Fast screening for children's developmental language disorders via comprehensive speech ability evaluation—using a novel deep learning framework

Xing Zhang[1#], Feng Qin[2#], Zelin Chen[3], Leyan Gao[4], Guoxin Qiu[2], Shuo Lu[4]

[1]College of Foreign Studies, Jinan University, Jinan, China; [2]Department of Neurosurgery, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China; [3]School of Data and Computer Science, [4]Department of Chinese Language and Literature, Sun Yat-sen University, Guangzhou, China

Contributions: (I) Conception and design: X Zhang, F Qin, S Lu; (II) Administrative support: G Qiu, F Qin; (III) Provision of study materials or patients: S Lu, F Qin; (IV) Collection and assembly of data: Z Chen, L Gao; (V) Data analysis and interpretation: Z Chen, S Lu, L Gao; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first author.

Correspondence to: Shuo Lu, PhD. Neurolinguistics Teaching Laboratory, Department of Chinese Language and Literature, Sun Yat-sen University, 135 Xingang Road, Guangzhou, China. Email: Lush3@mail.sysu.edu.cn.

Background: Developmental language disorders (DLDs) are the most common developmental disorders in children. For screening DLDs, speech ability (SA) is one of the most important indicators.

Methods: In this paper, we propose a solution for the fast screening of children's DLDs based on a comprehensive SA evaluation and a deep framework of machine learning. Fast screening is crucial for promoting the prevalence and practicality of DLD screening which in turn is important for the treatment of DLDs and related social and behavioral abnormalities (e.g., dyslexia and autism). Our solution is focused on addressing the drawbacks existing in the previous DLD screening methods which include test failure due to text-based inducing material design and illiteracy of most young children, incomplete language evaluation indicators, and professional-reliant evaluation procedures. First, to avoid test failure, a novel comprehensive inducing procedure (CIP) with non-text (i.e., audio-visual) stimulus materials was designed that could cover a large range of modalities to adequately explore the comprehensive SA of the subjects. Second, to address incomplete language evaluation, a set of comprehensive evaluation indicators with full consideration of the characteristics of the children's language acquisition is proposed; furthermore, to break the professional-reliant limitation, we specifically designed a deep framework for fast and accurate screening.

Results: Experimental results showed that the proposed deep framework is effective and professional with a 92.6% accuracy on DLD screening. Additionally, to provide a benchmark for the novel problem, we provide a CIP dataset with about 2,200 responses from over 200 children, which may also be useful for further DLD studies and insightful for the fast screening design of other behavioral abnormalities.

Conclusions: Fast screening of children's DLDs can be achieved at accuracy up to 92.6% by our proposed deep learning framework. For successful fast screening, an elaborated CIP with corresponding comprehensive evaluating indicators is necessary to be designed for children suspected to have DLDs.

Keywords: Developmental language disorders (DLDs); developmental language disorder indicators (DLD indicators); fast screening

Submitted Oct 13, 2019. Accepted for publication Apr 29, 2020.
doi: 10.21037/atm-19-3097
View this article at: http://dx.doi.org/10.21037/atm-19-3097

Introduction

Developmental language disorders (DLDs) are the most common developmental disorders in children with a prevalence of 5–8% in preschool (1,2). Most children with DLDs suffer various speech disorders (e.g., pronunciation or comprehension obstacles), which lead to other social and behavioral abnormalities (e.g., dyslexia, communication disorders, autism, and attention deficit disorder). Screening for DLDs in young children is extremely important and necessary to take precautionary measures and treat development disorders effectively.

For screening DLDs, we argue that a comprehensive evaluation of the children's speech ability (SA) plays a core role since the SA is the most important indicator for language development (3,4) and covers a broad range of language abilities for children. Specifically, SA is defined as proficiency in oral language, which includes the ability to repeat or retell properly, pronounce correctly and fluently, and express grammatical and logical content. Usually, SA is evaluated based on a collection of the subject's responses of some specific speech-inducing procedures prompted by a professional (5-9). To the best of our knowledge, few studies have focused on comprehensive SA evaluation for children's DLD screening, since it is incredibly challenging. In this paper, we attempt to address these difficulties and provide a feasible solution by designing a fast screening framework based on the deep learning technology.

