Acoustic Models for Speech Recognition in Children's Reading Miscue Detection

MTP Stage 2 Report

Submitted in partial fulfillment of the requirements for

Master of Technology

by

Shreeharsha B S

(18307R002 EE1)

Under the guidance of

Prof. Preeti Rao

Department of Electrical Engineering

Indian Institute of Technology Bombay

December 2020

Declaration

I declare that this written submission represents my ideas in my own words. I have adequately cited and referenced the original sources where necessary. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Shreeharsha B S
(18307R002)
Department of Electrical Engineering
IIT Bombay

Date: June, 2021

Acknowledgement

I express gratitude towards my guide, Prof. Preeti Rao, for her patience and guidance throughout the project. I also thank Nagesh Nayak and Kamini Sabu for their help and comments in this work. I would also like to thank Asha, Avinash and all other members of the DAP lab for their support. I also thank the members of the kaldi help group for answering the questions I had.

Shreeharsha B S

Abstract

Literacy is an essential skill for the betterment and prosperity of an individual. Improved literacy rates provide political, cultural, social and economic benefits. It acts as bridge by connecting people to the world and all the information available in it. In this work, the development of an automatic assessment system for evaluating the reading abilities of children (by analyzing audio recordings of literacy tests) in Hindi are examined.

The inherent difficulties involved in collecting children's data for research (which leads to a shortage of the data available) and the highly varying speaker characteristics and background noises of various types present obstacles for automatic assessment systems. These challenges are tackled in this work by building automatic speech recognition (ASR) systems using relatively more abundant adult speech datasets and adapting these systems to the specific use case. The ASR system used is the state of the art TDNN-HMM model in kaldi. Adapting the models to the target speech is done with the help of a limited amount of labelled target data using weight transfer techniques. Further, data augmentation methods are explored to both, improve the match between train and target data and to increase the adaptation data size. Preliminary experiments with denoising are also conducted. With these two techniques, improvements in the detection of reading miscues are obtained. This work also enlightens the road for further experimentation and future work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Literacy is an important measure of prosperity and also critical to the well being of an individual and his/her community. It has been found that Children at the Bottom of the Economic Pyramid (BOEP) have limited access to good education standards, particularly in the field of language learning and reading. Annual Status of Education Report (ASER) is literacy test and survey conducted by a branch of the NGO (Non-governmental organization) Pratham. It is carried out by trained volunteers/surveyors. In 2018, this survey covered nearly 5,50,000 children in different districts across India, belonging to the 3-16 year age group [5]. The results from the survey have some worrying implications are: 27.2% of class VIII students sampled were unable to read a text that is meant for a class II student. These percentages are worse than what they were 10 years ago, "In 2008, 84.8% of Class VIII students could read a text meant for Class II; by 2014, only 74.6% could do so" [5] [6]. A more complete review of literacy surveys, discussion and implications of technology use can be found in [7].

The above results concerning reading skills were obtained by manual assessments of children selected across many districts and villages. With the use of automatic assessment systems (which use speech recognition and signal processing tools) reading skills can be measured while reducing the drudgery involved in individual assessments and the much more difficult problem of providing individual feedback on the reading assessment can be resolved. Automatic speech recognition is used to convert speech to text and can be used to quickly perform an assessment of the child and his/her ability to read a reference text. These quick assessments can also act as an objective measure of learning intervention programs and their effectiveness (or lack thereof) by the changes in the reading ability of children before and after.

Building an ASR system for children's speech is a challenging task because of the high degree of variability in their speech patterns which is explained due to the 'unrefined' motor control, during speech production, of children which gets more refined with age [8]. These difficulties are not as persistent when building systems for adult speech. Modern ASR systems use statistical methods to predict the text output of a speech signal. These methods are heavily reliant on the type of speech data available and its amount. This also implies that with a decent representation of all types of children's speech much of the acoustic variability can be handled and accounted for. The crux of the ASR system that models this variability is called the acoustic model. Acoustic models, in general, find mappings and relationships between speech features and linguistic units (phones or characters). With advancements in the complexity and modeling power of state of the art acoustic models, many recent research tasks have focused on building ASR systems for children's speech.

Obtaining a representative set of children's speech data is a challenging task. Regional accents, gender and age play a big role in the variability of data. A great addition to the scarce children's speech dataset is the ASER dataset [9]. Although it does not contain the transcripts of what the child spoke, the reference texts read by the child, information about the fluency and number of mistakes committed is available through limited manual annotations. This dataset is a salient part of this work and is manually annotated for this work. It is an expanded version of the datasets used in the first stage [10].

1.1 Focus of this work

In this work, the main objective is to build an ASR system for the detection of oral reading errors using a limited amount of target speech. For this, state of the art neural network based acoustic models are examined. These acoustic models, trained on relatively easily accessible adult speech, are used as a baseline for the children's speech. For now, the data used in both the children's samples and adult samples come from the same language (Hindi). Then, techniques of modifying the adult speech data to better suit the target data and data augmentation techniques are examined. Apart from this, acoustic model adaptation methods which tune the baseline model to work better on the target speech using limited amounts of adaptation data including transfer learning techniques are also examined. A block diagram describing these is shown in Figure 1.1.

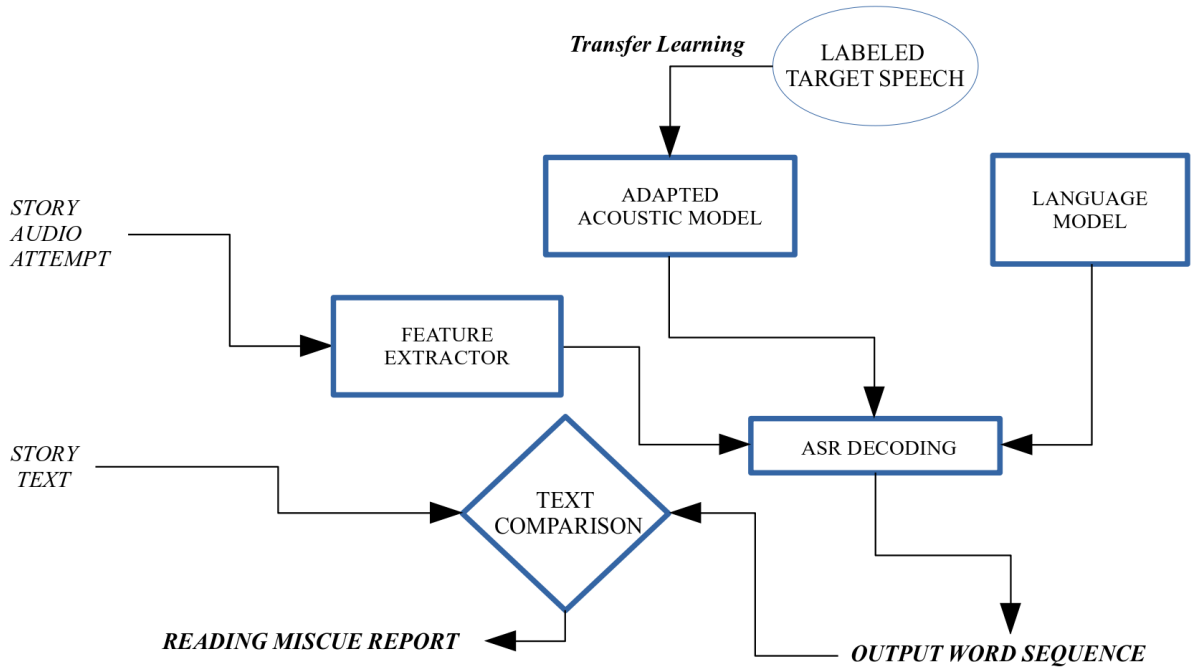

Figure 1.1: Block diagram of the proposed system

1.2 Report organization

There are seven more chapters in the rest of this report. Chapter 2 describes the training and adaptation datasets used in this work. Chapter 3 discusses the data augmentation strategies in use. Chapter 4 explains the acoustic model used in this work (kaldi TDNN model), its architecture and training methodologies. Chapter 5 details the transfer learning and adaptation techniques used in this work. Chapter 6 reports the experiments done, parameters tuned and the evaluation metrics used. Chapter 7 illustrates the results of the experiments and understandings developed. Chapter 8 concludes the work in this report along with future goals and directives.

Chapter 2

Dataset

As stated earlier, gathering and transcribing children's dataset is a challenging task. In this work a mixture of adult and children's speech data is used. Four datasets in Hindi are examined, three of which are children's speech and one consisting of adult speech:

2.1 IITM data

The IITM data set comprises of Hindi speech by volunteer adults who read predetermined segments chosen from newspapers. It was released as part of an (Automatic Speech Recognition) ASR challenge by the Speech Processing Lab of IIT Madras and was funded by the Ministry of Electronics and Information Technology (MeitY) [11]. Also provided are the corresponding transcriptions, a dictionary containing the lexicon and phone set and a recipe for building a baseline model. The text data covers a variety of genres like politics, sports, entertainment. Information about the gender, age and other characteristics of the speaker are unfortunately not available, however the region where the speaker came from is encoded as part of the name of the utterance recording (eg: cdg-Chandigarh, dli-Delhi, mum-Mumbai, pue-Pune).

All the recordings are sampled at 16 kHz sampling frequency. The recordings seem to be made in a closed room with very little noise in most of them, although a very few of the recordings of some speakers have a low frequency background hum (in the 2-3 Khz range) that is noticeable. The IITM data set consist of three parts that were released as part of an ASR challenge:
Train Set: 40 hours
Dev Set: 5 hours
Eval Set: 5 hours

The train set is used to build a baseline TDNN model used in the challenge. There are 418 unique speakers in train set. Each utterance in the dataset is actually short segments from a larger set of recordings. These segments consist of one to three sentences (sometimes single words) lasting anywhere between 0.3 seconds to 15 seconds. The majority of the utterances are greater than one second, with only 57 out of the 27131 utterance segments being less than one second. Only the train set is used in this work and its stats are summarized in Table 2.3.

2.2 ASER data

The ASER dataset [9] consists of recordings of children attempting the ASER (Annual Status of Education Report) literacy test which measures reading levels of children who are in the 6-14 years range collected.

2.2.1 2012 ASER data

The 2012 data are a collection of recordings obtained from children reading texts of varying complexity in Hindi, Marathi and English which was displayed on a custom mobile app and recorded using a headset. As more transcriptions became available, this 2012 set is an updated/expanded version of the ASER dataset used in the stage 1 report [10]. The text complexity varies from individual letters and words to paragraphs and stories being read out. A sample is shown in Figure 2.1. Along with the recordings a JSON file is also available which contains information about the standard the student is studying in, the number of mistakes made while attempting the test and whether the intended text was spoken correctly or not as estimated by a trained 'surveyor'. More information about the survey process and how volunteers are trained can be found on the ASER website [1].

The recordings are sampled at 16Khz. Some of the recordings do have foreground speech by speakers other than the child taking the test, which adversely affects automatic reading assessment systems. There is also a variety of noises (stationary and non-stationary) in the recordings ranging from mic pops and bursts to vehicles, wind, babble and other noise types.

A subset of this ASER data set, consisting of Hindi story and paragraph recordings of children from Uttar Pradesh (UP) and Rajasthan (RJ) is used in this work. Furthermore,

Figure 2.1: ASER sample survey [1]

with the help of a baseline DNN-HMM [12] the recordings are categorized, based on the WER between DNN-HMM output and the story/paragraph canonical text, and those recordings with WER less than 80% have been transcribed to be useful for the speech recognition based experiments in this work.

Furthermore, the recordings are separated into skill-based categories in order to investigate the performance of the system in two different contexts. After the transcription process, those recordings with miscue rates (WER between the transcribed output and the story/paragraph canonical text) less than or equal to 20%, called a 'Ratable' set or High proficiency recording (HPR), and those with miscue rates greater than 20% and less than 80%, called a 'Non-ratable set' or Low proficiency recording (LPR) are created and used for experimentation.

From the above two sets (HPR and LPR) three subsets, each of which have no speaker overlap with the others, are derived for the transfer learning experiments in this work:
i) Train set, used for weight transfer/adaptation of the baseline TDNN model. The training data sets are also split at the sentence level using the manual transcriptions and sentences that contain irrelevant speech (IR)(regions which have distinct and audible

speech sounds that are irrelevant to the story being read) as determined by the transcription process described in Section 2.2.3 are discarded.
ii) Validation set, used for determining decoding parameters for the test set and reporting results. It is also used as a diagnostic during weight transfer experiments to tune the retraining parameters based on validation loss.
iii) Test set, to report the results of the various experiments. Unlike the previous experiments in the first stage [10], the validation and test sets are not split at the sentence level anymore since they are not indicative of a real testing scenario.

These subsets are created, after transcription, on both ASER UP an ASER RJ sets and their duration, speaker information and other relevant characteristics are summarized in Table 2.1. Figures 2.2 and 2.3 show the distribution of the number of recordings over the stories and miscue rates. A similar representation of various proficiency levels and the stories/paragraphs spoken are present in each of the subsets.

