Assessing the Expressive Language Levels of Autistic Children in Home Intervention

Yueran Pan, Biyuan Chen, Wenxing Liu, Ming Cheng, Dong Zhang,
Hongzhu Deng, Xiaobing Zou and Ming Li, *Senior Member, IEEE*

***Abstract*—The World Health Organization (WHO) has established the Caregiver Skill Training (CST) program, designed to empower families with children diagnosed with Autism Spectrum Disorder the essential caregiving skills. The Joint Engagement Rating Inventory (JERI) protocol evaluates participants' engagement levels within the CST initiative. Traditionally, rating the Expressive Language Level and Use (EXLA) item in JERI relies on retrospective video analysis conducted by qualified professionals, thus incurring substantial labor costs. This study introduces a multimodal behavioral signal-processing framework designed to analyze both child and caregiver behaviors automatically, thereby rating EXLA. Initially, raw audio and video signals are segmented into concise intervals via voice activity detection, speaker diarization and speaker age classification, serving the dual purpose of eliminating non-speech content and tagging each segment with its respective speaker. Subsequently, we extract an array of audio-visual features, encompassing our proposed interpretable, hand-crafted textual features, end-to-end audio embeddings and end-to-end video embeddings. Finally, these features are fused at the feature level to train a linear regression model aimed at predicting the EXLA scores. Our framework has been evaluated on the largest in-the-wild database currently available under the CST program. Experimental results indicate that the proposed system achieves a Pearson Correlation Coefficient of 0.768 against the expert ratings, evidencing promising performance comparable to that of human experts.**

***Index Terms*—Autism spectrum disorder, Expressive language, Home intervention, Behavior signal processing**

This research is funded in part by the Guangdong Science and Technology Plan (2023A1111120012), Science and Technology Program of Guangzhou City (202007030011), National Natural Science Foundation of China (62171207, 62173353) and China Medical Board Open Competition Funding Program (82000-59063001/22-483). *(Corresponding author: Ming Li.)*

This study (including the use of the JERI-WHO-CCI Database and the CPEP-3 database) was approved by the Third Affiliated Hospital of Sun Yat-sen University Institutional Review Board (IRB No. [2020]02-060-01) and Duke Kunshan University Institutional Review Board (IRB No. 2022ML065,2023LM159). All patients were volunteers and signed consent forms for the study and publication. Access to this dataset remains restricted, but we are prepared to share and contribute towards collaborative efforts concerning the available data. Should there be any scope for modifying the IRB protocol and data management plan in a manner that aligns with the legal requirements, we stand ready to extend our full cooperation.

Yueran Pan, Wenxing Liu, Ming Cheng and Ming Li are with the School of Computer Science, Wuhan University, Wuhan, China, 430072, and the Digital Innovation Research Center, Duke Kunshan University, Kunshan, China, 215316, (e-mail: panyr.math@whu.edu.cn; ming.cheng@dukekunshan.edu.cn; wenxing.liu@dukekunshan.edu.cn; ming.li369@dukekunshan.edu.cn).

Biyuan Chen, Hongzhu Deng and Xiaobing Zou are with the Child Development and Behavior Center, Third Affiliated Hospital of Sun Yat-sen University, No.600 Tianhe Road, Guangzhou, China, 510630, (email: chenbiy2@mail.sysu.edu.cn; denghongzhu@foxmail.com; zouxb@mail.sysu.edu.cn)

Dong Zhang is with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: zhangd@mail.sysu.edu.cn).

I. Introduction

THE World Health Organization (WHO) has developed a Caregiver Skill Training (CST) program [1] to provide a family-centered intervention approach for families who have children diagnosed with Autism Spectrum Disorder (ASD). Evaluating participants' engagement states help improve the effectiveness of interventions, for which WHO has proposed The Joint Engagement Rating Inventory WHO Adaptation (JERI-WHO) scheme [2] as an evaluation guide. Expressive Language Level and Use (EXLA) is an item of rating engagement in JERI-WHO. It reflects a child's language useability of words, word combinations, and sentences by 7-level scores. Raters who reach the standard globally are qualified to evaluate engagement items according to JERI-WHO. It requires a lot of human resources and time to train raters and rate new videos. The scenarios of free-living activities in home interventions and the multi-level scoring standards present significant challenges for automated scoring.

In this study, we propose a Multimodal Behavior Signal Processing Framework, following the JERI-WHO standard, to automatically rate EXLA. We work on the JERI-WHO Caregiver-Child Interaction (JERI-WHO-CCI) database, currently the largest database under the CST program. Our framework comprises three modules. First, raw audio and video signals are split into role-aware active time segments by voice activity detection, speaker diarization and speaker age classification, removing irrelevant time intervals as well as labeling each audio segment with the corresponding speaker. Then we extract end-to-end audio embeddings, end-to-end video embeddings, and our proposed hand-crafted text features by modified state-of-the-art behavior signal processing methods. Finally, we fuse these features at the feature level and train a linear regression model for predicting the EXLA scores.

The main contributions are summarized as follows:

(1) We design a computational framework to predict EXLA scores from raw videos in free-living and in-the-wild scenarios. The framework fuses hand-crafted interpretable audio behavior features, end-to-end audio embeddings and end-to-end visual action embeddings to describe children's and caregivers' behavior. Our method enhances analytical capabilities when analyzing poor-quality audio-visual signals and provides a good reflection of children's performance.

(2) We propose a role-aware speech front-end module to detect and segment children's and caregivers' speech in order to analyze them separately and predict children's EXLA. We integrate robust speaker diarization and speaker age

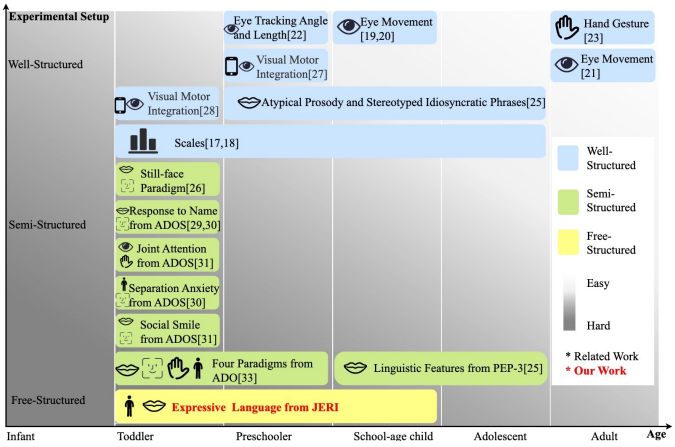

Fig. 1. **Selective Machine Learning Methods Applied in ASD.** ASD = Autism Spectrum Disorder, ADOS = Autism Diagnostic Observation Schedule, JERI = the Joint Engagement Rating Inventory, EXLA = the Expressive Language Level and Use.

classification methods to label audio segments with timestamps and corresponding speaker ID more accurately when data are not structured and not limited to a certain number of speakers.

(3) We introduce supplementary hand-crafted audio features derived from the results of multiple speech front-end modules to complement the primary audio core features defined in the rating manual under the free-living scenarios. It not only enhances interpretability but also outperforms end-to-end features in EXLA rating where only limited in-domain data is available. It could also assist medical doctors in comprehending the automated assessment in a more meaningful way.

II. Related Works and Motivations

A. Related Works

In this section and Figure 1, we introduce analysis methods of (1) ASD, (2) engagement: an aspect of ASD evaluation, and (3) expressive language: an item of joint engagement.

1) ASD: In autism diagnosis and family intervention, it is necessary for physicians to observe the child's responses in terms of both auditory and visual information concurrently. The behavior of caregivers also reflect the performance of children [3]–[5]. To systematically and quantitatively analyze a child's behavior in auditory and visual settings, medical doctors typically diagnose autism using rating scales. Symptom assessment is diagnosed with the widely used Autism Diagnostic Observation Schedule (ADOS) [6], DSM-5 [7] and CARS [8] protocols. Ability assessment is mainly used for rehabilitation intervention, with tools such as VB-MAPP [9] and CPEP-3 [10]. Rating scales requires professional skills, human resources and sometimes additional experiments with children in person. To enhance efficiency and objectiveness, researchers propose different behavior signal processing techniques to predict the rating scores and assist the assessment [11].

a) Unimodal Methods: include analyzing scales scores, visual processing, speech processing, etc. Filtering a minimal subset of items in scales by machine learning improves the efficiency of ASD screening [12], [13]. Unimodal behavior analyses usually study single modality signals, including eye movement patterns [14]–[17], hand gestures in imitation tasks [18], linguistic features [19], atypical prosody, stereotyped idiosyncratic phrases [20], etc.

b) Multimodal Methods: mainly process behavior signals under structured protocols or scenarios, which are designed to quantify stimuli and subjects' responses and evaluate subjects' development following scales. [21] analyzed audio and face of infants after mothers' stimuli. Smartphone apps could capture audio-visual data while providing visual motor integration activities [22], [23]. Protocols in ADOS, including response to name [24], [25], joint attention [26], social smile [27] and Separation Anxiety [25], [28], were designed as semi-structured experiments. Researchers generally proposed multi-step machine-learning-based assessments and utilized pre-trained models in a pipeline to handle the problem of lacking in-domain data [28]. Overall, the analysis of behavioral data in autism is categorized into three top-level experimental setups in Figure 1. Well-structured and semi-structured experimental setups typically require devices to be positioned at specific angles and close distances to participants, restricting the scope of activities. Free-structured setups allow activities without predefined themes or procedures, an area that needs further exploration.