The first challenge of fast DLD screening via SA evaluation is the elaborate design of inducing procedures and evaluation scales. For the design of speech-inducing procedures, the conventional stimulus materials that rely highly on reading texts (6,10,11) may lead to test failure (which denotes that the children give no responses for the stimulus materials) on young children, since they usually have an immature language system and nonstandard speech characteristics. For the design of evaluation scales to screening DLDs, the comprehensive test should evaluate an extensive range of linguistic aspects (e.g., pronunciation accuracy for the low-level aspect and logical consistency for the high-level aspect). Based on the above consideration of the first challenge, we propose a novel comprehensive inducing procedures (CIP) and the relevant comprehensive CIP indicators (scales) that are specifically designed for DLD screening. To avoid test failure and induce the children's full speech performance sufficiently and effectively, the stimulus materials are audio-visual in nature with no reliance on texts. The inducing procedures cover a broad range of difficulties from easy to hard by elaborately controlling the word choices, sentence length, grammar, and semantic complexity. For a comprehensive evaluation, not only the inducing procedures cover a full variety of speech modalities including repetition, restatement and free conversation, but also the relevant comprehensive CIP indicators cover an extensive range of linguistic aspects; i.e., pronunciation, expression efficiency, fluency grammar, semantic and logic.

The second challenge of fast DLD screening is the traditional high reliance on professionals, which is expensive, subjective and time-consuming. To achieve fast and accurate screening, we specifically designed a novel deep framework to automatically rate the responses of the inducing procedures in CIP according to the SA performance and created a CIP dataset with about 2,200 responses from over 200 children. The automated CIP rating is very challenging due to the comprehensive (multi-aspect) nature of evaluation and the large variance of responses (Var-Resp) caused by the broad range of modalities and the varied speech performances of the children. For the former (multi-aspect evaluation), a two-stream architecture was proposed, utilizing the complementary features from the pronunciation stream (for low-level aspect) and content stream (for high-level aspect). For the latter (Var-Resp), an excellent-comparison architecture was proposed, which evaluates the response by comparing it with the excellent responses (i.e., the response with the highest rating). The model is procedure-aware and excellent-aware, which alleviates the Var-Resp problem.

The main contributions of this paper are as follows:

❖ A novel and meaningful problem is addressed; i.e., the fast screening for children's DLDs via comprehensive SA evaluation;

❖ An elaborated CIPs with corresponding comprehensive evaluating indicators was especially designed for children suspected to have DLDs;

❖ A novel deep framework was designed for fast and automated screening, and achieved a remarkable performance, with a verification accuracy of 92.9% for DLD screening and a rapid analysis time of 1.07 s.

The study in this paper has been approved by the Ethics Committee of the Sun Yat-sen University, and conforms to the provisions of the Helsinki Declaration as revised in 2013 (available at: http://www.wma.net/en/30publications/10policies/b3/%20index.html). All the subjects' guardians have signed an informed consent form.

