Teacher-Student learning based Acoustic Models for Robust Speech Recognition in Naturalistic Childhood Classroom settings.

Prasanna V. Kothalkar, John H.L. Hansen[1]

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas, Richardson, Texas, USA
{prasanna.kothalkar, john.hansen}@utdallas.edu

Dwight Irvin, Jay Buzhardt

Juniper Garden's Children's Project (JGCP)
University of Kansas, Kansas, USA
{dwight.irvin, jay.buzhardt}@ku.edu

Abstract

Speech and language development are early indicators of overall analytical and learning ability in children. The preschool classroom is a rich language environment for monitoring and ensuring growth in young children by measuring their vocal interactions with both teachers and classmates. Early childhood researchers recognize the importance in analyzing naturalistic vs. controlled lab recordings to measure both quality and quantity of child interactions. Recently, large language model-based speech technologies have performed well on conversational speech recognition. In this regard, we assess performance of such models on the wide dynamic scenario of early childhood classroom settings. This study investigates an alternate Deep Learning-based Teacher-Student learning solution for recognizing adult speech within preschool interactions. Our proposed adapted model achieves the best F1-score for recognizing most frequent 400 words on test sets for both classrooms. Additionally, F1-scores for alternate word groups provides a breakdown of performance across relevant language-

[1]This work was funded by the National Science Foundation Grant 131602.

based word-categories. The study demonstrates the prospects of addressing educational assessment needs through communication audio stream analysis, while maintaining both security and privacy of all children and adults. The resulting child communication metrics from this study can also be used for broad-based feedback for teachers.

Keywords: Child-Adult Speech Interactions, Naturalistic Speech Recognition, End-to-end Speech Recognition, Transformer subword, Teacher-Student Learning, Preschool Classroom

1. Introduction

The diversity of language background, socio-economic conditions, development level, and potential communication disorders represent challenges in assessment of child speech and language skills [1]. The language environment of young children plays an important role in development of speech, language, vocabulary and thus, knowledge/learning ability. Taken collectively, these impact life prospects of the child. The quality and quantity of interaction given a rich language environment helps meet essential language development outcomes in early childhood [2]. Thus, early childhood researchers recognize the need for analyzing classroom interactions of preschool children to monitor and provide proactive support.

In classroom settings, teachers prompt interaction by asking questions that engage child's curiosity and experimentation, particularly in science-focused activities[3]. Therefore, tracking sentences with such WH-question words or WH-words [4, 5] (what, where, when, who, why, how) can help teachers review their interactions with the children. Furthermore, WH-words represent the questions, which can be analyzed in terms of frequency of occurrence within the classroom. Active learning [6] practices for preschool children recommends learning through direct interaction with practical, everyday objects that can be handled by the child. Adults are recommended to encourage children to manipulate such objects for discovering the relationship among them, by actively

engaging all senses and muscular utilization. Thus, tracking such object words from these audio files can help educators asses their interactions with children during such 'play' or 'object interaction' sessions, as well as keeping a record of objects introduced to the children. Finally, tracking words that promote responses from children in the form of 'repetition' or 'high word-count response' can help researchers understand qualitative engagement due to 'teaching style' [7], 'concerned activity' and/or 'topic of interest'.

For this purpose, the authors have collected multi-session dataset in a real preschool during their daily activities. A typical preschool is composed of separate areas for specific activities to be conducted during alternate times of the day as seen in Fig. 1. Due to the massive amounts of daylong audio recordings, it is not feasible for humans to manually perform analysis. Automatic speech recognition(ASR) offers the prospect of extracting text content from a conversational speech signal represented as the transcript of the speech. Recently, ASR[8, 9] and machine learning [10, 11] techniques have been utilized for automated processing and analysis of child-centered data. Previously, diarization[12, 13, 14] and children's speech recognition[9] have been explored on such data. Previous study of adult speech recognition [8] on this dataset focused on limited types of keywords in location-based fashion. This study will focus on acoustic and language analysis as well as automated solutions for word-detection and word-counting, for adult speech in naturalistic conditions of the preschool classrooms.

There are a range of potential real-time, portable, voice recording platforms as well as recent algorithms for word-counting [15, 16], speaker diarization, end-to-end (E2E) small footprint ASR models for word-detection and/or word-counting. However, these solutions for classroom conditions have not utilized state-of-the-art (SoTA) advancements in ASR models. Thus, small-footprint E2E ASR models with strong performance for word-detection and/or word-counting in naturalistic preschool classrooms are desired. The contributions of this study are stated as follows. First, we introduce the speech database for child-adult interaction based on a North American preschool classroom and analyze conversational acoustical variations for multiple classrooms. Next,

groundtruth transcriptions generated by the CRSS transcription team are utilized for performing classroom-specific qualitative analysis on vocabulary diversity and the language interactions between adults and children. This includes distribution word analysis based on parts-of-speech(POS) as well as statistics of adult words relevant for 'high child engagement' with teachers.

This study also provides insight on performance of SoTA end-to-end (E2E) ASR models on out-of-domain data in challenging preschool classroom conditions, as well as Teacher-Student (T/S) learning-based model adaptation to improve ASR performance. Classroom speech data is naturalistic, with far-field adult speech , where performance is measured in terms of Word Error Rate (WER).

The dataset is a classroom-independent scenario, such that pretrained ASR models adapted on one classroom condition, are used for testing on speech segments from the other classroom. This will be the first effort on this dataset to explore data splits from alternate classrooms, thus allowing for model adaptation performance comparison.

2. Outline

The following is an overview of this paper which starts with Section 3. Section 3 reviews the previous work on naturalistic speech recognition using E2E ASR models. Section 4 provides details of the dataset (including classrooms) and analysis of the audio (acoustic) and text transcriptions (language). Acoustical analysis of audio from alternate classrooms includes Signal-to-Noise Ratio (SNR) measures. Language analysis of audio transcripts includes word-specific child-response statistics in the alternate classrooms. Section 5 explains the procedure for producing the ASR output from raw audio followed by details of the ASR model and proposed T/S model adaptation strategy. Section 6 talks about the experimental design and the metrics used for evaluating the experiments, while we look and discuss the results in Section 7, followed by conclusions and future work in Section 8.

3. E2E ASR in Naturalistic conditions

E2E models [17] are preferred over traditional hybrid models (comprising of individual language, acoustic and lexical model components) due to serval advantages related to implementation, computation and deployment. E2E models use a single objective function, while traditional hybrid models optimize individual components separately, which cannot guarantee global optimization. Also hybird models are complicated, and require expert knowledge of many components to design such systems with each component having its own research sub-field. E2E models have been shown to outperform traditional hybrid models[18, 19]. E2E models directly output either a character sequence, word pieces [20] (subwords) or whole words, which greatly simplifies the ASR pipeline. Also, E2E models can be deployed on devices with high accuracy because a single network is used for ASR.

Although ASR performance has surpassed human-level recognition for standard speech recognition datasets in both hybrid as well as E2E speech recognition models, these models do not generalize well to unseen, out-of-domain data. This is especially, the case for naturalistic, noisy, far-field conditions, which are challenging for both hybrid and E2E models. Recently, self-supervised speech representations trained on massive amount, of unlabelled data are finetuned to specific data for E2E ASR, providing promising results. As an example, the Wave2Vec [21] family of models use a task similar to a Masked Language Model to pre-train a network using unlabelled speech before fine-tuning on the specific ASR task. This allows the network to learn contextual speech embeddings that can then be finetuned using parallel speech and transcripts corpora. Similarly, multi-domain supervised data has also been trained on large deep neural networks to develop a general model. In transfer learning, such general knowledge is transferred via fine-tuning on a downstream task, which is typically low resource. Thus, robust ASR in naturalistic conditions such as CHiME-6 dataset[22], or Fearless Steps Corpus[23] utilize large supervised models[24] trained on massive amounts of supervised data or self-supervised models for

current best performance by fine-tuning on specific dataset.