Table 2.1: Summary of the 2012 ASER UP and RJ datasets

Dataset	# of Recordings Total (HPR, LPR)	# of Unique speakers	# of non IR sentences	Duration (min)
2012 UP train	792 (644,148)	489	4923	374
2012 UP test	502 (441,61)	298	-	230
2012 UP valid	194 (143,51)	128	-	93
2012 RJ train	789 (692,97)	518	3696	333
2012 RJ test	500 (458,42)	361	-	220
2012 RJ valid	189 (155,34)	128	-	81

2.2.2 2016 ASER data

The 2016 data is another collection of recordings, similar to the 2012 data but only in Hindi, of children from five states: Chattisgarh (CG), Jharkhand (JH), Rajasthan (RJ), Maharashtra (MH), Uttarakhand (UK) recorded in 2016. This 2016 set of recordings can be further divided into two sets: those which have no common stories with the 2012 set and those having story overlap with the 2012 sets. The no story overlap subset is also further split into validation and test subsets (about a 60:40 split) with recordings from all regions present and similar miscue distributions in both the sets as shown in Table 2.2. These sets are used only for decoding with the acoustic models trained. Two salient features about this 2016 set is that:

1. The children reading here have made more mistakes compared to the 2012 set. This

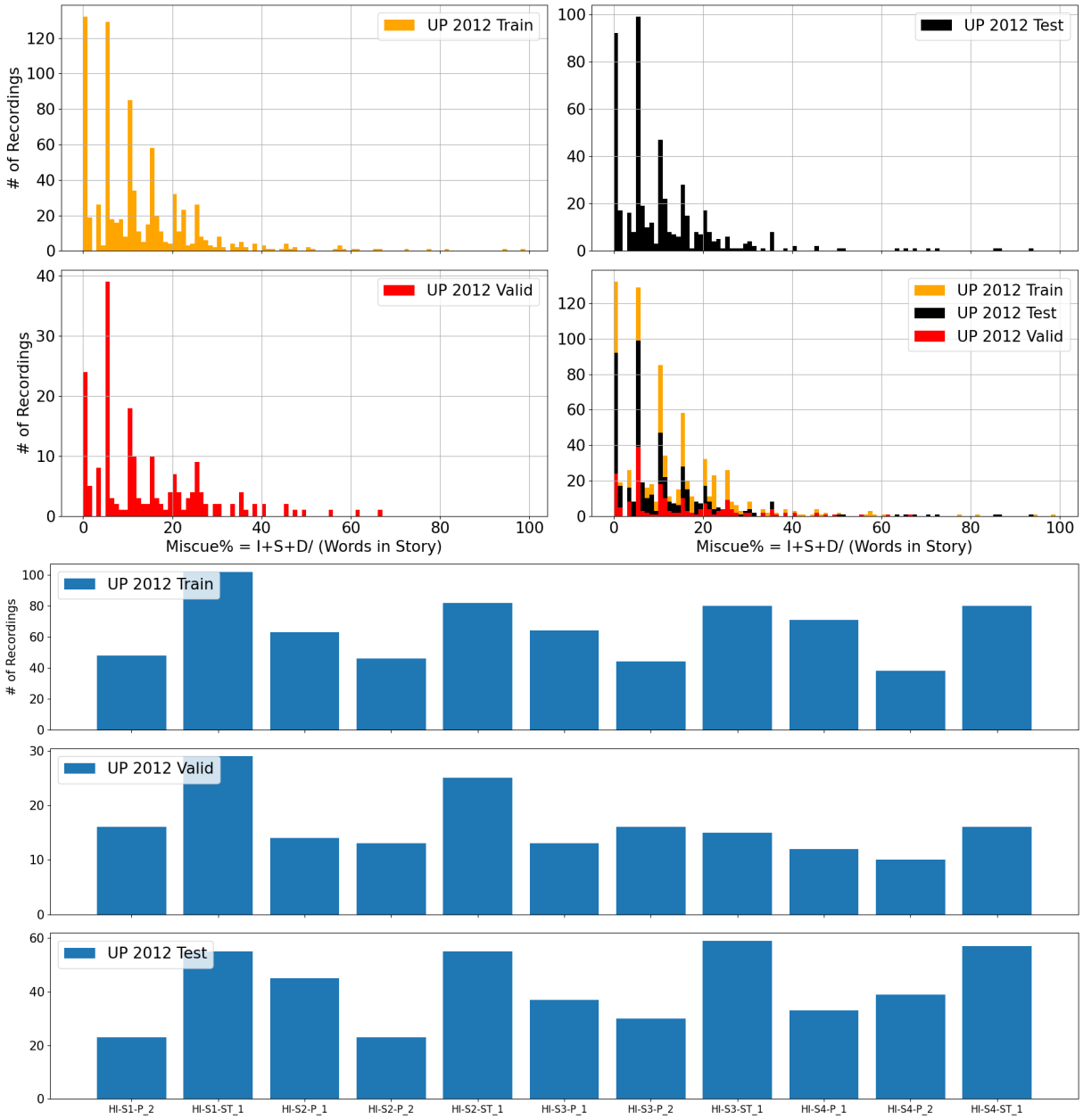

Figure 2.2: Distribution of UP recordings over miscue rate and Hindi story/paragraphs.

is evident from figures 2.4 and 2.5, which also has the 2012 UP test set in the last cell for comparison.

2. The 2016 sets are much more noisier compared to the 2016 sets.

The HPR, LPR threshold is also re-defined to be at 10% miscue rate because of the higher proportion of recordings with miscues in the 10-20% range

The following subsection goes into detail on the transcribing process in use and the other labels present in the transcription including the IR label described previously.

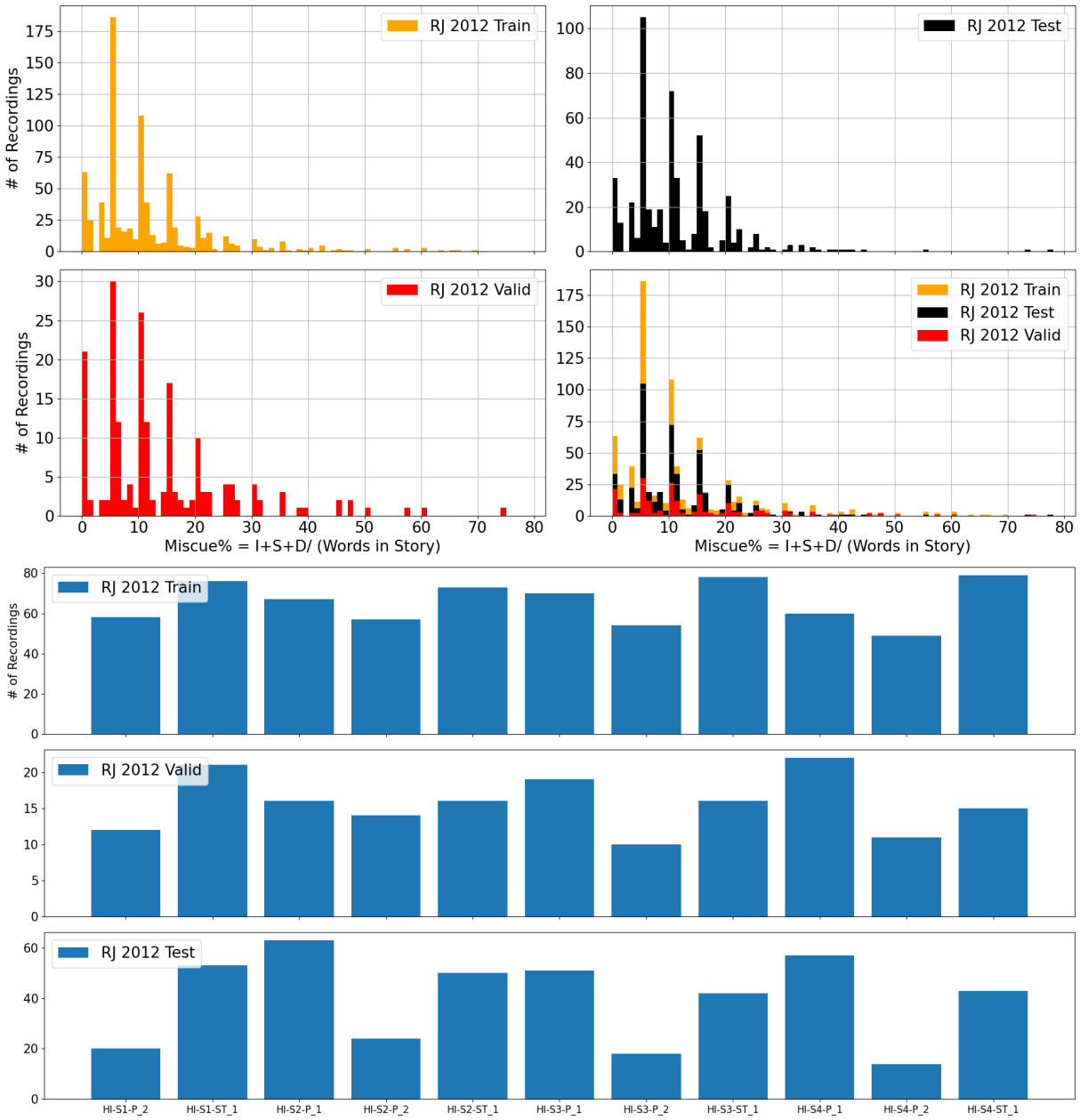

Figure 2.3: Distribution of RJ recordings over miscue rate and Hindi story/paragraphs.

Table 2.2: Summary of the 2016 ASER Hindi datasets

Dataset	# of Recordings Total (HPR, LPR)	# of Unique speakers	Duration (min)
2016 no story overlap valid	482 (182,300)	317	252
2016 no story overlap test	333 (113,220)	220	192
2016 with story overlap test	459 (143,51)	289	256

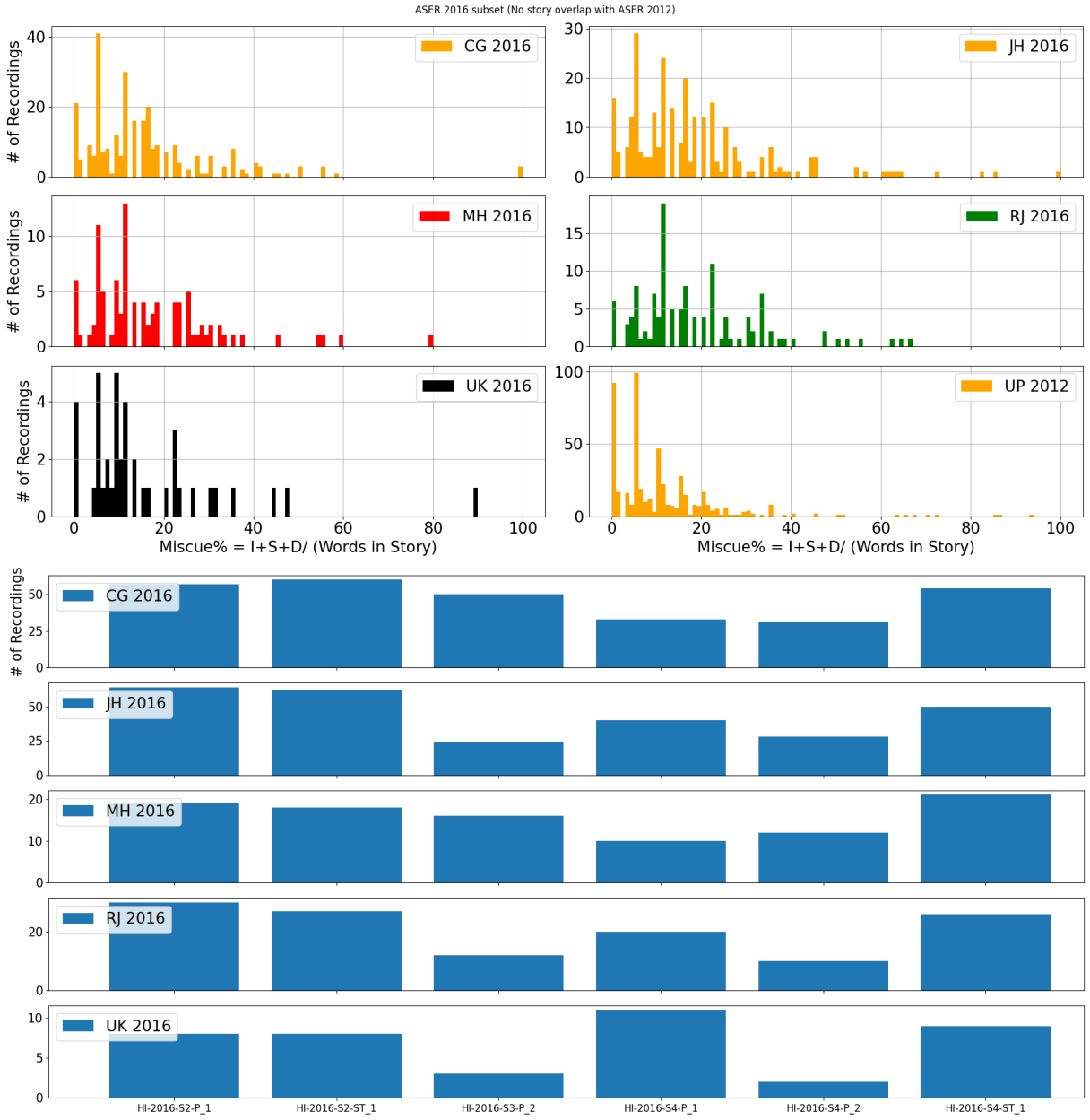

Figure 2.4: Distribution of 2016 ASER subset, which has no story overlap with 2012 set, over miscue rate and Hindi story/paragraphs.