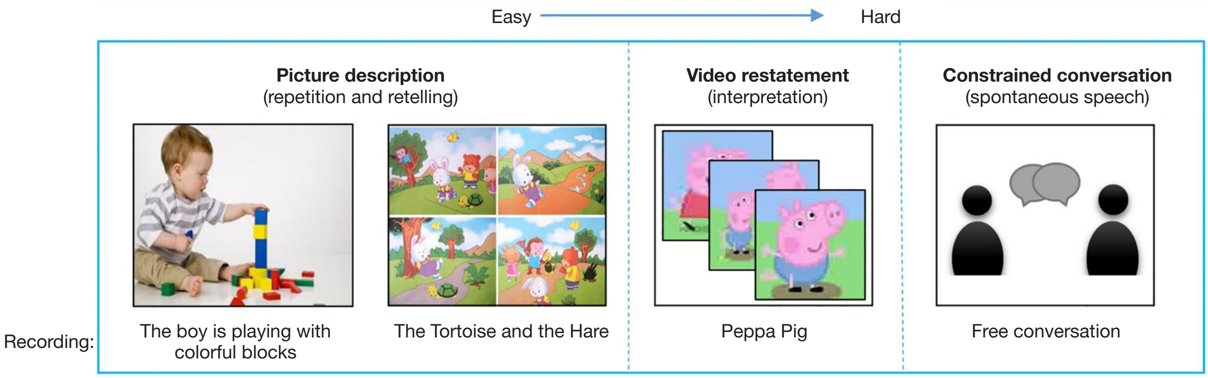


Figure 1 Examples of CIPs. CIP, comprehensive inducing procedures.

Children's DLD screening

DLDs are the most common developmental disorder in childhood. The early detection of language disorders has come to be recognized as a crucial step in achieving the best possible outcome for the affected children (12-14). For DLD screening, language evaluation scales are used (1,3,15-17). Most of these scales-based tests suffer the problems of insufficient SA evaluation. For example, in Rescorla (17), the language survey consists of a vocabulary checklist designed to screen DLDs, but it only focuses on one type of children's language disorder, namely, vocabulary development delay, ignoring many other disorders (e.g., pronunciation obstacles and communication disorders). Also, the checklist is completed by the parents, causing a large subjective bias or instability. The Clinical Evaluation of Language Fundamentals (CELF) is also a commonly used test battery for the diagnosis of DLDs only suitable for evaluating children aged five years and above (18). The scale-based tests also suffer from a professional-reliant limitation; i.e., they require speech and language therapists (SLTs) or speech-language pathologists (SLPs). Due to these obstacles, it is very challenging to distinguish a potential disorder from the normal course of development in young children, and there is often subjective bias and instability.

Different from the previous works, our solution of DLD screening is comprehensive (addressed by the CIP) and not reliant on professional experts (addressed by the proposed deep framework).

SA evaluation

The SA evaluation includes two parts: inducing procedures and evaluation indicators. Firstly, the subject is asked to give responses following a designed inducing procedure, and then the responses are evaluated by a professional based on the selected evaluation indicators (5,7,10,19).

Inducing procedures

Many SA inducing procedures have been proposed. Some (5-7) induce the speech based on texts (e.g., asking the subjects to read fixed texts). However, for young children aged 2 to 8, text-based procedures can lead to test failure, since most children at these ages are illiterate. Some (8,9) induce the speech based on word or sentence repeating tasks, which can be conducted uniformly. However, such tasks restrict the children's full language performance and fail to observe aspects of their high-level speech abilities such as semantics and logic, which are important for screening DLDs. To address the above challenges (test failure and incomplete evaluation) of the SA inducing procedures design, we propose the CIPs, as shown in *Figure 1*. The CIP stimulus is based on audiovisual materials (e.g., recordings, pictures, and videos) and is specifically designed according to children's cognitive abilities to avoid test failure. To avoid incomplete evaluation, the CIP covers a large range of speech modalities and difficulties, and can induce subject responses effectively from various linguistic aspects. A comparison of CIP with the commonly used SA inducing procedures is shown in *Table 1*.

Evaluation indicators

Most of the relevant literature (8,22-24) proposes SA evaluation indicators of fluency, pronunciation, and prosody, which focus on low-level speech information. For example,

Table 1 The comparison of the CIP with the commonly used SA inducing procedures

Comparison procedures	Automated	Comprehensive	Fast screening	Multi-modalities	Test success for DLDs
Sutherland *et al.* (20,21)	×	×	×	×	×
CELF-5 (8,18)	×	×	×	×	×
CIP (ours)	√	√	√	√	√

CIP, comprehensive inducing procedure; DLD, developmental language disorder; SA, speech ability.

