Acoustic and Language Modeling for Children's Read Speech Assessment

Hitesh Tulsiani, Prakhar Swarup, Preeti Rao
Department of Electrical Engineering,
Indian Institute of Technology Bombay, India
{hitesh26,prkhr,prao}@ee.iitb.ac.in

***Abstract*—Automatic speech recognition can be used to evaluate the accuracy of read speech and thus serve a valuable role in literacy development by providing the needed feedback on reading skills in the absence of qualified teachers. Given the known limitations of ASR in the face of insufficient task-specific training data, the selection of acoustic and language modeling strategies can play a crucial role in achieving acceptable performance in the task. We consider the problem of detecting mispronunciations in read-aloud stories in English (as a second language) by children in the age group 10-14. Multiple available datasets that separately capture the characteristics of children's speech and Indian accented English are used to train and adapt the acoustic models. A knowledge of the text together with the prediction of mispronunciation errors helps to define an effective language model. We present mispronunciation detection performance on a small test dataset of field recordings and discuss implications for further work.**

I. Introduction

It is well known that in India's large rural population, millions of children complete primary school every year without achieving even basic reading standards[1]. Since reading competence enhances overall learning by enabling the child to self-learn various subject material from the vast available text resources, the importance of imparting reading skills in early school cannot be overstated. Technology holds the promise of scalable solutions to alleviate the literacy problem.

It is the goal of the present work to consider scalable technology solutions that facilitate oral reading assessment in situations where access to language teachers is limited. We choose the specific context of second language English which is a curriculum subject across schools in rural India where the medium of instruction is primarily the regional language. We describe our preliminary efforts to assess the recorded read speech by children in English using an Automatic Speech Recognition (ASR) system.

The task of automatic assessment of reading ability is to assign an overall rating suggestive of reading ability. Ideally, overall rating should correlate with human ratings and should take into account multiple cues (such as pronunciations errors, prosody, fluency, and speech rate of the child) that human evaluators may use to assign the rating[2]. The task of automatic assessment of reading ability can be broadly classified into two categories:

- Suprasegmental-level assessment: Includes assessing children's reading ability based on prosody, fluency, and speech rate. It has been shown that the prosody of the student's reading, i.e. its phrasing and intonation, is an important predictor of comprehension[3].
- Word-level assessment: Here, assessment is concerned with detecting word-level mispronunciations. It is important to note here that mispronunciation includes the following errors:
  1) Substitution: Incorrectly pronouncing a word
  2) Omission: Not uttering a word
  3) Disfluency: Includes (i) Hesitation: Partially pronouncing a word (ii) Sound-out: Pronouncing a word as sequence of distinct rather than continuous sounds (iii) Mumbling or unintelligible speech.

In our present work, we restrict ourselves to the word-level assessment of read speech in the English language by rural Indian children in the age group of 10-14 years (grade 4-8). However, recognition of children's speech is in itself a very challenging task primarily because the spectral and temporal characteristics of children's speech are highly influenced by growth and other developmental changes [4]. In order to tackle these problems, various techniques such as Vocal Tract Length Normalization (VTLN) and feature Space Maximum Likelihood Linear Regression (fMLLR) have been proposed for GMM-HMM ASR system[5][6]. Over and above the difficulties involved in recognizing children's speech, any mismatch between training data and testing data degrades the performance of ASR. For our task where we try to assess the speech read by children in the English language, we do not have any readily available speech corpus in English spoken by Indian adults or children. Further, the performance of a word mispronunciation detection system is also heavily dependent on the Language Model (LM) used for speech recognition. Two major studies on assessment of children read speech, TBALL[7] and LISTEN[8], tried to explicitly model the mispronunciations by providing parallel paths to account for substitutions, omissions and repetitions in Finite State Grammar (FSG) LM. In this paper, we address two questions

1) How to make use of possibly mismatched datasets to come up with good acoustic models for the task.
2) How to design the LM to detect mispronunciations efficiently.

While ASR has been used previously in the objective assessment of language skills of children, the present work is targeted towards the more challenging scenario of continuous speech (rather than isolated words as in the extensive work by Alwan et al.[7]) comprising read-aloud stories. This makes the task of utterance segmentation very crucial, and a good LM that is task-specific but also flexible and not overly constrained is required.