2.2.3 Transcribing the ASER set

All audio recordings are first passed through a baseline DNN-HMM ASR [12] with a tri-gram language model trained on the specific story/paragraph text (along with a garbage model containing around 1500 common English words and all single phones) to get the decoded text output. Using the alignment of the decoded text with the canonical text, the number of miscues (insertions, deletions and substitutions) of the child are detected. This is then normalized by the number of words spoken to get the miscue rate. Using a threshold on the normalized miscues value, it is determined if a recording is Ratable (less

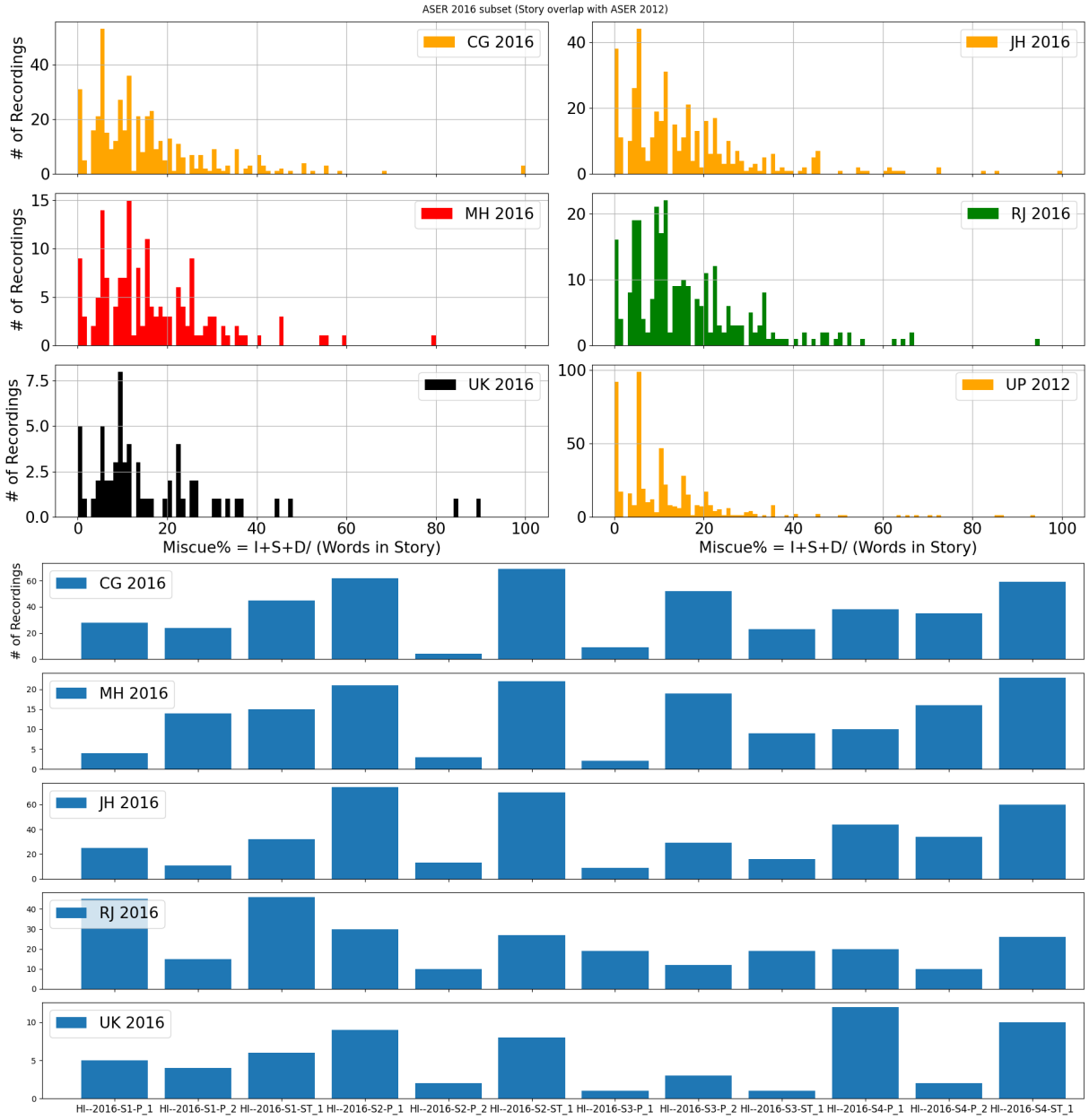

Figure 2.5: Distribution of 2016 ASER subset, which has story overlap with 2012 set, over miscue rate and Hindi story/paragraphs.

than or equal to 20%) or Transcribable (greater than 20% and less than or equal to 80%) or 'WeakReader' (greater than 80%). Recordings having miscue rates greater than 80% could also have blank recordings where nothing is spoken.

This decoded text acts as a first level automated transcript for the recording to help the transcriber. This first level transcript is saved as an Audacity label track which also contains the timestamps aligned to the transcript. The first level transcript is slightly modified before passing it to the transcriber by removing isolated phones which get decoded due to the garbage model except for the following phones which are replaced by

the tags mentioned below:

1. 1aa word in the decoded text(i.e. a single 'aa' phone) is replaced by FP (Filled pause).

2. 1s word in the decoded text (i.e. a single 's' phone) is replaced by BR (Breath Noise).

   The First level transcript label track and recording is split sentence wise based on its alignment with the canonical text. Then each line is aligned to the respective portion of its audio where the speaker started speaking the line to when the speaker stopped and is presented to the transcriber. Additionally, silences greater than 200ms are inserted as SIL labels as obtained from DNN-HMM CTM (time-marked conversation) outputs.

   The transcriber then modifies this first level transcript in the following ways in Audacity: The transcribers move the boundaries of sentences if needed. They modify the transcript using the Devanagari script in case an English word appears (because of the garbage model) or if they feel a slight modification to the decoded word is required or if they feel there is a mistake in the decoded text and a better transcription is suited. The transcriber also adds in the following labels to the transcript by listening to the audio and modifying the label track:

3. Breath (inhalations/exhalations), sniffling or 's' sound between sentences are marked as BR.

4. Other noises (ON) if appearing in isolation are tagged as ON. Common noises observed include birds, mobile, vehicles, bell, mic noise, etc. When speakers clear their throat, it is also marked as an ON label.

5. Filled pauses like 'uh', 'hmm', 'umm' etc. are marked as FP.

6. For regions of unintelligible words or indecipherable mumbling, an MB symbol is used.

7. In case a section contains child whispering a word (this has been observed when the child has difficulties decoding the word), a WH label is used. Even if the words are audible, they are not transcribed and a WH label is used.

8. Irrelevant speech usually present at the start or end of the recording is transcribed by a separate label IR. This usually happens if the facilitator gives instructions to the child before the test has commenced or after it is over. These regions are distinct and the audible speech sounds made are irrelevant to the story being read.

   In the case where some labels overlap, the dominant source of a label is used as the only label.

The transcriber also ensures that the Audacity label tracks are contiguous and that additional SIL/BR/ON tags are used at the start/end of the sentence wherever necessary. Different disfluencies like Elongations, Stalling or Hesitations are marked as an additional (HS) tag in brackets along with the actual transcript. On finishing the transcription, the audio and label tracks are exported. Additional comments like School Noise, Loud background speaker, etc. may be entered in another field.

A stage of QC (Quality control/check) follows this transcription process ensuring the labels are marked accurately and the transcripts are satisfactory. These detailed levels of transcription and tags are treated as words whose pronunciation is the SIL (silence) phone from the IITM phone set while creating the train, test and valid datasets because the IITM baseline model is not trained on these. In evaluating the experimental results and student reading errors, these tags are removed from the ground truth transcription while comparing with the decoded text. These tags are for future purposes and experiments where training separate phones for some of these tags will be quite useful for identifying and delivering specific feedback to the student speakers such as recognizing at what words hesitations (HS) commonly occurs or if the child whispers (WH) the word before pronouncing it with difficulty. It is also expected to help boost the ASR performance.

2.3 Campus School Hindi data

The Campus School (CS) Hindi data is a subset of a larger CS dataset which consists of children from the IIT Bombay campus school reading Hindi paragraph stories. They wear a headset while speaking and have the option to listen to a narrator reading the paragraph story presented to them on a tablet. The recordings are again sampled at 16Khz. The size of this dataset is very small compared to the other sets in this work (approximately 30 minutes in duration). These recordings have been transcribed from scratch unlike the

procedure described in section 2.2.3 (because it was used to build the system in [12]) and the tags used were also not as extensive.

Some recordings and their transcriptions examples, involving the different tags and labels discussed, can be found in [13].

The CS Hindi datasets used and their statistics have been summarized in Table 2.3:

Table 2.3: Summary of the IITM and CS Hindi datasets

Dataset	# of Utterances	# of Unique speakers	Duration (min)
IITM train	27131	418	2400
CS Hindi	695	11	30

The phone sets of the ASER dataset and the CS Hindi dataset are different from the IITM data phone set. A manual mapping is created between the phone sets and the ASER and CS Hindi lexicons are converted to the IITM phone set. More information can be found in Appendix A. In the next chapter, a review of recent data augmentation methods is done and methods of augmenting and modifying the IITM dataset to suit the target set scenario is presented.

Chapter 3

Data augmentation

Data augmentation is a method of applying certain ecologically valid transforms on to the training data available. The transformed data can then be used along with the original training set to increase the count of the training data available or the original training set can remain unused with only the transformed data used for training. These methods are intended to groom the model towards certain test scenarios.

VTLP (Vocal tract length perturbation) was introduced by Jaitley et al. [14] where random warp factors chosen randomly from a normal distribution were used to 'corrupt' each utterance in the training data and it resulted in a slight reduction in the Phone Error Rate (PER) on the TIMIT task [15]. In [16] two different augmentation methods and their combined effects were considered on Assamese and Zulu recordings with only ~10hrs of training data available:
i) VTLP warping of the training data done three and seven times on Assamese and Zulu respectively using randomly sampled factors in the range [0.8, 1.2].
ii) Semi-supervised training (where a big pool of unsupervised data is decoded using an existing decoder and its output is used as the transcript for further training procedures). They observed a decrease in Token Error Rate (TER) in all augmentation cases, however combining the two methods did not yield the best result for Assamese, while it did for Zulu.

SpecAugment and SpecSwap are two recent methods of Data augmentation introduced by Park et al. [17] and Song et al. [18] respectively. Both of these methods involve spectrogram modifications and are tested using end-to-end ASR models which use Transformer networks and the Listen, Attend and Spell (LAS) networks respectively. In SpecAugment [17], three augmentation techniques inspired by computer vision are employed:

1. Time warping (which deforms the spectrogram image along the time axis)
2. Time masking (where the entire spectrogram between certain time steps are made zero)
3. Frequency masking (where an entire band of frequencies are masked throughout the signal's length)

In SpecSwap, inspired by the SpecAugment techniques and previous work by Song et al. [19] (where speech features were permuted), time swapping and frequency swapping techniques are examined. This involves two non-overlapping contiguous blocks of features, along time and frequency respectively, being swapped [18]. Both SpecAugment and SpecSwap when used along with the original data gave improvements over the baseline end-to-end models.

Speed perturbation (by re-sampling the signal), Tempo perturbation (corrupting the speech rate of the signal i.e. changing the tempo of the signal while keeping the pitch and the envelope of its amplitude spectrum constant) and VTLP as data augmentation techniques were compared by Ko et al. [20]. They obtained improvements using all techniques and concluded that the speed perturbation (which both slows down and speeds up the audio at 0.9x and 1.1x the original speed) when used along with the original train set was the best technique of the three methods. Changing the pitch of the audio signal without changes to the speed or duration of the recording is called pitch perturbation. Direct signal processing methods pitch perturbation were examined in [21]. Other methods use Generative algorithms to generate synthetic examples for further increasing the training data size [22].

Noise augmentation is another popular augmentation strategy. Overlaying noise on top of the training data at various SNRs (signal to noise ratio) to create more training examples is the common method used in these scenarios. This augmentation, along with providing variety to the acoustic training data also provides robustness against noisy recordings while decoding.