Table 2 The summary of the proposed comprehensive linguistic indicators (see Sec. 2.2)

Linguistic level	Linguistic aspect	Indicator	Definition and rating method
Low	Pronunciation	Initial consonant	Number of errors on initial consonants
		Tone	Number of errors on tones
		Vowel	Number of errors on vowels
	Expression efficiency	Syllable count	Number of syllables
		Speech speed	The number of syllables produced in a second
		Pronunciation duration	The duration of the cumulative pronunciation
Medium	Fluency	Content restatement/replication	Restatements or repeated pronunciations
		Redundant articles	Particles like “uh, a, then, this, that” being used as a gap filler between phrases or sentences
		Pause count	Silence longer than 0.3 seconds counts as a pause
		Pause duration	Cumulative duration of all pauses longer than 0.3 sec
	Grammar	The wrong usage of grammar	Instances of incorrect grammar usage including of function words, grammatical construction, and word order
High	Semantic	Keywords missing	Number of keywords missing and redundant words conflicting with the materials
	Logic	Information organization	Number of incorrectly sequenced keywords

in Zechner *et al.* (25) and Xie *et al.* (26), the authors used indicators mostly in the fluency domain to build an automated scoring system for non-native speech. The previously proposed indicators were not suitable for our setting since the SA evaluation between the patient and the healthy subject is considerably different. Specifically, DLD children present SA evaluation with a unique challenge: the evaluators must differentiate between fundamental speech disorders and perceived difficulties resulting from an individual’s normal developing speech differences. Differences may be expressed in sentence structure, speech sound production, vocabulary, and pragmatics (27). To achieve this, we argue that a comprehensive evaluation is essential. Our evaluation indicators consider various levels of language features and speech aspects (see *Table 2*).

Automated speech evaluation

To the best of our knowledge, there are no proposals for Mandarin automated speech evaluation for children with DLDs. However, several studies (7,11,28,29) have proposed an automated speech evaluation system for other applications. Proenca *et al.* (11) proposed the features of speech rating, pronunciation measure, and repeating accuracy to evaluate oral proficiency. Loukina *et al.* (7) adopted lasso regression (30) to select the effective features for speech scoring. Yoon *et al.* (10) proposed the word-embedding base content features for automated scoring. Recently, to break the limitation of elaborated hand-crafted features, Chen *et al.* (29) proposed a bi-directional long short-term memory model (bi-LSTM) (31,32) with an attention-based architecture on the spoken content. Most methods above adopted automatic speech recognition (ASR) (33), which transforms the voice signal into text to extract features. However, for potential DLD children, the performance of conventional ASR is not promising (34), since most of the ASR models were specifically designed for the adults but not DLD children (i.e., the training data of the ASR models came from adults and could hardly be transferred to models for DLD children). To alleviate the less-than-promising performance of ASR, in our model, an additional pronunciation stream that directly extracts features from the voice signal is concatenated to the text-based feature. Moreover, the methods above are not very suitable for a CIP automated rating, since none of them consider the challenges in CIP rating, such as comprehensive evaluation and the Var-Resp problem.

Methods

In this section, we elaborate on the design of CIP (Sec. 2.1), the evaluation indicators of CIP (Sec. 2.2), and the CIP dataset (Sec. 2.3), and the model for fast screening (Sec. 2.4).

The design of CIPs

To avoid test failures with speech-inducing for children with DLDs and to collect their speech performance as quickly as possible, we propose a novel design of CIP, which is specifically designed for a fast screening of DLDs.