978-1-5090-5356-8/17/$31.00 © 2017 IEEE

In the next section, we describe the data collection and annotation details and follow it up with discussion on Acoustic and Language Modeling for mispronunciation detection task in sections III IV.

II. Dataset

High quality transcribed data obtained in the application scenario is important for the successful deployment of the ASR engine. The following subsections briefly describe our speech data collection and annotation methodologies.

A. Tablet application for data collection

Mobile tablets provide for a low cost, portable medium that can be easily handled by children. The screen space of a tablet is sufficient for the convenient display of text and pictures in story reading. For the proposed work, we adopt the SensiBol Reading Tutor app (2016)[9] for Android tablets due to the availability of customization for classroom use with multiple separate child accounts. The app allows a child to listen to a narrator reading out the story in a listening mode. The child can then use the record mode while reading aloud himself/herself. The stored recording synchronized with the video is available on the tablet for listening which encourages self-assessment and more practice. The SensiBol RT app also provides backend support where every registered child's audio recordings, together with metadata information such as the child's name, story name, date and time of recording can be archived. All recordings are made at 16 kHz with a headset mic to minimize background noise which can be very detrimental for ASR. The target text content is a selection of stories from BookBox[10], a readily available rich resource of illustrated text designed for child readers. We have, so far, been able to collect a total of 1000 recordings spanning 18 stories from 70 students in grades 4-8 of a tribal school in Aine, Mumbai where tablet based story reading is a scheduled and supervised activity conducted in the school hours as part of the Learn English Through Stories (LETS) project[11].

B. Annotation

Data annotation is carried out using a web-based ratings panel. The panel displays audio at the sentence level together with expected story text. This is obtained by segmenting the full story recording based on a combination of information from the video timings combined with the detection of long silences in the audio. The sentence-level audio is then labelled with reference to narrator audio.

Annotation is done at 3 levels: Word-level, Sentence-level, Story-level. We have labels for:

1) Mispronunciations : This type of marking is done at the word level. Each word is categorised into one of the five categories: 'correct', 'substituted', 'incorrect', 'missed', or 'disfluency before word'.
2) Noisy/Clean : This marking is done at all the 3 levels.

A word is marked as 'substituted' when the child has not pronounced the word correctly, but pronunciation is intelligible whereas a word is marked as 'incorrect' when the pronunciation is unintelligible. Here intelligible means that pronunciation is either another valid English word or decipherable enough to get its phone sequence. For 'substituted' words we also write down substitutions. 'Correct' label is assigned when the child has uttered an acceptable pronunciation of the word, and 'missed' label is assigned when the child has skipped the word. 'Disfluency before word' includes hesitation, sound-out and mumbling. If there is some overlapping noise with the word/sentence it is marked as 'noisy'.

C. Evaluation Dataset

For evaluation, we used a subset of the LETS[11] speech data, collected by us (as explained in section II-A). The details of the dataset used are described in the table I. We would like to mention the terminology used in rest of the paper:

- Sentence: It refers to a line in the story text (one per video frame; segmented using combination of video frame timings and detection of long pause). There can only be as many sentences as the number of unique lines in stories.
- Utterance: Utterance is a read out sentence. There can be many utterances of a sentence by different children, or at different times.

Dataset	
# Students	27
# Stories	5
# Sentence	115
# Utterance	961
# Duration (min)	43
# Sentence >10 utterance	59

TABLE I: Evaluation dataset statistics

We considered 961 utterances by 27 speakers which were noise free and only contained substituted, missed or correctly uttered words. Among all the mispronounced words, 53% were substitutions and 47% were omissions. Also, 40% of substituted words were unpredictable i.e. Out-of-Vocabulary (OOV) words. It is important to note here that all the speakers in the evaluation dataset hail from the same region and therefore have the same native language and dialect.

III. Acoustic Modeling

In order to build robust acoustic models for our automatic reading assessment task, we are faced with the following challenges:

- Insufficient application-specific training data: Ideally, ASR training and testing should be done with speech recorded under similar conditions. However, since data collection and transcription for our application is currently underway, we do not have enough labelled speech data to train acoustic models. To mitigate the effect of insufficient application-specific training data, we train acoustic models using possibly mismatched data and adapt with limited application-specific data. Maximum-aposteriori (MAP)[12] adaptation for GMM-HMM ASR systems has been used in the past under similar situations[13].
- Inherent difficulty in children's speech recognition: Recognition of children's speech is difficult due to multiple reasons[4]. Non-nativeness of the target population (i.e. children with Marathi as native language) in our case also adds to the challenge of robust acoustic modeling. To alleviate one major source of variability in children's speech (viz., physiological characteristics) we use Vocal Tract Length Normalization (VTLN) technique[6].

