FEBRUARY 08 2024

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; John H. L. Hansen ; Dwight Irvin; Jay Buzhardt

J. Acoust. Soc. Am. 155, 1198–1215 (2024)
https://doi.org/10.1121/10.0024353

JASA ARTICLE

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,[1] John H. L. Hansen,[1,a)] Dwight Irvin,[2] and Jay Buzhardt[3]

[1]*Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, Texas 75080, USA*

[2]*Anita Zucker Center for Excellence in Early Childhood Studies, University of Florida, Gainesville, Florida 32611, USA*

[3]*Juniper Garden's Children's Project (JGCP), University of Kansas, Kansas 66101, USA*

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 alternate deep learning-based lightweight, knowledge-distilled, diarization solutions for segmenting classroom interactions of 3–5 years 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 lightest CNN model achieves a best F1-score of ∼76.0% on data from two classrooms, based on dev and test sets of each classroom. It is utilized with automatic speech recognition-based re-segmentation 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.

https://doi.org/10.1121/10.0024353

(Received 7 May 2023; revised 21 December 2023; accepted 24 December 2023; published online 8 February 2024)

[Editor: B. Yegnanarayana] Pages: 1198–1215

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 speaker 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, 2024; 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 research questions that we attempt to solve with this study are as follows:

(1) How can we organize/combine well-researched deep neural networks from speech detection/ recognition literature, to develop robust child vs. adult speech diarization system for naturalistic audio streams recorded in preschool classrooms using mobile devices worn in

a)Email: John.Hansen@utdallas.edu

https://doi.org/10.1121/10.0024353

jackets by the children? Can shallow neural networks utilized for speech activity detection tasks in literature match the performance of deep neural networks investigated here?

(2) How accurately can we measure the interaction metrics of the children using the diarization system from step 1? Can audio stream diarization for a few hours of duration be visualized compactly using a software solution? Can it provide a visual assessment of the child-adult diarization system by analyzing subregions of such a diagram?

(3) Can the neural networks developed in step 1 be analyzed for regions of the input features that have greater contribution in terms of attention or saliency maps for this noisy, naturalistic audio dataset? Will the regions detected in these maps have any physical significance within the speech/audio domain?

(4) Can the performance of shallow neural networks developed in step 1, be improved using knowledge distillation (KD) from deeper neural networks developed in step 1?

The contributions of this study are stated as follows. First, we introduce the child-adult speech/speaker-type classification framework explained later for designing the scope of the speech-segment classification task. Next, standard convolutional neural network (CNN) 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. To improve performance of smaller CNN models, KD techniques [teacher-student (T/S) learning] of learning from advanced CNN architectures having larger number of parameters are applied to smaller CNNs with fewer parameters. The performance improvements are analyzed using attention and saliency maps for the input log-mel-spectrogram images. The speech/speaker-type detector is integrated with an automatic speech recognition (ASR) re-segmentation 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 DNN architecture embeddings for child vs adult speech-type classification. DNN 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 re-segmentation 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 mentioning the background including speaker characteristics and child-adult speech diarization. Section IV introduces our framework for end-to-end (E2E) child-adult speech/speaker-type classification which includes the assumptions and scope of our problem formulation. Section V provides details of the dataset. Section VI explains the procedure for producing the classification from raw audio including steps displayed in Fig. 1. 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 baseline CNN (CNN60) (Alam and Khan, 2020), spectro-temporal attention CNN (STACNN) (Lee *et al.*, 2020) and ResNet18 (He *et al.*, 2016) neural network used for segment classification. Section 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

i-Vectors (Dehak *et al.*, 2011b; Hansen and Hasan, 2015) are fixed length vectors that characterize speaker identity from arbitrary length sequential data (i.e., speech samples). They are standard features for speaker recognition (Dehak *et al.*, 2011b) and have been used extensively as a baseline system in recent studies. They have also been used for language recognition (Dehak *et al.*, 2011a), accent recognition (Bahari *et al.*, 2013), emotion recognition (Xia and Liu, 2012), etc. Alternatively, DNNs (McLaren *et al.*, 2015; Snyder *et al.*, 2018b; Snyder *et al.*, 2016) can be used to directly capture language or speaker characteristics. They achieve improved results over i-Vectors using mel-frequency cepstral coefficients (MFCCs) or log-mel-spectrograms as features. Here, log-mel-spectrograms can be

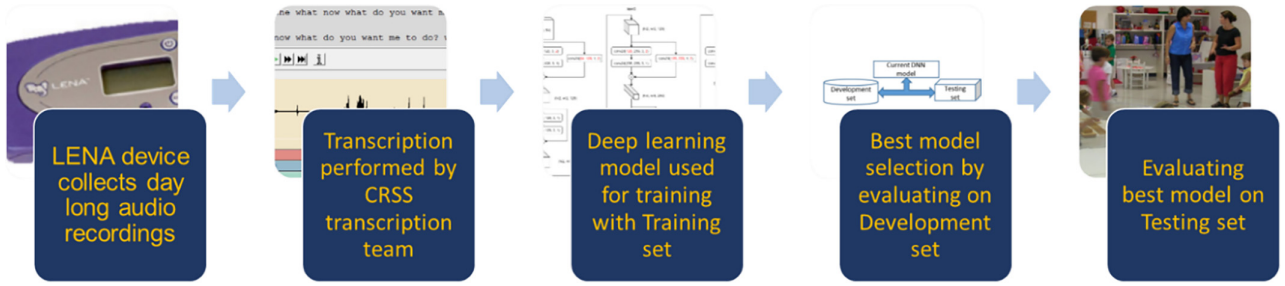

FIG. 1. (Color online) System diagram for child-adult speech-type classification system for alternate neural network architectures.

defined as the logarithm of the mel-spaced filterbanks. It is generated by the series of operations consisting of framing and windowing, followed by applying discrete or fast Fourier transform, then applying mel operation to the spectrogram and last logarithm operation. It helps in generating features where the frequencies are overlapped and non-uniformly spaced on the frequency axis such that the perceptual difference in frequencies stays the same for very high frequencies. Finally, applying discrete cosine transform (DCT) to log-mel-spectrograms generates MFCCs.

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 has utilized i-Vectors (Kothalkar *et al.*, 2019; Najafian *et al.*, 2016; Cristia *et al.*, 2018) and x-Vectors (Xie *et al.*, 2019a) as features for speaker classification. The SincNet-based speaker identification model has 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 s 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

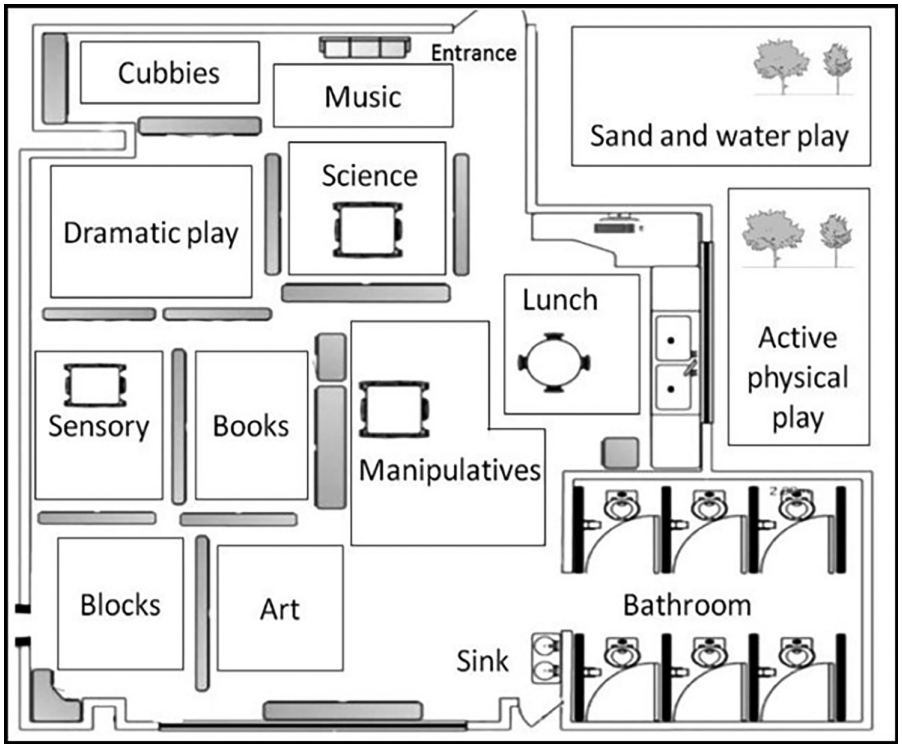

FIG. 2. Illustrative example of floor plan for child learning spaces within preschool classrooms (i.e., learning stations: Books/Reading, Science, etc.).

https://doi.org/10.1121/10.0024353

for real-time application in an E2E 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 and Laurent, 2021; Ozturk *et al.*, 2022; Radford *et al.*, 2022; Bain *et al.*, 2023) and possible E2E classification approach.