The inducing procedures include three types of tasks (i.e., picture description, video statement, and conversation). There are different rubrics under each task type differing in difficulties, content, and linguistic characteristics (e.g., syntactic structures, semantic reference prominence etc.). The content and topics of the materials use common scenes close to children's daily life (e.g., clothes, food, toys, and schools), which can avoid test failures caused by children's cognitive limits. The inducing procedures cover a large range of difficulties to meet the diversity of SA in young children. Additionally, some rubrics are presented with guide words, and some are not, which is to differentiate the difficulty and predictability of the expected response. Some examples of CIP are shown in *Figure 1*. A brief description of the task types and the rubrics are as follows.

Task type I: picture description

This task type mainly addresses the speech modality of repeating and retelling. The subject is asked to describe the content of the pictures. The pictures are presented together with a recording of guide words describing the picture content. Rubrics differ in difficulty by controlling the sentence length, syntactic, and semantic complexity. The easiest rubrics can be finished by simply repeating the heard record (e.g., "*I like eating apples*."). The hardest involves restating a short story according to four-frame comic pictures (e.g., "*The Race Between Hare and Tortoise*"), which requires high-level SA (such as adequate logic and coherence).

Task type II: video restatement

This task type evaluates the speech modality of interpretation and restatement. The children are presented two short videos. The first video is the popular children's animation "Peppa Pig" which shows a complete simple story for children. The second video is silent, displaying a man helping an old lady cross the street. To finish the two rubrics, children must first interpret and memorize the plot properly and generate their speech with well-formed logic and coherence while choosing adequate words and expressions.

Task type III: constrained conversation

This task type evaluates the speech modality of spontaneous communication. The children's responses to the questions about their familiar personal experience (e.g., family situation, and favorite games or cartoons) are recorded. The answer time is limited to 1 minute. All the responses obtained from the CIP are evaluated on 6 major SA aspects with 13 indicators.

Indicators for evaluation

To perform a comprehensive evaluation of the response, we chose 13 indicators along 6 major linguistic aspects: pronunciation, expression efficiency, fluency, grammar, semantics, and logic. The first two are low-level aspects, fluency and grammar are on the medium-level, while semantics and logic are high-level linguistic aspects. A brief description of the indicators is available in *Table 2*. The proposed indicators are decided with the consideration of the purpose of comprehensive evaluation and ontological linguistic knowledge (25,35-40). Below is a brief description of the indicators of 6 major linguistic aspects from low- to high-level.

Pronunciation (low-level): clear and correct pronunciation is the foundation of SA. In Mandarin, the pronunciation of characters usually includes an initial consonant, a vowel following the initial consonant, and the pronunciation of tone. For example, in the character “ wŏ ” (water), the initial consonant is “w”, the vowel is “o”, and the tone is “ŏ”. We evaluate pronunciation by counting the pronunciation errors in the initial consonant, vowel, and tone.

Expression (low-level): the expression efficiency is defined as the ability to produce meaningful sentences within a given unit of time. This is a necessary language aspect for SA, especially when evaluating spontaneous speech since it indicates the speech generation ability. Specifically, we introduce three indicators (syllable count, speech, and pronunciation duration) to evaluate the efficiency aspect, inspired by the Mandarin proficiency test (21,41), IELTS (42), and TOEFL oral test (43). The details are shown in *Table 2*.

Fluency (medium-level): speech fluency is used here to manifest a medium-level spoken language proficiency regarding the smoothness or flow with which sounds, syllables, words, and phrases are joined together when speaking (44). Fluency can be heavily damaged by meaningless elements in speech, such as pauses and redundant articles. Indicators are identified for evaluation speech fluency and evaluated by counting the occurrence of flaws, as shown in *Table 2*.

Grammar (medium-level): as an isolating language (languages that use little or no inflection to indicate grammatical relationships), the grammar of Chinese is relatively simple and loose (23,45,46). Functional words (words with little lexical meaning and express grammatical relationships among words within a sentence.), grammatical construction (any syntactic string of words ranging from phrasal structures to certain complex lexemes, such as verb-object constructions.), and word order are the main locales of grammar’s structural rules, so the grammar aspect is measured by counting the number of times these three rules are violated.