Recently, OpenAI released pretrained models referred to as 'Whisper' where 'WSPSR' acronym stands for Web-scale Supervised Pretraining for Speech Recognition. However, this family of models (of alternate sizes) are the largest supervised ASR models [25] in terms of training data for English (438k hours) as well as multilingual (680k hours). The study [25] utilizes 'LibriSpeech' as the reference dataset due to its widespread utilization in modern speech recognition research and the availability of many released models trained on it. This allows for characterizing robustness behaviors by studying out-of-distribution performance on 12 other speech recognition datasets. The Whisper models do not outperform the SoTA for ASR on the standard book reading 'Librispeech' dataset. However, even the smallest Whisper model is competitive with the high-performance SoTA model over 'Librispeech', on multiple other datasets including CHiME-6. Thus, Whisper model provides zero-shot competitive performance without additional finetuning and suggested to provide generalization. Hence, we utilize Whisper ASR model for zero-shot (i.e. without training) evaluation on test set of Kentucky preschool adult speech corpus. However, since there is significant scope for improvement for Whisper on naturalistic corpora like CHiME-6 or our Kentucky adult speech corpus (as shown in Results section), we perform adaptation to our naturalistic corpus using novel training strategies. The improvement gains are suggested to be due to the acoustical variations of the real classroom conditions, which are unlikely to be present in massive audio data collected over the internet for Whisper or other traditional ASR datasets. Apart from classroom acoustics, the novel acoustic variations to be learned by the model include 'Child sounds' during partially audible words from other children in the classroom or loud noises while the children are playing.

4. Data specifics

4.1. Data collection

Dataset used in this study consists of spontaneous conversational speech recorded with the LENA units attached to subjects in a high quality childcare learning center in the United States. Daylong audio recordings consist of 54 preschool daylong audio files across 3 days in 7 sessions in 2 classrooms (Class A or Class B).

4.2. Classroom details

Data collected using LENA recorders in two classrooms have multiple child work stations. These learning station activities include reading, blocks, play, singing, science etc (see Fig. 1). The physical dimensions and layout of the two classrooms are different, which may affect the recorded audio in terms of reverbration and noise.

Classroom A is 24 ft by 24 ft while Classroom B is larger with dimensions of 24 ft by 40 ft. (e.g. see floor plan in Fig. 1). Algorithm performance will be explored for alternate classroom scenarios by training on Class A and test on Class B, and vice versa.

4.3. Dataset distributions

Adult speech segments are extracted from daylong, naturalistic noisy audio recordings in preschool classrooms having 3 to 5 year olds accompanied by one or more adults (e.g. typically teachers). Most children wear LENA devices along with 1-3 adults in the classroom. Depending on distance of adult from the LENA recorder worn by the child, the audio files recorded by the corresponding LENA recorder can be near or far-field.

Audio files are divided into train, development and test sets following classroom-based division. Audio data corresponding to either Class A or Class B are used for training alternate models. Data from the alternate classroom is used for model development and test. During model development, a separate hold-out

development set, is used to find the best operational model during neural network training.

The training set has 11 hours and 32 minutes (21505 utterances) of adult speech from classroom A and 7 hrs and 20 minutes (14027 utterances) of adult speech from classroom B. The test set has 1 hour and 41 minutes (2076 utterances) of adult speech from classroom A and 1 hour and 53 minutes (3340 utterances) of adult speech from classroom B.

4.4. Dataset acoustic variability

To understand the dataset acoustic diversity for both classrooms, we measure SNR values of 1-sec segments from the complete audio and plot the probability density function (pdf) of audio segments Vs. classroom as seen in Fig. 2. It can be seen that there are a lot more segments with lower SNR in Classroom B as compared to Classroom A.

4.5. Dataset language variability

4.5.1. Average child word counts in response to adult words

By analyzing text transcripts, we calculate the number of words spoken by children in response to a given adult statement. We assign the total child word count in such responses, to each word in the adult statement. By taking the sum of child word counts per adult word, across multiple adult-child turns, we are able to compile the total and average child word count, in response to all words spoken by adults. The words from adult statements, followed by the highest average child word counts per turn, are listed in Tables 1 and 2 for Class A and Class B respectively. It helps us understand adult words that produce good engagement with children or words that are produced during high engagement interactions. These could reflect 'housekeeping' interaction words like wear, hold, talk, keep, look, play, find etc., encouragement words such as awesome, great, friend etc., or topic specific keywords such as rocks, sea, turtles, christmas, rat, dough etc.

4.5.2. Average child word counts in child responses with repeated adult words

In this study, we also analyze average child word counts per turn for child statements that contain a repeat of an adult word, from the preceding adult statement. Here, we calculate the total and average child word counts in same way as Sec. 4.5.1, but require the repetition of an adult word by the child. Tables 3 and 4 summarize such words in the test set of Class A and Class B respectively. These include cardinal digits such as one, two, three, five etc. , action words: play, take, go, please, thank, hurray, oustide or topic keywords: hand, quesadilla, paper, bubble, chain, monster, ketchup, pack, teachers etc.

5. Method

5.1. System pipeline

Fig. 3 presents the high-level system diagram for end-to-end adult speech recognition task. It begins with the data collection module using our LENA device in preschool classroom. Once collected, this data is transcribed by CRSS transcription team for use in system development for recognizing speech in this naturalistic learning space. After data segmentations, adult speech segments are used to adapt the pretrained SoTA ASR models using the training set from alternate classrooms. The ASR model adapted on training data for a given classroom are tested on the development set of the other class. The best performing epoch is then used for open evaluation using the test set for ASR performance assessment.

5.2. Multihead Attention-based Transformer model

The Attention Encoder-Decoder (AED) model was first introduced in [26] for neural machine translation. Without any conditional independence assumption as in CTC [27], AED was successfully applied to E2E ASR [28, 29] and has recently achieved superior performance to conventional hybrid systems [20].

AED directly models the conditional probability distribution $P(\mathbf{Y} \mid \mathbf{X})$ over the sequence of output tokens $\mathbf{Y} = \{y_1, \ldots, y_L\}$ given the sequence of input speech frames $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ as:

adult word	number of independent turns of child response	total responded child words across independent turns	avg. child words per turn
wear	5	32	6.4
words	5	31	6.2
talk	6	36	6.0
thank	21	113	5.4
listening	5	26	5.2
awesome	9	47	5.2
hold	17	81	4.8
great	7	33	4.7
time	9	42	4.7
us	8	37	4.6
friend	11	50	4.5
button	6	27	4.5

Table 1: Classroom A statistics of counts of words spoken by children, in response to words spoken by adults within sentences, as transcribed from audio in test set.

adult word	number of independent turns of child response	total responded child words across independent turns	avg. child words per turn
sitting	11	62	5.6
blocks	8	45	5.6
turn	6	33	5.5
keep	6	32	5.3
find	7	37	5.3
christmas	5	26	5.2
rat	5	26	5.2
minutes	5	25	5.0
hold	14	68	4.9
quiet	8	38	4.8
mat	6	29	4.8
dough	5	24	4.8
look	10	47	4.7
think	17	79	4.6
rocks	5	23	4.6
line	12	54	4.5
sea	6	27	4.5
turtles	6	27	4.5