2) Engagement: In the study of engagement, scholars have introduced specialized databases [29]–[32] (detailed in Table I) and analytical approaches. For short-duration video databases [29], typical methods are end-to-end based. [33], [34] take whole videos as input and use networks containing spatial and temporal architectures to sense engagement levels. Pre-trained modules are adopted to extract facial features due to insufficient data. For long-duration video databases [30]–[32], recent studies highlight two-stage methods because of lacking subject-level in-domain labeled data. Two-stage methods [35] involve utilizing intermediate features as a foundation and deploying back-end models (e.g., linear regression) for prediction. Intermediate features automatically extracted using Behavior Signal Processing (BSP) models trained by out-of-domain large-scale datasets (e.g., gaze extracted by OpenFace [36]) help capture important behavior patterns.

3) Expressive Language: Expressive language serves as a critical metric in developmental research, particularly concerning engagement [2], ASD [37], [38], Down syndrome [39], [40], and language delays [41]–[43]. Researchers mainly employ three methodologies to assess expressive language skills: norm-referenced instruments [6], [38], questionnaires [40], [41], unconstrained language samples [44], [45].

Existing research highlights a strong correlation between visual cues and vocal behaviors. In dialogues, unintentional and intentional gestures [46]–[49], head motions, facial expressions [47], lip movements and other movements [46] aid in speech interpretation [50], especially for unique populations like children with autism. After linguistic feature extraction, there are two types of common computational back-end methods. One [6] involves aggregating the scores based on a predetermined scale. [51] identifies language disorders by analyzing scores from assessments provided by caregivers. Another one extracts features from speech data and then employs machine learning algorithms such as decision trees and linear regression for

TABLE I
INFORMATION OF ENGAGEMENT DATABASES IN RELATED WORKS.

Database	Case Number	Video Length	Modal	Perspective	Objects	Method	Annotations
DAiSEE [29]	9068	5-60 seconds	Video	First-person	Adult-computer	End-to-end	**Four-level engagement labels**
Noxi [30]	84	$\geq$ 10 minutes	Audio-video	First-person	Adult-adult	Two-stage	Low-level social signals (e.g. gestures, smiles), functional descriptors (e.g. turn-taking, dialogue acts) and interaction descriptors (e.g. **engagement**, interest, and fluidity).
eHRI [31]	24	4-8 minutes	Audio-video	First-person	Adult-computer	Two-stage	**Connection events**, visual smiles, head nods and transcribed speech
FUTURE WORLDS [32]	85	2-12 minutes	Software log, video	First-person	Preadolescent-computer	Two-stage	Posture, gesture, facial expression, eye gaze, interaction trace logs, and **dwell time**
WHO-JERI	305	$\geq$ 12 minutes	Audio-video	Third-person	Child-adult	Two-stage	**Seven-level engagement labels**

predictive modeling [52], [53]. Here, speech features included spectrogram features, prosody features, frequency, energy, spectrum, atypical prosody transcripts and other embeddings extracted from pre-trained models [19]–[21], [54]–[56].

Although these works utilize speech processing methods, they did not fully explore the transcripts generated by an automatic speech front-end and were not directly focusing on in-the-wild EXLA score prediction. Predicting multi-level expressive language scores with raw audio-visual data captured in far-field third-person perspective remains unexplored.

B. Motivations

In our project, the aforementioned multi-step machine-learning-based frameworks inspire us to address the problem of lacking in-domain data. However, the methodologies and data from previous works cannot be directly applied due to several limitations and challenges: (1) Compared to end-to-end approaches for engagement detection, we think it is essential to study medically meaningful and interpretable features further for expressive language modeling. (2) Existing machine learning approaches for predicting expressive language levels primarily rely on manually transcribed text from separate close-talking microphones. We aim to automatically process speech data recorded in one channel by a far-field iPad. (3) Lacking in-domain data leads to insufficient domain adaptation and degraded performance when fine-tuning pre-trained end-to-end models with our limited in-domain data.

The JERI-WHO program involves evaluating free-living in-the-wild data. Three potential strategies are considered:

(1) Designing a pipeline to split the prediction into multiple steps and then utilize existing large-scale datasets and pre-trained models for analysis step by step. (2) Adding meaningful supplementary features on top of automatically generated transcripts to enhance the description of language expression ability. (3) Utilizing complementary video information and caregivers' behavior data to further enhance the performance.

III. DATABASE DESCRIPTION

A. JERI-WHO Caregiver-child Interaction Database

1) Data Source: The JERI-WHO-CCI Database comes from the Intervention effect of WHO-CST program on ASD study (Registration number: ChiCTR2000035176). It is currently the largest database under the CST program. It consists of multiple family intervention videos taken at home using a hand-held iPad. In the database, there are 305 videos from 179 children with ASD and their caregivers.

TABLE II
TOOLBOX CONTENT AND RELATED ACTIVIES

Toy	Number	Activity
Picture book	1	Telling stories
Animal model	8	Imagination
Sand hammer	4	Music games
Puppet	2	Imagination
Ball	2	Interactive games
Pen set, paper, painting templates	1	Painting
Cardboard box set	1	Stacking boxes

2) Recording Protocol: Videos were taken during each home visit. The recording scenes are shown in Figure 2(a)-(c). Both audio and video were collected by iPads to prevent children from being distracted by unfamiliar devices. During a recording, a caregiver and a child sit face-to-face and play together. Another person holds an iPad to record from approximately two meters. The caregiver and the child can play with toys freely to enhance children's social interaction skills. Caregivers primarily focus on providing game support and engaging with children in interactions. Their activities include following children's interests, imitating children's appropriate behaviors, expanding children's language, or demonstrating new skills [1].

The guidelines for shooting home intervention videos are as follows: (1) Paticipants use toys in the toolbox to play. (2) Each video is more than 12 minutes. (3) Children's faces and hand gestures are recorded as much as possible. (4) Videos capture caregivers' hands, faces and toys involved in games as much as possible. (5) The person recording the video does not participate in the interaction. (6) The person except for the child and the caregiver cannot appear in the video recording.

3) Collected Data: All participants speak Mandarin. They are from Guangdong province, China. The distribution of scores for item EXLA, the gender and age distribution of children in the database is shown in Figure 2(d)-(f). The ratio of boy-to-girl is close to 4:1, which is similar to the common gender ratio of ASD [57]. The JERI-WHO-CCI database aims to observe the level of a child's communication skills and joint engagement. Since videos are homemade, the original frame rate of each video is not fixed, ranging from 30 to 60. All original videos are downsampled to 18 FPS. Audios are recorded in one-channel format and converted to 16 kHz sampling rate.

Every JERI item is rated from 1 to 7 by a qualified rater who is confident in complying with the WHO standard. More

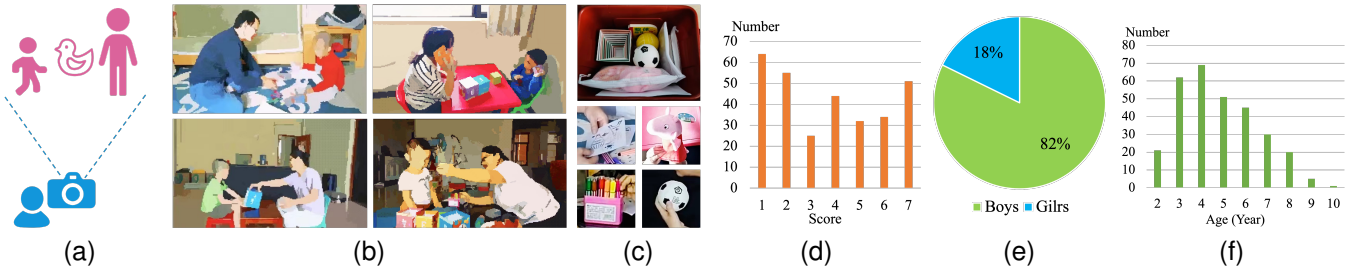

Fig. 2. **Recording Environment and Database for JERI-WHO. (a) A Recording Scene** A caregiver and a child are playing with toys together when a bystander holds up an iPad recording. **(b) Frame Examples** Four frame examples in videos, with blurred effects, to protect participants' privacy. **(c) Toys** A toolbox example and four toy examples. **(d)** The numbers of rating scores for EXLA. **(e)** The distribution of children's gender. **(f)** The numbers of children's ages.

