Child-adult speech diarization in naturalistic conditions of preschool classrooms using room-independent ResNet model and automatic speech recognition-based re-segmentation

Prasanna V. Kothalkar[2], John H.L. Hansen[1,2], Dwight Irvin[3], and Jay Buzhardt[3]

[2]*Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, Texas, USA*

[3]*Juniper Garden's Children's Project (JGCP), University of Kansas, Kansas, USA*

[1] John.Hansen@utdallas.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 teachers and classmates. Early childhood researchers are naturally interested in analyzing naturalistic vs. controlled lab recordings to measure both quality and quantity of such interactions. Unfortunately, present-day speech technologies are not capable of addressing the wide dynamic scenario of early childhood classroom settings. Due to the diversity of acoustic events/conditions in such daylong audio streams, automated speaker diarization technology would need to be advanced to address this challenging domain for segmenting audio as well as information extraction. This study investigates an alternate Deep Learning-based diarization solution for segmenting classroom interactions of 3-5 year old children with teachers. In this context, the focus on speech-type diarization which classifies speech segments as being either from adults or children partitioned across multiple classrooms. Our proposed ResNet model achieves a best F1-score of ~78.0% on data from two classrooms, based on dev and test sets of each classroom. It is utilized with Automatic Speech Recognition-based resegmentation modules to perform child-adult diarization. Additionally, F1-scores are obtained for individual segments with corresponding speaker tags (e.g., adult vs. child), which provide knowledge for educators on child engagement through naturalistic communications. 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 have been used for broad-based feedback for teachers with the help of visualizations.

KEYWORDS: Child-Adult Speech, Speech-type Diarization, End-to-end Diarization, ResNet-18, Multiclass classification, location-independent modeling.

PACS: 43.72.-p Speech processing and communication systems

I. Introduction

The diversity of language background, socio-economic conditions, development level, or potential communication disorders represents a challenge in assessment of child speech and language skills **(Rosenbaum and Simon, 2016)**. The language environment of young children plays an important role in the development of speech, language, vocabulary and thus, knowledge/learning ability. Taken collectively, these impact the life prospects of the child. The quality and quantity of interaction in a rich language environment helps to meet essential language development outcomes in early childhood **(Hart and Risley, 1995)**. Thus, early childhood researchers are interested in analyzing classroom interactions of preschool children to monitor and provide proactive support. As daylong recordings are collected on a regular basis, the amount of data to be analyzed keeps increasing at much a faster pace than what is practically feasible to review manually. Automated speech processing would be of great value for understanding and assessing the vast amounts of data in this early childhood domain. The preliminary task of analyzing such data environments involves Speaker Diarization (i.e., segmenting and tagging 'who spoke when') followed by Speech Recognition, Keyword Spotting, etc. In this study, Speaker group (or type) Diarization is performed on child-adult and child-child interactions of preschool children in naturalistic active learning environments. The audio data in this study was collected using LENA devices **(LENA; Ziaei et al., 2013)** worn by children in different classrooms at different days and times. The recordings continue while subjects move around during a typical school day and are paused only during nap time.

The contributions of this study are stated as follows. Firstly, we introduce the child-adult speech/speaker-type classification framework explained later for designing the scope of the speech-segment classification task. Next, standard Deep Neural Network (DNN) architectures are explored for this challenging task of distinguishing children's speech from adult speech and non-speech. Additionally, we analyze classifications of speech segments into alternate speech types in terms of F1-score. The speech/speaker-type detector is integrated with an Automatic

Speech Recognition (ASR) resegmentation module and provides diarized outputs based on different system configurations. Thus, the Diarization Error Rate (DER) is also provided, which helps in understanding the performance achieved by the different speech-type modeling techniques and system configurations. This study would be one of the first efforts for child-adult speech/speaker-type Diarization on a large North American English dataset of child-adult naturalistic recordings in diverse classroom conditions. Previous studies have considered the application of alternate Deep Neural Network architecture embeddings for Child vs. Adult speech-type classification. Deep Neural Network multi-label classification **(Lavechin et al., 2020)** has achieved segment-level classification of child or adult speech detection for diarization which included fine-grained labels like 'key child', 'other child' and generic labels like 'speech' for multitask learning as a general audio-tagging task. A single label for an audio segment can be useful for downstream speech tasks. Moreover, as we are testing on the segment-level audio, the output speech-type classification and ASR resegmentation can be performed in an online fashion **(Xue et al., 2021)** (i.e., every segment can be processed as it is recorded). This has advantages in classroom settings where immediate feedback for teachers/adults can be provided. For offline processing, the entire recording would need to be provided to generate any final output estimated knowledge of the speech segment type.

Additionally, we also divide the dataset in a classroom-independent scenario, such that models trained on one classroom condition are available for testing on audio from another classroom condition. This will be the first effort on this dataset to look at data splits with audio data from alternate classrooms, thus allowing for a statement on model generalization capability. Finally, we introduce a novel visualization diagram referred to as donut diagram which provides speech segment classifications over a period of time, as a feedback mechanism and practical evaluation of our proposed classification models.

II. Outline

The following is an overview of this paper which starts with Sec III mentions the Background including speaker characteristics and child-adult speech diarization. Sec IV introduces our framework for end-to-end child-adult speech/speaker-type classification which includes the assumptions and scope of our problem formulation. Sec V provides details of the dataset. Sec VI explains the procedure for producing the classification from raw audio including steps displayed in Fig 2. Within Sec VI of the method, Sec VI. A provides details on the system diagram based on Fig 2, Sec VI. B introduces data preprocessing which includes segment generation and labeling, Sec VI. C provides details about the Deep Learning architectures of Emphasized Channel Attention and Propagation -Time Delay Neural Network (ECAPA-TDNN **(Desplanques et al., 2020)**) and ResNet18 **(He et al., 2016)** used for segment classification. Sec VII talks about the experimental design and the metrics used for evaluating the experiments, while we look and discuss the results in Sec VIII, followed by conclusions and future work in Sec IX.

III. Background

A. Modeling speaker characteristics

For speaker modeling and recognition, i-Vectors **(Dehak et al., 2010; Hansen and Hasan, 2015)** are fixed length vectors that characterize speaker identity from arbitrary length sequential data (i.e., speech samples) and are traditional features for speaker recognition **(Dehak et al., 2010)**. They have also been used for language recognition **(Dehak et al., 2011)**, accent recognition **(Bahari et al., 2013)**, emotion recognition **(Xia and Liu, 2012)**, etc. Alternatively, DNNs **(McLaren et al., 2015; Snyder et al., 2018b, 2016)** can be used to directly capture language or speaker characteristics. They achieve improved results over i-Vectors using Mel-Frequency Cepstral Coefficients or Filterbank Coefficients as features.

The current standard framework consists of a discriminatively trained DNN that maps variable-length speech segments to embeddings called x-Vectors **(Snyder et al., 2018b)**. x-Vectors are deep speaker embeddings based on a Time-Delay Neural Network (TDNN)

architecture. This approach has achieved excellent results for speaker recognition **(Snyder et al., 2018b)**, diarization **(Sell et al., 2018)** and language recognition **(Snyder et al., 2018a)** with further advancements being actively researched. ECAPA-TDNN **(Dawalatabad et al., 2021)** were recently introduced and provide enhancements over TDNN **(Snyder et al., 2018b)** by introducing channel and context-dependent attention mechanism.

B. Child-Adult Speech Diarization

Previous work on child speech have utilized i-Vectors **(Kothalkar et al., 2019; Najafian et al., 2016)** and x-Vectors **(Xie et al., 2019a)** as features for speaker classification. The SincNet-based speaker identification model have been used in university classroom setting **(Dubey et al., 2019)** with effective results. Previous work on this dataset **(Najafian et al., 2016)** used much lesser data and fixed segments of length 1.5 seconds with a Support Vector Machine (SVM) backend for classification. A recent study **(Kothalkar et al., 2019)** with more data transcribed for the dataset, used DNN modeling with i-Vectors as features, and provided promising results. Since, we aim to perform classification for real-time application in an end-to-end diarization scenario, multiple pipelines of DNN models for speech activity detection (SAD) or voice activity detection (VAD), speech/speaker-type classification and ASR are combined for their strong performance in related studies **(Silero Team, 2021; Kim et al., 2021; Bredin et al., 2021; Ozturk et al., 2022; Radford et al., 2022; Bain et al., 2023)** and possible End-to-End classification approach.

C. End-to-end Child-Adult Speech Diarization

Recently studies have considered neural network-based classification systems trained for classifying child or adult speech/speaker-type. These utilize some form of fixed length embedding as input for another neural network for final classification of child or adult based on class posterior values **(Kolluguri et al., 2021; Kumar et al., 2020)** or traditional speaker clustering **(Krishnamachari et al., 2020)**. Alternately, such embeddings have also been utilized for child-adult speech/speaker-type diarization, where neural network training is formulated as a sequence

classification problem with output belonging to one of three classes: child speech, adult speech or silence. These solutions are effective in moderate noise conditions such as home environments with limited number of children and/or adults.