C. E2E 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 (Koluguri *et al.*, 2020; Kumar *et al.*, 2020) or traditional speaker clustering (Krishnamachari *et al.*, 2021). 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, and 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 2 s audio segments. This implied each segment could be reported as multiple voice-types resulting in multiple classes for downstream processing tasks like ASR or keyword spotting. 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 (CRNN), time delay neural networks (TDNN) (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 re-segmenting the initial speech segments generated from speech activity detection outputs. The IBM system (Huang *et al.*, 2008) 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 AND DIARIZATION

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 disease (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), CNN, recurrent neural networks (RNN), CRNN, and/or ResNet (He *et al.*, 2016) architectures have also been utilized for detection of speech (Silva *et al.*, 2017; Bai *et al.*, 2019b; Horiguchi *et al.*, 2021; Kwon *et al.*, 2021; Lin *et al.*, 2020a; Villalba *et al.*, 2019; Wang *et al.*, 2020; Braun and Tashev, 2021; Wilkinson and Niesler, 2021), speaker (Xie *et al.*, 2019b), music (Lee *et al.*, 2006), stuttering (Sheikh *et al.*, 2021, 2022), Parkinson's disease (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 CNN-based architectures. With the intent of utilizing noise-robust neural networks having lightweight architecture for potential real-time application, we propose to experimentally verify the detection of child and adult speech from non-speech in naturalistic audio using alternate types of CNNs having vanilla baseline (Alam and Khan, 2020), attention-based (Lee *et al.*, 2020), and ResNet18 architecture (He *et al.*, 2016) along with 2-dimensional (2D) input feature. Here, non-speech comprises silence, inaudible speech within crowd noise by adults or children, background music or electronic devices. Child-specific background non-speech further 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 logarithm of the mel-spaced spectrogram (log-mel-spectrogram) input features. The hypothesis is that regions of child/adult speech in the log-mel-spectrograms would be distinguishable by a DNN compared to regions of non-speech in both clean and noisy conditions. The outputs from these architectures are compared and combined with outputs from state-of-the-art speech activity detection systems for performance evaluations. Further verification and boundary refinement of the captured intelligible speech is performed using ASR re-segmentation of the detected speech segments.

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). Most of the LENA units record the data at 16 kHz sampling rate. Although some of the LENA units in classroom B have recorded the audio at 22 kHz, every audio segment is resampled at 16 kHz sampling rate before applying any signal processing technique such as feature extraction.

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. Both classrooms A and B have audio recorded from 4 adults in the distance from LENA devices worn by the children. In both the classrooms A and B, some of the audio sessions have one adult wearing the LENA recording device in a vest. Classroom A has 8 children wearing the LENA recorder device while classroom B has 9 children wearing the same.

The total audio from classroom A is of duration 61 h and 18 min and from classroom B is 63 h and 57 min. Thus, around 60 h of audio or approximately 230 000 segments of 1 s duration are used for training the classroom-specific models. For this dataset, an organized set of approximately 19 h of speech from classroom A and similar amount of speech from classroom B are established as the evaluation set for the corresponding classrooms.

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 data from classroom B, and tested on data for remaining timepoints from the same classroom B. Similarly, a model trained on data from 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

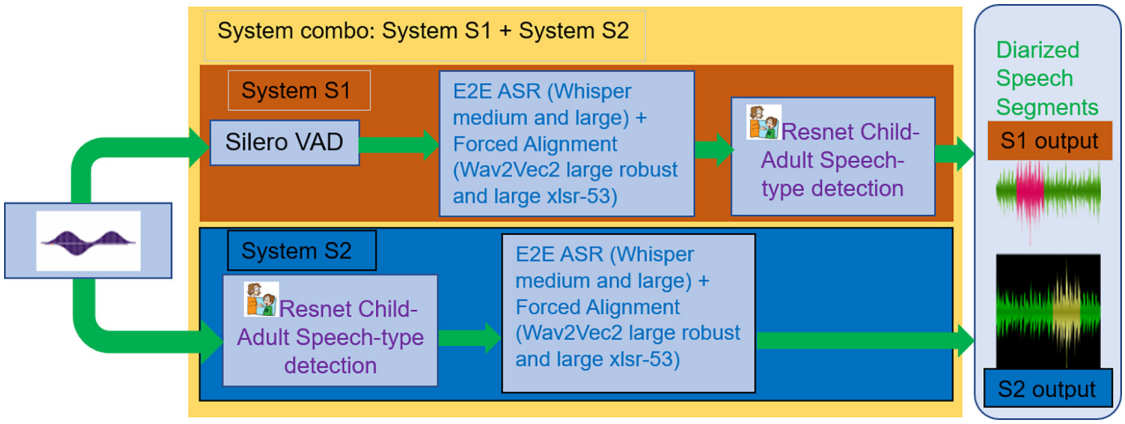

FIG. 3. (Color online) System configurations for child-adult diarization using ASR-based re-segmentation.

https://doi.org/10.1121/10.0024353

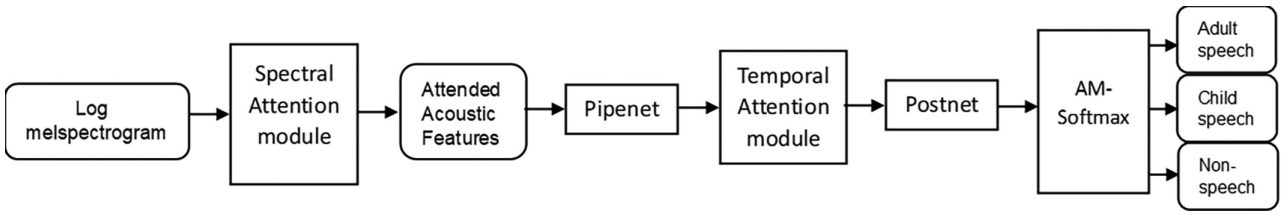

FIG. 4. Block diagram for STA CNN model.

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

Figure 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 is finally evaluated on the test set for speech/speaker-type classification as mentioned in the details of Sec. V C.

2. ASR re-segmentation for child-adult speech/speaker diarization

The ASR re-segmentation module consists of an 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 (Baevski *et al.*, 2020). 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, Facebook's wav2vec2-large-robust model finetuned on noisy conversational Switchboard speech data and XLSR-53 large model finetuned on English version of common voice for speech recognition were considered. The variations were based on the datasets utilized to fine-tune the base Wav2Vec2 model. The alternate configurations of the Speech-type classification and ASR re-segmentation modules are displayed in Fig. 3 and explained as follows:

a. System S1. System S1 consists of an industry-strength Silero (Silero Team, 2021) SAD system followed by an ASR-based re-segmentation module. The ASR-based re-segmentation module marks the start and end times of the ASR recognized segments from the SAD segmented audio files. The Silero SAD system consists of CNN and transformer-based architectures. Finally, if presence of child speech-type is detected by the speech-type detector ResNet module, the speech-type of the segment is marked accordingly. All combinations of Whisper E2E ASR models and Wav2Vec2 forced alignment models are utilized to produce multiple diarized segment system outputs for the entire test set.

b. System S2. System S2 consists of speech-type detector ResNet module followed by ASR-based re-segmentation module. Here, our speech-type detection module acts as an implicit speech activity detector with an additional class for detecting child speech. The ASR re-segmentation module performs the task of marking the timestamps of the ASR recognized speech-types. All combinations of Whisper E2E ASR models and Wav2Vec2 forced alignment models are utilized to produce multiple diarized segment system outputs for the entire test set.

c. System S1 + S2. In the combination system, we combine the multiple diarized segment outputs from systems S1 and S2.

3. Merging strategies of ASR re-segmentation module for child-adult speech/speaker diarization

Irrespective of the segment speech-type, for output segments with overlapping timestamps from any of the system outputs of system S1 and/or S2, the segments from the two

TABLE I. Configurations of all operators in ResNet-18 where I.C. represents input channel and O.C. represents output channel.