TABLE III
THE RATING ANCHOR FOR EXPRESSIVE LANGUAGE LEVEL AND USE

Score	Rating rule
1	The child does not have expressive language.
2	The child only uses 1 or 2 different words in the whole video.
3	The child can speak about 5 different words.
4	The child uses more than 10 single words with few phrases.
5	The child mainly uses 2-word utterances.
6	The child frequently uses 3-word utterances.
7	The child can use predominately sentences.

specifical rating rules for EXLA are shown in Table III. The standard is that a rater is considered qualified if the Root Mean Square Error (RMSE) between his/her ratings and the "ground truth" is less than 1 in assessments. Here the "ground truth" label is annotated by experienced professional clinicians. Given the need for data confidentiality and specialized project comprehension, all raters involved in this database are medical professionals and in compliance with IRB guidelines. We had three qualified raters who scored videos and an additional expert to perform the review and ensure the accuracy.

4) Analytical Difficulties: Unprofessional homemade recording leads to many problems: 1. Data quality is poor; 2. Some children hardly speak or speak in a low voice; 3. The camera position is not fixed, and the video screen shakes; 4. Many faces are obscured; 5. The range of human activities is random; 6. In free-living and in-the-wild environments, there are noises, sometimes even voices, from sudden intruding people.

B. CPEP-3 Database

The CPEP-3 database [58], featuring audio conversations involving children and medical doctors, is comparable to the scenarios in JERI-WHO-CCI. CPEP-3 serves to either pre-train or fine-tune the speech processing models discussed in Section 4 shown in Table V, improving the model's ability to extract robust and interpretable audio features specific to JERI-WHO-CCI.

The CPEP-3 database consists of audio recordings during a semi-structured CPEP-3 Performance Test [10], in which natural autism evaluation conversations between children, medical doctors and parents are collected. Many of the children included in the database exhibit limited speech or use only phrases. The microphone is placed at a distance from the child, and the environment is highly noisy. Audios related to the EXLA part are transcribed and annotated. Annotations include timestamps of each utterance, related speaker ID (tagged as a child, doctor 1, doctor 2, and parent), text content and the overall three-level expressive language score. In this dataset, 13.6 hours of audio feature children's voices, while others comprise adults' voices (medical doctors and parents).

IV. METHODS

We apply a multimodal framework on JERI-WHO-CCI to process the behavior signals, as shown in Figure 3. The framework we conducted is divided into three parts: 1. the audio system, 2. the visual system, and 3. fusion and prediction.

The audio system first extracts active speech segments with role-specific identities (child or caregiver) based on the combination of voice activity detection, speaker diarization and speech age classification. Then, an Automatic Speech Recognition (ASR) system generates text transcriptions and build interpretable hand-crafted features. Moreover, a ResNet-34-based model extracts end-to-end audio embedding features.

The visual system conducts active visual segment detection and feature extraction. The video is segmented according to speech timestamps detected by the audio system because EXLA score is highly related to speech. After refinement, we extract embeddings from active video segments using a Swin Transformer [59] based end-to-end video classification model.

The Final stage is a feature-level fusion of features from the aforementioned systems and predicting the EXLA scores. The reason for not directly using fully end-to-end audio and video classification methods to predict the EXLA scores is due to the lack of in-domain data.

A. The Audio System

1) Speech Segmentation: To extract children's and caregivers' corresponding speech segments in raw audio signals, we adopt a two-stage process consisting of a speaker diarization module and an age classification module. Speaker diarization is usually the task of predicting "Who Spoke When" in a given recording. However, it can only output the relative speaker identities without speaker role information. To tackle these problems, we introduce an age classification module to assign the "child" or "adult" labels to corresponding clustered speaker identities. Thus, each extracted speech utterance can be associated with a specific role (child or caregiver).

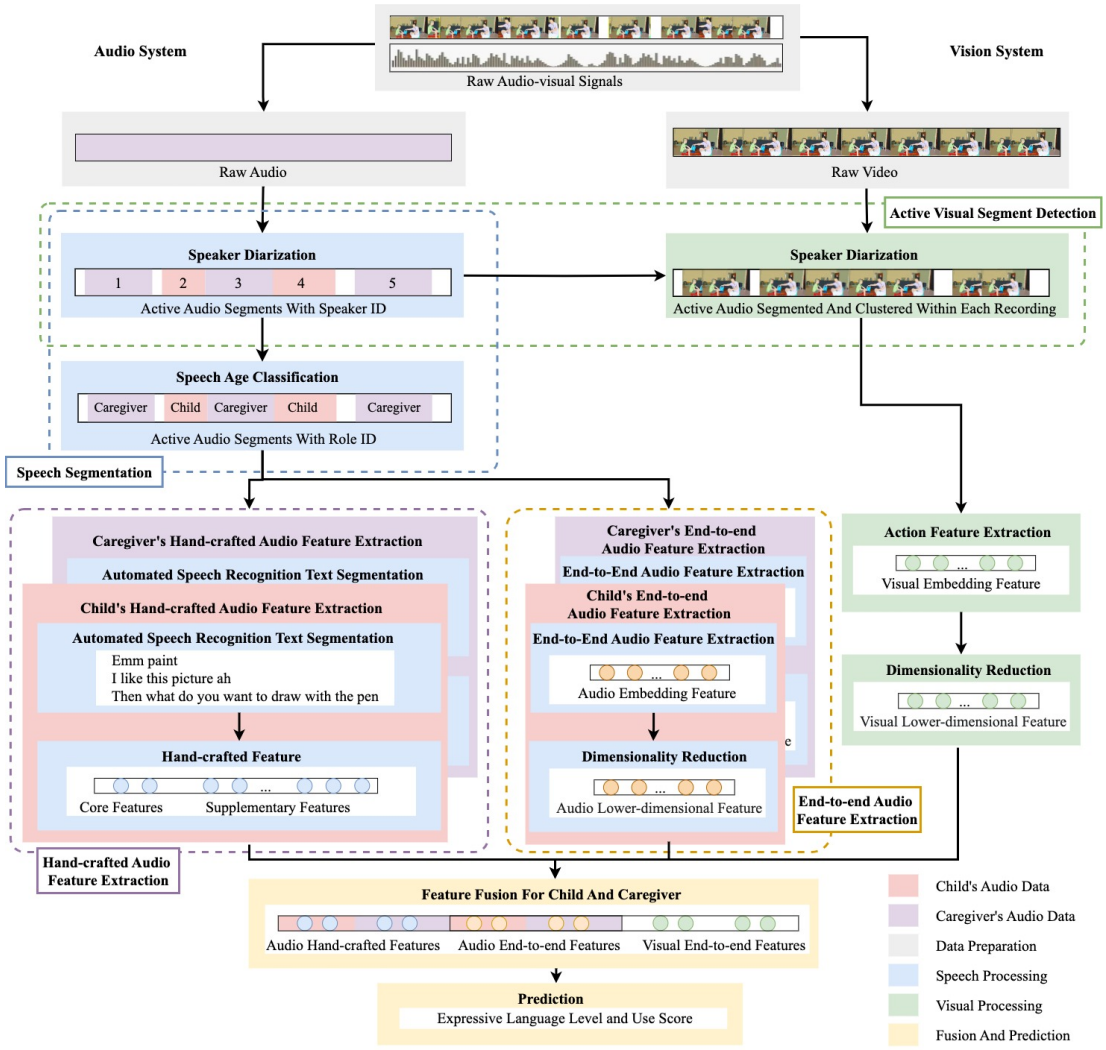

Fig. 3. **The Overview of the Automatic Expressive Language Level Assessment Framework.** The blue branch extracts interpretable hand-crafted linguistic features. The orange branch is for end-to-end audio feature extraction. The green branch is for the video system, which extracting end-to-end visual features. The yellow module is for feature-level fusion and prediction of the EXLA scores.