(Lavechin et al., 2020) formulated the child-adult Diarization task as a multi-label classification task using SincNet followed by Long-Short-Term-Memory (LSTM) layers for activating multiple voice types present in 2s audio segments. This implied each segment could be reported as multiple voice-types resulting in multiple classes for downstream processing tasks like Automatic Speech Recognition (ASR) or Keyword Spotting (KWS). Speech-type specific ASR models could be utilized for downstream recognition and analysis tasks, if such specific information can be extracted. Thus, multiple segment labels may not be optimal for extremely noisy data/scenarios with audible/intelligible speech from single unique speech/speaker-type.

Speech activity detection (SAD) and audio classification are similarly aligned tasks as our speech/speaker-type diarization and have achieved effective performance using single DNN multitask classification. A single DNN with multi-class classification has performed effectively for short duration audio on tasks such as SAD or audio classification. **(Hebbar et al., 2019)** utilized standard deep learning architectures for image classification tasks with ResNet for segment-based robust speech activity detection (clean, music, noise classes) with impressive performance. Apart from Convolutional Recurrent Neural Networks, Time Delay Neural Networks (TDNNs) **(Snyder et al., 2018b)** have been utilized to model long-term dependencies while performing SAD with advantage of overall lower computational costs.

D. ASR word alignments to refine Diarization results

In early works, ASR has been utilized in the context of diarization for resegmenting the initial speech segments generated from speech activity detection outputs. The IBM system **(Huang, 2007)** for RT07 evaluation incorporates word alignments from the speaker independent ASR system to refine the SAD outputs and reduce false alarms, thus resulting in better segment clustering output.

IV. FRAMEWORK FOR CHILD-ADULT SPEECH/SPEAKER TYPE CLASSIFICATION

The TDNN **(Snyder et al., 2018b)** architecture embeddings have been utilized for detection of speech **(Bai et al., 2019b; Ogura and Haynes, 2021)**, language **(Garcia-Romero and McCree, 2016)**, acoustic scene **(Bai et al., 2019a)**, Parkinson's **(Wodzinski et al., 2019)**, audio Session **(Raj et al., 2019)**, gender **(Raj et al., 2019)**, speaking rate **(Raj et al., 2019)**, words **(Raj et al., 2019)**, phoneme **(Raj et al., 2019),** utterance length **(Raj et al., 2019)** etc. Recently, ECAPA-TDNN **(Dawalatabad et al., 2021)** embeddings have provided state-of-the-art results for speaker recognition **(Chung et al., 2018)** and speaker diarization **(Dawalatabad et al., 2021)** tasks in noisy audio.

The posterior probabilities from the TDNN **(Snyder et al., 2018b)** and/or ResNet **(He et al., 2016)** architectures have also been utilized for detection of speech **(Bai et al., 2019b; Horiguchi et al., 2021; Kwon et al., 2021; Lin et al., 2020a; Villalba et al., 2019)**, speaker **(Xie et al., 2019b)**, music **(Lee et al., 2006),** stuttering **(Sheikh et al., 2021, 2022),** Parkinson's **(Wodzinski et al., 2019)**, spoken term **(Ram et al., 2019)**, dysarthria **(Gupta et al., 2021)**, intoxication **(Wang et al., 2019)** etc.

Based on the effectiveness in these studies, we pose the child-adult speech/speaker-type detection problem as a multi-class classification task using modern DNN architectures. Thus, we propose to experimentally verify the detection of child and adult speech from non-speech in naturalistic audio using a single deep neural network like ECAPA-TDNN **(Desplanques et al., 2020)** for 1D input raw audio feature and a deep neural network like ResNet for 2D input feature. Here, non-speech comprises silence, inaudible speech within crowd noise by adults or children, background music including vocals or electronic devices. Child-specific non-speech comprises laughs, cries, screams, breathing, burping, babbling, growling, squealing etc. Due to the pervasiveness of such noisy non-speech along with speech, for long periods of interaction in the preschool classroom, we prioritize capturing speech-types in clean as well as extremely noisy

conditions, by training a single model for distinguishing clean/noisy child-adult speech from non-speech.

To capture the minor variation in perceptual differences between intelligible speech from children and adults, in the presence of near-identical unintelligible adult noise or child non-speech sounds, we formulate it as a multiclass classification task, for a single neural network with logMelSpectrogram input features. The hypothesis is that regions of child/adult speech in the mel-spectrograms would be distinguishable by a DNN compared to regions of non-speech in both clean and noisy conditions.

V. DATA SPECIFICS

A. Data collection

The dataset in this study consists of spontaneous conversational speech recorded with the help of 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 (A or B).

B. Classroom details

Data collected using LENA recorders in two classrooms have multiple working stations. These learning station activities such as reading, blocks, play, singing, science etc. (see Fig 1). The dimensions of the two classrooms are different, which may affect the recorded audio in terms of reverberation. Classroom A is 24 ft. by 24 ft. in dimension. Classroom B is much larger with dimensions of 24 ft. by 40 ft. An illustration of a floor plan in a preschool classroom is shown in Fig 1. Thus, to understand the performance of our algorithms in diverse environmental conditions, it would be useful to have data from these classrooms in different sets for model training and test.

C. Dataset distributions

Audio for this study have children who are 3 to 5 years along with one or more adults (e.g., typically, teachers). Most children wear LENA devices as well as accompanying 1-3 adults are also wearing them.

The total audio from classroom A is of duration 61 hours and 18 minutes and from classroom B is 63 hours and 57 minutes. Thus, around 60 hours of audio or approximately 230,000 segments of 1 second duration are used for training the classroom-specific models. For this dataset, an organized set of 38.6 hrs of speech from classroom B and similar amount of speech from classroom B are established.

The audio segment files are divided into training, development and test sets following the classroom-based division such that there is no overlap of data between the sets. The audio data corresponding to classrooms A and B are used for training alternate models. Data from the other classroom is used for model development and testing. During model development, a separate hold-out set known as development data, is used in order to find the best performing model (based on training epoch) during neural network training.

For example, a model trained on data from classroom A, is used for model development on data from a given timepoint in classroom B, and tested on remaining timepoints from the same classroom B. Similarly, a model trained on classroom B, is used for model development on data from given timepoint in classroom A and tested on data from remaining timepoints in classroom A. Thus, training set is from alternate classroom compared to development and test sets. This provides an opportunity for a model developed on data from one classroom, to be evaluated on two subsets of data from other classrooms. Also, such a data split has practical application for new classroom scenarios where smaller, transcribed pilot data from new classroom can be used for model epoch selection and rest of the untranscribed data for testing. Even if transcription for new classroom data is not feasible, the current data split provides generalized models for testing based on train-development split.

VI. METHOD

A. System pipeline

1. Speech/Speaker-type Classification

Fig 2 explains the high-level system diagram for child-adult speech-type classification task. It starts with data collection using our LENA device in preschool classroom. This data is transcribed by the CRSS transcription team for recognizing the speech in this naturalistic audio. After data preprocessing steps, the modified data is used to train Deep Learning models using the training set. The best model on the training set is evaluated on the development set for model selection. The best performing model on the development set is finally evaluated on the test set for final speech/speaker-type classification.

2. ASR resegmentation for child-adult Speech/Speaker Diarization

The ASR resegmentation module consists of an end-to-end (E2E) ASR system for recognizing the text in the audio segment followed by another E2E ASR system for recognizing the timestamps as shown in Fig. 3. We utilize Whisper for recognizing the text in the speech segment due to its high-quality transcription performance in naturalistic conditions. This is followed by the forced alignment using another E2E ASR model known as Wav2Vec2. This combined system for forced alignment is implemented in the tool WhisperX **(Bain et al., 2023)**. For a given system alternate model variations of the two E2E ASR systems were utilized. For Whisper its medium and large models for English language were considered. For Wav2Vec2 ASR system, two variations of XLSR-53 large model (trained for speech recognition) were considered. The variations were based on the datasets utilized to finetune the base Wav2Vec2 model. The alternate configurations of the Speech-type classification and ASR resegmentation modules are displayed in Fig. 3 and explained as follows:

a. System S1

System S1 consists of an industry-strength SAD system Silero **(Silero team, 2021)** followed by an ASR-based resegmentation module to mark the start and end times of the speech within the daylong audio files. The Silero SAD system consists of Convolutional Neural Network and Tranformer-based architectures. Finally, if child speech-type is detected by the

speech-type detector ResNet module the presence of child speech is marked within the segment.

b. System S2

System S2 consists of speech-type detector ResNet module followed by ASR-based resegmentation module. Here, our speech-type detection module acts as an implicit speech activity detector with an additional class for detecting child speech. The ASR resegmentation module performs the task of marking the timestamps of the recognized speech-types.

c. System S1 + S2

In the combination system, we combine the final outputs from systems S1 and S2. Irrespective of the segment speech-type, for overlapping output segments from systems S1 and S2, the segments from the two systems are merged using following segment merging strategies:

1. If one segment completely bounds the other segment on the time axis, the smaller segment is removed.
2. For a given segment from System S1/S2, if it overlaps a segment from System S2/S1 to its right along the time axis, the segment from System S1/S2 is truncated to start of segment from System S2/S1.