Table 2: Classroom B statistics of counts of words spoken by children, in response to words spoken by adults within sentences, as transcribed from audio in test set.

adult word repeated by child	number of independent turns of child response	total responded child words across independent turns	avg. child words per turn
one	7	26	3.7
going	5	33	6.6
need	5	34	6.8
two	5	13	2.6
play	4	24	6.0
keep	3	21	7.0
got	3	15	5.0
take	3	12	4.0
hand	2	20	10.0
want	2	11	5.5
pack	2	7	3.5
outside	2	10	5.0
go	2	13	6.5
quesadilla	2	15	7.5
five	2	7	3.5
right	2	4	2.0

Table 3: Classroom A statistics of counts of words repeated by children, in response to words spoken by adults within sentences, as transcribed from audio in test set.

adult word repeated by child	number of independent turns of child response	total responded child words across independent turns	avg. child words per turn
play	5	29	5.8
chocolate	4	5	1.3
please	4	10	2.5
dough	4	22	5.5
high	3	12	4.0
hurray	3	5	1.7
paper	3	15	5.0
rat	3	18	6.0
little	3	13	4.3
line	3	6	2.0
bubble	3	9	3.0
thank	2	4	2.0
morning	2	6	3.0
minutes	2	16	8.0
teachers	2	8	4.0
chain	2	8	4.0
monster	2	10	5.0
ketchup	2	7	3.5

Table 4: Classroom B statistics of counts of words repeated by children, in response to words spoken by adults within sentences, as transcribed from audio in test set.

$$P(\mathbf{Y} \mid \mathbf{X}) = \prod_{l=1}^{L} P\left(y_l \mid \mathbf{Y}_{0:l-1}, \mathbf{X}\right).$$

Vaswani et. al. [30] proposed encoder-decoder architectures referred to as Transformers based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. At each step the model is auto-regressive, i.e. it consumes the previously generated tokens as additional input when generating the next most likely token. To achieve this, the Transformer follows the AED architecture using stacked six layers for both encoder and decoder with each layer having sublayers. The sublayer consists of Multi-Head Self-Attention (MHSA) mechanism and position-wise, fully connected feed-forward network. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs Multi-Head Cross-Attention (MHCA) over the output of the encoder stack.

The attention mechanism acts as a mapping function to generate the output vector which is a weighted sum of the values given a set of query, key and value vectors. It can be described as mapping a query and a set of key-value pairs to an output. The weight assigned to each value is computed as a compatibility function of the query with the corresponding key.

5.3. Whisper acoustic model

Basic Transformer model was utilized in Whisper to evaluate ASR system performance, irrespective of model enhancements and only due to availability of large amount of training data. Unlike other recent speech recognition models, the Whisper model is completely trained using supervised learning on weakly labelled data. The resulting network model was constructed based on 680k hours of multilingual speech and transcribed data from across 96 languages. The resulting network is an Encoder-Decoder transformer trained using multi-task learning with tasks that include transcription, translation and timestamp prediction. Here, a basic transformer is used to evaluate the ASR system performance solely due to availability of this large amount of training data, irrespective of model enhancements.

The 'medium' (769 M parameters) Whisper model trained on English-only corpora, has the best performance in terms of WER on the publicly-available naturalistic ASR corpus of CHiME-6, among the family of Whisper ASR models. Here we consider the 'small' (244M parameters) and 'medium' (769 M parameters) size network models trained for English language data to measure their zero-shot performance on adult speech for our preschool corpus. Both these models have similar performance with the 'small' version having 2.3% higher WER than 'medium' model. The 'small' model is approximately 2 GB in VRAM and is 6 times faster than the 'large' (1550 M parameters) model which needs 10 GB of VRAM. Thus, the 'small' model is selected based on a future goal of operation on small portable devices.

5.4. Teacher-Student Learning for adult speech recognition in preschool classroom

Simply finetuning Whisper models on our Kentucky corpus resulted in worse performance than using the models themselves. This was most likely due to catastrophic forgetting [31, 32] of well-trained transformer models when trained on comparatively smaller dataset of novel acoustic variations of the preschool classroom corpus. Functional approaches to catastrophic forgetting add a regularization term [33] to the objective that penalizes changes in the input-output function of the neural network. This takes a form of knowledge distillation such that predictions [34] (or final layer hidden activations [35]) of the previous task's neural network are encouraged to be similar to current network when trained on data from new task. Thus knowledge distillation has been utilized for improving training performance [36, 37] irrespective of model compression outcomes.

Knowledge distillation [38, 39] helps the training process of "student" networks by distilling knowledge from one or multiple well-trained "teacher" networks. The key here is to leverage the soft probability outputs, of teacher networks, where incorrect-class assignments reflect how a teacher network generalizes from previous training. By mimicking probabilities output, the student network is able to incorporate the knowledge that the teacher network discov-

ered earlier, allowing the performance of the student network to be better than if it were trained with labels only.

Hinton et al.[38] introduced "softmax temperature" function $\sigma_{\tau_s}(\cdot)$ to produce a softer probability distribution output when a large temperature τ_s (usually greater than 1) is picked. Since it takes logits from final layer as input, it decays to normal softmax function $\sigma(\cdot)$ when τ_s equals 1. The function value for the i^{th} instance from the dataset $(\mathbf{X}, \mathbf{Y})$ can be calculated as follows:

$$\sigma_{\tau_s}(x_i) = \frac{\exp(x_i/\tau_s)}{\sum_{x_j \in \mathbf{X}} \exp(x_j/\tau_s)} \tag{1}$$

Next, we present the formulation of Teacher-Student (T/S) learning for ASR on the preschool corpus. Given a pre-trained teacher network $f_{\theta_T}(\cdot)$ and a student network $f_{\theta_S}(\cdot)$, where θ_T and θ_S denote the network parameters, the goal of knowledge distillation is to force the output probabilities of $f_{\theta_S}(\cdot)$ to be close to that of $f_{\theta_T}(\cdot)$. Let (x_i, y_i) denote a training sample in dataset $(\mathbf{X}, \mathbf{Y})$ of sequence of L output tokens and $\mathbf{Y}$ containing sequence of N input speech frames. $P_{f_\theta}(x_i)$ indicate the logit response of x_i from $f_\theta(\cdot)$. The student network f_{θ_S} can then be learned by the following relation:

$$\begin{aligned} min_{\theta_S} \sum_{(x_i,y_i)\in\mathbf{X},\mathbf{Y}} & \alpha \times \tau_s^2 \times \mathbf{KL}\left(\sigma_{\tau_s}\left(P_{f_{\theta_T}}(x_i)\right), \sigma_{\tau_s}\left(P_{f_{\theta_S}}(x_i)\right)\right) \\ & + (1-\alpha) \times \mathbf{CE}\left(\sigma\left(P_{f_{\theta_S}}(x_i)\right), y_i\right), \end{aligned} \tag{2}$$

where $\mathbf{KL}(\cdot,\cdot)$ and $\mathbf{CE}(\cdot,\cdot)$ are the Kullback-Leibler divergence (K-L divergence) and cross-entropy loss, respectively. Another hyper-parameter α is utilized to balance the solution between T/S loss and cross entropy loss, which performs well when the weight for T/S loss is higher.

6. Experimental Design and Evaluation metrics

Evaluations are performed for both the development and test sets to allow for a comparison of the models as formulated in Sec. 4.3.