a) Speaker Diarization: We first remove non-speech regions from raw audio by a voice activity detection (VAD) model [60] to reduce the impact of background noise. Then we utilize a speaker diarization (SD) module [61] to detect when a speaker is active, complete with timestamps. VAD is used to determine when there are human voices present or not. VAD filters out sound clips in which people speak for further processing. The VAD model comprises a front-end ResNet-34 model, statistical pooling layer, and Bi-LSTM neural network as the back-end to output frame-level posterior probability of speech region [60] (code at [62]). The front-end ResNet-34 model extracts the frame-level feature map and robustly captures deep and discriminative features from data. The combination of ResNet-34 with a statistical pooling layer and subsequent BiLSTM layers captures both the temporal and spectral characteristics of speech, crucial for accurate voice activity detection. This setup enables the system to detect active voice and filter out irrelevant background voices efficiently. Then, filtered active audios are segmented by the SD module into homogeneous regions and labeled with a relative speaker ID [63]. The VAD module is pre-trained on a mixed training dataset, including AMI [64], ICSI [65], ISL, NIST, SPINE12, AISHELL-4 [66], DIHARD II [67] and DIHARD III [68]. The speaker embedding extraction module is pre-trained on the development set of VoxCeleb2 [69]. The dataset contains over 1 million utterances for 5994 different speakers, which is widely acknowledged for speaker embedding extraction. For similarity measurement, the Bi-LSTM model with spectral clustering [60], is pre-trained on the CALLHOME database [70] and fine-tuned on the CPEP-3-B1 subset. Fine-tuning on CPEP-3-B1 improves the performance on Diarization Error Rate (DER) from 19.95% to 16.61% on CPEP-3-B2 as shown in Table V. Since generated relative speaker IDs are not assigned with 'child' or 'caregiver', we would aggregate speech segments from each speaker together to detect the role information.

b) Age Classification: We introduce a ResNet-34 [71] model followed by a linear layer to perform utterance-level age classification. The model parameters are initialized by

TABLE IV
HAND-CRAFTED AUDIO FEATURES.THE FIRST TWO ITEMS ARE CORE FEATURES; THE NEXT TEN ARE SUPPLEMENTARY.

	Feature	Definition
1	Average number of words per sentence	It is calculated by $\frac{TNW}{Numberofsentences}$. It is for measuring a speaker's language variety per sentence.
2	Number of different words (NDW)	It counts the number of unique words in a video clip, reflecting a speaker's vocabulary diversity.
3	$\frac{1}{Number\ of\ sentences}$	It is a measure of how many sentences a person speaks.
4	$\frac{1}{Turn\ number}$	It is a measure of the count of how many times speakers switch and start to talk after other people's words.
5	Total number of words (TNW)	It counts all words a person speaks in a video.
6	Median word count per sentence	It is the median of word count per sentence.
7	Median word count of 5 longest sentences	It represents the median of 5 largest lengths of words in every sentence. It represents a speaker's best performance in language organization.
8	Average number of words per turn	It is calculated by $\frac{NDW}{Number\ of\ turns}$. It is for measuring a speaker's words length per speaker switch turn.
9	Average number of different words per sentence	It is calculated by $\frac{NDW}{Number\ of\ sentences}$. It is for measuring a speaker's language variety per sentence.
10	Median different word count per sentence	It is the median of the unique word count per sentence.
11	Median different word count of 5 longest sentences	It represents the median of 5 largest amounts of unique words in every sentence. It reflects a speaker's best performance in a diverse language organization.
12	Type token ratio (TTR)	It is calculated by $\frac{NDM}{TNW}$. Type token ratio shows how frequently a caregiver or a child adopts new vocabulary. It indicates a speaker's word variety.

TABLE V
PERFORMANCE OF AUDIO PRE-PROCESSING MODULES. DER = DIARIZATION ERROR RATE, CER = CHARACTER ERROR RATE. THE SIZE RATIO OF SUBSETS CPEP-3-B1 AND CPEP-3-B2 IS 1:1.

Module	Fine-tuning/ Training Subset	Evaluation Subset	Before	After
Speaker Diarization	CPEP-3-B1	CPEP-3-B2	DER 19.95%	DER 16.61%
Age Classification	80% of CPEP-3	20% of CPEP-3	Recall 80.53%, Precision 90.01% for children	
Automatic Speech Recognition	CPEP-3-B1	CPEP-3-B2	CER 43.17% for adults CER 87.15% for children CER 54.87% for all	CER 28.75% for adults CER 74.31% for children CER 36.05% for all

an open-source pre-trained speaker embedding model [62]. In accordance with diarization annotations, we partition the CPEP-3 dataset into individual audio clips, each containing the voice of a single person. These clips are then further segmented into 4-second intervals using a sliding window of 1 second and annotated as either 'child' or 'adult'. The whole dataset is split at $8:2$ for training and evaluation. The hyper-parameters of the pre-trained ResNet-34 can be found in [62], [71] Training results are shown in Table V.

2) Automatic Speech Recognition: We apply the standard recipe of the WeNet [72] toolkits to build a speech recognition model with the Conformer [73] architecture. The model is pre-trained on a 17000+ hours Mandarin Chinese corpus collection (WENETSPEECH [74],AISHELL-2-train [75], MAGICDATA [76], and Aidatatang_200zh [77]). We extract the 80-dimensional FBank features with a 0.025s window and 0.01s frameshift. We employ SpecAugment [78] with 2 frequency masks (F = 30) and 2 time masks (T = 50) and global CMVN technique. We use Adam optimizer with a maximum learning rate of 0.001. The Noam learning rate scheduler with 5k warm-up steps is used. The pre-trained model is fine-tuned on CPEP-3-B1 and tested on CPEP-3-B2 (shown in Table V) We also tested the ASR model before and after fine-tuning on the AISHELL-1-test dataset [79], which was not used for pre-training. The Character Error Rate (CER) results were 1.57% and 1.9%, respectively, indicating that our fine-tuned model did not overfit the small vocabulary of CPEP-3. We use the Jieba [80] toolkit with MDBG Chinese-English Dictionary [81] to perform text segmentation on top of the ASR transcripts.

3) Hand-crafted Audio Features: We are motivated by linguistic expressive language features in many aspects, including language diversity, talkativeness, complexity and speech disfluency [44], [45]. Global and local features of transcripts calculated from words and typical lexical units potentially capture children's language strengths and weaknesses [82]. Talkativeness can be reflected by Total Number of Words (TNW), while vocabulary diversity is related to Number of Different Words (NDW) [82], [83]. [82], [83]. Language complexity can be reflected by Mean Length of Utterance MLU [45]. MLU refers to the average number of morphemes per utterance. Mean Length of Utterance of the Longest Five Utterances (MLU5W) can indicate the best performance of language complexity. Type–Token Ratio (TTR) is affected by lexical diversity and speech disfluency [84].

According to the rating rules , the number of words a child speaks in a sentence and the number of different words a child says in a whole video define the rating anchor. Considering only two features to describe a wild-environment video are not sufficient, we adopt additional transcript-level hand-crafted features from [44], [83], [85] as audio supplementary features as well. In total, we list the following 12 features in Table IV, of which the first 2 are the audio core features defined according to the rating rules, and others are supplementary features we adopt here. All the hand-crafted features were measured based on the results of our automatic speech modules. Feature 3-5 are included to measure the partition of participants' speaking. Feature 5 and 12 supplement language diversity. Feature 6, 8-10 are refined from MLU to add information of complexity. We apply the Median word count/ different word count of the 5 longest sentences (Feature 7, 11) refined from MLU5w to reflect speakers' best performance.

All hand-crafted features are calculated from the role-aware text transcripts automatically generated by the audio front-end system. First, we split the ASR transcript of each video into two groups by role labels derived from the result of speaker diarization and age classification. Then, based on the ASR results of each speaker within a recording, we compute global

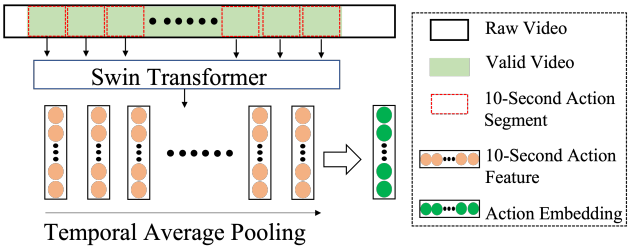

Fig. 4. **The Structure of the Action Feature Extraction.** The Raw Video is one of the videos in our database, which is over 12 minutes. The Valid Video indicates the range of valid activities in the video. The 10-second action segment is a valid video clip with a length of ten seconds. The 10-second action feature denotes the feature of the 10-second segment. The action embedding feature denotes the end-to-end embedding feature from the visual side.

features in Table IV. Finally, we concatenate all features of a child and a caregiver together. For a single participant, the dimension is 12, resulting in a dimension of 24 per audio.

4) End-to-end Audio Features: We also explore the potential of end-to-end audio-based embedding features for comparison and fusion. Using the ResNet-34-based speaker verification model [70], [71], we extract end-to-end audio-based embedding features. First, we divide the CPEP-3 database based on annotations into children's and adults' speech subsets. Then, we separately fine-tune a ResNet-34-based model [62] for each subset of CPEP-3 database, using associated three-level expressive language scores. Finally, we utilize the fine-tuned models (child and adult) to extract audio-based embedding features on JERI-WHO-CCI. Based on the previous speaker diarization and age classification steps, each recording is split into multiple utterances with role information. Therefore, the pre-trained child and adult end-to-end models are used to extract embeddings before the models' output layer separately. For each recording, all embeddings from the child and adult will be separately averaged to obtain their respective overall child and adult audio embedding features, respectively. Averaging the embedding vectors extracted from multiple utterances in an audio-visual recording provides a simple method for temporal aggregation, which summarizes the overall information of the recording and makes it a compact fixed dimension vector for the subsequent back-end modeling.

