building a first language model for code-switch arabic-english," *Procedia Comput. Sci.*, vol. 117, pp. 208–216, 2017.
[13] D. C. Lyu, R. Y. Lyu, Y. C. Chiang, and C. N. Hsu, "Speech recognition on code-switching among the Chinese dialects," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 1, 2006.
[14] T. Lyudovyk and V. Pylypenko, "Code-switching speech recognition for closely related languages," in *Proc. Spoken Lang. Technol. Under-Resourced Lang.*, 2014, pp. 188–193.
[15] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1764–1772.
[16] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *Proc. Deep Learn. Represent. Learn. Workshop*, 2014.
[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[18] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Aug. 2017, pp. 939–943.
[19] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4960–4964.
[20] G. Indra Winata, A. Madotto, C.-S. Wu, and P. Fung, "Towards end-to-end automatic code-switching speech recognition," 2018, *arXiv:1810.12620*. [Online]. Available: http://arxiv.org/abs/1810.12620
[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.
[22] A. Graves, "Sequence transduction with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn., Represent. Workshop*, 2012.
[23] H. Seki, S. Watanabe, T. Hori, J. L. Roux, and J. R. Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4919–4923.
[24] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Investigating End-to-end speech recognition for mandarin-english code-switching," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6056–6060.
[25] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, "Towards end-to-end code-switching speech recognition," 2018, *arXiv:1810.13091*. [Online]. Available: http://arxiv.org/abs/1810.13091
[26] K. Li, J. Li, G. Ye, R. Zhao, and Y. Gong, "Towards code-switching ASR for End-to-end CTC models," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6076–6080.
[27] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the End-to-End solution to mandarin-english code-switching speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Sep. 2019.
[28] S. Ganji, K. Dhawan, and R. Sinha, "IITG-HingCoS corpus: A hinglish code-switching database for automatic speech recognition," *Speech Commun.*, vol. 110, pp. 76–89, Jul. 2019.
[29] Y. Wu, X. Lu, H. Yamamoto, S. Matsuda, C. Hori, and H. Kashioka, "Factored language model based on recurrent neural network," in *Proc. COLING*, 2012, pp. 2835–2850.
[30] F. Grosjean, *Life With Two Languages: An Introduction to Bilingualism*. Cambridge, MA, USA: Harvard Univ. Press, 1982.
[31] L. Milroy and P. Muysken, *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching*. Cambridge, U.K.: Cambridge Univ. Press, 1995.
[32] 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. Symp. About Lang. Soc.*, 2001.