B. Data Preprocessing

Audio recordings from both classroom A and B are divided into audio segments using a sliding window of 1000ms duration with no overlap. Based on text transcripts from the data, ground-truth speaker-types are assigned as “adult” or “child” speech on the basis of greater talk time by either the adult or child speaker over each 1000ms audio segment respectively. This approach was motivated by an earlier study that also considered a different challenging diarization scenario **(Lin et al., 2020b)**. For segments with speech tags that occupy less than 12.5% of the total segment duration, these are marked as non-speech. The ability to set a

speech/silence threshold balance, achieving overall effective diarization robustness, has also been explored in other studies **(Hebbar et al., 2019)**.

C. Deep Learning Model Architectures

End-to-end deep learning systems for speech classification tasks consist of the following steps: i) frame-level feature extraction using DNNs, ii) temporal aggregation of frame-level features, and iii) optimization of classification loss. Most speaker verification/recognition systems have a base DNN architecture such as a 2D CNN with convolutions in both time and frequency domains such as ResNet **(He et al., 2016)** or a 1D CNN with convolutions only in the time domain such as ECAPA-TDNN **(Desplanques et al., 2020)**. Here the focus is to evaluate these for speaker/speech-type classification. Thus, looking at both 1D and 2D CNN architectures will help to evaluate features and architectures for systems that can perform well on child or adult speaker/speech-type detection from non-speech. The ECAPA-TDNN **(Desplanques et al., 2020)** performs better than the ResNet architecture for speaker recognition tasks, due to its ability to learn complex patterns that occur in any frequency region since 1D convolutions cover the complete frequency range of the input features. However, this leads to hardcoding **(Thienpondt et al., 2020)** of absolute frequency position of each input feature. Our hypothesis is that this may not translate to appropriate generic speech/speaker-type classifications due to differences in frequency variability within adult/child speakers. ResNet models are expected to benefit due to 2D convolutions with small receptive fields by exploiting the local speech-type frequency patterns that repeat for small frequency shifts, thus providing generality for modeling speakers within child/adult groups.

1. ECAPA-TDNN model

TDNN **(Snyder et al., 2018b)** model differs from a conventional DNN by introducing a multi-splicing concept that enables an efficient way of modelling the large temporal context. Multi-

splicing implies that feature frames and intermediate DNN-layer outputs are time delayed and stacked to form an input to an upstream neural network layer.

ECAPA-TDNN **(Desplanques et al., 2020)** is an enhanced version of the TDNN **(Snyder et al., 2018b)** model using novel blocks and modules for robust speaker embeddings. The pooling layer uses a channel and a context-dependent attention mechanism, which allows the network to 'attend' to different frames per channel. Here, the 1-dimensional Squeeze-Excitation (SE) blocks rescale the channels of intermediate frame-level feature maps to insert a global context information into the locally operating convolutional blocks. Also, 1-D Res2 blocks and Multi-layer Feature Aggregation (MFA) improves performance by using grouped convolutions and merging the complementary information respectively. MFA provides complementary information for statistics pooling by concatenating the final frame-level features with intermediate features of previous layers.

2. Input representation for ECAPA-TDNN

Here, 80-dim. log-Mel-Spectrograms are extracted over 25ms window lengths with 10ms skip rate from 1000ms audio segments as input features. Stacked frame blocks of 1000ms duration (100 frames) are used to generate the serialized input 2D features for the task of speech/speaker-type classification.

3. ResNet18 model

The ResNet model is used for training very deep networks with the help of residual learning which involves skip connections to help overcome the problem of vanishing gradient due to increase in depth. Configuration details for the ResNet18 **(He et al., 2016)** model is presented in Table I. ResNet is a block-based model which includes identity block and convolution block. Here identity block passes the original input to the output of the convolution block by skipping intermediate convolutional layers within the block. For convolutional block, the original input is passed through another convolutional layer to match the output dimensions of the convolutional

block during summation. This creates an alternate path for the vanishing gradient to pass through from deeper layers. This approach will allow the model to learn an identity function, which allows the higher layer in the model to perform as effectively as the lower layer. After initial convolution (Layer 0) and batch normalization and ReLU operations, there are always 4 blocks (Layer 1-Layer 4) with each block containing multiple convolutions, batch normalization and ReLU operations. Layer 0 represents the input layer and layers 1-4 are the residual blocks in the ResNet architecture with skip connections as summarized in Table I. The architecture finishes with a convolutional layer, flatten operation, average pool operation and output layers.

Name	Output size	I.C. size, O.C. size	Kernel size, Stride size
Layer0	99 × 80	3,64	7, 2
Layer1	50 × 40	64,64 64,64	3, 1 3, 1
Layer2	25 × 20	64,128 128,128	3, 2 3, 1
Layer3	13 × 10	128,256 256,256	3, 2 3, 1
Layer4	7 × 5	256,512 512,512	3, 2 3, 1

Avg. Pool	4 × 3	512,3	1, 1
Embedding	1 × 1	-	1, 1
Softmax	1 × 1		

TABLE I. Configurations of all operators in ResNet-18 where I.C. represents Input Channel and O.C. represents Output Channel.

4. Input representation for ResNet18

For this system, 80-dimensional log-Mel-Spectrograms are extracted over 25ms windows with 10ms skip rate as input features. Stacked frame blocks of 1000ms duration (100 frames) are used to generate serialized input 2D features for the task of speaker/speech-type classification.

VII. EXPERIMENTAL DESIGN AND METRICS

A. Experimental Design

For uniformity in system evaluation, both ECAPA-TDNN **(Desplanques et al., 2020)** and ResNet18 **(He et al., 2016)** models are trained with an Additive Margin-Softmax loss with margin=0.15 on input features for 40 epochs using the RMSprop algorithm with a learning rate of 0.001, $\alpha = 0.95$ and $\varepsilon = 1 \times 10^{-8}$. Each epoch consists of 800 batches of randomly selected segments of batch size 32. Figs. 4 and 5 highlight the block diagram for ECAPA-TDNN **(Desplanques et al., 2020)** model and ResNet18 **(He et al., 2016)** models respectively. Results are reported for both development and test sets for both models as explained in Sec V. C.

B. F1-score for speech type detection by model on testing dataset

To understand the child-adult speaker/speech-type detection, we test our models on classroom specific test data. Different metrics can assess model performance in terms of their ability to recall as well as precision of detection. 'Accuracy' is defined as the total number of samples that are predicted correctly. 'Precision' is the fraction of relevant instances among all

the detected instances. These would be the fraction of actual segments of speech/speaker type or non-speech type, among all such detected segments.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

where TP represents True Positives and FP represents False Positives.

'Recall' is defined as the fraction of the relevant instances that were actually detected. In our case, these would be the fraction of segments of particular speech/speaker or non-speech type that were predicted correctly.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

where TP represents True Positives and FN represents False Negatives.

F1-score is defined as harmonic mean of the precision and recall, and takes both precision and recall into account for providing an overall balanced assessment.

$$F1 - score = \frac{2\ X\ Precision\ X\ Recall}{Precision+Recall} \quad (4)$$

C. Diarization Error Rate

Diarization error rate (DER) can be defined as the sum of errors due to an incorrect speaker (E_{spkr}), missed speech (E_{miss}), false alarm speech (E_{FA}) and overlapping speakers (E_{ovl}) based on the predictions of the Diarization system. E_{ovl} and are not considered in this evaluation.

$$DER = E_{spkr} + E_{miss} + E_{FA} \quad (1)$$

In the literature, Speaker Confusion Error for audio streams is mostly reported as DER. However, we have reported DER comprised of speaker confusion error, false alarm error and

missed speech error. Missed speech error **(Kumar et al., 2020)**, are most important for follow-on downstream tasks of both speech analysis and ASR.

VIII. RESULTS AND DISCUSSIONS

A. F1-score and DER

Table II reports corresponding F1-scores for each of the speaker/speech types and non-sp. audio where non-sp. represents non-speech. Table III reports diarization error rate on the test subsets for classrooms A and B.