- Presence of noise and other disturbances: We attempt to minimize the effect of noise by using a close-talking microphone for recording stories. A significant amount of noise can nevertheless creep into our recorded data due to the school recording environment. Various methods such as filler models and multi-condition training have been studied in the past to tackle noise. However, in the present work we restrict ourselves to noise free test data.

Children using our reading assessment application are expected to be non-native speakers of English. Therefore we expect substitutions of unfamiliar English language phones with acoustically similar native language (i.e. Marathi) phones. Keeping this in mind we built a application-specific phone set (47 non silence phones and 1 silence filler phone) by combining phones from English and Hindi. We selected Hindi due to the availability of a transcribed database and its similarity to Marathi in the phone inventory. We next describe various training and adaptation datasets available and the strategy to combine these datasets in best possible way.

A. Training and Adaptation datasets

In the table II (below), statistics of various datasets available to us for training and adaptation is shown.

	TIFR[14]	**GIE**	**CSLU Kids'[15]***
Duration	1.3 h	43 min	14.8 h
Speakers	100	35	493
Utterances per speaker	10	20	45
Population	Adults	Adults	Children (Grades: 4-8)
Target text	Sentences	Sentences	Words and Sentences
Type of speech	Scripted	Scripted	Scripted and Spontaneous
Language/Dialect	Hindi	Indian English	American English

TABLE II: Details of training and adaptation speech corpora. *Statistics of a portion of CSLU Kids' dataset (relevant to the targeted age group) is mentioned.

The CSLU Kids' speech corpus appears to be the most suited training set for our task mainly because of two reasons, the first being similar target populations (children in both cases) and the second being its large size as compared to other datasets. However, the CSLU Kids' corpus is American English speech. The TIFR Hindi dataset is used for modeling the non-English phones which are not present in the CSLU Kids' corpus. The GIE corpus, recorded at our lab, is useful for incorporating phonetic realizations of English phones by Indian speakers uttering TIMIT[16] prompts. Ideally, we would want to train non-English phones and adapt using children's data but due to unavailability of such a dataset we resort to TIFR and GIE adult's datasets. However, the question which remains to be answered is how to incorporate the information present in TIFR and GIE datasets into an existing model trained on CSLU Kids' corpus. This is determined by experimenting with various training and adaptation methods, the results of which are described in the following section.

B. Experiments and Results

The Kaldi[17] speech recognition toolkit was used for all our experiments. We chose Phone Error Rate (PER) as the evaluation metric for assessing quality of acoustic models. This ensures that the language model part of an ASR system plays a negligible role in the evaluation of this part of the work. All results reported in the following sections use the combined phone set described in section III. Among the 48 phones, we have a separate filler model called 'SIL' for pauses. This represents a generic 3-state HMM which is trained on silences present at the start and end of each utterance in the training data. Another important aspect is the dictionary we use for converting word-level transcription into phone sequences for PER calculation. For American English, we use the CMU dictionary[18] as it is freely available and has been extensively used in literature. However, for Indian English we maintain a separate dictionary which contains acceptable Indian English pronunciations. This dictionary has been manually prepared by us keeping in mind the pronunciations we expect from a good English speaker from any part of India. We use phone-bigram LM, trained on ground-truth transcriptions of the test utterances.

PERs observed under different training/adaptation conditions are summarized in tables III, IV, V. Some important implementation details related to these experiments are:

1) Monophone: Context-independent models trained with 39-dimensional MFCC features.
2) Triphone: Context-dependent models (1000 senones and 16 Gaussians per senone) trained with 40-dimensional LDA features (+/- 3 context window spliced MFCC features reduced to 40 dimensional using LDA).
3) Triphone - VTLN: Context-dependent models trained with Vocal Tract Length Normalized (VTLN) features. This was done to normalize the effect of presence of both adult and child population in our training and adaptation datasets. At this stage due to insufficient speaker-specific data we do not try out Speaker Adaptive Training (SAT) but limit ourselves to VTLN adaptation only.
4) Due to the high level of acoustic mismatch between the available datasets, we have adapted all the Gaussian parameters (mean vectors, covariance matrices and mixture weights) through MAP adaptation.
5) N-fold adaptation refers to a typical leave-one-out adaptation strategy in which we assume 1 speaker as a test speaker and use all the data from the remaining N-1 speakers for MAP adaptation. In our case N = 27 i.e we do 27-fold MAP adaptation.