6.1. Experimental Design

The input to the Whisper models consist of 80-channel log-magnitude Mel-spectrogram representation, computed on 25-millisecond windows with a stride of 10 milliseconds. Prior to the transformer block, 2 convolution layers with a filter of width 3 are applied followed by GELU activation function and Sinusoidal position embeddings. Byte-level Byte Pair Encoding text tokenizer used in GPT-2 [40] is used as tokenizer for generating token ids from text transcripts as well as predicted ASR transcripts.

The 'small' size model is utilized for both the teacher and student models to evaluate WER performance as seen in Fig. 4. This is done to evaluate improvement solely due to adaptation procedure, rather than model size differences. Studies[36, 37] have confirmed T/S learning based improvement for same size teacher and student models. Such improvements are attributed to label-smoothing regularization.

In our case, every parameter from the decoder block consisting of MHA and feedforward network layers are adapted during the training procedure. The student model is utilized for evaluation on development and test sets and is referred to as 'small-adapt.' for the rest of the paper.

The adaptation procedure is carried our for 12 epochs having batch size 64, with best performing model on development set utilized for testing on evaluation set. Based on empirical performance, α value is set to be 0.9 and τ_s value is set to be 4.

Results are reported for both development and test sets for both models as explained in Sec. 7.1 and 7.2. The results are reported in terms of WER, F1-score and Root Mean Squared Error (RMSE). These metrics are explained as follows:

6.1.1. F1-score for word group detection

To understand word detection capability in topic groupings and diverse acoustic conditions, the proposed models are evaluated based on classroom specific test data. Performance 'accuracy' is defined as the total number of

occurrences of all the words in a given word group that are predicted correctly. 'Precision' is the fraction of relevant instances among all detected word occurrences. These would be the fraction of word predictions from ASR output for a given group, that have been predicted correctly on comparing with text transcripts.

$$Precision_{word-group} = \frac{TP_{word-group}}{TP_{word-group} + FP_{word-group}} \tag{3}$$

where TP represents True Positives and FP represents False Positives.

'Recall' is defined as the fraction of relevant instances that were actually detected. Here, these are the fraction of occurrences of words within a 'word group' from the text transcript, that were predicted correctly as per ASR predictions.

$$Recall_{word-group} = \frac{TP_{word-group}}{TP_{word-group} + FN_{word-group}} \tag{4}$$

where TP represents True Positives and FN represents False Negatives.

Next, the F1-score is defined as the harmonic mean of the precision and recall, and takes both precision and recall into account to provide an overall balanced score assessment,

$$F1 - score_{word-group} = \frac{2 \times Precision_{word-group} \times Recall_{word-group}}{Precision_{word-group} + Recall_{word-group}} \tag{5}$$

6.1.2. Root Mean Squared Error (RMSE) for word group detection

RMSE measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well a model is able to predict the target value. A lower RMSE value indicates a better model. For this study, RMSE is used to measure the difference between actual and predicted word counts for the alternate groups of words.

Consider a word group that has 'n' unique words. Let the actual word counts as per text transcriptions for the i-th word in the group be c_i while let the predicted word counts based on ASR predictions be $\hat{c}_i$. Hence, the RMSE

for the word group can be written as per equation 7 below,

$$RMSE_{word-group} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\frac{||c_i - \hat{c_i}||}{\sigma_i}\Big)^2} \tag{6}$$

Adapt on **Train** set of:	Evaluate on **Test** set of:	Model	Subs. Error Rate (%)	Ins. Error Rate (%)	Del. Error Rate (%)	WER **(%)**
No	Room B	small	25.4	**3.7**	17.7	46.8
No	Room B	medium	**22.6**	5.3	15.2	43.1
Room A	Room B	small-adapt	29.2	4.8	**9.0**	**43.0**
No	Room A	small	22.5	4.2	15.2	41.9
No	Room A	medium	**20.1**	**3.5**	13.8	**37.4**
Room B	Room A	small-adapt	25.4	4.9	**8.8**	39.1

Table 5: Word Error Rate results on testing subset recordings of classroom A and classroom B audio.

7. Results and Discussions

7.1. WER

Table 5 reports the WER of small, medium and small-adapt model while applied on test set of adult speech data from both the classrooms. The results for the small and medium model are utilizing the pretrained models of sizes 'small' and 'medium' to test on the test set. Table 5 shows the results in terms of WER for the alternate classrooms are achieved by the proposed 'small-adapt' model which is better than both 'small' and 'medium' size models. This can be analyzed to be due to lowest percentage of errors due to deletions compared to the pretrained models for both the classrooms. Insertion error rate increases marginally for models tested on test sets of both classrooms but is lower than medium-size model for Class A.

Adapt on **Train** set of:	Test on **Eval** set of:	Model	$F1$ (400 most frequent words)	$F1$ (WH words)	$F1$ (Object words)	$F1$ (High engagement words)	$F1$ (Child Repeated words)
No	Room A	small (λ_A)	57.9%	60.8%	62.3%	57.2%	59.5%
No	Room A	medium	**62.8%**	63.8%	**67.1%**	**64.2%**	64.4%
Room B	Room A	small-ad (θ_{BA})	60.3%	**67.4%**	63.2%	60.1%	**65.5%**
$F1$-score(θ_{BA}) - $F1$-score(λ_A):			2.4%	6.6%	0.9%	2.9%	6.0%

Table 6: F1-score results for detecting words in corresponding groups on test set utterances of classroom A audio.

Adapt on **Train** set of:	Test on **Eval** set of:	Model	$RMSE$ (400 most frequent words)	$RMSE$ (WH words)	$RMSE$ (Object words)	$RMSE$ (High engagement words)	$RMSE$ (Child Repeated words)
No	Room A	small (λ_A)	5.2	12.8	2.4	6.6	16.7
No	Room A	medium	4.5	9.5	**1.9**	4.3	14.2
Room B	Room A	small-ad (θ_{BA})	**3.6**	**8.7**	2.4	**3.4**	**7.3**
$RMSE$(θ_{BA}) - $RMSE$(λ_A):			-1.6	-0.4	0.0	-3.2	-9.4

Table 7: Root Mean Squared Error (RMSE) results for detecting words in corresponding groups on test set utterances of classroom A audio.

Adapt on **Train** set of:	Test on **Eval** set of:	Model	$F1$ (400 most frequent words)	$F1$ (WH words)	$F1$ (Object words)	$F1$ (High engagement words)	$F1$ (Child Repeated words)
No	Room B	small (λ_B)	51.5%	55.7%	43.7%	55.7%	51.3%
No	Room B	medium	**56.1%**	**60.2%**	48.7%	**61.6%**	55.3%
Room A	Room B	small-ad (θ_{AB})	54.8%	59.8%	**50.6%**	59.1%	**57.0%**
$F1$-score(θ_{AB}) - $F1$-score(λ_B):			2.0%	4.1%	6.9%	3.4%	5.7%

Table 8: F1-score results for detecting words in corresponding groups on test set utterances of classroom B audio.

Adapt on **Train** set of:	Test on **Eval** set of:	Model	$RMSE$ (400 most frequent words)	$RMSE$ (WH words)	$RMSE$ (Object words)	$RMSE$ (High engagement words)	$RMSE$ (Child Repeated words)
No	Room B	small (λ_B)	9.2	16.8	6.4	7.5	12.1
No	Room B	medium	7.5	**16.3**	5.6	**7.3**	**8.0**
Room A	Room B	small-ad (θ_{AB})	**7.3**	25.3	**5.1**	9.2	9.4
$RMSE(\theta_{AB})$ - $RMSE(\lambda_B)$:			-1.9	9.0	-1.3	1.7	-3.5

Table 9: Root Mean Squared Error (RMSE) results for detecting words in corresponding groups on evaluation subset utterances of classroom B audio.