B. The Visual System

We extract features from videos as a supplement to evaluate the EXLA scores of children. Videos contain action information, which helps reflect children's EXLA levels. In this module, we detect temporal active segments from videos and then use Swin Transformer [59] as the backbone network to extract features in an end-to-end manner. The detector and the feature extractor we chose are currently widely used in related fields. Swin transformer integrates the characteristic advantages of CNNs in vision with transformers' efficient and robust architecture, which is computationally very effective in dense prediction tasks. It performs well on action recognition [59], fitting our task well.

The active visual segment detection module includes speaker diarization and time-based segmentation using timestamps from the speaker diarization module. Specifically, if it is detected that a time segment at a certain timestamp involves someone speaking, then the segment in this period is considered active. Frames during this period are utilized for subsequent analysis. Conversely, if no one talks, frames are filtered out because our goal is to predict expressive language levels, focusing on time segments with active vocal activity.

Swin Transformer is pre-trained on the Kinetics-400 [86] database. It is a typical large vision model that used in multiple health informatics applications [87], e.g. pain assessment [88]. Pre-training the Swin Transformer model on the large-scale Kinetics-400 dataset significantly improves the model's proficiency in recognizing and encoding human actions and activities. It enhances the process of visual feature extraction. Given that our database is relatively small, this is of significant importance. The backbone extracts action features with approximately 300K 10-second action videos. In Figure 4, a raw video is filtered into a valid video according to the speaker diarization label, then multiple 10-second active segment videos are fed into the action feature extraction module. In detail, we extract a feature per 10-second video. Each video is split into 5 clips with 64 frames. And then we take 1 frame every 2 frames to get 32 frames. For the 32 frames, each frame is transformed into a $3 \times 224 \times 224$ array. So for each 10-second video, we have a $5 \times 32 \times 3 \times 224 \times 224$ array as an input to feed into Swin Transformer. And the size of the output is 1×768. For a valid video, there are multiple 10-second valid segments resulting in an action embedding feature sequence. Finally, We do temporal average pooling on the time channel and get a 1×768 action embedding representing the whole video.

C. Prediction and Fusion

For each input audio-visual recording sample, the audio system extracts 12-dimensional hand-crafted features and 256-dimensional end-to-end audio embeddings for the child and the caregiver separately. Meanwhile, the video system compresses the entire visual contents into a 768-dimensional embedding vector. We implement Latent Dirichlet Allocation (LDA) on audio and visual embedding features to reduce the dimension and try to avoid over-fitting as well as an imbalanced prediction system [89], [90]. We determine the number of principal components mainly based on two considerations: ensuring sufficient retained information and avoiding dimensions that could cause overfitting. Deciding the threshold is a trade-off between these two considerations. Given the small sample size of only 305 videos, it is necessary to significantly reduce the dimensionality of the features. This ensures the model can effectively learn without overfitting, despite the limited data. For end-to-end audio features, we applied Min-Max Scaling followed by Linear Discriminant Analysis (LDA) to reduce the dimensionality to 2 dimensions for each speaker. Similarly, for end-to-end visual features, after Min-Max Scaling, we used LDA to reduce the dimensionality to 31 dimensions. The total dimension of all fused variables is $(12 + 2) \times 2 + 31 = 59$.

We employ a linear regression model to predict the EXLA scores after dimensionality reduction. We choose regression rather than classification because the EXLA scores are categorized into seven levels of 1-7, indicating continuity and

TABLE VI
PREDICTION RESULTS OF DIFFERENT METHODS. MSE = MEAN SQUARE ERROR, RMSE = ROOT MEAN SQUARE ERROR, PCC = PEARSON CORRELATION COEFFICIENT, R^2 = COEFFICIENT OF DETERMINATION, ✓= SELECTED FEATURES.

Method	Hand-crafted Audio Features				End-to-end Audio Features		End-to-end Visual features	MSE	RMSE	PCC	R^2
	Core feature		Supplementary feature								
	Children	Caregivers	Children	Caregivers	Children	Caregivers					
1							✓	4.620	2.149	0.143	0.015
2					✓			3.509	1.873	0.502	0.252
3						✓		4.125	2.031	0.347	0.120
4					✓	✓		3.285	1.812	0.547	0.299
5					✓		✓	3.358	1.832	0.533	0.284
6						✓	✓	4.091	2.023	0.359	0.127
7					✓	✓	✓	3.185	1.785	0.567	0.321
8	✓							3.469	1.863	0.519	0.260
9		✓						3.409	1.846	0.523	0.273
10	✓	✓						3.052	1.747	0.595	0.349
11	✓		✓					2.528	1.590	0.681	0.461
12		✓		✓				2.731	1.653	0.651	0.418
13	✓	✓	✓	✓				2.251	1.500	0.723	0.520
14	✓						✓	3.324	1.823	0.548	0.291
15		✓					✓	3.390	1.841	0.527	0.277
16	✓	✓					✓	2.994	1.730	0.606	0.362
17	✓		✓				✓	2.461	1.569	0.693	0.475
18		✓		✓			✓	2.715	1.648	0.655	0.421
19	✓	✓	✓	✓			✓	2.174	1.474	0.734	0.536
20	✓	✓	✓	✓	✓	✓		1.991	1.411	0.760	0.575
21	✓	✓	✓	✓	✓	✓	✓	**1.939**	**1.392**	**0.768**	**0.587**

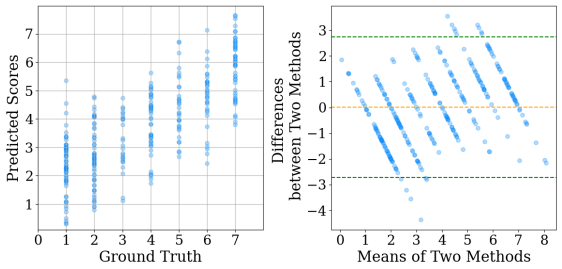

Fig. 5. **(a)Scatterplot of Prediction Results for Our Method.** The Scatterplot corresponds to Method 21 in Table VI. The x-axis is for ground truth, and the y-axis is for predicted scores. Results are plotted as translucent blue scatters. Dark blue appears close to the ground truth, reflecting effective prediction.**(b)Bland-Altman Plot of Method 21 Prediction.** Each scatter point represents the mean and the difference between a pair of the ground truth score and our predicted score. The orange dashed line shows the mean bias. The green dashed lines show the limits of agreement.

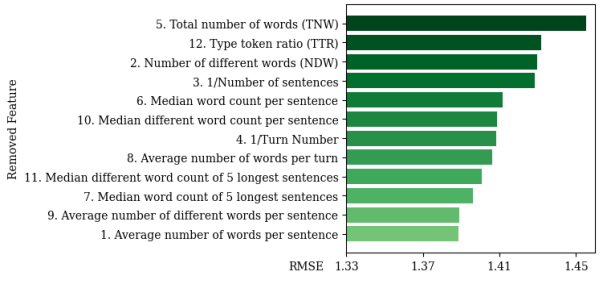

Fig. 6. **Root Mean Square Error (RMSE) for Feature Ablation.** This plot represents the performance of methods that remove features individually from the hand-crafted system (Method 13).

TABLE VII
PERFORMANCE OF FULLY END-TO-END METHODS AND OUR PROPOSED TWO-STAGE METHODS IN 2-FOLD CROSS VALIDATION. E2E = END-TO-END, MSE = MEAN SQUARE ERROR, RMSE = ROOT MEAN SQUARE ERROR, PCC = PEARSON CORRELATION COEFFICIENT.

Method	MSE	RMSE	PCC
Visual Fully E2E	7.51	2.74	0.24
Audio Fully E2E	4.83	2.20	0.37
Method 1	4.64	2.15	0.14
Method 4	3.33	1.82	0.54
Method 21	**3.28**	**1.81**	**0.63**

TABLE VIII
F-TEST RESULTS FOR ERROR VARIANCES OF DIFFERENT METHODS.