Train on **Train** set of:	Test on **Test** set of:	Model	$F1_{child}$ **(%)**	$F1_{adult}$ **(%)**	$F1_{non\text{-}sp.}$ **(%)**	$F1_{overall}$ **(%)**
Room A	Room B	ECAPA-TDNN	71.0%	68.5%	74.3%	71.5%
		ResNet18	**79.0%**	**74.4%**	**79.8%**	**77.9%**
Room B	Room A	ECAPA-TDNN	69.0%	73.4%	75.7%	72.7%
		ResNet18	**77.4%**	**82.1%**	**84.3%**	**81.3%**

TABLE II. F1-score results on testing subset recordings of classroom A and classroom B audio where non-sp. represents non-speech.

Train on **Train** set of:	Test on **Test** set of:	System combination with Resnet model	E_{spkr} **(%)**	E_{FA} **(%)**	E_{MISS} **(%)**	DER **(%)**
Room A	Room B	System S1		6.2	55.4	76.7

			15.1			
		System S2	**3.7**	**1.2**	54.2	59.1
		System S1+S2	12.1	7.3	**31.4**	**50.8**
Room B	Room A	System S1	16.8	3.1	48.0	67.9
		System S2	**4.2**	**1.2**	54.3	59.7
		System S1+S2	13.5	4.3	**28.1**	**45.9**

TABLE III. Diarization Error Rate results on testing subset recordings of classroom A and classroom B audio.

The largest improvement by ResNet model is for segments containing child speech in terms of the F1-score as seen in Table II for test subset. Specifically, F1-score for child speech provides absolute improvement of +8.4% for test data from classroom A, and absolute improvement of +8.0% for test data from classroom B. For all results in Table II, the best F1-scores are for non-speech segments, for test sets of both classrooms A and B. We hypothesize the lower F1-scores for all the speech-types in test subset of classroom B to be due to the more challenging environmental noise conditions of classroom B Vs. classroom A. The highest F1-scores across all models and classrooms for non-speech type audio can be attributed to the disproportionate amount of non-speech present in these audio files, and therefore the distribution in the test segments.

As can be seen from Table III, System S2 outperforms System S1 significantly for speaker confusion error rate, false alarm error rate, and overall DER on the test set for both classrooms A and B. However, the best overall DER on the test set for both classrooms A and B is by System S1+S2. The relative improvements by System S1+S2 Vs. System S1 on classroom A test audio data are +19.6% for speaker confusion error rate, 41.5% for missed speech error rate, and

+32.4% for overall DER. Relative improvements by System S1+S2 Vs. System S1 on classroom B test audio data are +19.9% for speaker confusion error rate, 43.3% for missed speech error rate, and +33.8% for overall DER.

Thus, System S1+S2 provides improvement in overall DER Vs. Systems S1 due to relatively improved error rate for missed speech by 42-43% on test set for both classrooms A and B. System S1+S2 also provides improvement in overall DER Vs. System S2 due to relatively improved error rate for missed speech by 41-42% on test set for both classrooms A and B. It can be observed from Table III that the false alarm error rate and speaker confusion rate for both the models on test sets of both the classrooms increase for System S1+S2 Vs. System S2. This can be attributed to the drastic drop in missed speech rate for system S1+S2 on test subsets of both the classrooms. Detecting more speech segments while improving the DER is more important than a lower false alarm rate for this dataset in order to perform analytics on the recognized conversational speech.

Thus, our speech/speaker-type classifier trained on classroom domain-specific data in conjunction with ASR models trained on massive amounts of audio data can match performance with Silero VAD and ASR models- both trained on massive amounts of audio data. In combination with Silero VAD our ResNet-based speech/speaker-type classifier can improve the missed speech error rate and thus, the overall child-adult diarization performance.

Although ECAPA-TDNN model performs better than a ResNet variant for speaker recognition **(Desplanques *et al.*, 2020)** and diarization **(Dawalatabad *et al.*, 2021)** tasks, certain ResNet variants perform better than ECAPA-TDNN for short-duration utterance speaker verification **(Thienpondt et al., 2020)**. Also, some ResNet variants perform better than TDNN variants for far-field speaker recognition (**Gusev et al., 2020**) using short duration test utterances. Thus, our results presented here, are along the line of results (of ResNet variant being better than ECAPA-TDNN) achieved for similar short-duration, noisy and near as well as far-field audio for speaker recognition/ verification.

B. Visualization of speech-type density and turn-taking using donut diagrams

Also, we present the speaker/speech-type density and turn-taking with a visualization tool known as ”donut diagram” that reflects the speech density per speaker over different times of a session. It begins in the east-most section of the donut and displays times along an anti-clockwise direction until time is complete, reaching the same point 360 degrees later.

Figs. 6 and 7 represent the actual and predicted (using ResNet **(He *et al.*, 2016)** model) talktimes for a session in classroom A with a child wearing the LENA device. We see the percentage difference between predicted and actual talktimes differ between 2.6% (child) and 3.1% (adult). Although child and adult speech is predicted more than in reality, the density of speech-type and change in speech-types in alternate sections are captured well and offers an excellent high-level assessment of child-adult conversational engagement. For example, the left half of the diagram with multiple interactions between children and adults is useful for further analysis. The mapping between dense regions of child speech (thick segments of pink) and adult speech (thick segments of green) are also matched closely between Figs. 6 and 7, where thick segments would have speech for a single type for significant duration.

For example, certain thick green segments are matched at 85 degrees and between 150 and 210 degrees. Similar, thick pink segments are between 180 and 210 degrees. Figs. 8 and 9 represent the actual and predicted (using ResNet model) talktimes for a session in classroom B with a child wearing the LENA, resulting in much more recorded adult speech. Approximately, 10% of child speech is missed in this predicted donut diagram, and approximately a similar amount of non-speech is misclassified. However, regions with significant child or adult communication-as represented by thick segment of single color (green or pink) - interspersed with the speech type are present and well matched in both figures. For example, presence of thick green segments between approximately 260-300 degrees-representing significant adult talk during that time of the session, along with child speech in between in classroom A with a child wearing the LENA device.

IX. CONCLUSIONS AND FUTURE WORK

In this study, a child-adult speech-type diarization system for recognizing speech/speaker type from day long audio recordings was developed. State-of-the-art Deep Learning models renowned for speaker recognition were utilized for predicting speech-type activity. Specifically, ECAPA-TDNN models provided good and consistent results in terms of F1-scores for all speech activity types recognized based on the posterior probabilities. However, ResNet model with 80-dim. Log-Mel-spectrogram inputs have outperformed ECAPA-TDNN model in terms of F1-scores of all speech activity types as well as DER. These models were trained on audio data from one classroom and tested on audio data from a separate classroom, which proves the generalization of our models for alternate classroom conditions. The predicted segments of fixed duration 1s, were visualized with novel visualizations referred to here as donut diagrams. These were shown to be an effective method for detecting continuous child and/or adult speech segments over a period of time, providing visual feedback of child-adult interactions. Thus, the diagrams can provide feedback to teachers/adults on their communication metrics with children during different times of the session. The child-adult speech-type predicted outputs are combined with an ASR resegmentation module in various configurations to provide multiple child-adult diarization systems. A combination of two such child-adult diarization systems provides the best performance in terms of diarization error rate. For future work, we suggest training and testing multi-class classification tasks for attention-based ResNet models for smaller duration segments. Also we would like to utilize more advanced ASR resegmentation modules that have been customized to speech data from preschool classroom domain. Since the scope of this work involved classroom-independent diarization evaluation, future work could also include performance evaluation of the proposed diarization system for downstream speech technology tasks including ASR and Keyword Spotting.

X. ACKNOWLEDGEMENTS

This work was supported by the grant NSF Grant #1918032 (UT Dallas CRSS) (PI: Hansen) from the National Science Foundation. The authors would also like to thank all the teachers and children who participated in the data collection sessions.

REFERENCES

"https://www.lenafoundation.org" (last accessed Aug. 22, 2022)

Bahari, M. H., Saeidi, R., Van Leeuwen, D. et al. (**2013**). "Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp. 7344–7348.Bai, H., Chen, H., and Yan, Y. (**2019**a). "Audio scene classification with discriminatively-trained segment-level features," in 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, pp. 354–359.

Bai, Y., Yi, J., Tao, J., Wen, Z., and Liu, B. (**2019**b). "Voice activity detection based on time-delay neural networks," in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, pp. 1173–1178.

Bain, M., Huh, J., Han, T., & Zisserman, A. (**2023**). "WhisperX: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*.

Bredin, H., & Laurent, A. (**2021**). "End-to-end speaker segmentation for overlap-aware resegmentation," In *Interspeech 2021*.