In this work, the combined effects of speed perturbation, VTLP warping techniques of data augmentation are examined while training the baseline model. Since the aim of this work is to improve ASR systems for the target domain speech, the adult speech in the IITM data set is transformed in the following ways with the understanding that children tend to have higher formants than adults (which implies more energy concentration in

higher frequency bands compared to adults) as shown in Huber et al. [23] and other works [24], [25]:

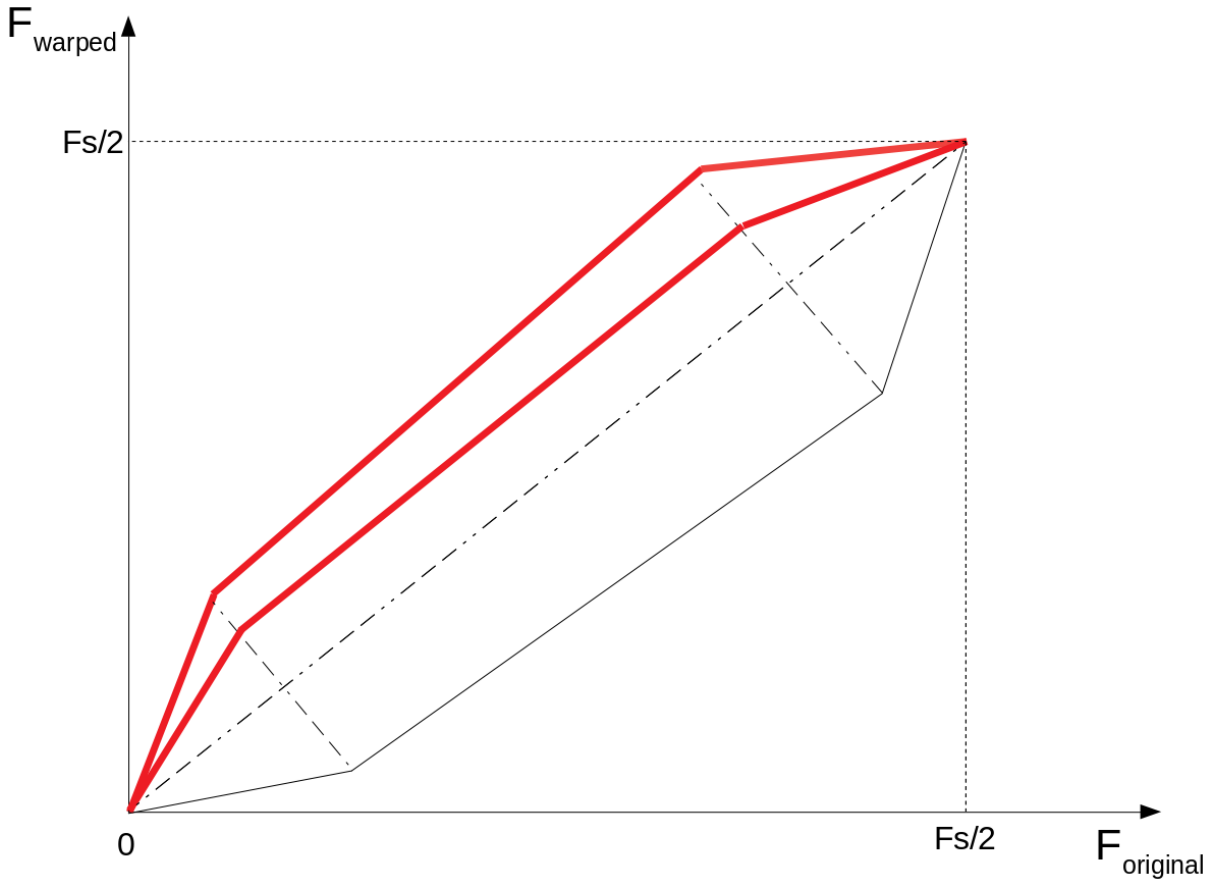

Figure 3.1: Representation of the VTLP warping function that maps frequencies to a new scale. In this work, the maps within the red perimeter are considered (corresponding to $0.8 <$ warp factors < 0.9).

1) VTLP warp factors ($F_{original}/F_{warped}$) are chosen (from a uniform distribution for each utterance) such that their action maps frequency bins to only higher values. This mapping is a piece-wise linear map as shown in Figure 3.1.

2) The speed of each utterance is increased, at two levels, and used along with the original speed. This is done to both the VTLP warped and original IITM dataset. Intuitively, it is observed that sped up audios sound 'higher' than their normal counterparts. This is done because increase in the speed of a signal shifts both pitch and formants to higher values.

For the transfer learning datasets, speed perturbation and pitch perturbation at 0.9x and 1.1x the original value, noise augmentation and SpecAugment methods are used in tandem. Together, the amount of retraining data used is enhanced considerably. The specifics of these augmentations are present in Section 6.

Having understood the reasons behind data augmentation and utilizing the above two augmentation strategies, these modified datasets are used in training acoustic models. The training procedure and architecture of TDNN acoustic models in kaldi are discussed in the next chapter.

Chapter 4

Kaldi TDNN chain models

Kaldi is a well known open source ASR toolkit introduced by Povey et al. [26] in 2011. Acoustic models in kaldi can be realized using Gaussian Mixture Models (GMMs) or other neural network based architectures which model the emission probabilities of a Hidden Markov Model (HMM). 'nnet3' models are a class of models that support the use of complex neural networks like Recurrent architectures (RNNs and LSTMs) and Time-delay Neural Networks (TDNNs) unlike the older 'nnet' setup which supported only simple deep feed-forward architectures. Chain models are a further subclass of these acoustic models which have innovative and unconventional HMM design and topologies. In this work, the baseline TDNN chain model recipe provided by the IITM ASR challenge organizers [11] is examined, which itself is a slight modification of the TDNN chain model recipe used in the WSJ (Wall Street Journal corpus) recipe provided by kaldi. Links to the recipes can be found here[1] [2]. First, a brief overview of the ASR problem is presented and then the TDNN chain model is described.

4.1 Overview of an ASR system

The broad goal of an automatic speech recognition system is to find a sequence of words that very closely matches the speech data present in an audio signal. It is defined as finding that sequence of words $\tilde{W}$, given a sequence of observation vectors O obtained from the audio signal, which maximizes the probability $P(W|O)$ where W is an arbitrary sequence of words. To put it mathematically, find $\tilde{W}$ such that

[1] WSJ recipe: https://git.io/JTH3n

[2] IITM baseline: https://git.io/JT9MZ

$$\tilde{W} = \arg\max_{W} P(W|O) \tag{4.1.1}$$

Using Bayes' rule and ignoring $P(O)$ since it is the same for all observation vectors this becomes

$$\tilde{W} = \arg\max_{W} P(O|W)P(W) \tag{4.1.2}$$

This product is now evaluated using two different modeling approaches. $P(O|W)$, the likelihood of observing O given that sequence of words spoken was W, is evaluated using an acoustic model. $P(W)$ (called a prior) models how likely the sequence W itself is irrespective of other information and is evaluated using a language model. Most modern day ASR systems use HMMs to do the Acoustic modeling excluding end to end models. A tutorial on HMMs and their role in training (learning the parameters of the HMM given a set of observation vectors and its transcripts) and decoding (as defined in Equation 4.1.1) for ASR applications can be found in [27]. Each HMM can be thought of as modeling a single phone or a context dependent phones (triphones).

The emission probabilities of an HMM can be modeled using a mixture of gaussians (GMM-HMM model) or a neural network. One conventional approach for building an ASR system is to first build a GMM-HMM system. Estimating the GMM's parameters is done simultaneously along with the HMM training using either the Baum-Welch algorithm or the faster Viterbi training. This GMM-HMM system act as a buffer for the neural network training by providing frame level phone labels for the neural network to train on. The objective function optimized in the neural network might be a frame level objective function (like the cross entropy between the HMM hidden state (triphone) and the observed feature vector for each frame).

This method of neural network training is an MLE (Maximum Likelihood Estimation) approach which maximizes the likelihood of the correct word sequence:

$$F_{MLE}(\theta) = \sum_{r=1}^{R} \log P_\theta \left(O_r \mid M_{w_r}\right) \tag{4.1.3}$$

where θ represents the parameters of the model, R is the total number of training utterances, w_r is an individual utterance's word transcript (which might be a sentence or two), M_{w_r} represents the hidden state sequence of the HMMs for that particular word sequence w_r. O_r represents the input feature sequence for that word sequence w_r.

Another well known approach which directly tries to maximize the posterior probability in Equation 4.1.1 is called the MMI (Maximizing Mutual Information) estimation

(which is part of other discriminative training approaches like MPE (Minimum Phone Error) and MWE (Minimum Word Error) [28].

The MMI objective function is:

$$F_{MMI}(\theta) = \sum_{r=1}^{R} \log \frac{P_\theta\left(O_r \mid M_{w_r}\right) P\left(w_r\right)}{\sum_{\hat{w}} P_\theta\left(O_r \mid M_{\hat{w}}\right) P(\hat{w})} \tag{4.1.4}$$

where the symbols have the same meaning as before in Equation 4.1.3 and $\hat{w}$ represents an arbitrary word transcript summed over all possible word transcripts. Unlike the MLE approach which only maximizes the likelihood of the correct word sequence from the training transcripts, MMI tries to maximize that likelihood while at the same time also minimizing the likelihood of all wrong word sequences.

In the next section the nnet3 TDNN chain model recipe is discussed which uses both the MMI training criterion and Cross entropy training.

4.2 TDNN chain model

The baseline TDNN chain model recipe from the IITM ASR challenge [11] is examined in this section. A GMM-HMM model provides lattices that align the training data with its transcripts which is required for the TDNN training. The TDNNs used in the baseline recipe use a 140D (dimensional) input vector for each frame (100D i-vectors and 40D LDA MFCCs) and have additional constraints on its layers. They are a factored form of TDNNs i.e. each matrix corresponding to a TDNN layer is factored into a product of two matrices, with one of the factors constrained to be semi-orthogonal [29]. The size of the factored matrices can also be adjusted using variables, called 'bottleneck-dim', 'big-dim' and 'small-dim', in the kaldi code depending on the type of layer.

The TDNN is trained using both the MLE estimation of cross entropy labels for each time frame and the MMI criterion between input feature sequence and output word sequence, discussed in section 4.1, on two different output layer blocks. This is conventionally referred to as multi-task training, although the cross entropy output (output-xent) block is unused while decoding. The output layer trained on cross entropy has a different learning rate, scaled to the MMI output learning rate, which can be set by the user. A snippet of this multi-output network's code can be seen in Figure 4.1, with another example in Figure 4.2 showing a final TDNN block which feeds two output layers. The MMI output layer has no softmax i.e. it's probabilities are not normalized and the costs are directly used.

```
prefinal-layer name=prefinal-chain input=prefinal-l $prefinal_opts big-dim=1024 small-dim=192
output-layer name=output include-log-softmax=false dim=$num_targets $output_opts

prefinal-layer name=prefinal-xent input=prefinal-l $prefinal_opts big-dim=1024 small-dim=192
output-layer name=output-xent dim=$num_targets learning-rate-factor=$learning_rate_factor $output_opts
```

Figure 4.1: Kaldi code snippet of the two output layers used in chain models. The input to both of them is the same.

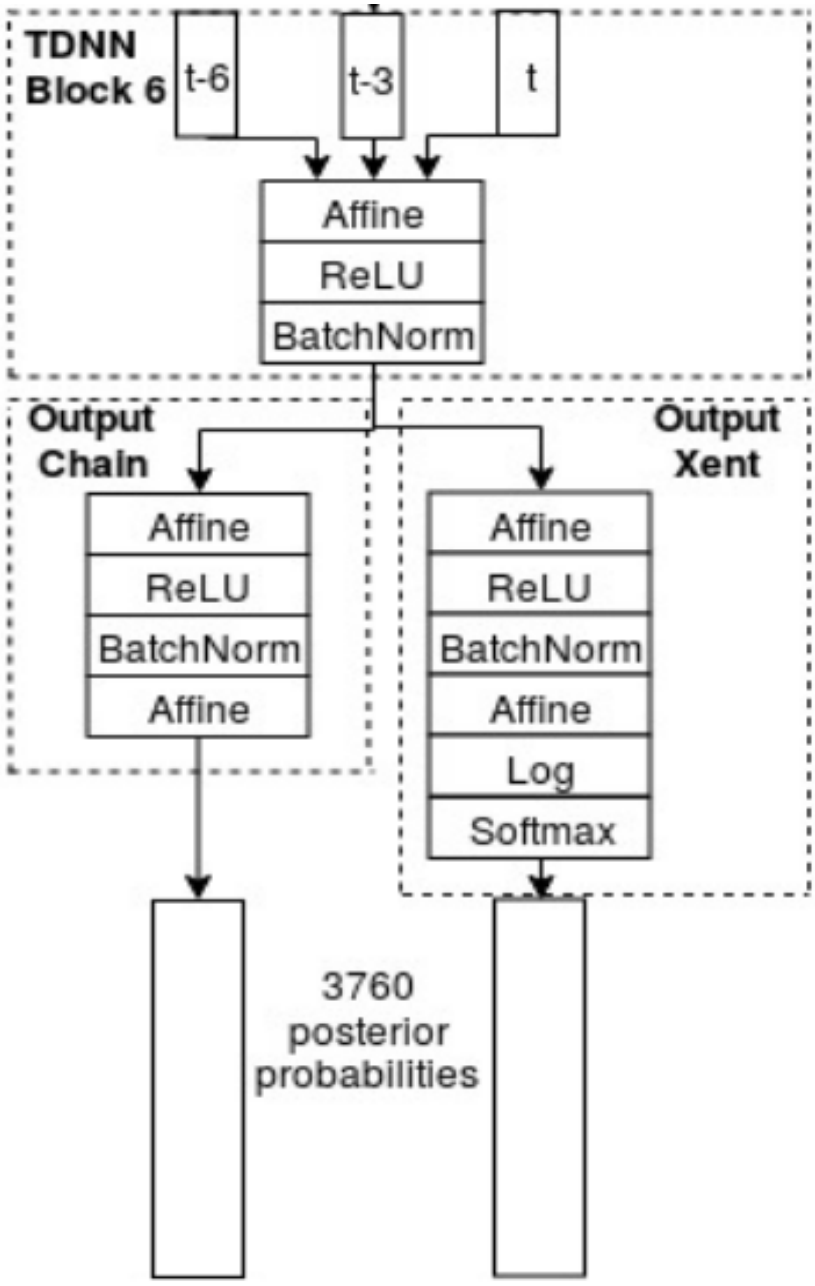

Figure 4.2: Final blocks of a TDNN architecture, representing the two output layers. [2]

In the following sections the TDNN architecture is discussed first, then the details of the chain model training (with its unconventional HMM topologies) which uses the TDNN outputs for training are discussed. Of course, TDNNs can be used to model

emission probabilities of conventional HMMs as well.