Comparison	F-statistics	P-value	Significance
Method 1 vs Method 21	18.34	7.53×10^{-47}	***
Method 4 vs Method 21	9.91	2.37×10^{-27}	***
Method 7 vs Method 21	10.02	1.02×10^{-25}	***
Method 8 vs Method 21	94.06	7.54×10^{-124}	***
Method 13 vs Method 21	403.45	3.83×10^{-133}	***
Method 20 vs Method 21	4.00	8.19×10^{-3}	**

TABLE IX
TRANSCRIPT EXAMPLES OF FAILURE CASES. TRUE SPEECH TEXT AND ID ARE MANUALLY LABELED. PREDICTED RESULTS ARE GENERATED BY OUR METHOD. SPEECH TEXT IS TRANSLATED INTO ENGLISH.

Case	True Speech Text	Predicted Speech Text	Ture ID	Predicted ID
1	Hey	Hey	Caregiver	Caregiver
	This is a dinosaur	This is a dinosaur	Caregiver	Caregiver
	And	And	Caregiver	Caregiver
	This is a pig	This is a pig	**Caregiver**	**Child**
2	Look	Look	Caregiver	Caregiver
	What's it	What's it	Caregiver	Caregiver
	Apple	**Null**	**Child**	**Null**

an ordinal relationship. In this case, regression provides more meaningful outcomes than classification. Given the decision to proceed with regression, we choose a linear method rather than a non-linear method, considering the overfitting issue.

TABLE X
PERFORMANCE OF THE PROPOSED HAND-CRAFTED FEATURES (METHOD21 IN TABLE VI) WITH DIFFERENT BACK-END MODELS

Backend Model	MSE	RMSE	PCC	R2
Linear Regression	**1.939**	**1.392**	**0.768**	**0.587**
Lasso Regression (alpha=0.1)	2.157	1.469	0.735	0.540
Ordinal Regression	2.010	1.418	0.756	0.571
Random Forest	1.979	1.407	0.760	0.578
Decision Tree	3.902	1.975	0.576	0.168
KNN	2.422	1.556	0.696	0.483
SVM	2.143	1.464	0.737	0.543

First, we conduct normalization and a feature-level fusion by concatenating audio and visual features. Then, we train a back-end regression model on fused features. Given that the sample size (305 videos) is only slightly larger than the variable dimensions in our proposed framework (Method 21 in Table VI), we employ leave-one-video-sample-out cross-validation (LOOCV) for all predictions. Each time, we leave one audio-video sample as the validation set and fit the final stage (a linear regression model) using the rest of the samples. For feature ablation, we iteratively remove one feature from our proposed system (Method 21) for both the child and caregiver and calculate the new RMSE.

Additionally, we compare the performance of a visual fully end-to-end method and an audio fully end-to-end method against our proposed feature+back-end two-stage methods (Method 1, 4, 21) using 2-fold cross-validation. In the two end-to-end systems, our main structure is the same as the 2-stage approach, but the output has been modified to produce a 7-class prediction. We use Swin Transformer for visual [59] shown in Figure 4 and ResNet-34 for audio [62] as foundational backbones and fine-tune them on each fold of JERI-WHO-CCI to obtain the final models.

V. EXPERIMENTAL RESULTS

Our framework is designed to rate the EXLA scores automatically from in-the-wild data, integrating audio-visual signal processing methods. Thus, we compare 21 methods in Table VI to evaluate our proposed method from three aspects:

(1) Effectiveness of adopting hand-crafted audio features.
(2) Effectiveness of fusing visual information.
(3) Effectiveness of including caregivers' behavior patterns.

a) Best Method: Our proposed interpretable audio-visual fusion system, Method 21, performs the best. It attains the best RMSE at 1.392 on a 1-7 scale and the best PCC at 0.768. As mentioned in Section III-A3, the certification standard for medical doctors typically requires an RMSE of less than 1.0. While our RMSE does not fully meet this standard, it demonstrates a competitive result given the inherent complexity of predicting ordinal labels on a 7-level scale. Furthermore, as shown in Figure 5(a), the scatter plot highlights that our predicted scores closely align with the diagonal, indicating a strong positive correlation between our model's outputs and the standard scores. The agreement between the prediction results and the ground truth is shown in Figure 5(b). The difference that is inconsistent towards different levels illustrates a certain level of heteroskedasticity crossing different rating scores.

b) Hand-crafted Audio Featrues: Audio supplementary features improve the effectiveness of a basic rating system with just audio core features, help achieving PCC>0.7 in Method 13, 17, 19, 20, 21. This is because our proposed hand-crafted features have good robustness. RMSE, PCC and R^2 all improve substantially by adding supplementary features (Method 8 vs. 11, 10 vs. 13, 14 vs. 17,16 vs. 19). Using children's core features, as in the definition of EXLA, cannot provide a comprehensive description. In Figure IV-C for feature ablation, Total number of words (TNW), Type token ratio (TTR), Number of different words (NDW) and Number of sentences play important roles in predicting EXLA scores.

c) End-to-end Audio Features: The proposed hand-crafted audio features describe EXLA better than the end-to-end audio features (Method 2 vs. 11, 4 vs. 13). However, without supplementary features, only core features cannot compete with end-to-end audio features (Method 2 vs. 8).

d) Caregivers' Features: The four systems fusing both children's and caregivers' data (Method 13, 19-21) achieve a highly positive correlation with PCC over 0.7, enhancing the performance. Compared to relying solely on caregivers' features (Methods 3, 6, 9, 12, 15, 18), methods using only children's audio features perform better.

e) Visual Features: The comparisons (Method 2-3 vs. 5-6, 8-9 vs. 14-15, 13 vs. 19) show that visual information can be used as an effective supplement to audio information. It reflects that visual activity is complementary, especially when the audio quality is very low. This is because speaking events usually happen with some simultaneous body movements. For example, in the upper right sub-pictures of Figure 2(b), the child raised his hand when answering the caregiver.

f) Fully End-to-end Systems: Table VII shows performances of the audio/visual fully end-to-end systems without any back-end model and two-stage methods (Method 1, 4, 21) in a 2-fold cross-validation setup. Two-stage methods outperform the fully end-to-end methods, which might be due to the limited amount of fine-tuning data provided in our setup.

g) Different Back-end Models: For linear methods, we have experimented linear regression, lasso regression and ordinal regression for the fusion of all hand-crafted and end-to-end features. As shown in Table X, our proposed linear regression model performs the best.

h) Statistical Validation and Limitation: To further validate the improvement of supplement and fusion, we conduct F-tests, detailed in Table VIII, to statistically evaluate the enhancements. Method 21 generally has a lower error variance compared to other methods at high significance. These results suggest that the fusion of supplementary hand-crafted features, end-to-end audio features and visual features helps improve prediction.

VI. DISCUSSION

Our framework shows promising discrepancy (RMSE at 1.392) and good consistency (PCC at 0.768) with ground truth rating scores from human experts. Due to the small sample size and in-domain evaluation based on cross-validation, these findings should be regarded as preliminary. Demonstrating a level of performance that, while not yet on par with

human experts, shows promise for further improvement and validation. Our proposed hand-crafted audio features show better robustness than end-to-end audio features, given limited in-domain training data. Supplementary hand-crafted features can support the EXLA scores and provide interpretable features and intermediate results for medical doctors, and the visual data and caregivers' behaviors improve EXLA ratings as well. Our proposed audio supplementary features, visual system and caregivers' behavior data compensate for the poor quality and wild environments of the raw data, improving RMSE from 1.863 to 1.392 and PCC from 0.519 to 0.768. Although the improvement of end-to-end modules is limited, it might be because of insufficient in-domain training data. Our applied pre-trained model and well-designed speech processing pipeline efficiently segment children's audio, although many of them have language disorders.

Compared to previous works, we analyzed longer and more free-living style data, which provides more potential in real home intervention scenarios. We provide multi-level score predictions for a specific item that aligns with the JERI protocol well. We follow the mainstream approach of using a pre-trained models-based pipeline to process limited data and introduce role-aware speech segmentation and hand-crafted features fusion to improve. Our hand-crafted features and multimodal fusion address the robustness of unstructured experimental data, while previous pipelines mainly deal with instructional and semi-structured experiments. In our work, manual manuscript annotating is replaced by automatic processing from in-domain pre-trained models. According to our proposed features, caregivers can make targeted improvements in home intervention. In practice, not only can we provide an EXLA score but also a list of linguistic features to clinicians and parents. With these features, clinicians and parents can more easily explain the levels of children's language expression deficits.

Another significance of our work is providing an objective EXLA scoring tool. It can potentially be used for longitudinal assessments of children's abilities on a weekly basis, producing a curve that illustrates changes over time. This future work will offer medical doctors and caregivers new capabilities for objective, longitudinal evaluations along a timeline.

In Figure 5(a), there are a few outliers in the top left corner and lower right regions. These anomalies predominantly arise from the constraints of speech processing modules, as illustrated in Table IX. Specifically, in Case 1, the caregiver and child exhibit nearly identical tonal qualities, making it challenging even for a human to discern the difference. In Case 2, the child's speech volume is insufficient for detection. These cases present significant obstacles to accurate speech recognition. Future research will focus on enhancing both speaker diarization and speech recognition under far-field and noisy environments while also exploring the incorporation of additional visual cues (e.g., eye gaze, lip movement) in free-living scenarios.