Chung, J. S., Nagrani, A., and Zisserman, A. (**2018**). "Voxceleb2: Deep speaker recognition," Proc. Interspeech 2018 1086–1090.

Cristia, A., Ganesh, S., Casillas, M., & Ganapathy, S. (**2018**). "Talker diarization in the wild: The case of child-centered daylong audio-recordings." In Interspeech 2018 (pp. 2583-2587).

Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., and Na, H. (**2021**). "ECAPA-TDNN embeddings for speaker diarization," ISCA INTERSPEECH-2021 3560–3564.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (**2010**). "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798.

528 Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., and Dehak, R. (**2011**). "Language
529 recognition via i-vectors and dimensionality reduction," ISCA INTERSPEECH-2011, pp.
530 857-860.
531 Desplanques, B., Thienpondt, J., and Demuynck, K. (**2020**). "ECAPA-TDNN: Emphasized
532 channel attention, propagation and aggregation in TDNN based speaker verification,"
533 ISCA INTERSPEECH-2020, pp. 3830–3834.
534 Dubey, H., Sangwan, A., and Hansen, J. H. L. (**2019**). "Transfer learning using raw waveform
535 SincNet for robust speaker diarization," IEEE ICASSP-2019: Inter. Conf. on Acoustics,
536 Speech and Signal Proc., pp. 6296– 6300.
537 Garcia-Romero, D., and McCree, A. (**2016**). "Stacked long-term tdnn for spoken language
538 recognition.," ISCA INTERSPEECH-2016, pp. 3226–3230.
539 Gupta, S., Patil, A. T., Purohit, M., Parmar, M., Patel, M., Patil, H. A., and Guido, R. C. (**2021**).
540 "Residual neural network precisely quantifies dysarthria severity-level based on
541 short-duration speech segments," Neural Networks **139**, 105–117.
542 Hansen, J. H. L., and Hasan, T. (**2015**). "Speaker recognition by machines and humans: a
543 tutorial review," IEEE Signal Processing Magazine **32**(6), 74–99.
544 Hart, B., and Risley, T. R. (**1995**). *Meaningful differences in the everyday experience of young*
545 *American children.* (Paul H Brookes Publishing).
546 He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition,"
547 in *Proceedings of the IEEE Conf. on computer vision and pattern recognition*, pp.
548 770–778.
549 Hebbar, R., Somandepalli, K., and Narayanan, S. (**2019**). "Robust speech activity detection
550 in movie audio: Data resources and experimental evaluation," in *ICASSP 2019-2019*
551 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,
552 IEEE, pp. 4105–4109.

553 Horiguchi, S., Yalta, N., Garcia, P., Takashima, Y., Xue, Y., Raj, D., Huang, Z., Fujita,
554 Y.,Watanabe, S., and Khudanpur, S. (**2021**). “The Hitachi-JHU DIHARD-iii system:
555 Competitive end-to-end neural diarization and x-vector clustering systems combined by
556 dover-lap,” arXiv preprint arXiv:2102.01363.
557 Huang, J., Marcheret, E., Visweswariah, K., & Potamianos, G. **(2008)**. The IBM RT07 evaluation
558 systems for speaker diarization on lecture meetings. In *Multimodal Technologies for*
559 *Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007,*
560 *Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers* (pp. 497-508). Springer
561 Berlin Heidelberg.
562 Kim, M., Ki, T., Anshu, A., & Apsingekar, V. R. (**2021**) North America Bixby Speaker Diarization
563 System for the VoxCeleb Speaker Recognition Challenge 2021.
564 Koluguri, N. R., Kumar, M., Kim, S. H., Lord, C., & Narayanan, S. (**2020**). “Meta-learning for
565 robust child-adult classification from speech.” In ICASSP 2020-2020 IEEE International
566 Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8094-8098).
567 IEEE.
568 Kothalkar, P. V., Irvin, D., Luo, Y., Rojas, J., Nash, J., Rous, B., and Hansen, J. H. L. (**2019**).
569 “Tagging child-adult interactions in naturalistic, noisy, daylong school environments
570 using i-vector based diarization system,” in *Proc. SLaTE 2019: 8th ISCA Workshop*
571 *on Speech and Language Technology in Education*, pp. 89–93.
572 Krishnamachari, S., Kumar, M., Kim, S. H., Lord, C., & Narayanan, S. (**2021**). “Developing
573 Neural Representations for Robust Child-Adult Diarization.” In 2021 IEEE Spoken
574 Language Technology Workshop (SLT) (pp. 590-597). IEEE.
575 Kumar, M., Kim, S. H., Lord, C., & Narayanan, S. (**2020**). “Improving speaker diarization for
576 naturalistic child-adult conversational interactions using contextual information.” The
577 Journal of the Acoustical Society of America, **147**(2), EL196-EL200.

578 Kwon, Y., Heo, H. S., Huh, J., Lee, B.-J., and Chung, J. S. (**2021**). “Look who’s not talking,” in
579 *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp. 567–573.
580 Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., and Cristia, A. (**2020**). “An open-source
581 voice type classifier for child-centered daylong recordings,” arXiv preprint
582 arXiv:2005.12656.
583 Lee, J.-W., Park, S.-B., and Kim, S.-K. (**2006**). “Music genre classification using a time-delay
584 neural network,” in *International Symposium on Neural Networks*, Springer, pp. 178-187.
585 Lin, Q., Li, T., and Li, M. (**2020**a). “The dku speech activity detection and speaker identification
586 systems for fearless steps challenge phase-02.,” in *INTERSPEECH*, pp. 2607–2611.
587 Lin, Q., Cai, W., Yang, L., Wang, J., Zhang, J., & Li, M. (**2020**b). DIHARD II is Still Hard:
588 Experimental Results and Discussions from the DKU-LENOVO Team}}. In Proc.
589 Odyssey 2020 The Speaker and Language Recognition Workshop (pp. 102-109).
590 McLaren, M., Lei, Y., and Ferrer, L. (**2015**). “Advances in deep neural network approaches
591 to speaker recognition,” in *2015 IEEE international conference on acoustics, speech and*
592 *signal processing (ICASSP)*, IEEE, pp. 4814–4818.
593 Najafian, M., Irvin, D., Luo, Y., Rous, B. S., and Hansen, J. H. L. (**2016**). “Automatic
594 measurement and analysis of the child verbal communication using classroom acoustics
595 within a child care center.,” in *WOCCI*, pp. 56–61.
596 Ogura, M., and Haynes, M. (**2021**). “X-vector based voice activity detection for multi-genre
597 broadcast speech-to-text,” arXiv preprint arXiv:2112.05016 .
598 Ozturk, M. Z., Wu, C., Wang, B., Wu, M., & Liu, K. R. (**2022**). Beyond Microphone:
599 mmWave-Based Interference-Resilient Voice Activity Detection. In Proceedings of the
600 *1st ACM International Workshop on Intelligent Acoustic Systems and Applications*
601 *(pp. 7-12).*
602 Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (**2022**). “Robust
603 speech recognition via large-scale weak-supervision arXiv preprint arXiv:2212.04356.

Raj, D., Snyder, D., Povey, D., and Khudanpur, S. (**2019**). “Probing the information encoded in x-vectors,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, pp. 726–733.

Ram, D., Miculicich, L., and Bourlard, H. (**2019**). “Multilingual bottleneck features for query by example spoken term detection,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, pp. 621–628.

Rosenbaum, S., and Simon, P. (**2016**). *Speech and Language Disorders in Children: Implications for the Social Security Administration’s Supplemental Security Income Program.* (ERIC).

Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S. *et al.* (**2018**). “Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge.,” in *Interspeech*, pp. 2808–2812.

Sheikh, S. A., Sahidullah, M., Hirsch, F., and Ouni, S. (**2021**). “Stutternet: Stuttering detection using time delay neural network,” in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 426–430.

Sheikh, S. A., Sahidullah, M., Hirsch, F., and Ouni, S. (**2022**). “Introducing ECAPA-TDNN and wav2vec2.0 embeddings to stuttering detection,” arXiv preprint arXiv:2204.01564.

Smith, K. (**2011**). *Acoustics* (Springer, New York), (in press, 2016).

Silero Team, **(2021)**. “Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier,” https://github.com/snakers4/silero-vad.

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., and Khudanpur, S. (**2018**a). “Spoken language recognition using x-vectors.,” in *Odyssey*, pp. 105–111.