4.2.1 TDNN

Time-delay neural networks were introduced to the Kaldi ecosystem in Peddinti et al. [3]. Figure 4.3 shows an example of a TDNN architecture. TDNNs are used to model long term contexts in the data using features calculated in the short term without explicitly modifying the features. The TDNN architecture in Kaldi has some other special properties and is similar to a 1D CNN (Convolutional neural network) in many ways. Each node (represented by the rectangular box) in a subsequent layer is computed using only the inputs within a given context from a previous layer. Activations at subsequent layers are not found at all time frames since there is a correlation between two adjacent nodes. So a sub-sampling of the nodes at each layer is done. The lines and boxes in red represent the sub-sampling, characteristic of the network. The blue lines represent a conventional TDNN. This sub-sampling leads to a huge reduction in training time compared to conventional TDNNs.

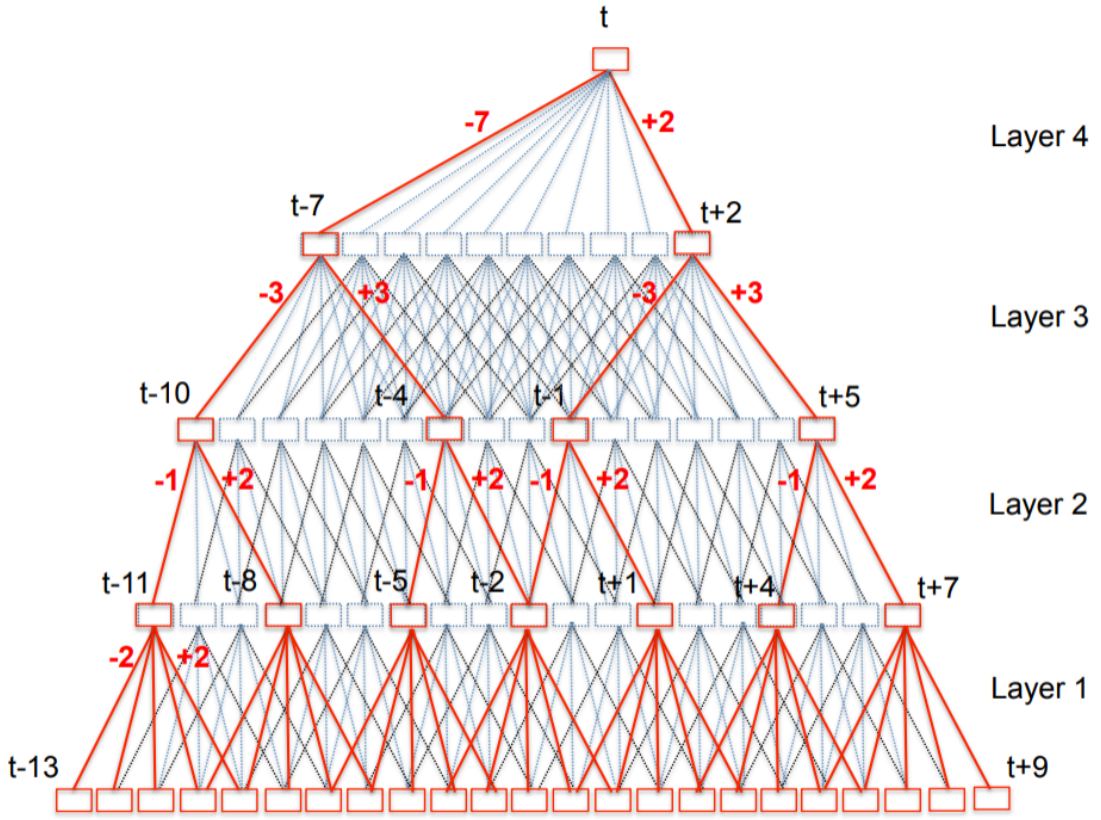

Figure 4.3: Example of a TDNN architecture [3]

Furthermore, as the network gets deeper it was found that using wider and wider

contexts was beneficial. This can be seen with the increased context range in Layer 3 compared to Layer 2. The context range was explored as a hyper-parameter in [3] and it was found that after increasing the contexts beyond a certain range, the WERs were adversely affected. All these observations seem to be in agreement with certain intuitive notions held about speech and phones. Also, it can be seen that the left context (past) is usually larger than the right context (future) which doesn't exceed three frames in the baseline model. This was done to support real time decoding used in other recipes. The baseline TDNN model used in this work has 12 TDNN layers followed by one linear layer (along with two linear output layers and one input layer) and the contexts used are: one frame for the first three layers, no context on the fourth layer and three frames of context for the remaining layers.

With the TDNN architecture described, the next section describes how a TDNN is trained in kaldi using the MMI and MLE criteria. This is referred to as chain training.

4.2.2 Chain training

Training the TDNNs involves the use of a previously trained conventional GMM-HMM model that provides the alignments of the TDNN training data with its transcripts. The GMM-HMM model uses a LDA + MLLT + SAT tri-phone GMM model (with 2750 tri-phones and approximately 50 gaussians per tri-phone) trained on the IITM data set. More information about GMM-HMM model training can be found in [12]. The input to the TDNN is a 140D vector (concatenating the 100D i-vector and the 40D LDA MFCCs). Considering Equation 4.1.4, two terms i.e. the numerator and denominator terms have to be evaluated. They are treated as graphs during training.

The denominator term is a constant for all utterances and need only be evaluated once and composed with every new training utterance. It is evaluated on a graphics processing units (GPU) as a phone graph, which is very similar to the construction of decoding graphs, using a 4gram phone LM (with no back-off) obtained from all training transcripts. There is no lexicon FST (L) here so the denominator is an HCP FST (Finite state transducer) containing only the HMM states (H), the context dependency (C) and the phone LM (P). More information about the determinization, minimization and other phone LM characteristics can be found in [30]. Even with these adjustments of using a phone graph the model size becomes prohibitively large, so another sub-sampling approach is taken up where the TDNN outputs are computed every 30 milliseconds (ms) and then

passed to be composed with the denominator FSTs, which is then evaluated to find the denominator in the loss function.

However, because the TDNN outputs are at 30 ms hops, the older HMM topolgies of three states per phone cannot be used. So a new topology where an HMM can be traversed in one transition between two states is used. The first state outputs the context dependent tri-phone and the second state emits a blank symbol. This was inspired by the work of Sak et al. [31]. It is expected that the emission probabilities and the MMI objective function make up for this crude topology. No lattices are constructed in the denominator and instead the 4gram phone LM models all possible utterance transcripts, hence the name lattice-free MMI.

Next, the numerator term is also evaluated as an FST. The numerator term gives the probability of the utterances' feature given its correct transcript. It is evaluated as an FST using the GMM-HMM alignments of the training data with its transcripts converted into lattices (containing alternate phone pronunciations for each word as well). Even for a single pronunciation, the lattice alignments used (instead of a single best forced alignment) while training offer extra variability, a simple example can be an alignment where there are only three feature frames present in an utterance and if the corresponding word W of that utterance has a phone transcript x-y, then possible alignments are x,x,y or x,y,y [32]. These lattice alignments are relaxed a bit; allowing a phone to occur on either side of its alignment by adding 50ms of slack. It gives the chain model more 'freedom' while training, as expressed in the kaldi code comments. However since the G FST (Language model) is just the words in the transcript for that utterance it is just composed as an acceptor. The HCLG fst is evaluated on the central processing unit (CPU) [33].

4.2.3 Chunk width

Another optimization used to reduce the size of the memory occupied while evaluating both FSTs is to split the training utterances into chunks of 1400ms i.e. (140 frames). These chunks are obtained by splitting the utterances and also using the lattice alignment information to modify the phone transcripts and get 1.4 second chunks for the numerator FST with a little tolerance. The same splits are used for the denominator FSTs. More information and optimizations involved in the FSTs of the split chunks can be found in [33], [32]. Codes that create the denominator FST[3] and numerator FST[4] with explanations in

[3]denominator.fst: https://git.io/JT78E

[4]numerator.fst: https://git.io/JT74f

the comments can be found in the footnote links.

With the numerator and denominator FSTs computed, their difference in log domain gives the MMI loss function. This loss is then minimized by the TDNN training by updating the TDNN parameters after every mini-batch. A mini-batch is a tensor (i.e a collection of C x 140 matrices) which is sent through a network forward (for computing the MMI loss) and backward (for updating the network parameters) where C is the chunk-width of 1400ms and the other 140 refers to the feature vector dimension (i-vector(100)+LDA MFCCs(40)). Network parameters are not updated after every training example has passed through the network but after every mini-batch. This is mini-batch gradient descent (GD) unlike vanilla GD, the parameters do not update for every training example. Each mini-batch consists of some random subset of all C x 140 matrices from the training data. The mini-batch size, R, is tunable (set to R=64), so each mini-batch is a tensor of size: R x C x 140.

No early stopping criteria is provided to stop the training at certain epochs. Instead, two subsets (300 utterances each) of the TDNN training data are formed called i) validation diagnostics (which has no utterance overlap with the actual data used to train the TDNN) and ii) train diagnostics (which has utterance overlap). The loss functions on each of these subsets at each training iteration can be used to determine whether underfitting or over-fitting occurred. The model parameters are also averaged over the final few iterations when creating the final TDNN model.

Once the TDNN training is done, an HCLG FST graph can be made using compatible language models and lexicons and the TDNN itself is available as a final.mdl file. Usual kaldi decoding procedures can then be followed with some slight modifications to the acoustic weight parameter [34]. Reducing beam-widths while creating decoding lattices can also significantly reduce decoding time with little to no degradation in WER. The TDNN outputs during decoding are also computed at 30ms hops and provide this improved decoding speed.

With the IITM adult speech trained TDNN chain model in hand, the next focus is on adapting these models to target domain speech. Works on Transfer learning and Adaptation are reviewed next.

Chapter 5

Transfer learning and Adaptation

The aim of this work is to explore the baseline TDNN model (trained on adult speech) and to improve its decoding ability on target speech. To this end, techniques of adapting and modifying the baseline model for the ASER data sets are explored.

5.1 Acoustic model adaptation

In this section, acoustic model adaptations explored in [12] are reviewed. Batch supervised adaptation techniques are considered since the requirements of this work right now do not need real time outputs (so online adaptations for each new utterance is not done) and the transcriptions obtained for the speech are manually done by a transcriber. There are broadly three adaptation methods:

1. **Feature Adaptation**

   In feature adaptation, the input features used in the ASR system are modified to either remove or control speaker effects. Common methods include Vocal Tract Length Normalization (VTLN) which tries to curtail the effects of the different vocal tract length the speakers have [35]. VTLP warping of the feature set can also be considered a form of adaptation although it is unguided and not optimal like VTLN techniques. fMLLR (feature maximum likelihood linear regression) methods find affine transforms of the features to remove variability and normalize across speakers. More information on applying feature transforms in kaldi can be found in [36]. Identity vector (i-vector) features are commonly used in neural network systems to capture information about the speaker identity and channel acoustics for speaker verification tasks but have also found use in speech recognition systems

[37], although in recent ASR systems they are an essential feature that are almost always used.