'Small-adapt' model improves over the 'small' model in terms of deletion error rate by 6.4% and 8.7% for test set of classrooms A and B respectively. 'Small-adapt' model improves over the 'small' model in terms of insertion error rate by 0.7% for test set of classroom A and increases by 1.1% for test set of classroom B. Thus, the adapted model is able to recognize the presence of words in noisy and far-field conditions of the preschool classroom better than the original model of same size. However, the substitution error rate for the adapted model increases by 2.9% over the test set for classroom B and by 3.8% for classroom A, compared to the baseline 'small' model. Thus, although the presence of words in possibly far-field, noisy audio segments is detected accurately, it may detect the complete word or a part of the word (subword) inaccurately. Hence, detection of words with relevance in measuring child-adult interactions, will be of greater importance.

Since the major focus for our task of ASR in challenging preschool classroom scenario is detection of relevant groups of words as highlighted before, we aspire for better detection performance for these words, rather than evaluating WER performance only. To understand the performance achievements for these relevant words, we calculate F1-scores for detecting these groups of words and RMSE scores of the word counts over these groups of words.

7.2. F1-score and RMSE

The alternate word groups to be detected include top 400 most frequently occurring words (excluding stop words), WH-Question words, object words, high engagement words and child-repeated words. Tables 6 and 7 report on the test data from Class A for 'word detection' (in terms of F1-score) and word-counting (in terms of RMSE) respectively, within the alternate word groups . The 'small-adapt' model provides improvement over the baseline 'small' sized model for F1-score metrics on all of the word groups. This is demonstrated by the absolute increase in F1-score values (F1-score(θ_{BA}) - F1-score(λ_A)) in Table 6, for each of the word groups. A lower RMSE implies lower error between the predicted word count and the actual word count. Absolute decrease ($RMSE(\theta_{BA})$ - $RMSE(\lambda_A)$) in RMSE values in Table 7 for majority of the word groups indicates

performance improvement for word-counting as well. The improvement in F1-scores for test set of Class A varies from +0.9% to +6.6% and is better than the 'medium' model for some of word groups as seen in Table 6. Similar improved performance in terms of 'RMSE' is observed for majority of the word groups in the test set of Class A, with error reduction ranging from +0.4% to +9.4% as seen in Table 7.

Tables 8 and 9 report on the test data from Class B for word-detection (in terms of F1-score) and word-counting (in terms of RMSE) respectively, within the alternate word groups. It is seen that the 'small-adapt' model provides improvement over the baseline 'small' model for 'word detection' on all the word groups. The increase in 'F1-score' values (($F1\text{-score}(\theta_{AB})$ - $F1\text{-score}(\lambda_B)$)) and decrease in 'RMSE' values (($RMSE(\theta_{AB})$ - $RMSE(\lambda_B)$) are seen in Tables 8 and 9 respectively for test set of Class B. The improvement in F1-scores for test set of Class B varies from +2.0% to +6.9% and is near or better than the 'medium' sized model for a majority of the word groups as seen in Table 8. Similar improved performance observation in terms of 'RMSE' is also observed for most word groups (except for WH-words) in the test set of Class B, with error reduction ranging from +1.9% to +3.5% as seen in Table 9.

Thus, the adapted version of the 'small' model i.e. proposed 'small-adapt' model is an improved word-counting and relevant word-detection solution with improved scores for both the tasks, on test data from both the classrooms. The performance of 'small-adapt' model is superior to 'medium' sized model on atleast two word groups for both the scores (i.e. F1-score and RMSE) on test data from both the classrooms.

7.3. Parts of Speech Recognition

Percentage distribution of parts of speech (POS) are also reported for the test sets in both the classrooms as per the transcription and the ASR predictions. Figs. 5 and 6 demonstrate the POS tags for text from Class A as per ground truth and ASR prediction respectively. Here, nouns and verbs have error of 2% missing and 2% gain respectively, while preposition and digits have 1% gain and 1% missing respectively based on ASR transcript predictions. Figs. 7 and 8 demonstrate the POS tags for text from classroom B as per ground truth and ASR prediction respectively. Here, nouns and verbs constitute POS for majority of the words in text. Out of this, the maximum error of 4% missing and 3% gained is for nouns and verbs respectively. The remaining 1% is gained by prepositions. Thus, we can see the text from both the

classrooms has a near similar distribution of nouns (52%-53%), verbs (24%-27%), adjectives (6%), prepositions (4%), digits (6%-8%) and adverbs (5%). Also the predicted POS distribution percentages have a maximum error of 4%.

7.4. Audio tags as observable correlates of error rate improvement

In order to uncover attributes related to improvement in significant error types (substitutions and deletions) with largest contributions to word error rate, we extract audio tags for each utterance. In following analyses, references of error rate are composed of substitution and deletion errors. A deep learning model [41] trained on Google's audioset, is utilized as a tool for predicting the most likely audio tags out of 527 audio labels. The outputs predict multiple labels with confidence values (probability) greater than zero.

Relevant labels related to specific aspects of the input audio are style of speech (conversational, narration), speaker-type (male, female, child), singer-type (male, female, child), animal sounds, sources of noise (chatter, crowd, zipper (clothing), writing, television, music, tools, air conditioning, utensils, animal, bird, radio etc.), environmental attributes (inside small room, inside large room, outside in urban environment, outside in rural environment etc.) or actions (shuffling cards, chopping food etc.).

Next, audio labels are investigated within speaker-type, environmental-attribute and noise-source categories across the utterance segments. Utterances that were improved by the small-adapt. model for substitution and/or deletion errors, showed higher percentages of utterances (minimum confidence threshold=0.1) of the test set.

The 'high occurring' audio labels in the improved utterance list along with their corresponding presence for Class A include 'Inside, small room' (+56.3%), 'Child speech' (+23.4%), 'Music' (+8.7%), 'Outside, urban or manmade' (+16.3%) and 'Children playing' (+14.9%). The improved utterances represent +18.3% (for 'Inside, small room' audio label) to +21.1% (for 'Outside, urban or manmade' audio label) of all utterances with labels. Thus, assuming a worst case complete overlap of audio label predictions, +56.3% of utterances with improved error types are indicative of being present with label 'Inside, small room' representing a minimum of +18.3% of the utterances belonging to the corresponding audio labels. Hence, some form of 'Child sounds' are present in 15%-23% of the error-improved utterances.

The 'high occurring' audio labels in the improved utterance list along with respective percentage of presence in test set of Class B includes 'Inside, small room'

(+64.9%), 'Inside, large room or hall' (+34.3%), 'Child speech' (+42.5%), 'Children playing' (+34.7%), 'Outside, urban or manmade' (+21.2%) and 'Music' (+81.7%). The improved utterances included +21.6% (for 'Music' audio label) to +22.3% (for 'Children Playing' audio label) of utterances with these labels. Thus, assuming a worst case complete overlap of these audio label predictions, +64.9% of utterances with these audio labels are present in the utterance list, with improved error types and a minimum of +21.6% of utterances having one of these audio labels show error improvement. So, some form of 'Child sounds' are present in 35%-43% of the error-improved utterances. Also, 'Music' is detected in 82% of error-improved utterances.

Thus, 'Child sounds' are present in higher percent of error-improved utterances of class B. Also, approximately 22% of utterances having 'Child sounds' in test set of Class B show improvement in errors Vs. approximately 20% of utterances having 'Child sounds' in test set of Class A. Thus, Class B has higher percentage of utterances with 'Child sounds' labels that show improvement. This could be due to more noisy conditions based on activities in Class B. Thus, despite the WER for 'small-adapt' model on test set of Class B not showing improvement better than the 'medium' sized model, the improvements are impactful for the unique acoustic attributes of the dataset for given classroom conditions.