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	3, 1
		64, 64	3, 1
Layer2	25×20	64, 128	3, 2
		128, 128	3, 1
Layer3	13×10	128, 256	3, 2
		256, 256	3, 1
Layer4	7×5	256, 512	3, 2
		512, 512	3, 1
Avg. Pool	4×3	512, 3	1, 1
Embedding	1×1	—	1, 1
Softmax	1×1	—	—

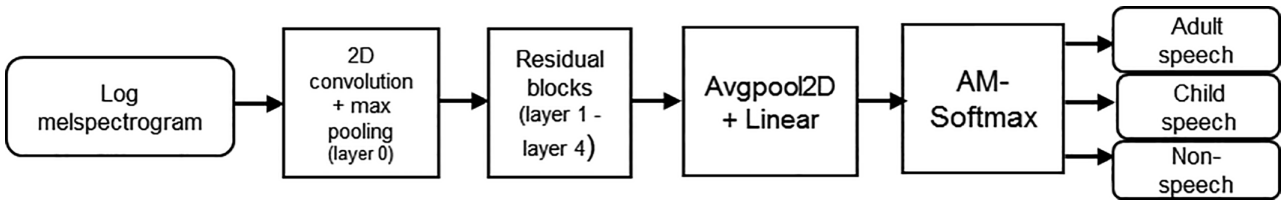

FIG. 5. Block diagram for E2E ResNet18 model.

systems are merged using following segment merging strategies:

(1) If a given segment from any of the system outputs of system S1/S2, has the same start or end time as that of a segment from any of the other system outputs, the segment with smaller talktime is discarded.
(2) If a given segment from any of the system outputs of system S1/S2, has the same start and end time as that of a segment from any of the other system outputs, and has presence of child speech/speaker-type in one of the speech segments, the segment is assigned with child speech/speaker-type class.
(3) If one segment from any of the system outputs of system S1/S2 completely bounds a segment from any of the other system outputs on the time axis, the smaller segment is discarded.
(4) For a given segment from any of the system outputs of system S1/S2, if it overlaps a segment from any of the other system outputs to its right along the time axis, the given segment from system S1/S2 is truncated to start of overlapping segment on its right along the time axis.

B. Data preprocessing

Audio recordings from both classroom A and B are divided into audio segments using a sliding window of 1000 ms duration with no overlap. Based on text transcripts from the data, ground-truth speaker-types are assigned as "adult" or "child" speech because of greater talk time by either the adult or child speaker over each 1000 ms 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 and input feature type

E2E 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). Here, the focus is to evaluate these for speaker/speech-type classification. Thus, looking at 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.

TABLE II. F1-score results on testing subset recordings of classroom A and classroom B audio.

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	CNN60	73.7%	72.9%	78.0%	74.9%
		STACNN	76.1%	77.5%	79.8%	77.9%
		ResNet18	76.9%	**81.7%**	80.0%	79.7%
		KD-CNN60	73.8%	76.4%	78.4%	76.3%
		KD-STACNN	**78.7%**	81.6%	**80.6%**	**80.4%**
Room B	Room A	CNN60	75.1%	77.7%	78.3%	76.9%
		STACNN	77.7%	78.5%	78.9%	78.4%
		ResNet18	**80.9%**	**82.6%**	**80.3%**	**81.3%**
		KD-CNN60	76.0%	77.7%	78.4%	77.4%
		KD-STACNN	79.9%	82.5%	80.2%	80.9%

https://doi.org/10.1121/10.0024353

TABLE III. Diarization error rate results on testing subset recordings of classroom A and classroom B audio.

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	20.7	14.7	39.4	74.8
		System S2	**1.3**	**3.5**	53.2	58.0
		System S1+S2	11.9	7.8	**21.6**	**41.3**
Room B	Room A	System S1	20.7	12.7	40.7	74.1
		System S2	**4.1**	**1.9**	40.3	46.3
		System S1+S2	11.3	7.0	**20.8**	**39.1**

1. CNN60 model

The CNN60 model is our baseline CNN system for the task of child-adult speaker-type detection with approximately 60 000 trainable parameters. It is composed of three convolution and pooling layers followed by two fully connected layers. The first convolution and pooling layer each use a (5×5) kernel. Second and final convolution layer use a (3×3) kernel and each is followed by a (2×2) max pooling layer. The three convolution layers consist of 32, 48, and 64 filters, respectively. The first fully connected layer has 64 hidden units, and the second fully connected layer has 3 hidden units corresponding to the speaker-types of child, adult, or non-speech.

2. STACNN model

The STACNN model as depicted in Fig. 4, consists of spectral and temporal attention modules that comprise of blocks of convolutional layers and multi-head attention layers, respectively. Specifically, the spectral attention layer is used to attend to speech features in the acoustic space and provides the robustness for the noisy data task of speaker-type detection.

The spectral attention module consists of T blocks with each block composed of a pair of convolutional layers and one-dimensional max pooling layer. The pipe-net contains two fully connected layers, each with N_p units, which acts as an information bridge between the spectral attention module and the temporal attention module. The temporal attention module attends to the most important positions from several neighboring input features using multi-head self-attention module. Temporal attention module is followed by two fully connected layers, each with N_p hidden units. The final linear layer has three hidden units for the three speaker-types to be detected. The logit output of the classification layer is passed through AM-Softmax loss function for CNN optimization.

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 the 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 as seen in the block diagram for ResNet model in Fig. 5.

4. Input representation for CNN60, STACNN and ResNet18

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

D. Knowledge distillation

Knowledge distillation (Hinton *et al.*, 2015; Gou *et al.*, 2021) 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 can incorporate the knowledge that the teacher network discovered earlier, allowing the performance of the student network to be better than if it were trained with labels only.

Let (x_i, y_i) denote a training sample in dataset (X, Y) where x_i contains a sequence of N input speech frames and y_i is the predicted speaker-type class. Hinton *et al.* (2015) introduced "softmax temperature" function $\sigma_{\tau_s}(.)$ 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 $\sigma(.)$ when τ_s equals 1. The softmax function value for instance x_i can be calculated as

$$\sigma_{\tau_s}(x_i) = \frac{\exp(x_i/\tau_S)}{\sum_{x_j \in X} \exp(x_j/\tau_S)}. \quad (1)$$

KD loss is defined as the sum of the KL-divergence between logits of teacher network output with the student network output and the cross-entropy of the dataset (X, Y). 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 $f_{\theta_T}(\cdot)$. $P_{f_\theta}(.)$ indicate the logit response of x_i from f_θ (.). The student network f_{θ_S} can then be learned by the following relation with the parameters of the teacher model f_{θ_T}:

$$min_{\theta_s} \sum_{(x_i,y_i)\in(X,Y)} \left(\alpha \times \tau_s^2 \times KL\left(\sigma_{\tau_s}\left(P_{f_{\theta_T}}(x_i)\right), \sigma_{\tau_s}\left(P_{f_{\theta_S}}(x_i)\right)\right)\right.$$
$$\left. + (1-\alpha) CE\left(\sigma_{\tau_s}\left(P_{f_{\theta_S}}(x_i)\right), y_i\right)\right), \quad (2)$$

where $KL(\cdots)$ and $CE(\cdots)$ are the Kullback-Leibler (KL-divergence) divergence and cross-entropy loss, respectively. Another hyperparameter α is utilized to perform the weighting between T/S loss and cross-entropy loss and performs well when the weight for T/S loss is higher.

In our case, we utilize the ResNet18 model to be the teacher for teaching the speaker-type detection task to CNN60 and STACNN models through the KD loss. Thus, KD loss comprising of T/S loss and cross-entropy loss is used in addition to the AM-Softmax loss. The models generated through KD training procedure are from now referred to as KDCNN60 and KD-STACNN for CNN60 and STACNN models, respectively.

VII. EXPERIMENTAL DESIGN AND METRICS

A. Experimental design

For uniformity in system evaluation, all CNN architectures including 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. Figures 4 and 5 highlight the block diagram for STACNN 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.

For the KD procedure, hyperparameters $\alpha = 0.9$ and $\tau_s = 4$ are set based on empirical observations. It ensures that the T/S loss receives much higher weightage compared to cross-entropy loss.

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

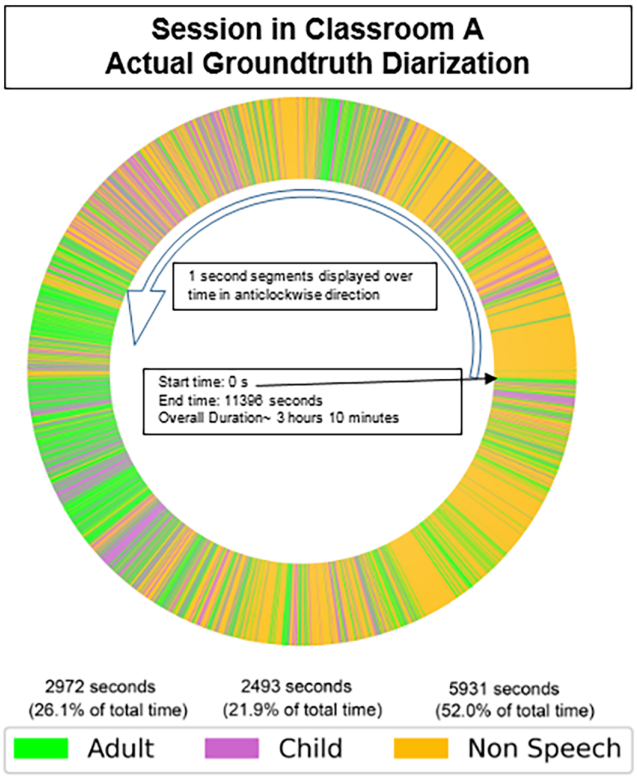

FIG. 6. (Color online) Actual talktime for child and adult speech as represented by a donut diagram for a session in classroom A with a child wearing the LENA device.