2. **Model space adaptation**

   Model space adaptations as the name suggests modify the model parameters instead of the features to suit the adaptation data and its speakers. Maximum a posteriori (MAP) adaptation is a well known adaptation for GMM-HMM models. It maximizes the likelihood of the adaptation data by modifying the parameters of the GMM-HMM model [38]. Maximum likelihood linear regression (MLLR) of GMM parameters is similar to fMLLR but the affine transform operates on the model parameters (instead of the features) to maximize likelihood adaptation data. These transforms can be found for specific pronunciation variations and can be applied for specific gaussian components (regression classes) as done by Yoo Rhee Oh et al. [39].

   For neural network models re-training specific layers of the network with the adaptation data gives improvements. Transfer learning is an extension of this and is explored deeply in Section 5.2.

3. **Speaker space adaptation**

   In speaker space adaptations, separate acoustic models are trained for each speaker type. Speaker types are estimated using clustering techniques or is done manually (for example building gender/age dependent models). Once the speaker clustering is done and different sets of parameters for each model are available, new speakers are estimated as weighted means of the cluster models and decoding is done. A generalization of this is to find the eigenspace of the speakers which reduces model size and complexity [40].

The above acoustic model adaptations are used in the baseline GMM-HMM model training which is used to provide phone level alignments of the training data for the TDNN training.

5.2 Transfer Learning

Transfer learning is a model-space adaptation method as discussed in Section 5.1. This is the main focus of experimentation in this work. Transfer learning methods in kaldi

were investigated extensively by Ghahremani et al. [41] which will be the reviewed in this section. Other kaldi transfer learning works like Manohar et al. [42] mainly refer to this.

Transfer learning methods are used when there is a small amount of domain specific data (target speech) compared to larger amounts of out of domain data. It involves using the domain specific data to improve a model trained on the out of domain data which will be called the 'source' model. This is achieved by using the source model and either modify its training mechanisms or re-train the model using the domain specific data to obtain a 'target' model that performs better on test sets from the domain specific data. The alignments of the training/valid data with its transcript, required for weight transfer, are obtained from the source TDNN chain model as well.

Lee et al. [43] showed that the intermediate layers of a network are not task specific by using a pre-trained model using some of the bottom layers and fine tuning them for a variety of audio classification tasks including speaker identification, phone classification, speaker gender classification and separately for music genre and music artist classification. This gives credence to the idea that transfer learning can be used extensively when the domain specific data is in shortage. In many cases of multi-lingual DNNs it was found that language similarity (between source and target) and the amount of data also played a part [44] [45]. A general block diagram of transfer learning is shown in Figure 5.1.

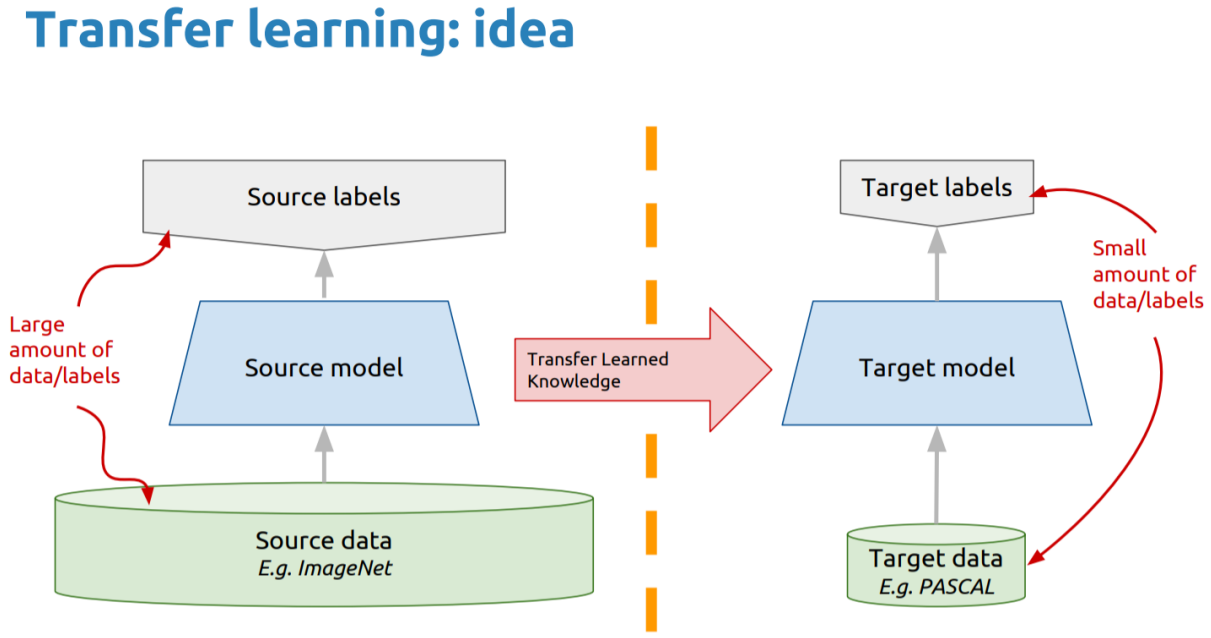

Figure 5.1: A general block diagram of transfer learning [4]

In Ghahremani et al. [41], two methods of transfer learning are investigated:

1. **Joint multi-task training:**

In this method, the initial layers of the network are trained on data from all domains in parallel, and specific final layers (all of which share the output of the last initial layer) are trained only on their respective domain specific data. Scaling the gradients from the domain specific data depending on their size has been found to help.

2. **Weight transfer:**

   Weight transfer closely mirrors the work of Lee et al. [43]. A source model trained on out of domain data is taken and the first 'x' number of source layers are retained. Then either new task specific layers are added as required and the model is re-trained using the domain specific data. This re-training also allows different strategies. In single stage tuning the new task specific layers are trained at higher learning rates compared to the source layers, whereas in two-stage training the source layers are frozen and the task specific layers are trained on the domain specific data followed by fine tuning both source and task specific layers using a small learning rate for a very small number of epochs.

From these experiments, Ghahremani et al. [41] concluded that:

1. Joint multi-task training is preferable to weight transfer, although both outperform the baseline and the multi-tasking takes more time to train.

2. Single-stage tuning is preferable to two stage tuning in weight transfer experiments

3. When the phone sets, type of speech (spontaneous or read speech) of the domain specific and out of domain data are the same, transferring all layers from the source model was more effective than some other partial transfer that does not include all layers.

During single-stage tuning, where the transferred source layers are trained at smaller learning rates compared to the task specific layers, it was found that increasing the ratio between the learning rates of the source and global learning rate (α) helped reduce the WER on a test set up to a degree.

In Zhang et al. [46], an alternative school of thought in transfer learning was presented which freezes task specific layers and trains lower layers. The output layer and top most one, two or three BLSTM (Bidirectional Long Short-Term Memory) layers of an end to end CNN-BLSTM model (with four BLSTM layers on top of the CNN) were frozen and the lower layers are retrained using the adapted data. This is justified with the premise

that although the intermediate layers of a network aren't task specific (they capture a useful representation of the input), the final layers are task specific, so retraining the intermediate layers while keeping the final layers frozen modifies the representation learnt from the target domain data. The best result they obtained was for freezing the top two BLSTM layers and the output layers. They also examine scaling the learning rates of the topmost layers by two factors of 0.1 and 0.5 with the global learning rate. The best results were obtained for the factor 0.5.

In Gergiou et al. [47], a synthesis of the above two retraining methods were examined. The middle layers of the network are frozen and the top and bottom layers are retrained because:

1. Acoustic variability in the features is captured by the layers near input of the model
2. Pronunciation variability is seen only at the layers near the output of the model.

They examine two retraining techniques:

1. Keeping weights of the middle hidden layers fixed and allow the top-most and bottom-most layers to update simultaneously,
2. Dis-jointly and alternately training the various layers (top and bottom) until convergence"

And found that simultaneous retraining is better when there is a reasonable amount of adaptation data available (at least 2hrs of adaptation data for a baseline model trained on 206 hrs). It was also found that retraining fewer than 3 top and bottom layers was most beneficial if the adaptation data was no more than 25hrs [47].

5.3 Parameters in transfer learning

The parameters tuned in transfer learning are explained here:

1. **Regularization factors:**

   In kaldi, it was empirically found (through experiments on various corpora) that a form of multi-task training i.e. having two or more output layers trained independently on different objective functions (cross entropy and MMI) helped decrease the WER. These output layers are called 'output-xent' (Short for cross(x) entropy(ent)) and 'output' (chain). The 'output' layer is the one that is finally used while decoding

while the 'output-xent' is only useful while training. The 'output-xent' layer also acts as a form of regularization [30]. Another regularization method used is the L2 norm regularization. The squared L2 norm of the 'output' layer of the TDNN (not 'output-xent') is penalized. This involves adding $-0.5c * ||y||_2$ to the loss function, where y is the output of any node in the 'output' layer. The variable 'c' is tuned and determines the degree of regularization. [30]. A separate term that scales the output of the 'output-xent' layer also exists called chain.xent-regularize. These two parameters can help tune the degree of regularization.

2. **Number of epochs:**

   The number of epochs is the total number of times the network sees the training data. It also determines the total number of training iterations along with the mini-batch size. The dataset is divided into mini-batches before passing it through the network. Each mini-batch consists of 'R' utterances where R is the mini-batch size. These mini-batches are constructed using the 1.4 second kaldi training chunks as explained in Chapter 4.2.2. These two parameters together determine the number of iterations which is roughly the (number of epochs)/(mini-batch size).

3. **Number of jobs:**

   This determines whether multiple GPUs are used which can do simultaneous training on different subsets of the training data. Although no significant differences are expected, due to the size of the dataset using one GPU is optimal and recommended to reduce the effects of model averaging (combining the training/model parameters across GPUs after training).

4. **Global initial and final effective learning rates:**

   These rates determine the learning rate of the network at each iteration. The learning rate starts at the initial learning rate on the first iteration and decreases at each iteration until it reaches the final learning rate. Together, the rates control the rate at which the training loss changes.

5. **Differential learning rates:**

   Differential learning rates are used to train different layers of the baseline model at different rates. The differential learning rates are represented like so: "5(4)-0(8)-0.625(3)*1e-6". In this example the first 4 TDNN layers have initial learning

rate 5e-6, the next 8 layers are frozen and the final 3 layers (2 output and 1 linear layer) have initial learning rate 0.625*e-6. In all experiments in this work, the final effective learning rate is 1/10th the initial learning rate.

6. **Chunk width/frames per chunk:**

   As discussed before, each mini-batch during training consists of an R x C x 140 tensor. C is usually set to 1.4 seconds in most kaldi recipes. The real advantage of changing C comes in when computing the MMI loss function for retraining. Since the loss function can be computed on any sequence of frames, the network can be made to learn text contexts of arbitrary length C, along with the acoustic characteristics of the speakers. In general, it becomes better at identifying text contexts of length C. Generally, during transfer learning, C is not changed from the value used in baseline training.

   When transfer learning needs to be evaluated on datasets which have no story overlap with the retraining set (i.e. when learning the text contexts is not as useful as learning speaker characteristics), it is desirable to reduce the chunk width. The retrained model learns both speaker characteristics as well as the text. However, in scenarios where there is a limited number of speaker types and unseen text contents, transfer learning might not work well without reducing the chunk width.

. The other parameter not fiddled with is the mini-batch size at 64, which is kept fixed across all experiments. In the next section, the experiments carried out based on the understandings from these parameters are explained.

Chapter 6

Experiments

The following observations and experiments are done, pertaining to the children's ASER test data, for the adaptation and data augmentation experiments in this work:

1. The VTLP warped and original IITM dataset are considered for training the baseline model. Along with this the speed perturbation factors are also changed from the default baseline used in kaldi (0.9x and 1.1x) to higher speeds which also shift the data to higher frequencies (1.1x and 1.2x) as discussed in Chapter 3.

2. For the transfer learning experiments, the UP and RJ train sets in Table 2.1 are combined and used for re-training the baseline model. This retraining data is further augmented by first overlaying various noises (exam noise recordings, wind, babble, rain, traffic, insects, children playing) on top of the retraining data at various SNR (5, 10 and 15dB). This results in 2 versions of the retraining data (original+noisy). Both the original and noisy versions are then sent once, through a SpecAugment module. This creates 2 more new versions. Next, these 4 versions are speed perturbation at 0.9x and 1.1x, pitch perturbation at 0.9x and 1.1x using the sox tool. In total, there are 20 versions of the acoustic retraining data used for transfer learning.

3. For the initial transfer learning experiments where the number of layers to be frozen are decided, only speed perturbation and noise augmentation (6 versions) is used. This is to help decrease the time for each retraining experiment and quickly tune the other retraining parameters. Also, retraining experiments with denoised 2012 retraining data are performed with only the above augmentations in hopes of getting improvements on the noisy 2016 set.

4. After deciding which layers to freeze and which to keep, further retraining experi-

ments are done using all augmented versions and results are obtained on the 2016 and 2012 validation and test sets.

6.1 Evaluation metrics and Decoding parameters

This section outlines the evaluation metrics used to measure the ASR system performance keeping in mind that it inherently measures (or rather should measure) the reading skills of the child when used as a decoder.