We will improve the JERI assessment in our future works and develop more objective assessment toolkits:

(1) We will improve the quantity and quality of JERI-WHO data. The recording setup lacks standardization, and signal quality could be enhanced by employing multi-view HD cameras and close-talking microphones instead of far-field data capturing. Acquiring high-quality audio-visual signals is essential for accurate automatic coding. After collecting large-scale in-domain data, we can better study end-to-end methods with different network structures and loss function design.

(2) Other items in JERI involving complex video understanding need to be explored and assessed as well. We believe that large-scale video understanding models or multimodal large language models can potentially provide high-quality features.

VII. Conclusion

In this paper, we introduce a computational framework designed to predict the expressive language aspect of children's development in real home intervention scenarios. This framework combines interpretable, hand-crafted audio-based features with end-to-end audio-visual features. This framework distinctly combines two primary components: interpretable, hand-crafted audio features and end-to-end audio features and visual features. Hand-crafted features are derived from the role-aware speech recognition transcripts of the audio data. End-to-end features are extracted from both audio and video inputs, respectively, using deep learning models.

The fusion of hand-crafted and end-to-end learned features not only underscores the methodological innovation of our approach but also highlights its practical implications. We thoroughly evaluate the proposed system on the JERI-WHO-CCI Database, the most extensive database to date under the CST program, demonstrating its application and effectiveness. The system achieves a PCC of 0.768 when compared against evaluations from human experts,indicating a strong correlation and thereby validating the framework's predictive accuracy. To the best of our knowledge, this represents the inaugural effort to implement an audio-visual automated assessment method for EXLA rating in real-world free-living scenarios. The system holds significant promise for providing instantaneous feedback to caregivers and medical doctors, thereby facilitating early intervention efforts.

References

[1] E. Salomone, L. Pacione, S. Shire, F. L. Brown, B. Reichow, and C. Servili, "Development of the who caregiver skills training program for developmental disorders or delays," *Frontiers in Psychiatry*, vol. 10, p. 769, 2019.

[2] L. B. Adamson, R. Bakeman, and K. Suma, "The joint engagement rating inventory (JERI)," Technical report 25: Developmental Laboratory, Department of Psychology, Georgia State University, Tech. Rep., 2016.

[3] A. C. Gulsrud, L. B. Jahromi, and C. Kasari, "The co-regulation of emotions between mothers and their children with autism," *Journal of Autism and Developmental Disorders*, vol. 40, no. 2, pp. 227–237, 2010.

[4] A. Kaale, L. Smith, A. Nordahl-Hansen, M. Fagerland, and C. Kasari, "Early interaction in autism spectrum disorder: Mothers' and children's behaviours during joint engagement," *Child: Care, Health and Development*, vol. 44, no. 2, pp. 312–318, 2018.

[5] S. Y. Patterson, L. Elder, A. Gulsrud, and C. Kasari, "The association between parental interaction style and children's joint engagement in families with toddlers with autism," *Autism*, vol. 18, no. 5, pp. 511–518, 2014.

[6] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity," *Journal of Autism and Developmental Disorders*, vol. 37, no. 4, pp. 613–627, 2007.

[7] F. Edition *et al.*, "Diagnostic and statistical manual of mental disorders," *Am Psychiatric Assoc*, vol. 21, pp. 591–643, 2013.

[8] E. Schopler, R. J. Reichler, R. F. DeVellis, and K. Daly, "Toward objective classification of childhood autism: Childhood autism rating scale (cars)." *Journal of Autism and Developmental Disorders*, 1980.

[9] M. L. Sundberg, *VB-MAPP Verbal Behavior Milestones Assessment and Placement Program: a language and social skills assessment program for children with autism or other developmental disabilities: guide*. Mark Sundberg, 2008.

[10] E. Coonrod and L. Marcus, *Psychoeducational Profile – Revised (PEP-3)*. Springer New York, 2013, pp. 2439–2444.

[11] D. Bone, T. Chaspari, and S. Narayanan, "Behavioral signal processing and autism: Learning from multimodal behavioral signals," in *Autism Imaging and Devices*. CRC Press, 2017, pp. 335–360.

[12] J. Kosmicki, V. Sochat, M. Duda, and D. Wall, "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning," *Translational psychiatry*, vol. 5, no. 2, pp. e514–e514, 2015.

[13] D. Bone, S. L. Bishop, M. P. Black, M. S. Goodwin, C. Lord, and S. S. Narayanan, "Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion," *Journal of Child Psychology and Psychiatry*, vol. 57, no. 8, pp. 927–937, 2016.

[14] W. Liu, X. Yu, B. Raj, L. Yi, X. Zou, and M. Li, "Efficient autism spectrum disorder prediction with eye movement: A machine learning framework," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction (acii)*, 2015, pp. 649–655.

[15] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.

[16] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *Proceedings of the ieee international conference on computer vision*, 2017, pp. 3267–3276.

[17] J. Li, Y. Zhong, J. Han, G. Ouyang, X. Li, and H. Liu, "Classifying asd children with lstm based on raw videos," *Neurocomputing*, vol. 390, pp. 226–238, 2020.

[18] B. Li, A. Sharma, J. Meng, S. Purushwalkam, and E. Gowen, "Applying machine learning to identify autistic adults using imitation: An exploratory study," *PloS one*, vol. 12, no. 8, p. e0182652, 2017.

[19] H.-y. Lee, T.-y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao, and T.-L. Pao, "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition." in *Proceedings of Interspeech*, 2013, pp. 215–219.

[20] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, and X. Zou, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Computer Speech & Language*, vol. 56, pp. 80–94, 2019.

[21] C. Tang, W. Zheng, Y. Zong, N. Qiu, C. Lu, X. Zhang, X. Ke, and C. Guan, "Automatic identification of high-risk autism spectrum disorder: a feasibility study using video and audio data under the still-face paradigm," *ieee transactions on neural systems and rehabilitation engineering*, vol. 28, no. 11, pp. 2401–2410, 2020.

[22] R. Zhang, J. Chen, G. Wang, R. Xu, K. Zhang, J. Wang, and W. Zheng, "Towards a computer-assisted comprehensive evaluation of visual motor integration for children with autism spectrum disorder: a pilot study," *Interactive Learning Environments*, vol. 31, no. 7, pp. 4083–4098, 2023.

[23] J. Hashemi, G. Dawson, K. L. Carpenter, K. Campbell, Q. Qiu, S. Espinosa, S. Marsan, J. P. Baker, H. L. Egger, and G. Sapiro, "Computer vision analysis for quantification of autism risk behaviors," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 215–226, 2018.

[24] W. Liu, T. Zhou, C. Zhang, X. Zou, and M. Li, "Response to name: A dataset and a multimodal machine learning framework towards autism study," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 178–183.

[25] W. Liu, "Identifying children autism spectrum disorder via machine learning based behavior analysis," Ph.D. dissertation, Carnegie Mellon University, 2022.

[26] J. Liu, Z. Wang, H. Qin, Y. Wang, J. Deng, H. Li, Q. Xu, X. Xu, and H. Liu, "Social recognition of joint attention cycles in children with autism spectrum disorders," *IEEE Transactions on Biomedical Engineering*, 2023.

[27] Y. Pan, K. Cai, M. Cheng, X. Zou, and M. Li, "Responsive social smile: A machine learning based multimodal behavior assessment framework towards early stage autism screening," in *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2021, pp. 2240–2247.

[28] M. Cheng, Y. Zhang, Y. Xie, Y. Pan, X. Li, W. Liu, C. Yu, D. Zhang, Y. Xing, X. Huang *et al.*, "Computer-aided autism spectrum disorder diagnosis with behavior signal processing," *IEEE Transactions on Affective Computing*, 2023.

[29] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," *arXiv preprint arXiv:1609.01885*, 2016.

[30] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. T. Torres, C. Pelachaud, E. André, and M. Valstar, "The noxi database: Multimodal recordings of mediated novice-expert interactions," in *Proceedings of 19th ACM International Conference on Multimodal Interaction*, 2017, p. 350–359.

[31] E. Kesim, T. Numanoglu, O. Bayramoglu, B. B. Turker, N. Hussain, M. Sezgin, Y. Yemez, and E. Erzin, "The ehri database: a multimodal database of engagement in human–robot interactions," *Language Resources and Evaluation*, pp. 1–25, 2023.

[32] A. Emerson, N. Henderson, J. Rowe, W. Min, S. Lee, J. Minogue, and J. Lester, "Early prediction of visitor engagement in science museums with multimodal learning analytics," in *Proceedings of the International Conference on Multimodal Interaction*, 2020, pp. 107–116.