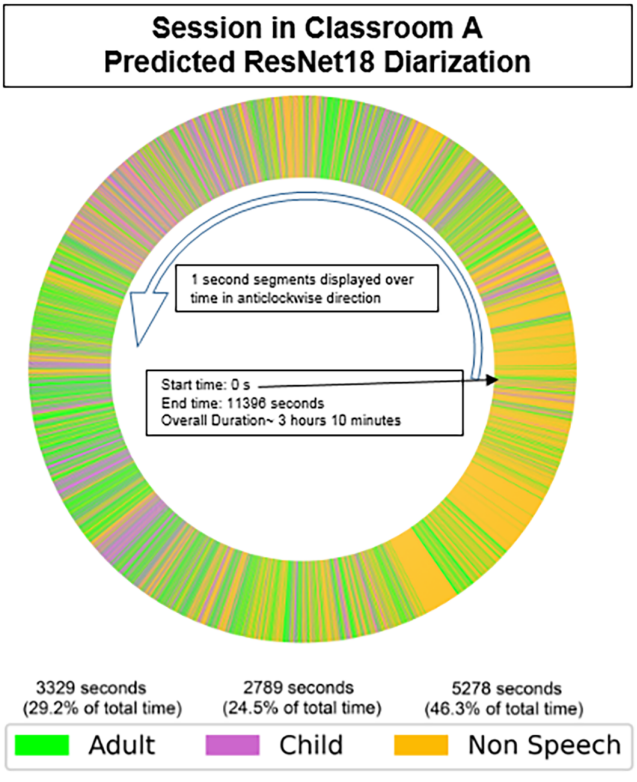

FIG. 7. (Color online) Predicted talktime for child and adult speech as represented by a donut diagram for a session in classroom A with a child wearing the LENA device.

https://doi.org/10.1121/10.0024353

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,

$$Precision = \frac{TP}{TP + FP}, \tag{3}$$

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,

$$Recall = \frac{TP}{TP + FN}, \tag{4}$$

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.

$$F1_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{5}$$

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,

$$DER = E_{spkr} + E_{miss} + E_{FA}. \tag{6}$$

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.

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 an absolute improvement of +8.4% for test data from classroom A, and an 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 +45.4% for speaker confusion error rate, +48.9% for missed speech error rate, and +47.2% for overall DER. Relative improvements by system S1+S2 vs system S1 on classroom B test audio data are +42.5% for speaker confusion error rate, +45.2% for missed speech error rate, and +44.8% 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 45%–49% 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 ~59% 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 of the combination of Silero VAD and ASR models. 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. Thus, models trained on multi-condition, massive speech corpora for multiple speech tasks are hypothesized to provide complementary information in terms of acoustical environmental conditions to models trained on domain-specific speech data for focused task of child-adult speech/speaker-type diarization.

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 a "donut

diagram" that reflects the speech density per speaker over different times of a session. The donut diagrams present an easy way to visualize the missed speech and false alarms at the segment level along with performance comparison in a temporal manner. 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° later.

Figures 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. Here, segment-level false alarms can be recognized between 250° and 300° based on the thickening of adult speech segments in that region of the diagram for Fig. 7 compared to Fig. 6. We see the percentage difference between predicted and actual talktimes differ between 2.6% (child) and 3.1% (adult). 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) is also matched closely between Figs. 6 and 7, where thick segments would have speech for a single type for significant duration.

Figures 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. Here, segment-level false alarms can be recognized between 150° and 200° on account of the empty spaces in that region of the diagram for Fig. 9 when compared with Fig. 8. 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 [which is 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° represents significant adult talk during that time of the session, along with child speech in between in classroom B with a child wearing the LENA device.

For example, certain thick green segments are matched at 85° and between 150° and 210°. Similar, thick pink segments are present between 180° and 210°.

C. Visualization of attention maps over input spectrogram images for the predicted label based on the output of ResNet, STACNN, and KD-STACNN models

In Fig. 10, we have presented the log-mel-spectrograms of four random audio segments containing adult speech with corresponding attention map outputs of the four ResNet blocks (denoted as g0, g1, g2, g3) in the first four rows. These are followed by log-mel-spectrograms of four random audio segments containing child speech with corresponding

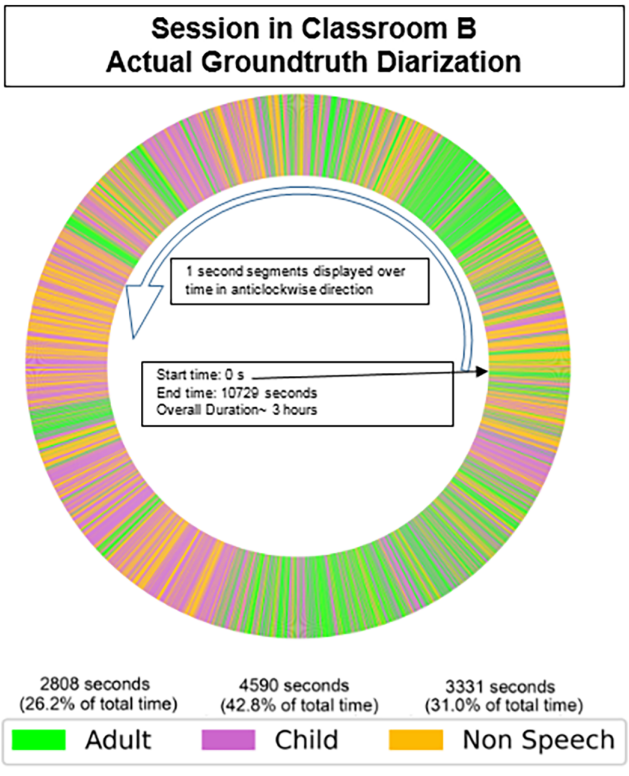

FIG. 8. (Color online) Actual talktime for child and adult speech as represented by a donut diagram for a session in classroom B with a child wearing the LENA device.

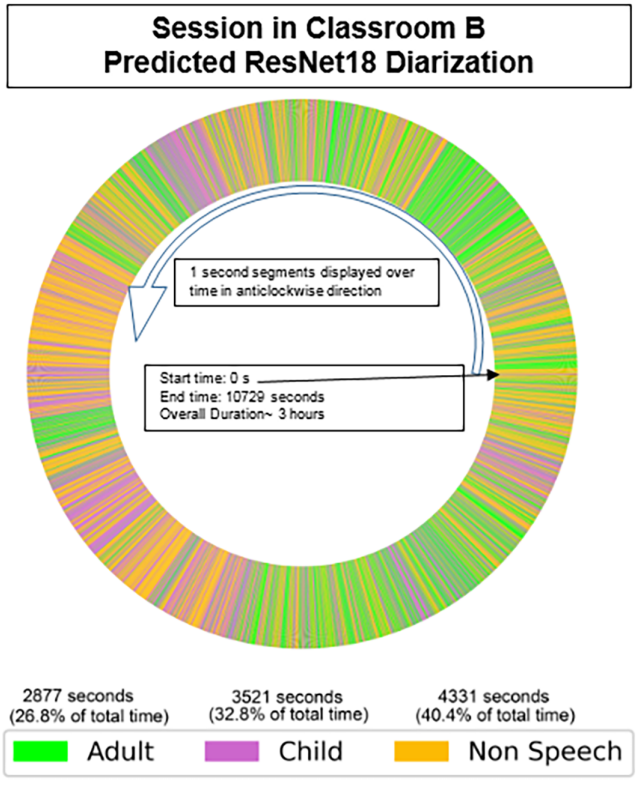

FIG. 9. (Color online) Predicted talktime for child and adult speech as represented by a donut diagram for a session in classroom B with a child wearing the LENA device.

https://doi.org/10.1121/10.0024353

attention map outputs of the four ResNet blocks in the last four rows of Fig. 10.