8. Conclusions and Future work

In this study, a T/S learning strategy for end-to-end speech recognition on adult speech segments of preschool classrooms was proposed. Initial data analysis was performed for measuring SNR for audio files from Class A and Class B. The analysis showed more audio segments from Class B to have lower SNR compared to Class A. Next, text transcripts of test subsets in both classrooms were analyzed. Words contributing to child-adult engagement and/or learning were grouped to characterize conversational interactions through their statistics of occurrence. Pretrained 'Transformer' models, renowned for state-of-the-art speech recognition performance on out-of-distribution and noisy data, were employed on evaluation test data from two classrooms. T/S learning-based adaptation strategies provided models with improved performance in terms of WER. Recognition of words belonging to distinct categories and corresponding word-counts from them, showed improved performance for the adapted model Vs. pretrained model of the same size, for alternate A Vs. B classroom con-

ditions. Thus, the improved WER performance in terms of deletions, resulted in improvement in performance for important groups of words in our analysis, counteracting any loss in performance due to substitution errors. For future work, it is suggested to explore further the T/S learning strategy with 'medium' size models and perform knowledge distillation to a 'small' size model to evaluate any further performance improvements. Since the scope of this study involved classroom-independent ASR, future work could also include performance evaluation of a diarization system along with ASR. While full recognition of adult-child speech within daily naturalistic classroom settings has been a major challenge in the field, these advancements have shown great promise in providing effective quantitative speech and language metrics for teacher-child conversational engagement. In addition, if there is greater concern regarding privacy, especially for at-risk children, the category based word-counting can offer rich feedback for teachers.

References

[1] S. Rosenbaum, P. Simon, Speech and Language Disorders in Children: Implications for the Social Security Administration's Supplemental Security Income Program., ERIC, 2016.

[2] B. Hart, T. R. Risley, Meaningful differences in the everyday experience of young American children., Paul H Brookes Publishing, 1995.

[3] C. Eberbach, K. Crowley, From seeing to observing: How parents and children learn to see science in a botanical garden, Journal of the Learning Sciences 26 (4) (2017) 608–642.

[4] M. L. Rowe, K. A. Leech, N. Cabrera, Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development, Cognitive science 41 (2017) 162–179.

[5] P. J. Yoder, B. Davies, K. Bishop, L. Munson, Effect of adult continuing wh-questions on conversational participation in children with developmental disabilities, Journal of Speech, Language, and Hearing Research 37 (1) (1994) 193–204.

[6] M. Hohmann, D. P. Weikart, A. S. Epstein, Educating young children: Active learning practices for preschool and child care programs, High Scope Press Ypsilanti, MI, 1995.

[7] H. Wood, D. Wood, Questioning the pre-school child, Educational Review 35 (2) (1983) 149–162.

[8] P. V. Kothalkar, S. Datla, S. Dutta, J. H. Hansen, Y. Seven, D. Irvin, J. Buzhardt, Measuring frequency of child-directed wh-question words for alternate preschool locations using speech recognition and location tracking technologies, in: Companion Publication of the 2021 International Conference on Multimodal Interaction, 2021, pp. 414–418.

[9] R. Lileikyte, D. Irvin, J. H. L. Hansen, Assessing child communication engagement via speech recognition in naturalistic active learning spaces, in: Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, 2020, pp. 396–401.

[10] H. Kaya, O. Verkholyak, M. Markitantov, A. Karpov, Combining clustering and functionals based acoustic feature representations for classification of baby sounds, in: Companion publication of the 2020 international conference on multimodal interaction, 2020, pp. 509–513.

[11] J. Sinclair, E. E. Jang, F. Rudzicz, Using machine learning to predict children's reading comprehension from linguistic features extracted from speech and writing., Journal of Educational Psychology (2021).

[12] M. Najafian, D. Irvin, Y. Luo, B. S. Rous, J. H. L. Hansen, Automatic measurement and analysis of the child verbal communication using classroom acoustics within a child care center., in: WOCCI, 2016, pp. 56–61.

[13] P. V. Kothalkar, D. Irvin, Y. Luo, J. Rojas, J. Nash, B. Rous, J. H. L. Hansen, Tagging child-adult interactions in naturalistic, noisy, daylong school environments using i-vector based diarization system, in: Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education, 2019, pp. 89–93.

[14] P. Kothalkar, J. H. Hansen, J. Buzhardt, D. Irvin, B. Rous, Child vs adult speaker diarization of naturalistic audio recordings in preschool environment using deep neural networks, in: ASEE 2021 Gulf-Southwest Annual Conference, 2021.

[15] J. Gilkerson, J. A. Richards, The lena natural language study, Boulder, CO: LENA Foundation. Retrieved March 3 (2008) 2009.

[16] Y. Wang, M. Hartman, N. A. A. Aziz, S. Arora, L. Shi, E. Tunison, A systematic review of the use of lena technology, American Annals of the Deaf 162 (3) (2017) 295–311.

[17] J. Li, et al., Recent advances in end-to-end automatic speech recognition, APSIPA Transactions on Signal and Information Processing 11 (1) (2022).

[18] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, T. Hayashi, Hybrid ctc/attention architecture for end-to-end speech recognition, IEEE Journal of Selected Topics in Signal Processing 11 (8) (2017) 1240–1253.

[19] J. Li, R. Zhao, Z. Meng, Y. Liu, W. Wei, S. Parthasarathy, V. Mazalov, Z. Wang, L. He, S. Zhao, et al., Developing rnn-t models surpassing high-performance hybrid models with customization capability, arXiv preprint arXiv:2007.15188 (2020).

[20] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al., State-of-the-art speech recognition with sequence-to-sequence models, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4774–4778.

[21] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised pre-training for speech recognition, Proc. Interspeech 2019 (2019) 3465–3469.

[22] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, et al., Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings, in: CHiME 2020-6th International Workshop on Speech Processing in Everyday Environments, 2020.

[23] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, A. Sangwan, The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio, ISCA INTERSPEECH-2019 (2019).

[24] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, M. Norouzi, Speechstew: Simply mix all available speech recognition data to train one large neural network, arXiv preprint arXiv:2104.02133 (2021).

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, arXiv preprint arXiv:2212.04356 (2022).

[26] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[27] A. Graves, Connectionist temporal classification, in: Supervised sequence labelling with recurrent neural networks, Springer, 2012, pp. 61–93.

[28] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, Advances in neural information processing systems 28 (2015).

[29] W. Chan, N. Jaitly, Q. V. Le, O. Vinyals, Listen, attend and spell, arXiv preprint arXiv:1508.01211 (2015).

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[31] I. J. Goodfellow, M. Mirza, A. Courville, Y. Bengio, An empirical investigation of catastrophic forgetting in gradient-based neural networks, stat 1050 (2014) 6.

[32] L. Fu, X. Li, L. Zi, Z. Zhang, Y. Wu, X. He, B. Zhou, Incremental learning for end-to-end automatic speech recognition, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021, pp. 320–327.

[33] F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence, in: International Conference on Machine Learning, PMLR, 2017, pp. 3987–3995.

[34] Z. Li, D. Hoiem, Learning without forgetting, IEEE transactions on pattern analysis and machine intelligence 40 (12) (2017) 2935–2947.