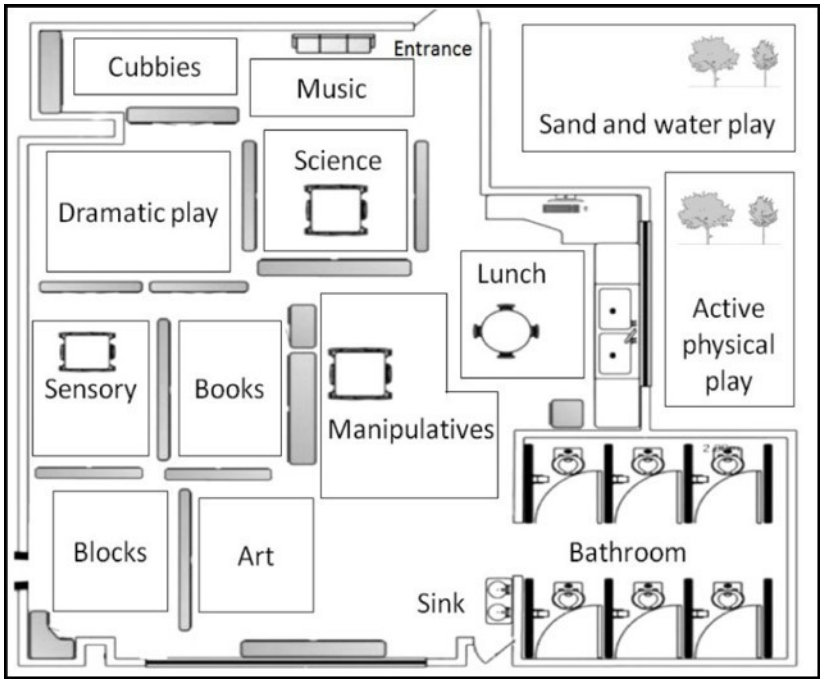
Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (**2018**b). “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5329–5333.

630 Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur,
631 S. (**2016**). “Deep neural network-based speaker embeddings for end-to-end speaker
632 verification,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp.
633 165–170.
634 Thienpondt, J., Desplanques, B., and Demuynck, K. (**2020**). “The idlab voxceleb speaker
635 recognition challenge 2020 system description,” arXiv preprint arXiv:2010.12468 .
636 Villalba, J., Garcia-Romero, D., Chen, N., Sell, G., Borgstrom, J., McCree, A., Snyder, D.,
637 Kataria, S., Garcıa-Perera, P., Richardson, F. *et al.* (**2019**). “The jhu-mit system
638 description for nist sre19 av,” in *NIST SRE19 Workshop*.
639 Wang, W., Wu, H., and Li, M. (**2019**). “Deep neural networks with batch speaker normalization
640 for intoxicated speech detection,” in *2019 Asia-Pacific Signal and Information*
641 *Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, pp.
642 1323–1327.
643 Wodzinski, M., Skalski, A., Hemmerling, D., Orozco-Arroyave, J. R., and N¨oth, E. (**2019**).
644 “Deep learning approach to parkinsons disease detection using voice recordings and
645 convolutional neural network dedicated to image classification,” in *2019 41st Annual*
646 *International Conference of the IEEE Engineering in Medicine and Biology Society*
647 *(EMBC)*, 564 IEEE, pp. 717–720
648 Xia, R., and Liu, Y. (**2012**). “Using i-vector space model for emotion recognition,” in
649 *Thirteenth Annual Conference of the International Speech Communication Association*.
650 Xie, J., Garc´ıa-Perera, L. P., Povey, D., and Khudanpur, S. (**2019**a). “Multi-plda diarization on
651 children’s speech.,” in *Interspeech*, pp. 376–380.
652 Xie, W., Nagrani, A., Chung, J. S., and Zisserman, A. (**2019**b). “Utterance-level aggregation
653 for speaker recognition in the wild,” in *ICASSP 2019-2019 IEEE International*
654 *Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.
655 5791–5795.

Xue, Y., Horiguchi, S., Fujita, Y., Watanabe, S., Garc´ıa, P., and Nagamatsu, K. (**2021**). “Online end-to-end neural diarization with speaker-tracing buffer,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp. 841–848.

Ziaei, A., Sangwan, A., and Hansen, J. H. L. (**2013**). “Prof-life-log: Personal interaction analysis for naturalistic audio streams,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 7770–7774

663 **FIGURES**

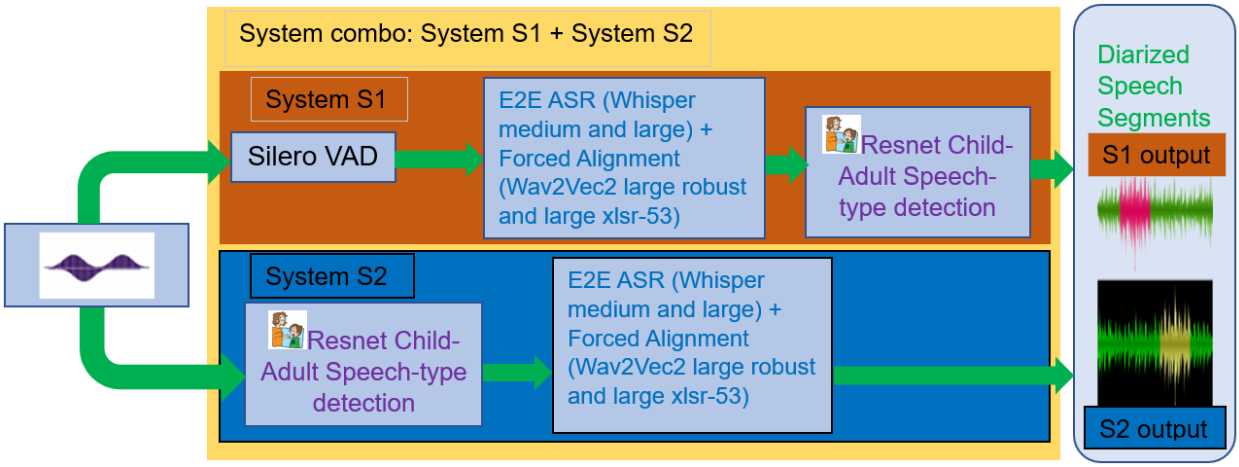


664

665


LENA device collects day long audio recordings
Transcription performed by CRSS transcription team
Deep learning model used for training with Training set
Best model selection by evaluating on Development set
Evaluating best model on Testing set

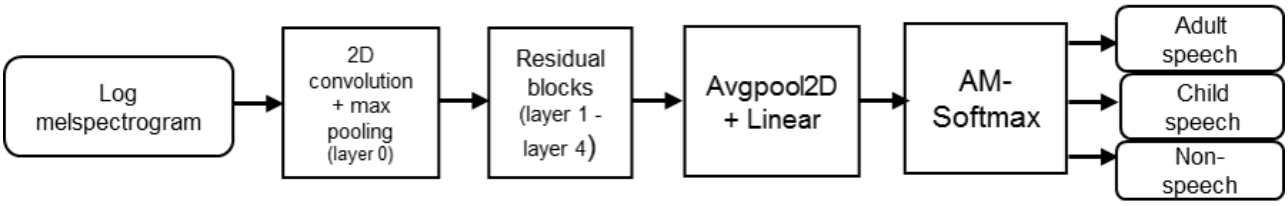
666


System combo: System S1 + System S2
System S1
Silero VAD
E2E ASR (Whisper medium and large) + Forced Alignment (Wav2Vec2 large robust and large xlsr-53)
Resnet Child-Adult Speech-type detection
System S2
Resnet Child-Adult Speech-type detection
E2E ASR (Whisper medium and large) + Forced Alignment (Wav2Vec2 large robust and large xlsr-53)
Diarized Speech Segments
S1 output
S2 output

667

668

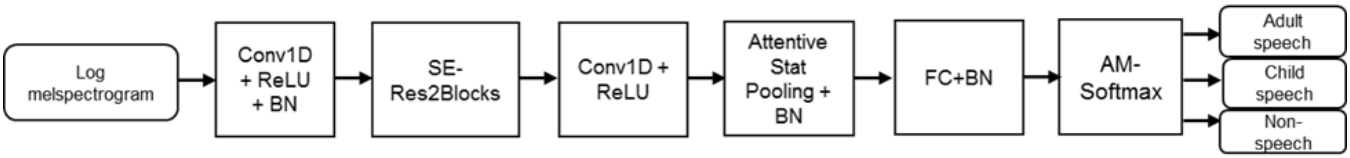
669 .