Here, the second column (denoted as g0) represents the output of the first ResNet18 block and provides detailed view of the regions corresponding to the image, that have greater contribution for the ResNet model inference score from the forward operation of the CNN. It is clearly visible by comparing columns 1 and 2 of Fig. 10 that high energy regions of the attention maps from column 2 have similar shape and location as the high energy formant frequency

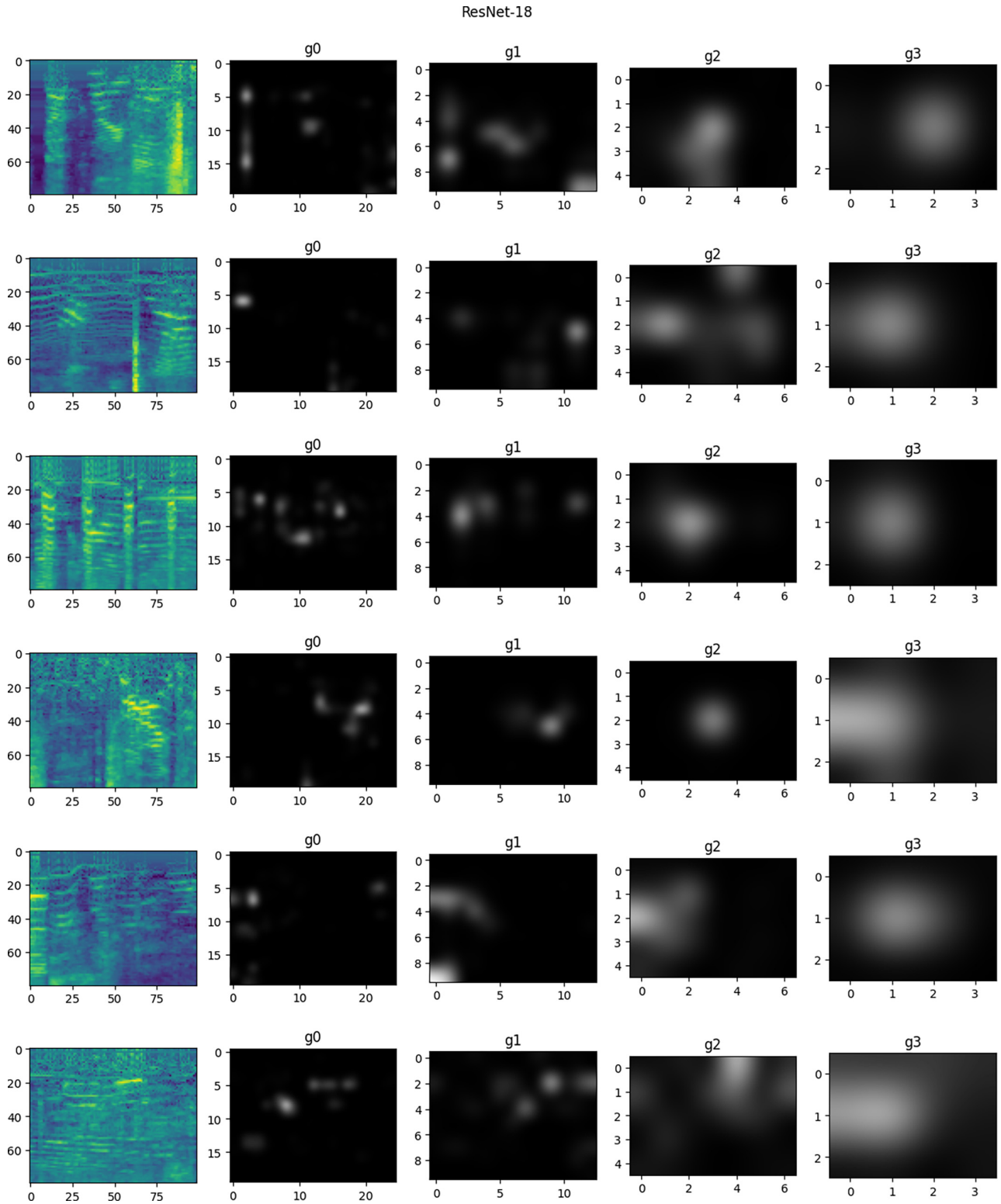

FIG. 10. (Color online) Log-mel-spectrograms of eight random audio files containing adult (top four) and child (bottom four) speech along with the attention map outputs from the four blocks of the ResNet model.

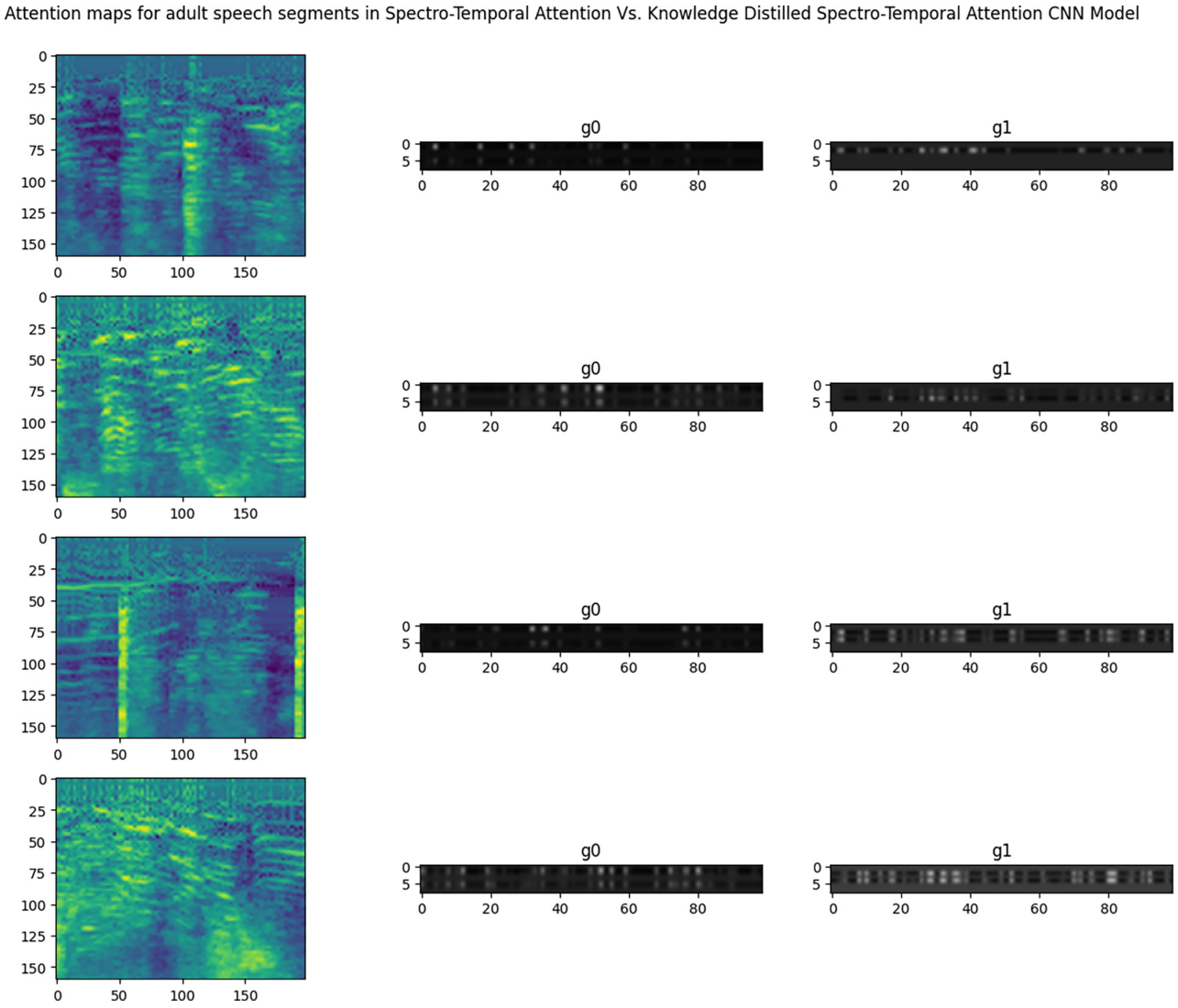

FIG. 11. (Color online) Log-mel-spectrograms of four random audio segments containing adult speech that show improvement due to KD along with the corresponding attention map outputs of multihead attention layer in the spectro-temporal attention CNN model.

contours of the log-mel-spectrograms in column 1 of Fig. 10. Thus, the ResNet model can detect regions of high format frequencies for predicting the speaker-type class.

The output of the deeper layers of the ResNet18 model in terms of the outputs for the second, third, and fourth ResNet blocks are presented in the third, fourth, and fifth columns of Fig. 10. Since the block size reduces for deeper layers of the model with corresponding increase in channel size, a rescaled version of the output image displays the regions of focus after application of ResNet blocks. Since deeper layers of the CNN learn high-level features for a given classification task, the outputs are as per expectations of standard procedure for inference in CNNs.

Figure 11 presents log-mel-spectrograms of four random audio segments containing adult speech that show improvement due to KD along with the corresponding attention map outputs of multihead attention layer in the STACNN model. Here, the second column (denoted as g0) presents the output of the multihead attention layer of the STACNN model, and the third column (denoted as g1) presents the output of the multihead attention layer of the KD-STACNN model.