[35] H. Jung, J. Ju, M. Jung, J. Kim, Less-forgetting learning in deep neural networks, arXiv preprint arXiv:1607.00122 (2016).

[36] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, A. Anandkumar, Born again neural networks, in: International Conference on Machine Learning, PMLR, 2018, pp. 1607–1616.

[37] L. Yuan, F. E. Tay, G. Li, T. Wang, J. Feng, Revisiting knowledge distillation via label smoothing regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3903–3911.

[38] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, stat 1050 (2015) 9.

[39] J. Gou, B. Yu, S. J. Maybank, D. Tao, Knowledge distillation: A survey, International Journal of Computer Vision 129 (6) (2021) 1789–1819.

[40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.

[41] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, Panns: Large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2880–2894.

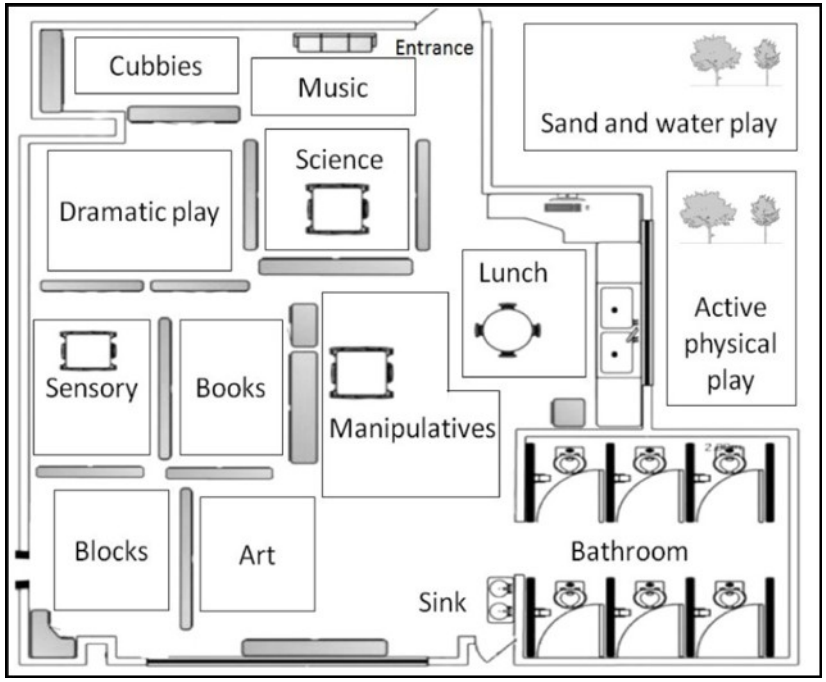

Figure 1: Illustrative example of floor plan for child learning spaces within preschool classrooms. (i.e. learning stations: Books/Reading, Science etc.

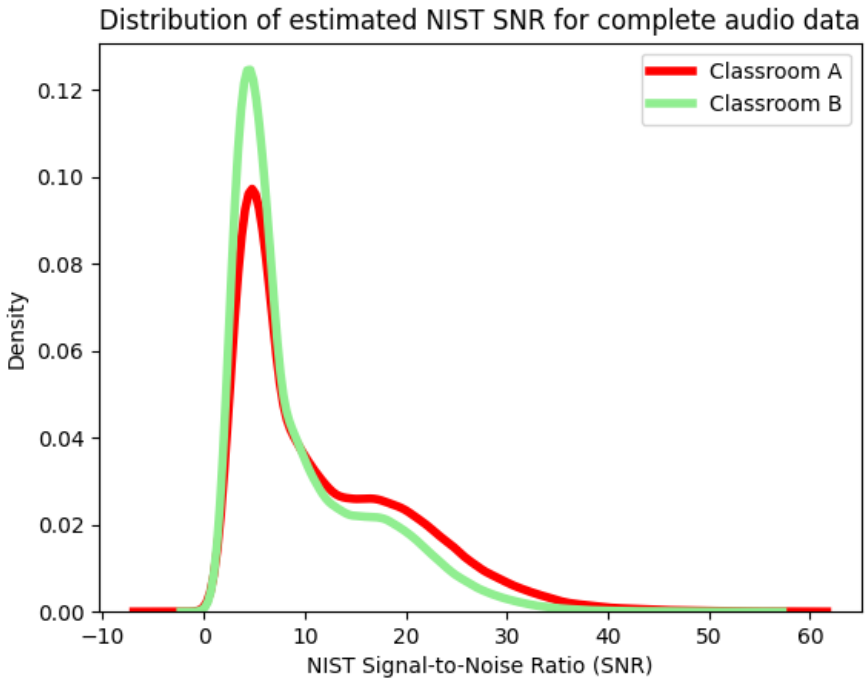

Figure 2: Probability density function of NIST SNR measured over 1 second segments for all audio within classrooms A and B.

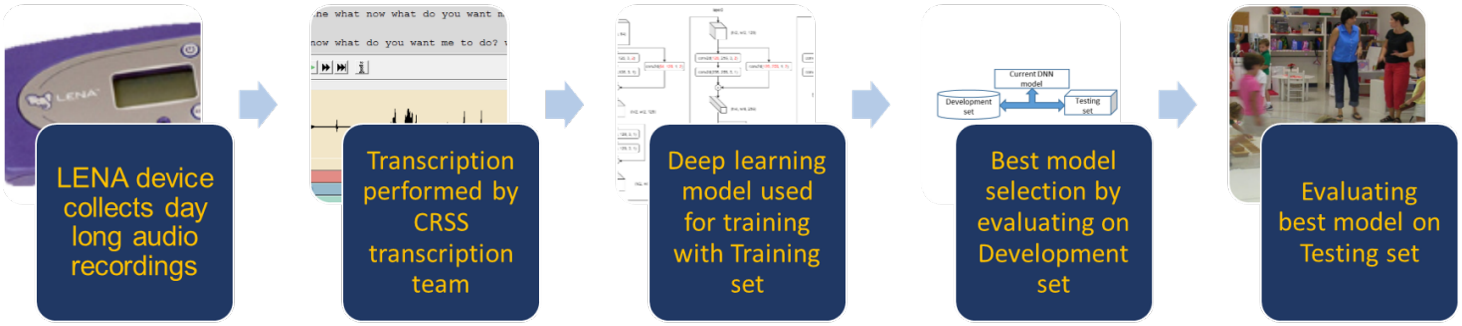

Figure 3: System diagram for end-to-end adult speech recognition system performing model adaptation on the training set, followed by evaluation on test set based on model selection on development set.

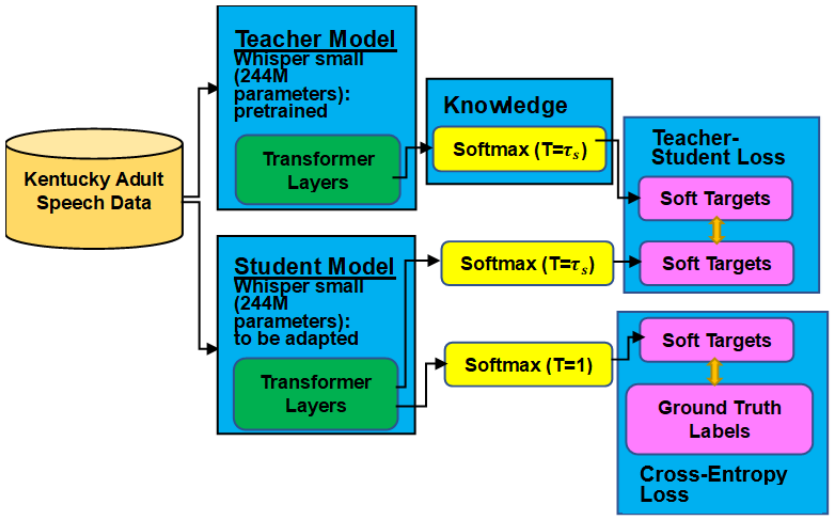

Figure 4: Block Diagram for Teacher-Student Learning on Kentucky adult speech corpus using Whisper small-size models.