Log melspectrogram
2D convolution + max pooling (layer 0)
Residual blocks (layer 1 - layer 4)
Avgpool2D + Linear
AM-Softmax
Adult speech
Child speech
Non-speech

670

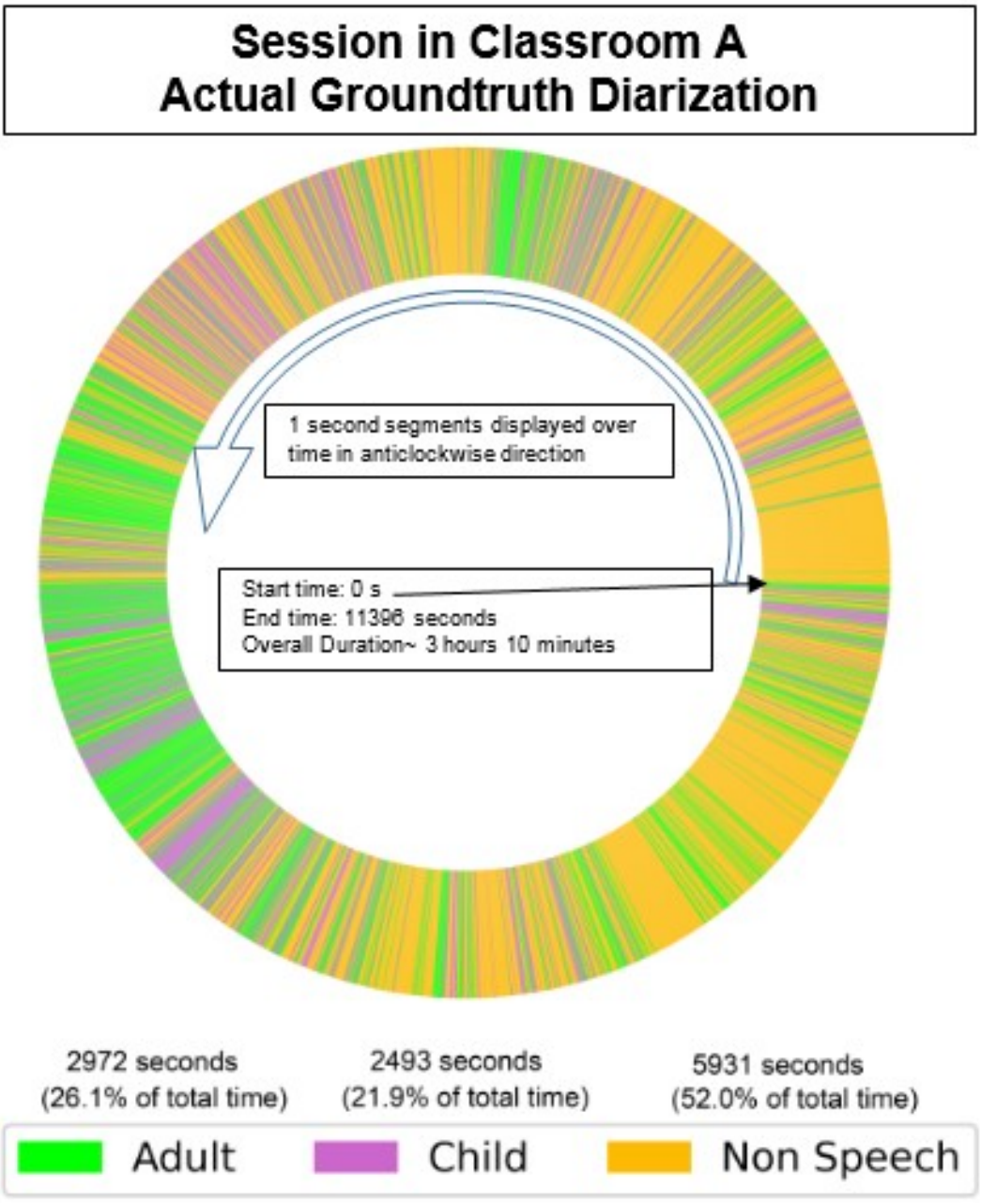
671

672


Log melspectrogram
Conv1D + ReLU + BN
SE-Res2Blocks
Conv1D + ReLU
Attentive Stat Pooling + BN
FC+BN
AM-Softmax
Adult speech
Child speech
Non-speech

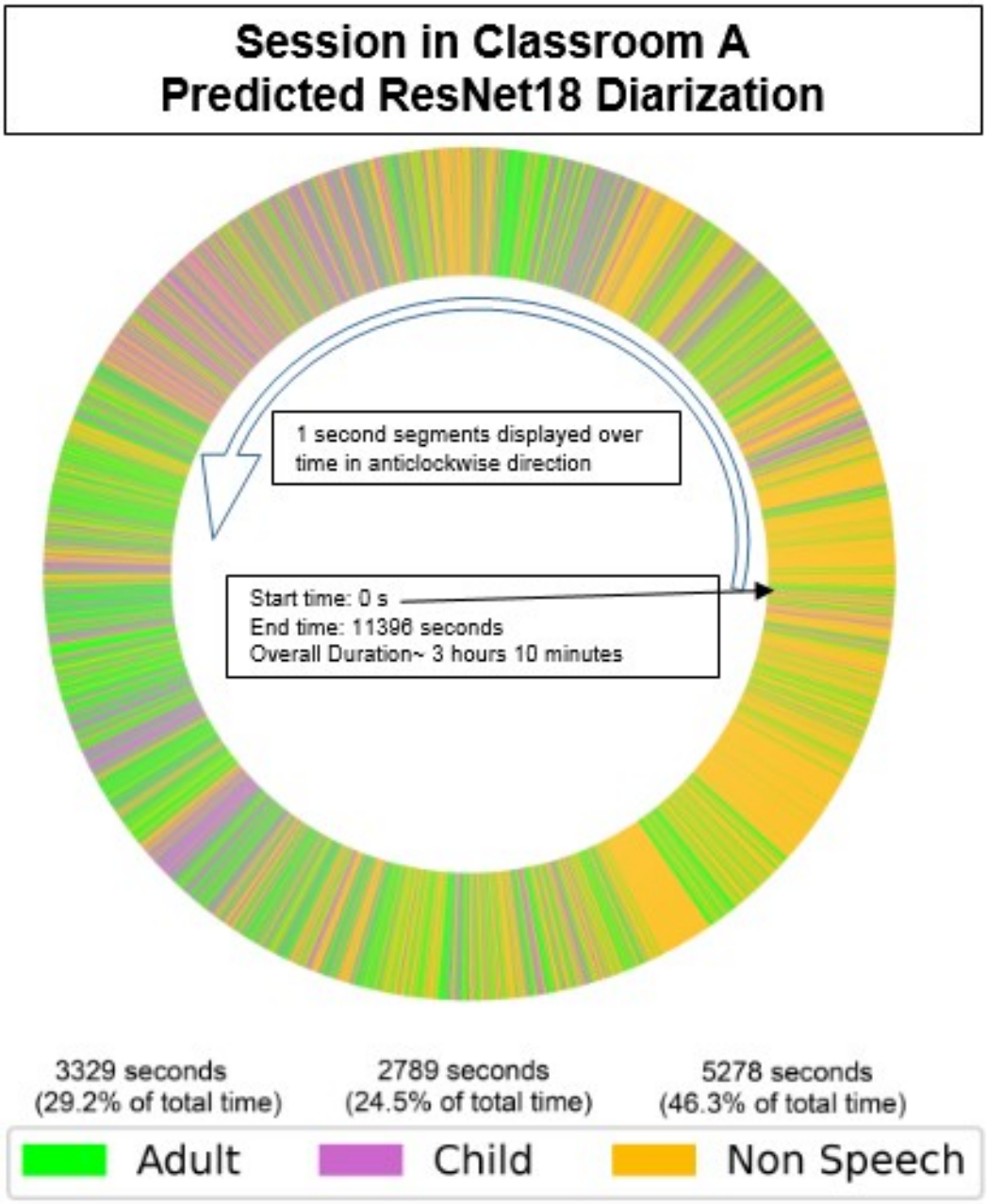
673

674


Session in Classroom A
Actual Groundtruth Diarization
1 second segments displayed over time in anticlockwise direction
Start time: 0 s
End time: 11396 seconds
Overall Duration~ 3 hours 10 minutes
2972 seconds (26.1% of total time)
2493 seconds (21.9% of total time)
5931 seconds (52.0% of total time)
Adult
Child
Non Speech