Train Set	**Monophone PER(%)**	**Triphone PER(%)**	**Triphone-VTLN PER(%)**
CSLU	74.93	71.30	69.80
CSLU + TIFR	72.00	63.94	63.41
CSLU + GIE	74.12	65.93	64.29

TABLE III: PER(%) using various train set combinations on LETS test set

Train Set	**MAP Adaptation Set**	**Triphone PER(%)**	**Triphone-VTLN PER(%)**
CSLU	**TIFR**	65.48	65.68
CSLU	**GIE**	65.89	65.19
CSLU + TIFR	**GIE**	63.17	61.17
CSLU + GIE	**TIFR**	64.19	61.97

TABLE IV: PER(%) with various combinations of train and adaptation set on LETS test set

C. Observations

1) Incorporating TIFR/GIE with CSLU training dataset helps in improving the phone recognition accuracy. This was expected as discussed in section III-A.

Train Set	MAP Adaptation Set	N-fold MAP Adaptation Set	Triphone PER(%)	Triphone-VTLN PER(%)
CSLU + TIFR	**GIE**	**LETS**	49.27	51.99

TABLE V: PER(%) on LETS test set with TIFR + CSLU train set and GIE + LETS adaptation set

2) MAP adaptation using TIFR/GIE datasets provides improvement over acoustic models trained on CSLU corpus only. Considering that the TIFR Hindi dataset provides the additional phones, it should be added to the training set. This training data combined with GIE data for MAP adaptation gives the lowest phone error rate in Table IV.
3) N-fold adaptation results in a marked improvement in PER as shown in Table V. This implies that bringing in our application-specific data through MAP adaptation is beneficial for the performance of a GMM acoustic model trained on mismatched data. In the present scenario, we also benefit from the same LETS text in both adaptation and test data.
4) As expected, acoustic models trained with VTLN feature seem to outperform models trained without VTLN feature normalization (Tables III,IV). However, same improvement is not observed when task-specific LETS data is used for N-fold MAP adaptation (Table V). This counter intuitive result needs further investigation.

We can thus conclude from these experiments that using TIFR dataset in training along with CSLU Kids' data models the non-native phones not in English. MAP adaptation of these models with GIE enables us to capture phonetic realizations of English phones by Indian speakers whereas LETS data helps to bridge the gap between train and test conditions.

IV. Language Modeling

In the context of read-aloud speech, the language model essentially aligns the utterance with the known canonical text. Correct alignments can help us evaluate the prosody of the utterance with reference to the expected prosody (related to phrasing, sentence-ending and prominent words) apart from identifying specific word-level pronunciation errors. A well-designed LM should take into account following observed phenomena in children's read-aloud speech with regard to the given text:

1) A given word can either be uttered or omitted (missed).
2) A given word can be substituted by another valid English word or an invalid word (i.e. an Out-of-Vocabulary word) or by unintelligible speech.
3) Possibility of the presence of silence or noise between two words.
4) Possibility of disfluency (hesitation, sound-out or mumbling) before any word in a sentence.

For the present work, we restrict our attention to utterances that don't have noise or disfluencies as per our transcriptions. We use OpenFST toolkit[19] to build Finite State Transducer based LM.

A. Basic Framework

Our approach is similar to that adopted for TBALL[7]. Since we aim to detect word-level mispronunciations from sentence-level utterances (of stories), we build a Finite State Grammar (FSG) LM for each sentence in the story text. The idea is best understood by first analyzing word-level LM.

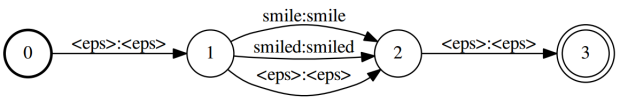

Fig. 1: LM for the word 'smiled' with parallel paths for detecting mispronunciations.