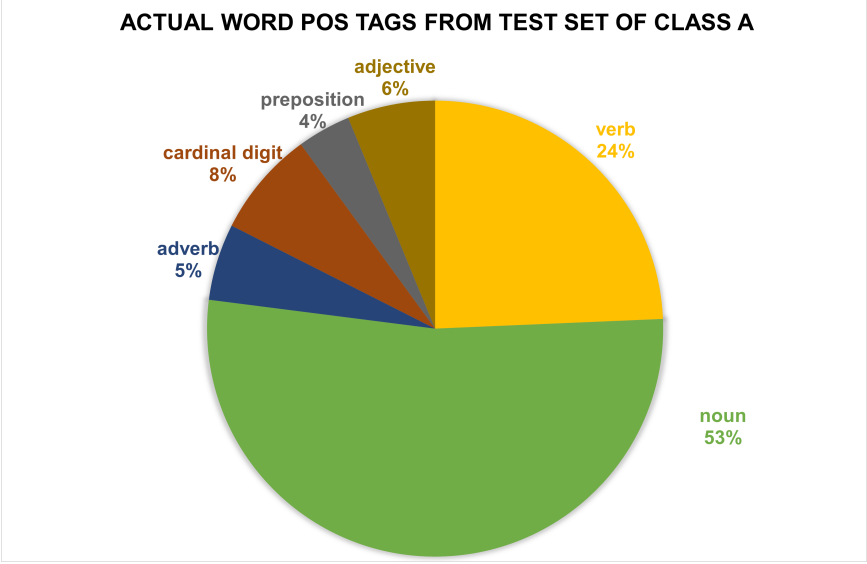

Figure 5: Actual distribution of parts of speech from adult talk transcript as represented by a pie chart for a session in classroom A with a child wearing the LENA device.

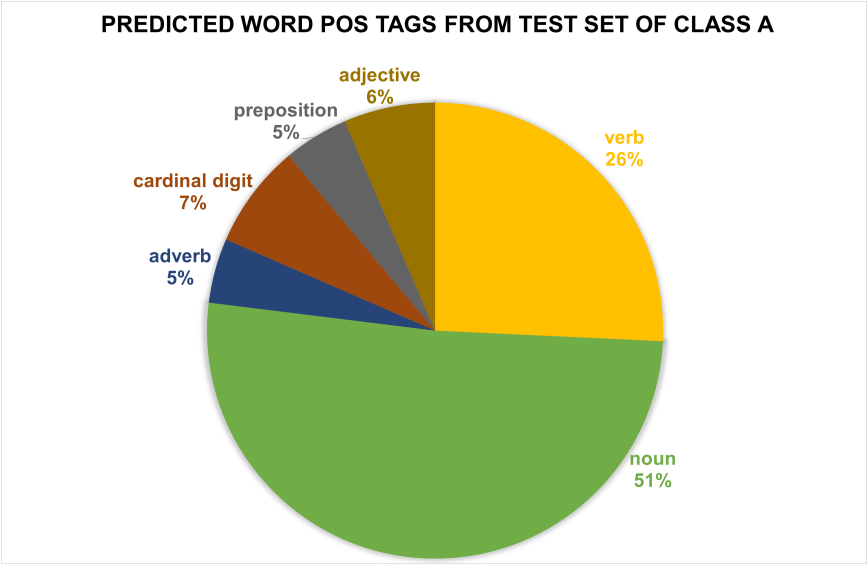

Figure 6: Predicted distribution of parts of speech from adult talk ASR predictions as represented by a pie chart for a session in classroom A with a child wearing the LENA device.

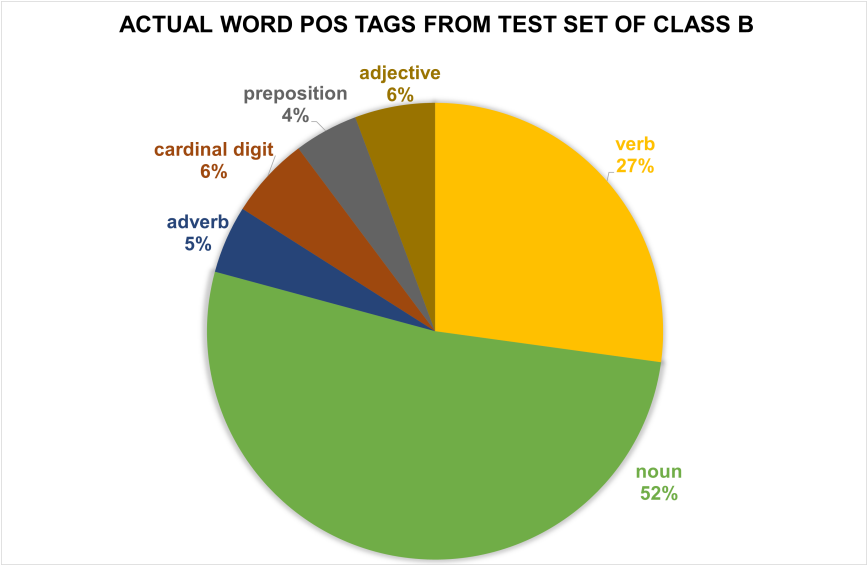

Figure 7: Actual distribution of parts of speech from adult talk transcript as represented by a pie chart for a session in classroom B with both children and adults wearing the LENA device.

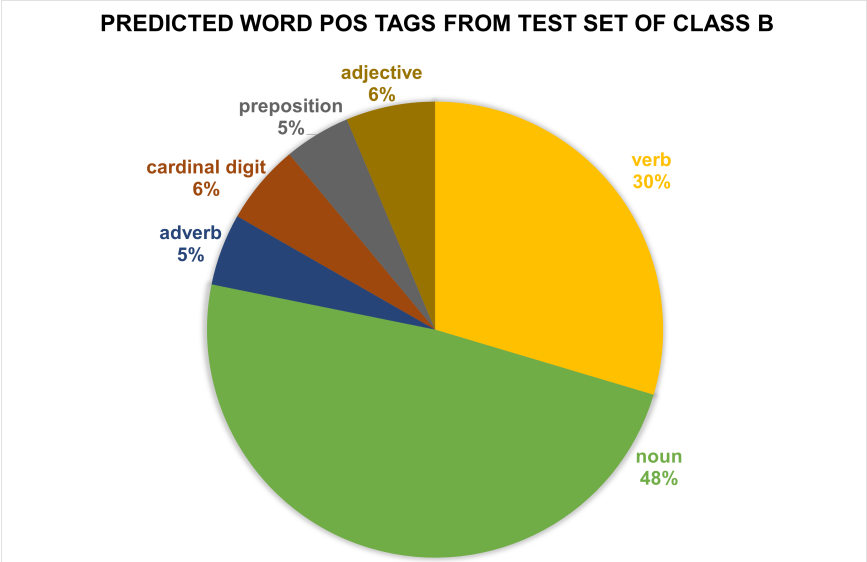

Figure 8: Predicted distribution of parts of speech from adult talk ASR predictions as represented by a pie chart for a session in classroom B with both children and adults wearing the LENA device.