While decoding, two parameters are involved in determining the decoded text. These parameters influence the cost associated with each word. The decoding cost, consisting of an acoustic cost, language model cost and an insertion penalty is:

$$Total\_decoding\_cost = \log P(O|W) + LMWT * \log(P(W)) + WIP * |W| \quad (6.1.1)$$

$|W|$ represents the total number of words in an utterance. LMWT represents the language model weight and WIP represents word insertion penalty. In all experiments LMWT is an integer ranging from 10 to 50 and the WIP values used are -0.5,0.0,0.5,1.0.

6.1.1 Evaluation metrics

The experiments made are evaluated using the WER metric and the F-score of the correct words between the ASR output and ground truth text. To calculate the WER on the test sets, a tri-gram language model (LM) which is trained on the canonical text of all the 2012 and 2016 ASER Hindi stories is used when decoding the 2012 sets and 2016 validation and test sets respectively . In addition to this, a uni-gram garbage model with a fixed cost = 0.2 (tuned from other experiments) is appended to the language model to provide words from outside the stories' text. The garbage model used when decoding the 2012 sets contain all unique words from the transcriptions of the recordings of the UP 2012 set that occur at least twice in the train and validation set. While decoding the 2016 set, the garbage model contains, in addition to the words mentioned before, all unique words from the transcriptions of the 2016 validation set that occur at least twice. It also contains all single phones present in the lexicon and words from the Hindi Barakhadi. This slightly restrictive LM is expected to provide a good performance for speakers who are good readers (because of the restrictive text), while the garbage model allows other words to come in. As stated before in Section 2.2.3, the other tags like ON, FP, BR etc are removed from the ground truth while evaluating these metrics.

Since the ultimate goal of this work is in identifying reading miscue detections, another metric introduced is the F-score of the correct words as identified in the GT. This F-score is the harmonic mean of the precision and recall of the correct words, defined as follows:

1. Precision (P): Of the total number of correct words identified by the ASR system how many were actually correctly spoken by the child in the GT.

2. Recall (R): Of the total number of correctly spoken words by the child according to the GT, how many were identified by the ASR system.

Here, a correct word is a word that is present in the canonical text and is correctly spoken by the child. The F-score is then $(2 * P * R)/(P + R)$. One way to support the use of this measure is that it is of slightly less import as to what the exact mistake a child made while reading as long as the mistake is identified. By reducing the effect of the deficiencies of the ASR system, this metric is useful for quantifying literacy levels.

6.1.2 Selection of decoding parameters

The WER of the test sets on all experiments is reported using the best performing LMWT and WIP found on the validation set. The decoded texts corresponding to these parameters are also used to report the F-score of the correct words on the test sets. These are the metrics on which improvement is desired.

The next chapter presents the results of these experiments using the WER and the F-score metric defined above.

Chapter 7

Results

Firstly, the experiments involving the freezing of various layers/blocks of the baseline TDNN model and retraining the remaining with the noise augmented and speed perturbed ASER 2012 UP+RJ training sets are presented below in Table 7.1. For all retraining combinations, after tuning the regularization parameters, number of epochs and differential learning rate factors, the best retraining parameters obtained were:

–chain.xent-regularize 0.001

–chain.l2-regularize 0.0

–trainer.num-epochs 4

–egs.chunk-width 140

The performance of the various combinations is measured using the validation loss on the 2012 UP+RJ combined validation sets. Higher the validation loss, better the model.

Table 7.1: Experiments involving freezing of various layers/blocks of the TDNN

Retraining parameter setting	Diagnostic 2012 Train UP+RJ MMI loss	2012 UP+RJ Validation MMI loss
5(1)-0(12)-0.625(2) * 1e-6	0.4443	0.369501
5(2)-0(10)-0.625(3) * 1e-6	0.444404	0.369544
5(3)-0(8)-0.625(4) * 1e-6	0.444407	0.369603
5(4)-0(8)-0.625(3) * 1e-6	0.444351	0.369564
5(4)-0(6)-0.625(5) * 1e-6	0.44442	0.36959

From Table 7.1, it is clear that training the first 3 TDNN layers and the last TDNN layer, 1 linear and 2 output layers gives the best validation loss. Repeating the best retraining combination with all 20 augmented versions as described in Chapter 6 the

validation loss improves further to 0.379159 from 0.369603. The number of epochs here is also now reduced to 1 since, naturally, the model sees many versions of the data due to the large amount of data augmentation. Increasing the number of epochs beyond 1 reduces the validation loss and causes over-fitting.

Next, the improvements obtained in WER and F-score of the correct words on the 2012 validation and test sets are displayed in Table 7.2. All the models in Table 7.2 are trained with chunk width of 1.4 seconds. The retraining data used and the data augmentations applied is mentioned as well. Clearly, improvements in WER of around 6-8% are obtained, while the baseline model also has a very good F-score and minor improvements are obtained on top of it.

Table 7.2: Improvements obtained on 2012 data

LM: 3 gram 2012 canonical ASER stories	**GM**: UP train+valid words (count>=2)						
Acoustic Model	Validation (UP+RJ) WER%	Validation F-score (P, R)	UP test WER%	UP test F-score (P, R)	RJ test WER%	RJ test F-score (P, R)	lmwt, wip
(VTLP warped + original) IITM Baseline; sp 1.1x 1.2x	23	0.96 (0.966, 0.954)	17.32	0.975 (0.976, 0.973)	19.63	0.97 (0.969, 0.97)	27, 1.0
5(3)-0(8)-0.625(4)*1e-6; (2012)UP+RJ+hindi_CS (noise_aug+sp)	14.68	0.971 (0.966, 0.975)	11.53	0.982 (0.978, 0.985)	12.71	0.97 (0.969, 0.97)	28, 0.0
5(3)-0(8)-0.625(4)*1e-6; (2012)UP+RJ+hindi_CS (noise_aug+sp+pp+SpecAug)	14.55	0.973 (0.966, 0.979)	11.45	0.982 (0.978, 0.985)	12.87	0.97 (0.969, 0.97)	24, 0.0

Next, experiments involving the ASER 2016 subset which has no story overlap with the 2012 sets are shown in Table 7.3. The 2016 data has higher WER and lower F-scores compared to the 2012 data on the raw undenoised recordings. This is expected because of the noisy nature of the 2016 data. To counter this, two denoising experiments are run using the facebook demucs denoiser [48], specifically a pretrained model called DNS64 that was trained on the DNS set. Other pretrained models were too aggressive and erased regions of actual child speech from a few listening observations. First, 2016 data is denoised using DNS64 model. No improvements are obtained here. Next, the 2012 retraining data is itself denoised and used for retraining the baseline model. Then this model is used to decode the undenoised and denoised 2016 sets. Again, no improvements are obtained in any scenario over the retraining and decoding without denoising. So the denoiser is abandoned for future experiments.

With all the augmentations in place further improvements in WER are obtained on

the 2016 dataset and reducing the chunk-width to 50 further improves the WER on the no story overlap subset. This is because the reduced effect of the text contexts during retraining is helpful when there is no story overlap between the training and validation sets. However, the 50 chunk-width model has a slightly higher WER on the 2012 validation sets (15.1% compared to 14.55% obtained in Table 7.2). Also of importance is the fact that the retrained models tend to have a higher proportion of deletions and a lower proportion of insertions compared to the baseline model results irrespective of the lmwt, wip. These are discussed in detail in the next section, but it is, in short, due to the phone mappings made from ON-SIL and other such mappings. No improvement is obtained in the F-score of the 2016 no story overlap subsets, although the drop off is quite small (0.2-0.6%) with a decrease in WER of around 8% in the best model.

Table 7.3: Improvements obtained on 2016 no story overlap with 2012 subset

LM: 3 gram 2016 canonical ASER stories

GM: UP train+valid words+ASER 2016 valid (count>=2)

Acoustic Model	Undenoised WER% (valid,test)	Undenoised F-score (P, R) (valid, test)	lmwt, wip	DNS64 denoised WER% (valid, test)	DNS64 denoised F-score (P, R) (valid, test)	lmwt, wip
(VTLP warped + original) IITM Baseline; sp 1.1x 1.2x	**30.27** [5350 / 17677, 2372 ins, 412 del, 2566 sub] **34.50**[4376 / 12684, 1930 ins, 308 del, 2138 sub]	0.953 (0.965, 0.942) 0.949 (0.960, 0.938)	35, 1.0	**31.59** [5585 / 17677, 2584 ins, 381 del, 2620 sub] **36.38** [4614 / 12684, 2098 ins, 272 del, 2244 sub]	0.952 (0.967, 0.937) 0.947 (0.961, 0.933)	36, 1.0
5(3)-0(8)-0.625(4)*1e-6; (2012)UP+RJ+hindi_CS (noise_aug+sp) chkwidth=140	**24.70** [4367 / 17677, 505 ins, 1381 del, 2481 sub] **28.52** [3618 / 12684, 390 ins, 1160 del, 2068 sub]	0.926 (0.974, 0.883) 0.915 (0.972, 0.865)	38, -0.5	**26.60** [4702 / 17677, 675 ins, 1103 del, 2924 sub] **30.74** [3899 / 12684, 548 ins, 933 del, 2418 sub]	0.920 (0.977, 0.869) 0.91 (0.975, 0.853)	31, 0.0
5(3)-0(8)-0.625(4)*1e-6; Denoised (2012)UP+RJ+hindi_CS (noise_aug+sp) chkwidth=140	**28.43** [5025 / 17677, 445 ins, 2065 del, 2515 sub] **33.34** [4229 / 12684, 326 ins, 1831 del, 2072 sub]	0.902 (0.974, 0.84) 0.884 (0.976, 0.807)	36, -0.5	**24.85** [4393 / 17677, 547 ins, 1237 del, 2609 sub] **29.31** [3718 / 12684, 449 ins, 1097 del, 2172 sub]	0.927 (0.973, 0.885) 0.914 (0.970, 0.864)	38, -0.5
5(3)-0(8)-0.625(4)*1e-6; (2012)UP+RJ+hindi_CS (noise_aug+sp+pp+SpecAug) chkwidth=140	**22.52** [3980 / 17677, 649 ins, 963 del, 2368 sub] **26.63** [3378 / 12684, 541 ins, 815 del, 2022 sub]	0.942 (0.973, 0.913) 0.933 (0.971, 0.897)	31, 0.0	–	–	–
5(3)-0(8)-0.625(4)*1e-6; (2012)UP+RJ+hindi_CS (noise_aug+sp+pp+SpecAug) chkwidth=50	**22.02** [3893 / 17677, 853 ins, 779 del, 2261 sub] **25.99** [3297 / 12684, 704 ins, 655 del, 1938 sub]	0.951 (0.973, 0.931) 0.943 (0.968, 0.919)	35, 0.0	–	–	–

Lastly, the results on the 2016 subset which has story overlap with the 2012 sets are shown in Table 7.4. Improvements are obtained in all performance indicators and similar trends to the 2016 no story overlap subset are observed here as well (higher proportion of deletions in the retrained model, higher WER compared to the 2012 sets).

Table 7.4: Improvements obtained on 2016 subset which has story overlap with 2012 set

LM: 3 gram 2012 canonical ASER stories **GM**: UP train+valid words (count>=2)

Acoustic Model	WER%	F-score (P, R)	lmwt, wip from 2016 no story overlap experiment
(VTLP warped + original) IITM Baseline; sp 1.1x 1.2x	35.90 [6772 / 18863, 2739 ins, 331 del, 3702 sub]	0.931 (0.968, 0.898)	35, 1.0
5(3)-0(8)-0.625(4)*1e-6; (2012)UP+RJ+hindi_CS (noise_aug+sp+pp+SpecAug) chkwidth=140	26.62 [5022 / 18863, 1089 ins, 761 del, 3172 sub]	0.940 (0.967, 0.913)	31, 0.0

The nature of these results and a few examples that explain them are presented in the next section.

7.1 Discussion of results

The WER on the 2016 sets is worse compared to the 2012 sets. This is due to two reasons:

1. The higher miscue rates of the children in the 2016 set plays an obvious role. More mistakes the children make, more difficult it is for the ASR to identify them, especially because of the lower miscue rates in the 2012 retraining set compared to 2016 test sets.

2. Comparing the 2012 and 2016 ASER sets, the test sets and validation sets in the 2016 set are found to have a much higher concentration of other tags including irrelevant speech (IR) and other noise (ON), described in Section 2.2.3. Simply counting the number of such tags in the transcriptions:

   (a) There are 8031 ON, 750 IR tags in 444 minutes of ASER test+valid 2016 data which comes to about 18 ON tags/min and 1.7 IR tags/min

   (b) In contrast to this, there are 8855 ON, 504 IR tags in 624 minutes of 2012 (UP+RJ test+valid) data which comes to about 14 ON tags/min and 0.8 IR tags/min.

These noises and irrelevant speech sections lead to a higher amount of insertions, substitutions and higher WER. A few substitutions and insertions observed were also due to GT words that are absent in both the LM and GM. The main difference in WER between the baseline and the retrained model is in the insertions and deletions count. There are a lot

more insertions of words in the results of the baseline model compared to 2012 datasets and the results from the retrained model. This is a result of the noisiness of the 2016 dataset which causes many insertions to pop up in the noisy regions compared to the retrained model, which has been trained to ignore these regions (because of the phone mapping ON-SIL) and treat them as SIL. The baseline model is especially stumped by the ON regions, while both the models have a little trouble with the IR regions.