[33] 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, and W. Pickett, "A cross-national profile of bullying and victimization among adolescents in 40 countries," *Int. J. Public Health*, vol. 54, no. S2, pp. 216–224, Sep. 2009.
[34] A. Dey and P. Fung, "A hindi-english code-switching corpus," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2014, pp. 2410–2413.
[35] T. Solorio and Y. Liu, "Part-of-speech tagging for english-spanish code-switched text," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2008, pp. 1051–1060.
[36] D. Imseng, H. Bourlard, H. Caesar, P. N. Garner, G. Lecorve, and A. Nanchen, "MediaParl: Bilingual mixed language accented speech database," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2012, pp. 263–268.
[37] 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 *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2016.
[38] 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. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2010.
[39] H. Cao, P. C. Ching, T. Lee, and Y. T. Yeung, "Semantics-based language modeling for cantonese-english code-mixing speech recognition," in *Proc. 7th Int. Symp. Chin. Spoken Lang. Process.*, Nov. 2010, pp. 246–250.
[40] E. van der Westhuizen and T. Niesler, "A first South African corpus of multilingual code-switched Soap Opera speech," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2018.
[41] A. C. Chandola, "Some linguistic influences of English on Hindi," in *Anthropological Linguistics*. 1963, pp. 9–13.
[42] S. Malhotra, "Hindi-english, code switching and language choice in urban, uppermiddle-class indian families," *Kansas Work. Papers Linguistics*, vol. 5, no. 2, pp. 39–46, Jan. 1980.
[43] 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. 1st Workshop Comput. Approaches to Code Switching*, 2014, pp. 116–126.
[44] Ö. Çetinoğlu, S. Schulz, and N. T. Vu, "Challenges of computational processing of code-switching," in *Proc. 2nd Workshop Comput. Approaches Code Switching*, 2016.
[45] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of multilingual ASR using End-to-end LF-MMI," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6061–6065.
[46] 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, and H. A. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *Proc. ISCA Workshop Speech Synth.*, 2013.
[47] B. M. L. Srivastava and S. Sitaram, "Homophone identification and merging for code-switched speech recognition," in *Proc. Interspeech*, Sep. 2018, pp. 1943–1947.
[48] K. Demuynck, T. Laureys, D. V. Compernolle, and H. V. Hamme, "Flavor: A flexible architecture for LVCSR," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2003.
[49] K. Demuynck and D. Van Compernolle, "Robust phone lattice decoding," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, vol. 4, 2006, pp. 1622–1625.
[50] G. Zweig and J. Nedel, "Empirical properties of multilingual phone-to-word transduction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4445–4448.
[51] G. Sreeram and R. Sinha, "Exploiting Parts-of-Speech for improved textual modeling of code-switching data," in *Proc. Twenty 4th Nat. Conf. Commun. (NCC)*, Feb. 2018, pp. 1–6.
[52] S. Ganji and R. Sinha, "A novel approach for effective recognition of the code-switched data on monolingual language model," in *Proc. Interspeech*, Sep. 2018, pp. 1953–1957.
[53] K. Kirchhoff, J. Bilmes, and K. Duh, "Factored language models tutorial," Dept. Elect. Eng., Univ. Washington, Seattle, WA, USA, Tech. Rep. UWEETR-2007-0003, 2007.
[54] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Cernocky, "RNNLM–Recurrent neural network language modeling toolkit," in *Proc. ASRU Workshop*, 2011, pp. 196–201.
[55] J. Gebhardt, "Speech recognition on English-Mandarin code-switching data using factored language models," M.S. thesis, Dept. Inform., Karlsruhe Institute of Technology, 2011.
[56] H. Adel, N. T. Vu, F. Kraus, T. Schlippe, H. Li, and T. Schultz, "Recurrent neural network language modeling for code switching conversational speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8411–8415.
[57] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[58] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5528–5531.
[59] V. Renkens. *Nabu: An End-to-End Speech Recognition Toolkit*. Accessed: Sep. 20, 2019. [Online]. Available: https://vrenkens.github.io/nabu/

GANJI SREERAM received the B.Tech. degree in electronics and communication engineering from the J. B. Institute of Engineering and Technology, Hyderabad, India, in 2012, and the M.Tech. degree in signal processing from IIT Guwahati, Guwahati, India, in 2015, where he is currently pursuing the Ph.D. with the Department of Electronics and Electrical Engineering. He has published research articles in various reputed journals and conferences. His current research interests include signal processing, machine learning, code-switching, speech recognition, and language modeling.

ROHIT SINHA (Member, IEEE) received the B.E. degree in electronics engineering from the University of Gorakhpur, Gorakhpur, India, in 1990, and the M.Tech. and Ph.D. degrees in electrical engineering from IIT Kanpur, Kanpur, India, in 1999 and 2005, respectively. In 1994, he began his career as a Lecturer with the Department of Electronics and Communication Engineering, Madan Mohan Malaviya Engineering College (currently MMM Technical University), Gorakhpur. From 2004 to 2006, he was a Postdoctoral Researcher with the Machine Intelligence Laboratory, Cambridge University, Cambridge, U.K. Since 2006, he has been with IIT Guwahati, India, where he is currently a Full Professor with the Department of Electronics and Electrical Engineering. Since November 2017, he has also been chairing the Head of the Department position. He has published more than 100 research articles in reputed journals and conferences. He has supervised several research students and involved with some sponsored research projects. His research interests include machine learning, pattern recognition, automatic speech recognition, and language modeling, speaker and language identification, and noise-robust speech and image processing.