The image above (Fig 1) shows a word-level LM for recognizing the target word 'smiled'. Here the path annotated as 'smiled' represents an acceptable pronunciation whereas the path annotated as 'smile' allows for an expected substitution. There may be more than one acceptable pronunciation of a word in the lexicon used for ASR. Also, '<eps>' annotated path allows for possible omission. It is to be noted that this LM will allow the unexpected substitutions of word 'smiled' to go undetected. To address this, we use a phone loop in parallel as explained in section IV-D.

Once we get the word-level LM, we can easily build sentence-level LM by concatenating word-level LMs for every word in the sentence text. This LM, in theory, allows us to detect mispronunciations (except for unpredictable substitutions). One of the drawbacks of such an LM is that it gives equal probability to all the parallel paths i.e. it assumes a child is as likely to pronounce the word correctly as he is to mispronounce it. But from the observations in our dataset, we found that a child is more likely to mispronounce a word in his initial reading attempts and correctly pronounce the words after a sufficient number of readings of the same story. We take care of this discrepancy by associating with each path a probability value.

LM for the sentence 'The moon smiled' is shown in Fig 2. Here numbers after '/' represent probabilities and self-loop of 'SIL' after every word indicates that there can be silence of arbitrary duration between two words. Allowing silence between two words can help us obtain accurate word-level alignment. There are two challenges in building this LM:

- How to obtain the probability of each path?
- How to determine expected substitutions of every word?

The answer to both the questions lies in DATA collected from field recordings.

B. Obtaining Expected Substitutions

To obtain expected substitutions of every word, we make use of the information about word substitutions gathered during data annotation. As explained in section II-B, if a child mispronounces a word we mark it as substituted and also note down the uttered pronunciation. Once we get all the observed substitutions for every word in the dataset we manually filter out, for every word, those substitutions that are not phonetically close to the target word. The remaining predictable substitutions, of the target word, are added to our lexicon (dictionary building) and as a parallel path in LM.

C. Learning Probabilities

To assign probabilities to various parallel paths, we compare ground truth transcription with the canonical transcription (text that was supposed to be read) to determine how many times

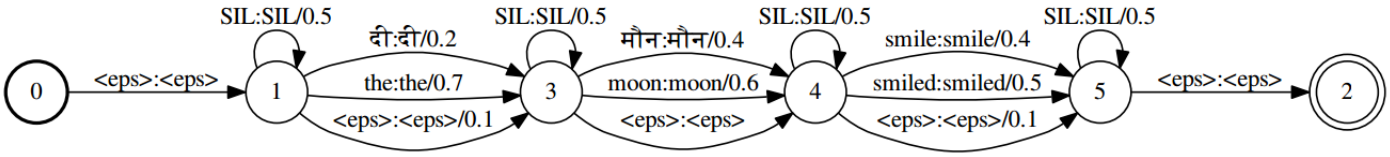

Fig. 2: LM with different parallel path probabilities for the sentence 'The moon smiled'. Number after '/' represents probability of the path.

a particular target word was omitted, correctly pronounced, or substituted with another word. For example, if the word 'cap' has occurred 20 times in canonical text and has been correctly pronounced 13 times, omitted 3 times, and substituted as 'caps' 4 times then the probabilities of various parallel paths of target word 'cap' would be:

- cap : 13/20
- <eps>: 3/20
- caps : 4/20

In our current work we explore two different ways to learn probabilities from data:

1) Entire data: Here, for learning the probabilities, we consider all the occurrences of the target word in all the utterances without regard to sentence i.e. we do not take into account the context in which target word has occurred in a sentence.
2) Same sentence: Here, to learn probabilities, we consider only those occurrences of target word that have the same context. To be precise, we learn the probabilities of various parallel paths only from utterances of the same story sentence. The idea here is that the child may make pronunciation errors depending on the context in which the word occurs and learning the probabilities from the utterances of the same sentence will help us capture context specific word errors accurately.

D. LM with Out-of-Vocabulary (OOV) modeling

To account for unexpected substitutions of a word, we used a phone loop in parallel with paths that allowed for correct, expected substitutions and omission of a word. A topological constraint of allowing a minimum of two phones and a maximum of five phones was imposed on phone loop path to ensure that OOV words are between two to five phones in length[20]. This constraint was imposed after observing that the average length of OOV words in our dataset was 4 phones.