Some excerpts that support this:

1. **biju_aser_cg_S3-P_2**

   (a) *Ground Truth:* SIL ON IR ON आज मामा आए

   (b) *Baseline model decoded text:* हँस चा बड़ी पर लौट रंग की गया आज मामा आए

   (c) *Retrained model decoded text:* पढ़ने लौट है आज मामा आए

   Fewer insertions in the retrained model text. When evaluating WER, GT is considered as "आज मामा आए " which results in the baseline WER having more insertions.

2. **jiyan_aser_mh_S4-P_2**

   (a) *Ground Truth*: ON IR मेरे चाचा की शादी है

   (b) *Baseline model decoded text:* घास बच्चे मेरे चाचा की शादी है

   (c) *Retrained model decoded text:* मेरे चाचा की शादी है

Another observation of interest is the lower F-score of correct words from the retrained models compared to the baseline model in the 2016 set even though there is a good improvement in WER. This is mainly because of cases where there are a few correct words in between a few ON tags or words on either side of an ON tag (i.e. the child spoke something amid some other noise), where nothing/SIL gets decoded. Two worst case examples are shown below:

1. **mushkaan_aser_uk_S4-ST_1**

   (a) *Ground Truth:* लंबी दौड़ लगाने ON अच्छ लगाना ON अच्छा लगता था अच्छा लगता था FP वे तीनों ON रोज़ साथ साथ ON मजे ON म मौज मस्ती करते थे ON IR ON

   (b) *Baseline model decoded text:* लंबी दौड़ लगाना अच्छा लगाना अच्छा लगता था वे वे तीनों रोज़ साथ साथ मज़े ला मौज-मस्ती करे

(c) *Retrained model decoded text:* लंबी दौड़ लगाना लगाना था दिनों मज़े

2. **mukesh_aser_uk_S4-P_2**

(a) *Ground Truth:* ON IR ON मोर ON मोर चाचा की MB ON सादी हुई ON IR ON सबको ON नई ON ON IR FP ON IR

(b) *Baseline model decoded text:* हर साथ्यों मौज चाँद सुरन की में एक लगाए हुई आती ही मोर खाकर रही थी सब को नी मंगाकर यह गाय है

(c) *Retrained model decoded text:* की सब को नहीं

In the above two cases, even though there would be improvements in WER with fewer insertions detected, because of the correct words getting lost in the sea of noise, the F-score does not improve over the baseline model results.

The above examples and understandings help explain the relatively poor WER and F-score of the correct words on the 2016 set compared to the 2012 set and other results obtained.

Chapter 8

Conclusion

With the goal of building an automatic assessment system for children taking literacy tests a few challenges arose. These included the vast variety in speaking style and speech characteristics of children, the scarcity of labeled data available. Since, there are a large number of adult speech datasets and ASR systems available, modifications to these systems and datasets were considered a viable opportunity. A state of the art TDNN ASR system was discussed and two strategies for improving them were considered. Data augmentation techniques were examined to increase the quantity of training data and also modify it to suit particular test scenarios for the target speech. Furthermore, with a small amount of labelled target speech data, weight transfer techniques were examined to fine tune the system for the target domain speech. Changing the chunk width during retraining was an important part of this. Both these techniques gave improvements in detecting reading miscues on a variety of test sets.

8.1 Future Work

Some points for future work that can be explored for further improvements to the system are:

- Since the initial denoising experiments using pre-trained models yield poor results after retraining and/or decoding, training a denoiser on ASER samples could be more beneficial since a better noise profile will be obtained. This action could perhaps help reduce the number of deletion errors from the retrained models.
- Similar transfer learning experiments done on Indian English adult speech baseline models adapted with children's English datasets have shown the importance of re-

ducing the chunk width. In fact, no improvements could be obtained with many retraining parameter tuning experiments without reducing the chunk width to 50 frames in the English transfer learning case. This was mainly due to a smaller speaker set and no story overlap between retraining and test sets. Further experiments can be done here as well.

- Cross language transfer learning experiments can also be performed. Initial experiments, on Marathi datasets i.e by retraining the baseline model with a reduced chunk width on ASER Hindi data and using Marathi validation sets to tune retraining parameters have yielded mixed results. There is a decrease in WER on the validation set but not on the larger more diverse Marathi test set.
- Further exploration of improvements that can be made to the LM and GM.

References

[1] ASER Centre. ASER Survey Process. http://www.asercentre.org/p/231.html, last accessed: October 2020.

[2] Alexandru-Lucian Georgescu, Horia Cucu, and Corneliu Burileanu. Kaldi-based dnn architectures for speech recognition in romanian. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE, 2019.

[3] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] Kevin McGuinness. 2nd Workshop on Deep Learning for Multimedia, Insight Dublin City University . https://github.com/telecombcn-dl/2018-dlmm/raw/master/D2L02_Transfer.pdf, Last accessed October 2020.

[5] ASER. ASER Report for the year 2018. http://img.asercentre.org/docs/ASER2018/ReleaseMaterial/aser2018nationalfindingsppt.pdf, Last accessed October 2020.

[6] The Hindu. Basic literacy, numeracy skills of rural Class VIII students in decline. https://www.thehindu.com/news/national/basic-literacy-numeracy-skills-of-rural-class-viii-students-on-a-decline-aser-2018/article26004114.ece, Last accessed October 2020.

[7] Shreeharsha B S. Spoken language assessment on mobile device. https://www.ee.iitb.ac.in/student/~shreeharsha/mng630_report_shreeharsha.pdf, 2019.

[8] Raymond D Kent. Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of speech and hearing Research*, 19(3):421–447, 1976.

[9] Dolly Agarwal, Jayant Gupchup, and Nishant Baghel. A dataset for measuring reading levels in india at scale. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9210–9214. IEEE, 2020.

[10] Shreeharsha B S. Acoustic models for speech recognition in children's reading miscue detection. `https://github.com/shreeharsha-dap/reports-and-stuff/blob/main/Shreeharsha_MTP_Stage_1_report.pdf`, 2020.

[11] Speech Processing Lab IIT Madras. Hindi ASR challenge. `https://sites.google.com/view/asr-challenge`, last accessed: September 2020.

[12] Prakhar Swarup. Acoustic model training and adaptation for children's read speech recognition. *MTP report, Dept. of Electrical Engineering, IIT Bombay*, 2017.

[13] Nagesh Nayak. Transcription Examples of tags and labels used. `https://docs.google.com/presentation/d/1Id54pZMuGPpDf7LbSWViF42_yKGYmo92IF2HZq-nfIY/`, last accessed: October 2020.

[14] Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, volume 117, 2013.

[15] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[16] Anton Ragni, Katherine Knill, Shakti P Rath, and Mark Gales. Data augmentation for low resource languages. In *Interspeech*, 2014.

[17] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[18] Xingchen Song, Zhiyong Wu, Yiheng Huang, Dan Su, and Helen Meng. Specswap: A simple data augmentation method for end-to-end speech recognition. In *Interspeech*, 2020.

[19] Xingchen Song, Guangsen Wang, Zhiyong Wu, Yiheng Huang, Dan Su, Dong Yu, and Helen Meng. Speech-xlnet: Unsupervised acoustic model pretraining for self-attention networks. *arXiv preprint arXiv:1910.10387*, 2019.

[20] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] S Shahnawazuddin, Waquar Ahmad, Nagaraj Adiga, and Avinash Kumar. In-domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7554–7558. IEEE, 2020.

[22] Peiyao Sheng, Zhuolin Yang, and Yanmin Qian. Gans for children: A generative data augmentation strategy for children speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 129–135. IEEE, 2019.

[23] Jessica E Huber, Elaine T Stathopoulos, Gina M Curione, Theresa A Ash, and Kenneth Johnson. Formants of children, women, and men: The effects of vocal intensity variation. *The Journal of the Acoustical Society of America*, 106(3):1532–1542, 1999.

[24] Suco Eguchi. Development of speech sounds in children. *Acta Otolaryngol*, 257:1–51, 1969.

[25] James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler. Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97(5):3099–3111, 1995.

[26] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[27] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[28] Daniel Povey. *Discriminative training for large vocabulary speech recognition*. PhD thesis, University of Cambridge, 2005.

[29] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747, 2018.

[30] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence trained neural networks for asr based on lattice free mmi (author's manuscript). Technical report, The Johns Hopkins University Baltimore United States, 2016.

[31] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays. Fast and accurate recurrent neural network acoustic models for speech recognition. *arXiv preprint arXiv:1507.06947*, 2015.

[32] Desh Raj. Blog on lattice free mmi and chain models. `https://desh2608.github.io/2019-05-21-chain/`, last accessed: January 2020.

[33] Kaldi. Chain Model Doc. `https://kaldi-asr.org/doc/chain.html`, last accessed: October 2020.

[34] Kaldi. Decoding using a chain model Doc. `https://kaldi-asr.org/doc/chain.html#chain_decoding`, last accessed: October 2020.

[35] Luís Felipe Uebel and Philip C Woodland. An investigation into vocal tract length normalisation. In *Sixth European Conference on Speech Communication and Technology*, 1999.

[36] Kaldi. Applying feature transforms. `https://kaldi-asr.org/doc/transform.html#transform_apply`, last accessed: October 2020.

[37] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 55–59. IEEE, 2013.

[38] J-L Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE transactions on speech and audio processing*, 2(2):291–298, 1994.

[39] Yoo Rhee Oh and Hong Kook Kim. Mllr/map adaptation using pronunciation variation for non-native speech recognition. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 216–221. IEEE, 2009.

[40] Patrick Nguyen, Roland Kuhn, Jean-Claude Junqua, Nancy Niedzielski, and Christian Wellekens. Eigenvoices: a compact representation of speakers in model space. In *Annales des télécommunications*, volume 55, pages 163–171. Springer, 2000.

[41] Pegah Ghahremani, Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur. Investigation of transfer learning for asr using lf-mmi trained neural networks. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 279–286. IEEE, 2017.

[42] Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur. Jhu kaldi system for arabic mgb-3 asr challenge using diarization, audio-transcript alignment and transfer learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 346–352. IEEE, 2017.

[43] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104, 2009.

[44] Reza Sahraeian and Dirk Van Compernolle. Using weighted model averaging in distributed multilingual dnns to improve low resource asr. *Procedia Computer Science*, 81:152–158, 2016.

[45] František Grézl, Ekaterina Egorova, and Martin Karafiát. Study of large data resources for multilingual training and system porting. *Procedia Computer Science*, 81:15–22, 2016.

[46] Shucong Zhang, Cong-Thanh Do, Rama Doddipatla, and Steve Renals. Learning noise invariant features through transfer learning for robust end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7024–7028. IEEE, 2020.

[47] Prashanth Gurunath Shivakumar and Panayiotis Georgiou. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer speech & language*, 63:101077, 2020.

[48] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020.

Appendix A

Miscellaneous information about transfer learning experiments

To use the IITM baseline model for retraining experiments, the phone sets used in the baseline model and the ASER UP and RJ transcription data must be mapped. A mapping between the ASER transcription phone set to the IITM phone set is derived manually by looking at examples in both the lexicons. Many of the mappings are straightforward but for some phones in the ASER phone set multiple phones in the IITM phone set had to be used. For example, 'oy' (as used in the word 'Voice' with pronunciation 'w oy s') in the ASER phone set is mapped to two phones in the IITM phone set 'ou y'.

Another change made to the lexicon is providing alternate pronunciations to nasalized words. This is done because the IITM phone set has two nasal phones 'q' and 'n'. To make use of all the phone information in the IITM phone set and by extension the baseline model, nasalized words are made to have two pronunciations. This was done automatically by identifying the unicode values of nasalized words/vowels in the lexicon. The complete mappings can be found here[1].

The other labels and tags made in the transcriptions of the ASER set are treated as words in the lexicon with SIL (silence) phone as its pronunciation. This is to provide better alignments between the ASER retraining sets and their transcriptions. There are 47 speech phones in the ASER transcription lexicon all of which are mapped to the 57 speech phones in the IITM phone set. The non-speech phones in the ASER transcription (ON, FP, BR etc.) are mapped to the single SIL phone in the IITM set which is it's only non-speech phone. While decoding all these phones are filtered out.

[1]Mapping of phones: https://rb.gy/jcjr90

Another point in decoding is treating the LPR and HPR recordings as separate data folders in kaldi. This optimizes the LMWTs and WIPs for both the sets and provides better results. The LMWT of the HPR recordings is usually much higher compared to the LMWT of the LPR recordings. To tune the retraining parameters while doing transfer learning, the validation loss is used. The parameters (number of epochs, learning rates, regularization factors) are adjusted by getting the best possible loss with the setting of freezing the 13 lower layers and tuning the final two output layers (0.0(13)-1.0(2)). Then, further experiments are carried out by modifying the differential learning rates only.