Semantic (high-level): the semantic aspect focuses on the accuracy of the content and is evaluated by counting the missing items of key target information spoken by children. Key information includes key words predefined according to the content of the inducing material.

Logic (high-level): the content should be expressed logically and coherently. Organizational ability is measured by counting the difference between the ideal sequence of keywords and the actual keyword sequencing found in the speech.

To make the scores of the same indicator across different rubrics be comparable and all the scores of the indicators satisfy the logic of “the higher, the better”, we normalized the scores of the indicators among each rubric. Specifically, for the indicator in the linguistic aspects of pronunciation, fluency, grammar and logic, the scores were normalized as follows:

$$x' = \frac{\max\{X\} - x}{\max\{X\} - \min\{X\}} \qquad [1]$$

where $x \in X \subset R^N$, X denotes all the scores (of different subjects) of a specific indicator of a rubric and N is the number of the subjects. For the indicators in the linguistic aspects of expression efficiency and semantic, we normalized the scores among each rubric as follows:

$$x' = \frac{x - \min\{X\}}{\max\{X\} - \min\{X\}} \qquad [2]$$

CIP dataset

To demonstrate the effectiveness of the CIP and the CIP indicators and establish an effective fast screening algorithm, a CIP dataset was collected, which included about 2,200 audio responses via CIP from 284 children (140 boys). Also, the DLD designations of the children were provided by a test conducted by experts on children’s language acquisition. Furthermore, to check the reliability of the evaluation based on CIP, 20 age-balanced children were randomly selected to conduct a second test following CIP two weeks after their first tests (retest reliability). The correlation coefficient of the rates of the two test times was calculated, showing high reliability (r>0.9).

Details of the dataset

The CIP dataset was collected from a kindergarten and a private rehabilitation institution for children. All the children were aged 2 to 8 years and gave their responses in Mandarin the first time for of all the inducing procedures. In total, 168 children (82 boys) DLD designations were available, with 36 DLDs (17 boys) and 132 non-DLDs. Some children’s DLD designation remained unknown, as they did not finish the DLD test. Since the CIP covers a broad range of difficulty, some failed responses to the more difficult procedures occurred. We excluded the failed responses from the dataset, and 2,198 responses ultimately remained. All the responses were evaluated according to the proposed CIP indicators as shown in *Table 2*. Additionally, since automatically learning the SA rating without supervision is particularly challenging, we further employed professors to rate each response according to the

Table 3 The scale of the response rating

Score	Description
0	No meaningful syllables or intelligible utterances
1	No complete sentences but individual words
2	A few short sentences with deficit in grammar or semantics
3	Some dysfluency sentences with intact semantics
4	Successful idea expression with obvious flaws in fluency or logic
5	Normal speech with only occasional flaws in pronunciation or fluency
6	Fluent, clear, complete, and standard speech

SA performance to provide a rough SA supervision for our deep model. The evaluation scales are shown in *Table 3*.

Evaluation protocol

For DLD screening evaluation, the subjects with DLD designations were randomly split into the training set and the test set with 133 and 35 subjects respectively, and the distribution of the DLD subjects in the training set and the test set were even. For SA evaluation, the responses of each inducing procedure in CIP were randomly split into a training set and test set in a ratio of 4:1. The training set contained 1,756 responses, and the test set contained 442 responses.

Model for fast screening

For fast DLDs screening, we first trained the deep model with the CIP responses' rating, and then leveraged the trained model to extract the audio feature for DLD screening. To this end, we first needed to address the main challenges in CIP rating; i.e., multi-aspect nature of evaluation and the large variance among different responses (Var-Resp) caused by various modalities and difficulties of the inducing procedure.