Multiple heads are active for the input audio spectrograms and across multiple timestamps for KD-STACNN model vs one or two attention heads and lesser timestamps for the same input audio spectrograms.

Figure 12 presents log-mel-spectrograms of four random audio segments containing child speech that show improvement due to KD along with the corresponding attention map outputs of multihead attention layer in the STACNN model. Here, the second column (denoted as g0) presents the output of the multihead attention layer of the STACNN model, and the third column (denoted as g1) presents the output of the multihead attention layer of the KD-STACNN model.

Multiple heads are active for the input audio spectrograms and across multiple timestamps for KD-STACNN model vs one or two attention heads and lesser timestamps for the same input audio spectrograms.

https://doi.org/10.1121/10.0024353

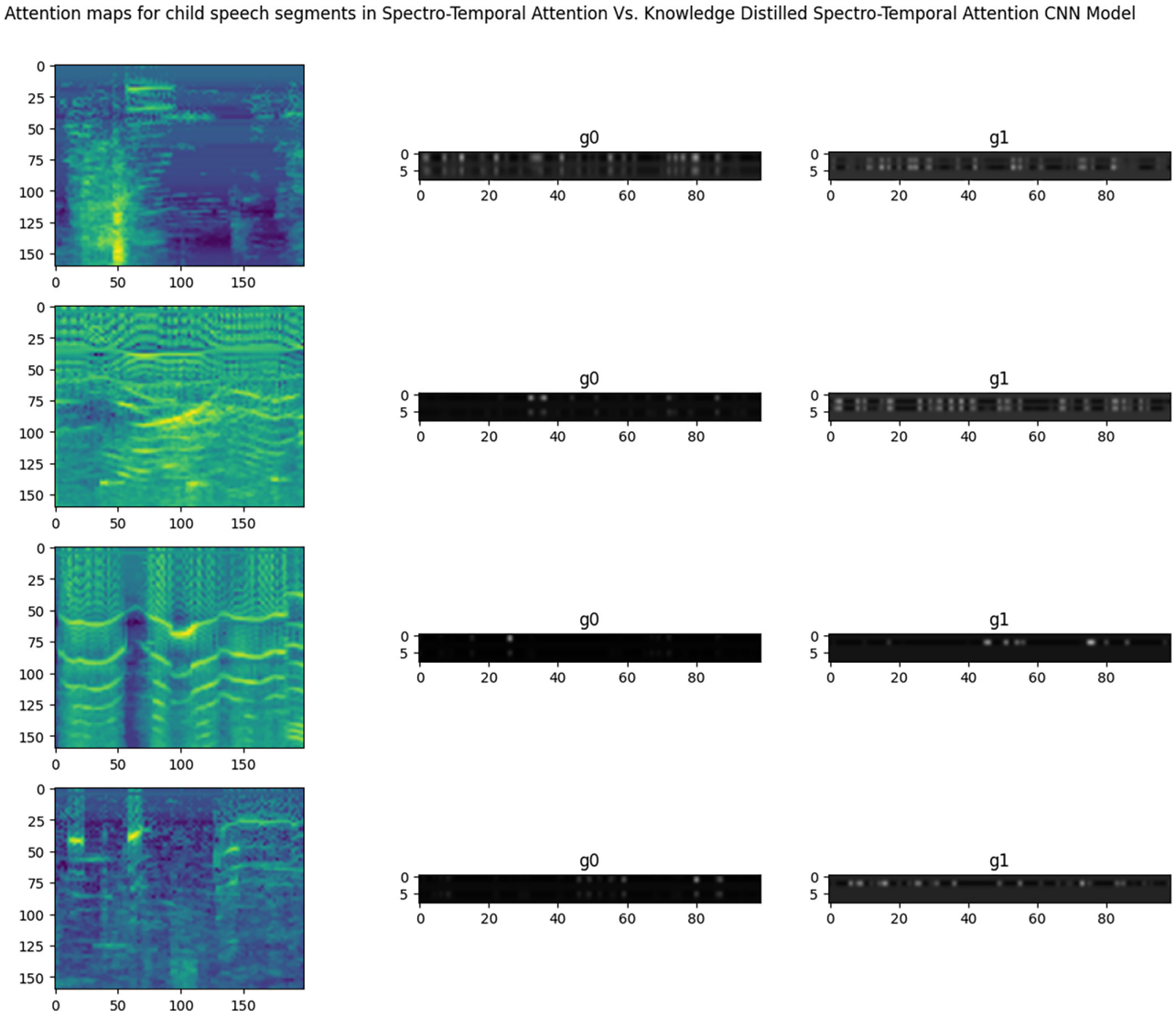

FIG. 12. (Color online) Log-mel-spectrograms of four random audio segments containing child speech that show improvement due to KD along with the corresponding attention map outputs of multihead attention layer in the spectro-temporal attention CNN model.

D. Visualization of saliency maps over input spectrogram images for the predicted label based on the output of STACNN and KD-STACNN models

The saliency map generation is inspired by the basics of backpropagation algorithm, which states that the deltas obtained at a layer L equal the gradient of the loss incurred by the subgraph of the CNN below L with respect to the outputs at L. Thus, backpropagating till the input data layer will yield us the gradient of the loss incurred by the whole CNN with respect to the input itself, thereby providing us the importance/saliency over the input image. Thus, saliency map for an image comprises the important pixels in the image that influence class score of the network prediction.

In Fig. 13 we have presented the log-mel-spectrograms of four random audio segments containing adult speech that show improvement due to KD along with the corresponding saliency map outputs in the spectro-temporal attention CNN model without KD (second column) and with KD (third column) KD. The second and third columns are denoted as g0 and g1, respectively. In the second column of Fig. 13, patterns that have strong contribution for prediction of adult speech display brighter colors (bright red to yellow) while lower contribution regions are marked black in color (dark red to black) as per the color map. Regions that show an increase in brightness in the third column of Fig. 13 vs the second column of Fig. 13 are marked in orange boxes. Another notable difference between the saliency maps of the two columns of Fig. 13 is that pixels with higher contribution towards the prediction score in the third column occur in consecutive locations and a definite pattern, like the contours of formant frequencies in log-mel-spectrogram images of the first column. Even certain pixels of lower score contribution that are completely missing in the second column of Fig. 13 are present in the third and fourth rows of the third column in Fig. 13.

Similarly, we have presented the log-Mel-Spectrograms of four random audio segments containing child speech that show improvement due to KD along with the corresponding