[33] A. Abedi and S. S. Khan, "Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network," in *2021 18th Conference on Robots and Vision (CRV)*, 2021, pp. 151–157.

[34] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," *Appl Intell*, vol. 51, pp. 6609–6621, 2021.

[35] S. Dermouche and C. Pelachaud, "Engagement modeling in dyadic interaction," in *Proceedings of the 2019 International Conference on Multimodal Interaction*. Suzhou: ACM, 2019, pp. 440–445.

[36] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2018.

[37] C. Kasari, N. Brady, C. Lord, and H. Tager-Flusberg, "Assessing the minimally verbal school-aged child with autism spectrum disorder," *Autism Research*, vol. 6, no. 6, pp. 479–493, 2013.

[38] A. B. Ellawadi and S. E. Weismer, "Using spoken language benchmarks to characterize the expressive language skills of young children with autism spectrum disorders," *American Journal of Speech-Language Pathology*, vol. 24, no. 4, pp. 696–707, 2015.

[39] A. J. Thurman, J. O. Edgin, S. L. Sherman, A. Sterling, A. McDuffie, E. Berry-Kravis, D. Hamilton, and L. Abbeduto, "Spoken language outcome measures for treatment studies in down syndrome: Feasibility, practice effects, test-retest reliability, and construct validity of variables generated from expressive language sampling," *Journal of Neurodevelopmental Disorders*, vol. 13, pp. 1–17, 2021.

[40] L. del Hoyo Soriano, J. C. Villarreal, A. Sterling, J. Edgin, E. Berry-Kravis, D. R. Hamilton, A. J. Thurman, and L. Abbeduto, "The association between expressive language skills and adaptive behavior in individuals with down syndrome," *Scientific Reports*, vol. 12, no. 1, p. 20014, 2022.

[41] S. Sachse and W. Von Suchodoletz, "Early identification of language delay by direct language assessment or parent report?" *Journal of Developmental & Behavioral Pediatrics*, vol. 29, no. 1, pp. 34–41, 2008.

[42] G. Conti-Ramsden and K. Durkin, "Language development and assessment in the preschool period," *Neuropsychology review*, vol. 22, pp. 384–401, 2012.

[43] J. E. Dockrell, "Assessing language skills in preschool children," *Child Psychology and Psychiatry Review*, vol. 6, no. 2, pp. 74–85, 2001.

[44] B. MacWhinney, "The CHILDES project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database," 2000.

[45] J. Miller and R. Chapman, "Systematic analysis of language transcripts (salt)," Madison, WI, 2008.

[46] L.-J. MA and J.-J. ZHANG, "Do speech-associated gesture and speech share the same communication system?" *Advances in Psychological Science*, vol. 19, no. 7, p. 983, 2011.

[47] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, 2002, pp. 396–401.

[48] M. W. Alibali, "Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information," *Spatial Cognition and Computation*, vol. 5, no. 4, pp. 307–331, 2005.

[49] L. Dipper, M. Pritchard, G. Morgan, and N. Cocks, "The language–gesture connection: Evidence from aphasia," *Clinical Linguistics & Phonetics*, vol. 29, no. 8-10, pp. 748–763, 2015.

[50] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2516–2520.

[51] L. M. Justice, W.-Y. Ahn, and J. A. Logan, "Identifying children with clinical language disorder: an application of machine-learning classification," *Journal of Learning Disabilities*, vol. 52, no. 5, pp. 351–365, 2019.

[52] A. Parola, I. Gabbatore, L. Berardinelli, R. Salvini, and F. M. Bosco, "Multimodal assessment of communicative-pragmatic features in schizophrenia: a machine learning approach," *NPJ Schizophrenia*, vol. 7, no. 1, p. 28, 2021.

[53] A. Gurrutxaga and I. Alegria, "Combining different features of idiomaticity for the automatic classification of noun+ verb expressions in basque," in *Proceedings of the 9th Workshop on Multiword Expressions*, 2013, pp. 116–125.

[54] B. Ashwini, V. Narayan, and J. Shukla, "Spasht: Semantic and pragmatic speech features for automatic assessment of autism," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[55] R. Lahiri, T. Feng, R. Hebbar, C. Lord, S. H. Kim, and S. Narayanan, "Robust self supervised speech embeddings for child-adult classification in interactions involving children with autism," *arXiv preprint arXiv:2307.16398*, 2023.

[56] A. Xu, R. Hebbar, R. Lahiri, T. Feng, L. Butler, L. Shen, H. Tager-Flusberg, and S. Narayanan, "Understanding spoken language development of children with asd using pre-trained speech embeddings," *arXiv preprint arXiv:2305.14117*, 2023.

[57] H. E. Nag, A. Nordgren, B.-M. Anderlid, and T. Nærland, "Reversed gender ratio of autism spectrum disorder in smith-magenis syndrome," *Molecular Autism*, vol. 9, no. 1, p. 1, 2018.

[58] L. Li, Y. Su, W. Hou, M. Zhou, Y. Xie, X. Zou, and M. Li, "Expressive language profiles in a clinical screening sample of mandarin-speaking preschool children with autism spectrum disorder," *Journal of Speech, Language, and Hearing Research*, pp. 1–22, 2023.

[59] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 009–12 019.

[60] W. Wang, D. Cai, Q. Lin, L. Yang, J. Wang, J. Wang, and y. Li, Ming journal=arXiv preprint arXiv:2109.02002, "The dku-dukeece-lenovo system for the diarization task of the 2021 voxceleb speaker recognition challenge."

[61] Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, "LSTM Based Similarity Measurement with Spectral Clustering for Speaker Diarization," in *Proceedings of Interspeech*, 2019, pp. 366–370.

[62] FFSVC, "Ffsvc2022 baseline system," 2022, gitHub repository. [Online]. Available: https://github.com/FFSVC/FFSVC2022_Baseline_System

[63] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[64] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[65] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[66] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu *et al.*, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," *arXiv preprint arXiv:2104.03603*, 2021.

[67] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.

[68] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third dihard diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[69] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[70] W. Wang, D. Cai, X. Qin, and M. Li, "The dku-dukeece systems for voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2010.12731*, 2020.

[71] X. Qin, M. Li, H. Bu, W. Rao, R. K. Das, S. Narayanan, and H. Li, "The INTERSPEECH 2020 Far-Field Speaker Verification Challenge," in *Proc. Interspeech 2020*, 2020, pp. 3456–3460.

[72] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "WeNet: Production Oriented Streaming and Non-Streaming End-to-End Speech Recognition Toolkit," in *Proceedings of Interspeech 2021*, 2021, pp. 4054–4058.

[73] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proceedings of Interspeech 2020*, 2020, pp. 5036–5040.

[74] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6182–6186.

[75] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[76] L. Magic Data Technology Co., "Magicdata mandarin chinese read speech corpus," http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101",05/2019, 2019.

[77] L. Beijing DataTang Technology Co., "aidatatang_200zh, a free chinese mandarin speech corpus," www.datatang.com, 2020.

[78] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proceedings of Interspeech*, 2019, pp. 2613–2617.

[79] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, Nov 2017, pp. 1–5.

[80] J. Sun, "Jieba," https://github.com/fxsjy/jieba, 2022.

[81] MDBG, "Cc-cedict," https://www.mdbg.net/chinese/dictionary?page=cc-cedict, 2022.

[82] J. F. Miller, *Assessing Language Production in Children: Experimental Procedures*. Baltimore, MD: University Park Press, 1981.

[83] M. J. Pezold, C. M. Imgrund, and H. L. Storkel, "Using computer programs for language sample analysis," *Language, Speech, and Hearing Services in Schools*, vol. 51, no. 1, pp. 103–114, 2020.

[84] A. J. Owen and L. B. Leonard, "Lexical diversity in the spontaneous speech of children with specific language impairment: Application of d," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 5, pp. 927–937, 2002.

[85] Y. Pan, J. Wu, R. Ju, Z. Zhou, J. Gu, S. Zeng, Y. Lynn, and M. Li, "A multimodal framework for automated teaching quality assessment of one to many online instruction videos," in *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2022.

[86] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[87] M. Benavent-Lledo, D. Mulero-Pérez, D. Ortiz-Perez, J. Rodriguez-Juan, A. Berenguer-Agullo, A. Psarrou, and J. Garcia-Rodriguez, "A comprehensive study on pain assessment from multimodal sensor data," *Sensors*, vol. 23, no. 24, p. 9675, 2023.

[88] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao *et al.*, "Large ai models in health informatics: Applications, challenges, and the future," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[89] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[90] J. F. Hair Jr, W. C. Black, B. J. Babin, and R. E. Anderson, "Multivariate data analysis," in *Multivariate data analysis*, 2010, pp. 785–785.