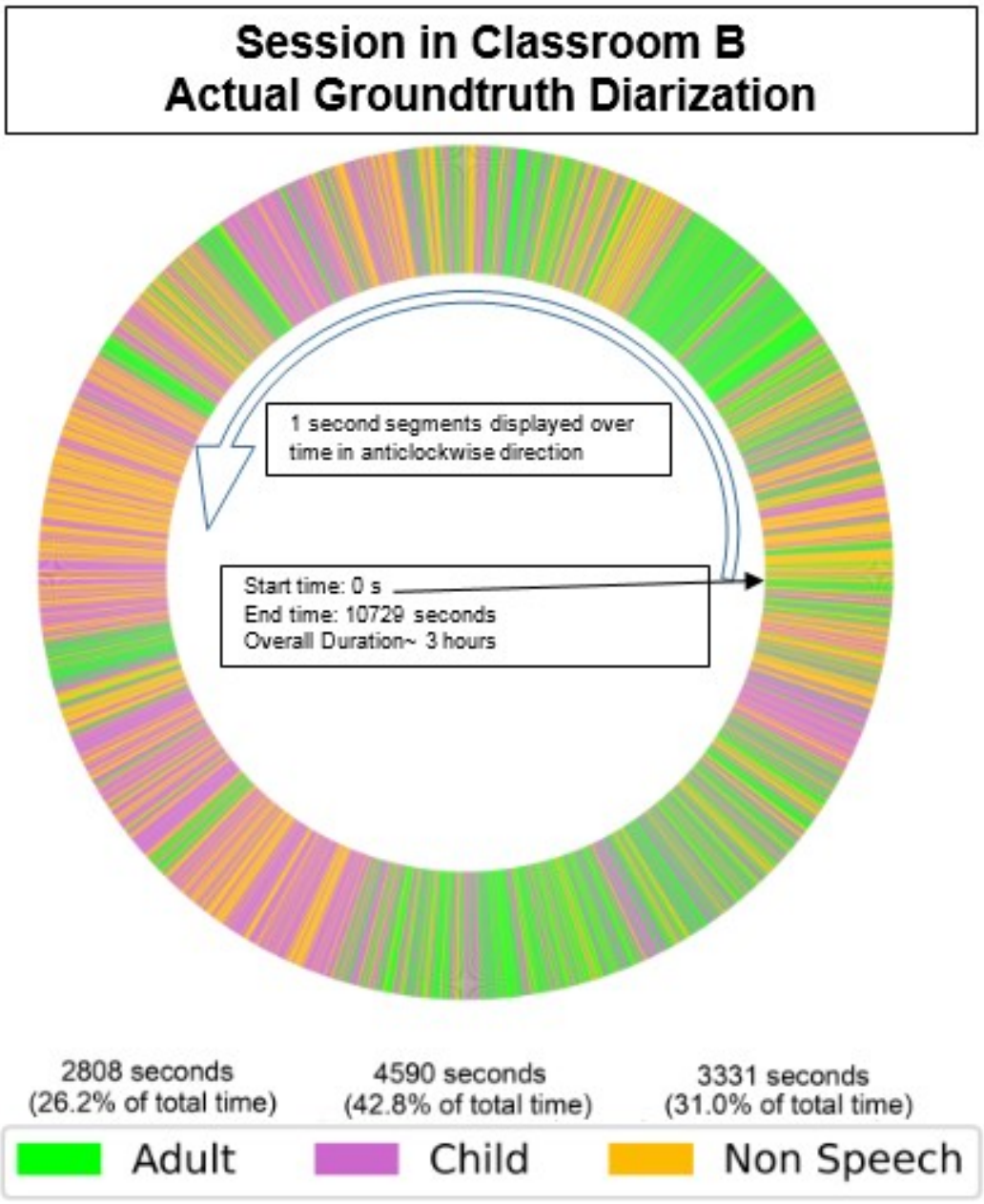
675

676


Session in Classroom A
Predicted ResNet18 Diarization
1 second segments displayed over time in anticlockwise direction
Start time: 0 s
End time: 11396 seconds
Overall Duration~ 3 hours 10 minutes
3329 seconds (29.2% of total time)
2789 seconds (24.5% of total time)
5278 seconds (46.3% of total time)
Adult
Child
Non Speech

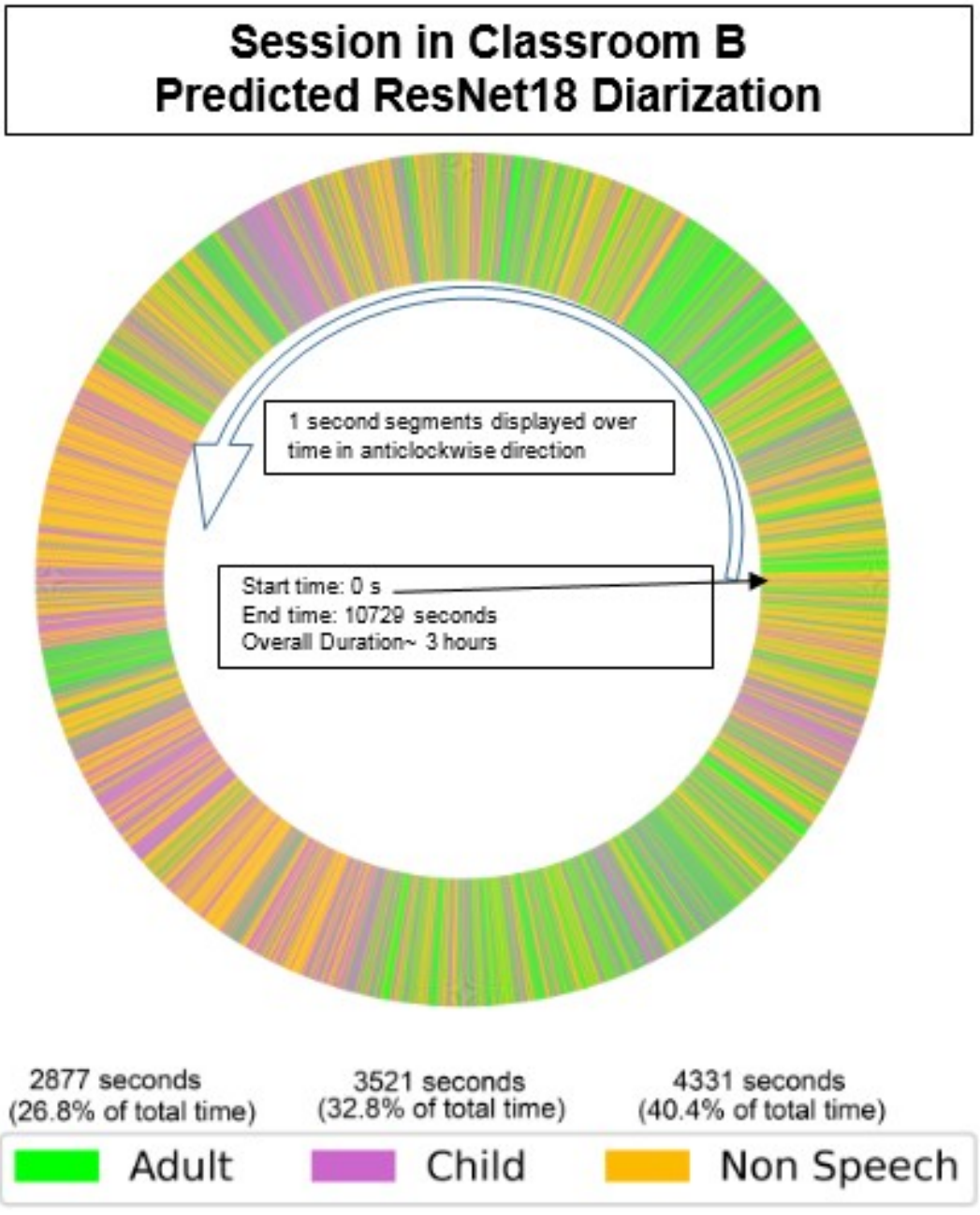
677

678


Session in Classroom B
Actual Groundtruth Diarization
1 second segments displayed over time in anticlockwise direction
Start time: 0 s
End time: 10729 seconds
Overall Duration~ 3 hours
2808 seconds (26.2% of total time)
4590 seconds (42.8% of total time)
3331 seconds (31.0% of total time)
Adult
Child
Non Speech

679

680


Session in Classroom B
Predicted ResNet18 Diarization
1 second segments displayed over time in anticlockwise direction
Start time: 0 s
End time: 10729 seconds
Overall Duration~ 3 hours
2877 seconds (26.8% of total time)
3521 seconds (32.8% of total time)
4331 seconds (40.4% of total time)
Adult
Child
Non Speech

681

682 **FIGURE CAPTIONS**

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

685 Figure 2: System diagram for child-adult speech-type Classification system.

686 Figure 3: System configurations for child-adult Diarization using ASR-based resegmentation.

687 Figure 4: Block diagram for End-to-End ECAPA-TDNN model.

688 Figure 5: Block diagram for End-to-End ResNet18 model.

689 Figure 6: Actual talktime for child and adult speech as represented by a donut diagram for a
690 session in classroom A with a child wearing the LENA device.

691 Figure 7: Predicted talktime for child and adult speech as represented by a donut diagram for a
692 session in classroom A with a child wearing the LENA device.

693 Figure 8: Actual talktime for child and adult speech as represented by a donut diagram for a
694 session in classroom B with a child wearing the LENA device.

695 Figure 9: Predicted talktime for child and adult speech as represented by a donut diagram for a
696 session in classroom B with a child wearing the LENA device.