E. Experiments and Results

In this section, we describe the experiments carried out to validate our LM. Since our task is mispronunciation detection, we report the results in terms of detection rate/recall (DR) and false alarm rate (FAR) of mispronunciations ([7][8]). We also use Word Error Rate(WER) as an evaluation metric. WER accounts for insertions, substitutions, and deletions and we have been strict enough not to consider two different but phonetically similar mispronunciations of a word as identical.

For all our experiments, we use GMM-HMM acoustic models trained on CSLU kids speech and TIFR datasets and adapted with GIE and LETS data (N-fold). This configuration was found to be best in terms of PER in section III-B. Also, we learn the probabilities of various parallel paths in LM (as explained in section IV-C) in leave-one-out (LOO) cross-validation mode i.e. we exclude the utterance under consideration when learning the probabilities.

We report results on two subsets of LETS test data (discussed in section II-C):

1) Sentence with more than 10 utterances in data (Set-A): This subset of 661 utterances across 59 sentences was chosen to ensure that the probabilities learned using 'Same sentence' technique are statistically significant.
2) All sentences in the dataset (Set-B): This set comprises of 961 utterances across all 115 sentences.

WER results using the two above sets for learning probabilities of various parallel paths using 'Same sentence' and 'Entire data' techniques (without OOV modeling) are shown in the table VI. The comparison of these two techniques of learning against assigning equal probability to all the paths (referred as 'Equiprobable' from here on) and Word bigram LM (learned from annotated text in LOO CV mode) is also shown. We also report our preliminary WERs obtained using LM with OOV modeling where probabilities of parallel paths are learned using 'Entire data' technique. This is referred to as 'Entire data - OOV' in table VI.

	WER(%)	
	Set-A	**Set-B**
Equiprobable	29.21	29.30
Entire data	24.00	23.51
Same sentence	24.15	24.05
Word bigram	30.54	32.58
Entire data - OOV	31.44	31.56

TABLE VI: WER comparison using 3 methods of assigning probability (LM without OOV modeling): (i) 'Equiprobable' (ii) 'Entire data' (iii) 'Same sentence' and 'Word bigram' and 'Entire data - OOV' (LM with OOV modeling)

Detection rate (DR) and False alarm rate (FAR) of mispronunciations using various techniques to learn probabilities in FSG type of LM are shown in table VII. We do not compute DR and FAR on Set-B using 'Same Sentence' technique since the probabilities learned are not statistically reliable. It is important to note that we do not differentiate between two mispronunciations of a word for computing detection and false alarm rate of substitutions.

F. Observations

1) As can be seen from the table VI, WER obtained using 'Entire data' and 'Same sentence' are almost equal. This shows that learning the probabilities taking word context into account ('Same sentence') doesn't provide any additional gains especially since it is a smaller size data. Also, 'Same sentence' and 'Entire data' techniques perform significantly better than 'Equiprobable' and 'Word bigram' which tells us that learning probability from data in FSG type of LM is beneficial. WER obtained using 'Entire data - OOV' is significantly higher than 'Entire data' indicating that many of the correctly uttered

	Set-A						Set-B					
	Mispronunciations (Substitutions and Omissions)		Substitutions		Omissions		Mispronunciations (Substitutions and Omissions)		Substitutions		Omissions	
	DR(%)	FAR(%)	DR(%)	FAR(%)	DR(%)	FAR(%)	DR(%)	FAR(%)	DR(%)	FAR(%)	DR(%)	FAR(%)
Equiprobable	73.49	20.47	58.09	13.86	82.64	4.11	74.09	20.67	57.21	13.80	82.17	4.21
Entire data	59.25	11.16	41.72	6.61	74.88	3.11	60.37	10.52	41.35	6.21	76.04	2.86
Same sentence	56.81	10.48	41.18	6.12	70.54	3.00	-	-	-	-	-	-
Entire data - OOV	77.59	26.11	65.64	19.04	70.09	4.44	78.46	27.01	63.76	19.30	70.61	4.89

TABLE VII: Detection rate and False alarm rate using various techniques (with and without OOV modeling in LM) to assign probabilities: (i) 'Equiprobable' (ii) 'Entire data' (iii) 'Same sentence' (iv) 'Entire data - OOV'

words are classified as mispronounced when LM with OOV modeling is used.