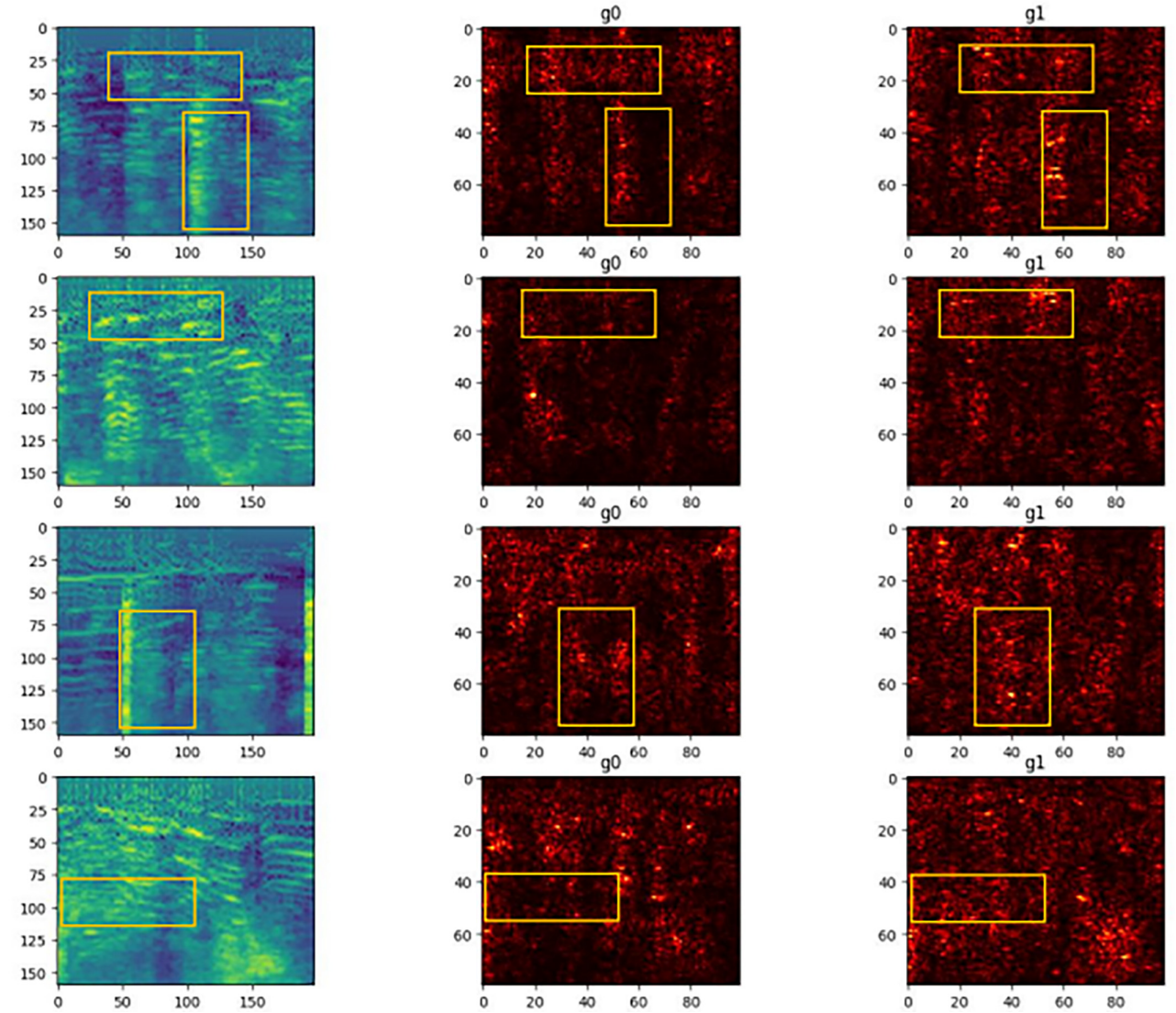

FIG. 13. (Color online) Log-mel-spectrograms of four random audio segments containing adult speech that show improvement due to KD along with the corresponding saliency map outputs of multihead attention layer in the spectro-temporal attention CNN model.

saliency map outputs in the spectro-temporal attention CNN model without KD (second column) and with KD (third column) KD in Fig. 14. The second and third columns are denoted as g0 and g1, respectively. Here, the first, second, and third rows of the third column can be seen containing consecutive pixels providing the greatest contribution to the CNN forward inference score due to the presence of yellow, formant shaped contours within the orange boxes.

Thus, we can detect the presence of child speech or adult speech better after application of KD to STACNN models due to improved detection of presence of relevant formant contours as observed from the saliency maps for log-mel-spectrograms.

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, STACNN models provided good and consistent results in terms of F1-scores for all speech activity types recognized based on the posterior probabilities. However, a ResNet model with 80-dimensional log-Mel-spectrograms inputs have outperformed STACNN model in terms of F1-scores of all speech activity types as well as DER. Knowledge distillation-based approaches were applied to CNN60 and STACNN models which improved their performance for the speaker-type classification task on the evaluation set. Also, the performance of STACNN model was very close to ResNet18 model in terms of F1-score for evaluation set of classroom A and better than performance of ResNet18 model for evaluation set of classroom B. Thus, KD-STACNN models can be substituted for ResNet18 models when smaller model sizes are desired such as for real-time application of speaker-type

https://doi.org/10.1121/10.0024353

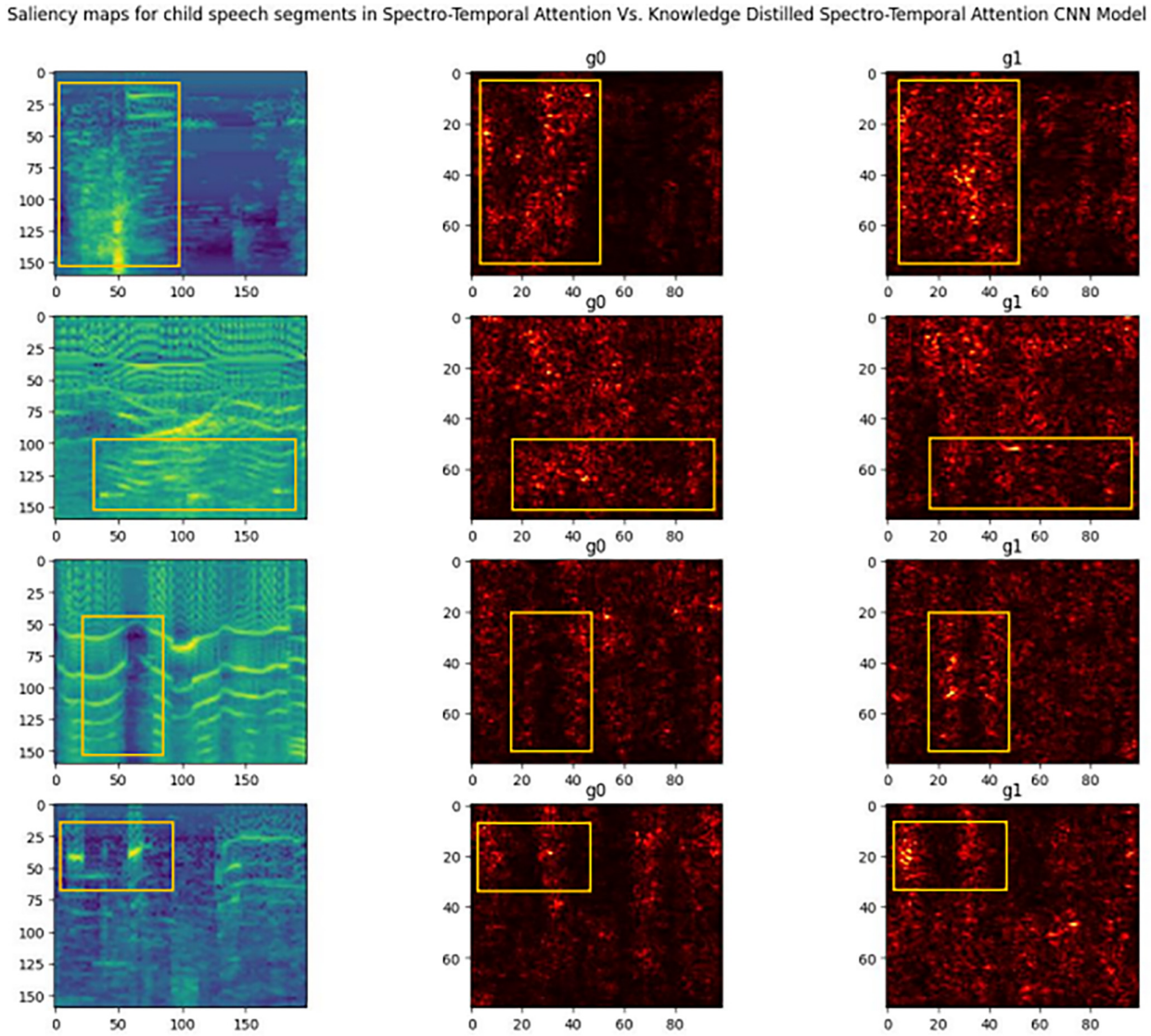

FIG. 14. (Color online) Log-mel-spectrograms of four random audio segments containing child speech that show improvement due to KD along with the corresponding saliency map outputs of multihead attention layer in the spectro-temporal attention CNN model.

detection. 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 1 s 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, 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. Regions or pixels of input log-mel-spectrograms contributing for speaker-type prediction were discovered using attention maps from gradients of model predictions for the corresponding input audio segments. Similar attention maps were also presented for STACNN models for the multihead attention layer. The improvements achieved in KD-STACNN models over STACNN model were tracked in attention and saliency map outputs for model inference over input log-mel-spectrogram images of the audio segments. The child-adult speech-type predicted outputs are combined with an ASR re-segmentation module in various configurations to provide multiple child-adult diarization systems. A specific combination of these 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 alternate ASR re-segmentation modules including those 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.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) under Grant No. 1918032 (UT Dallas CRSS) (PI: J.H.L.H.), and partially by the University of Texas at Dallas (UTDallas) from the Distinguished University Chair in Telecommunications Engineering held by J.H.L.H. The authors declare no conflict of interest given NSF support and no company/commercial interests. Also, no ethics issues/approval are needed with this study, and speech data used by CRSS-UTDallas followed an approved IRB protocol. Data from this study would be shared, however, there are potentially some distribution constraints regarding actual child audio data from classroom settings. The authors would also like to thank all the teachers and children who participated in the data collection sessions.

Alam, T., and Khan, A. (**2020**). "Lightweight CNN for robust voice activity detection," in *International Conference on Speech and Computer* (Springer International Publishing, Cham), pp. 1–12.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (**2020**). "wav2vec 2.0: A framework for self-supervised learning of speech representations," Adv. Neural Inf. Process. Syst. **33**, 12449–12460.

Bahari, M. H., Saeidi, R., Van hamme, H., and Van Leeuwen, D. (**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, New York), pp. 7344–7348.

Bai, H., Chen, H., and Yan, Y. (**2019a**). "Audio scene classification with discriminatively-trained segment-level features," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (IEEE, New York), pp. 354–359.

Bai, Y., Yi, J., Tao, J., Wen, Z., and Liu, B. (**2019b**). "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, New York), pp. 1173–1178.

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

Braun, S., and Tashev, I. (**2021**). "On training targets for noise-robust voice activity detection," in *2021 29th European Signal Processing Conference (EUSIPCO)* (IEEE, New York), pp. 421–425.

Bredin, H., and Laurent, A. (**2021**). "End-to-end speaker segmentation for overlap-aware re-segmentation," in *Interspeech 2021*.

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

Cristia, A., Ganesh, S., Casillas, M., and 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," in *ISCA INTERSPEECH-2021*, pp. 3560–3564.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (**2011a**). "Front-end factor analysis for speaker verification," IEEE Trans. Audio. Speech. Lang. Process. **19**(4), 788–798.

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., and Dehak, R. (**2011b**). " Language recognition via i-vectors and dimensionality reduction," in *ISCA INTERSPEECH-2011*, pp. 857–860.

Desplanques, B., Thienpondt, J., and Demuynck, K. (**2020**). "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *ISCA INTERSPEECH-2020*, pp. 3830–3834.

Dubey, H., Sangwan, A., and Hansen, J. H. L. (**2019**). "Transfer learning using raw waveform SincNet for robust speaker diarization," *IEEE ICASSP-2019: International Conference on Acoustics, Speech and Signal Processing*, pp. 6296–6300.

Garcia-Romero, D., and McCree, A. (**2016**). "Stacked long-term TDNN for spoken language recognition," in *ISCA INTERSPEECH-2016*, pp. 3226–3230.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (**2021**). "Knowledge distillation: A survey," Int. J. Comput. Vis. **129**, 1789–1819.

Gupta, S., Patil, A. T., Purohit, M., Parmar, M., Patel, M., Patil, H. A., and Guido, R. C. (**2021**). "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," Neural Netw. **139**, 105–117.

Hansen, J. H. L., and Hasan, T. (**2015**). "Speaker recognition by machines and humans: A tutorial review," IEEE Signal Process. Mag. **32**(6), 74–99.

Hart, B., and Risley, T. R. (**1995**). *Meaningful Differences in the Everyday Experience of Young American Children* (Brookes, Baltimore, MD).

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern* "*recognition*," pp. 770–778.

Hebbar, R., Somandepalli, K., and Narayanan, S. (**2019**). "Robust speech activity detection in movie audio: Data resources and experimental evaluation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York), pp. 4105–4109.

Hinton, G., Vinyals, O., and Dean, J. (**2015**). "Distilling the knowledge in a neural network," Statistics **1050**, 9.

Horiguchi, S., Yalta, N., Garcia, P., Takashima, Y., Xue, Y., Raj, D., Huang, Z., Fujita, Y., Watanabe, S., and Khudanpur, S. (**2021**). "The Hitachi-JHU DIHARD-III system: Competitive end-to-end neural diarization and x-vector clustering systems combined by dover-lap," arXiv:2102.01363.

Huang, J., Marcheret, E., Visweswariah, K., and Potamianos, G. (**2008**). "The IBM RT07 evaluation systems for speaker diarization on lecture meetings," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, Baltimore, MD, May 8-11, 2007, Revised Selected Papers (Springer, Berlin), pp. 497–508.

Kim, M., Ki, T., Anshu, A., and Apsingekar, V. R. (**2021**).). "North America Bixby speaker diarization system for the VoxCeleb speaker recognition challenge 2021."

Koluguri, N. R., Kumar, M., Kim, S. H., Lord, C., and Narayanan, S. (**2020**). "Meta-learning for robust child-adult classification from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York), pp. 8094–8098.

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

Krishnamachari, S., Kumar, M., Kim, S. H., Lord, C., and Narayanan, S. (**2021**). "Developing neural representations for robust child-adult diarization," in *2021 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, New York), pp. 590–597.

Kumar, M., Kim, S. H., Lord, C., and Narayanan, S. (**2020**). "Improving speaker diarization for naturalistic child-adult conversational interactions using contextual information," J. Acoust. Soc. Am. **147**(2), EL196–EL200.

Kwon, Y., Heo, H. S., Huh, J., Lee, B.-J., and Chung, J. S. (**2021**). "Look who's not talking," in *2021 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, New York), pp. 567–573.

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., and Cristia, A. (**2020**). "An open-source voice type classifier for child-centered daylong recordings," arXiv:2005.12656.

Lee, J.-W., Park, S.-B., and Kim, S.-K. (**2006**). "Music genre classification using a time-delay neural network," in *International Symposium on Neural Networks* (Springer, Berlin), pp. 178–187.

Lee, Y., Min, J., Han, D. K., and Ko, H. (**2020**). "Spectro-temporal attention-based voice activity detection," IEEE Signal Process. Lett. **27**, 131–135.

LENA (**2024**). https://www.lenafoundation.org (Last accessed August 22, 2022).

Lin, Q., Cai, W., Yang, L., Wang, J., Zhang, J., and Li, M. (**2020b**). "DIHARD II is still hard: Experimental results and discussions from the DKU-LENOVO team," in *Proceedings of Odyssey 2020 the Speaker and Language Recognition Workshop*, pp. 102–109.

Lin, Q., Li, T., and Li, M. (**2020a**). "The DKU speech activity detection and speaker identification systems for fearless steps challenge phase-02," in *INTERSPEECH*, pp. 2607–2611.

McLaren, M., Lei, Y., and Ferrer, L. (**2015**). "Advances in deep neural network approaches to speaker recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York), pp. 4814–4818.

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

Ogura, M., and Haynes, M. (**2021**). "X-vector based voice activity detection for multi-genre broadcast speech-to-text," arXiv:2112.05016.

Ozturk, M. Z., Wu, C., Wang, B., Wu, M., and Liu, K. R. (**2022**). "Beyond microphone: MmWave-based interference-resilient voice activity detection," in *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, pp. 7–12.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (**2022**). "Robust speech recognition via large-scale weak-supervision," 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, New York), 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, New York), 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, Washington, DC).

Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S., and Khudanpur, S. (**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, New York), 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:2204.01564.

Silero Team (**2021**). "Silero VAD: Pre-trained enterprise-grade Voice Activated Detector (VAD), number detector and language classifier," https://github.com/snakers4/silero-vad (Last viewed January 22, 2024).

Silva, D. A., Stuchi, J. A., Violato, R. P. V., and Cuozzo, L. G. D. (**2017**). "Exploring convolutional neural networks for voice activity detection," in *Cognitive Technologies* (Springer, Berlin), pp. 37–47.

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

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

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur, S. (**2016**). "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, New York), pp. 165–170.

Thienpondt, J., Desplanques, B., and Demuynck, K. (**2020**). "The IDLAB VoxCeleb speaker recognition challenge 2020 system description," arXiv:2010.12468.

Villalba, J., Garcia-Romero, D., Chen, N., Sell, G., Borgstrom, J., McCree, A., Snyder, D., Kataria, S., Garcia-Perera, P., Richardson, F., and Torres–Carrasquillo, P. A. (**2019**). "The JHU-MIT system description for NIST SRE19 AV," in *NIST SRE19 Workshop*.

Wang, M., Huang, Q., Zhang, J., Li, Z., Pu, H., Lei, J., and Wang, L. (**2020**). "Deep learning approaches for voice activity detection," in *Cyber Security Intelligence and Analytics* (Springer International Publishing, New York), pp. 816–826.

Wang, W., Wu, H., and Li, M. (**2019**). "Deep neural networks with batch speaker normalization for intoxicated speech detection," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, New York), pp. 1323–1327.

Wilkinson, N., and Niesler, T. (**2021**). "A hybrid CNN-BiLSTM voice activity detector," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York), pp. 6803–6807.

Wodzinski, M., Skalski, A., Hemmerling, D., Orozco-Arroyave, J. R., and Nöth, E. (**2019**). "Deep learning approach to Parkinsons disease detection using voice recordings and convolutional neural network dedicated to image classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, New York), pp. 717–720.

Xia, R., and Liu, Y. (**2012**). "Using i-vector space model for emotion recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*.

Xie, J., García-Perera, L. P., Povey, D., and Khudanpur, S. (**2019a**). "Multi-PLDA diarization on children's speech," in *Interspeech*, pp. 376–380.

Xie, W., Nagrani, A., Chung, J. S., and Zisserman, A. (**2019b**). "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York), pp. 5791–5795.

Xue, Y., Horiguchi, S., Fujita, Y., Watanabe, S., García, P., and Nagamatsu, K. (**2021**). "Online end-to-end neural diarization with speaker-tracing buffer," in *2021 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, New York), 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, New York), pp. 7770–7774.