2) Observing the DR and FAR of substitutions (table VII), it is clear that none of the techniques, for LM without OOV modeling, can accurately detect substitutions. This is because we did not account for unforeseeable substitutions (treated as OOV) of the target word (approximately 40% of substitutions) in LM. For the case where we accounted for OOV in LM, we get a high DR for substitutions as well as overall mispronunciations but at the cost of high FAR.
3) Results in table VII show that we can accurately detect omissions (high DR and low FAR) for all the techniques. This will ensure that we get accurate word-level alignments, which can be used for prosody analysis.

V. Conclusion

In the current work, we focused on building robust acoustic models from possibly mismatched datasets and designing LM to detect mispronunciations effectively for the task of automatic reading assessment. We showed that incorporating Hindi dataset in training along with American English dataset helped us to capture nonnative phones not in English. Also, adaptation of models with Indian English dataset allowed us to capture phonetic realizations of English phones by Indian speakers. In addition, adaptation with the task-specific dataset helped to bridge the gap between train and test conditions. Apart from building robust acoustic models, we explicitly modeled mispronunciations in Language Model. Two different ways to learn probabilities of parallel paths in FSG type of LM were explored. Both the proposed techniques performed equally good, in terms of detection rate and false alarm rate of mispronunciations, and had significantly lower WER as compared to equal probability to all the parallel paths. We observed that our LM was able to detect omissions efficiently but not substitutions. Also, our preliminary efforts to model OOV in LM using a phone loop with topological constraints gave us better detection rate of substitutions at the cost of increased false alarm rate.

In future, we would like to thoroughly investigate and improve upon the shortcomings of our LM to detect OOVs. We would also like to deal with presence of noise in our data by using filler models and better front end processing of utterance.

Acknowledgements

We thank Prof. Alka Hingorani and her team at IDC, I.I.T. Bombay for initiating the LETS project and organizing the field study. We would also like to thank Ankita Pasad and Kamini Sabu for their valuable feedback at various stages of the current work.

References

[1] "ASER: The Annual Status of Education Report (rural)," http://img.asercentre.org/docs/Publications/ASER%20Reports/ASER_2012/fullaser2012report.pdf, ASER Centre, 2012.

[2] P. Rao, P. Swarup, A. Pasad, H. Tulsiani, and G. Das, "Automatic assessment of reading with speech recognition technology," in *Proceedings of the 24th International Conference on Computers in Education (to appear)*, Mumbai, India, 2016.

[3] J. Miller and P. Schwanenflugel, "A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children," *Reading Research Quarterly*, vol. 43, no. 4, pp. 336–354, 2008.

[4] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of Acoustic Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[5] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[6] L. Lee and R. Rose, "Speaker normalisation using efficient frequency warping procedures," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, USA, 1996.

[7] P. Black, J. Tepperman, and S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1015–1028, 2011.

[8] J. Mostow, S. Roth, A. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proceedings of the National Conference on Artificial Intelligence*, Washington, USA, 1994.

[9] "Sensibol reading tutor app (2016)," http://sensibol.com/readingtutor.html, SensiBol Audio Technologies Pvt. Ltd.

[10] "Bookbox: A book for every child in her language (2016)," www.bookbox.com.

[11] "LETS : Learn English Through Stories (2016)," http://www.tatacentre.iitb.ac.in/15mobitech.php, Tata Centre for Design and Technology at IIT Bombay.

[12] C. Lee and J. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minnesota, USA, 1993.

[13] R. Bippus, A. Fischer, and V. Stahl, "Domain adaptation for robust automatic speech recognition in car environments," in *Proceedings of EUROSPEECH*, Budapest, Hungary, 1999.

[14] K. Samudravijaya, P. Rao, and S. Agrawal, "Hindi speech database," in *Proceedings of International Conference on Spoken Language Processing*, Beijing, China, 2000.

[15] K. Shobaki, J. P. Hosom, and R. Cole, "CSLU Kids' speech version 1.1 LDC 2007S18: Web download," 2007.

[16] J. Garofolo, "TIMIT Acoustic-phonetic continous speech corpus LDC 93S1: Web download," 1993.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, Hawaii, USA, 2011.

[18] "The CMU Pronouncing Dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, Carnegie Mellon University.

[19] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *International Conference on Implementation and Application of Automata*, Prague, Czech Republic, 2007.

[20] I. Bazzi, "Modelling out-of-vocabulary words for robust speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.