For the former challenge (multi-aspect nature of evaluation), a two-stream encoder was proposed to evaluate the responses from two main aspects: pronunciation and content presentation. Specifically, of the two streams, one is to extract audio pronunciation feature (pronunciation stream), and the other is to extract the feature of audio content presentation (content stream). Moreover, we additionally learned other auxiliary evaluating tasks rather than learn the rating task only to let the model be aware of the more evaluative aspects. For the latter challenge (Var-Resp), since the Var-Resp problem leads to a large variance of criteria among the different procedures, the deep model cannot be aware of the specific criterion of the evaluating response without addressing it. Hence, to help the deep model be aware of the specific criterion, an excellent-comparison architecture was proposed. In the excellent-comparison architecture, the network evaluates the response according to the excellent response (with the highest rating) of the same procedure, and the final rating is acquired by comparing the evaluating response with the excellent one. The overview of the proposed model is shown in *Figure 2*. Additionally, considering that the distribution of the rating levels is nonuniform, and the rating range of the response is only several contiguous integral points with values of 0, 1, 2, 3, 4, 5 and 6, we considered the rating task as a classification task rather than a regression task to force the network to give an integral rating and learn the distribution of the rating levels automatically.

Before elaborating the technical details of our model, we first give a brief introduction of transformer encoder (47) that is adopted in the two-stream encoder and excellent-comparison architecture. A brief overview of transformer encoder is shown in *Figure 3*. The transformer architecture mainly includes a feed forward layer and a multi-head attention layer.

Feed forward layer

Let the input $I \in \mathbb{R}^{T \times d}$ (T is the timesteps and d is the feature dimension of each timestep), and $H = FF(I)$ denotes the feed forward function as follows,

$$H = FF(I) = \sigma([I_1 W, \cdots, I_T W]) \quad [3]$$

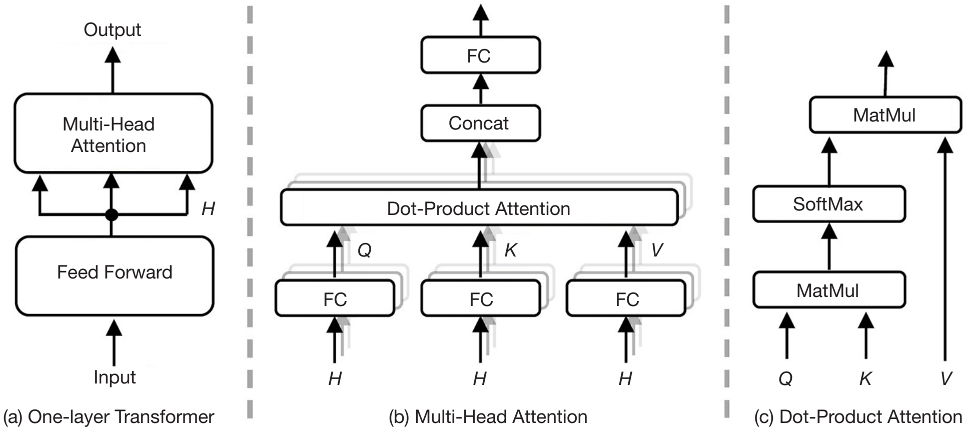
where $\sigma(\cdot)$ denotes activation function such as $relu(\cdot)$, I_* denotes the feature of timestep $* \in \{1, \cdots, T\}$ and $W \in \mathbb{R}^{d \times d}$ is learnable parameter shared across time-step.

Multi-head attention layer

Let $MHA(H)$ denotes the multi-head attention function as follows,

$$\begin{aligned} &MHA(H) = \sigma([DPA_1(Q_1, K_1, V_1), \cdots, DPA_N(Q_N, K_N, V_N)]W^A) \\ &DPA_*(Q_*, K_*, V_*) = SoftMax(Q_* K_*^T)V_* \quad [4] \\ &[Q_*, K_*, V_*] = [\sigma(HW_*^Q), \sigma(HW_*^K), \sigma(HW_*^V)] \end{aligned}$$

where $\sigma(\cdot)$ denotes activation function such as $relu(\cdot)$, $W^A \backslash W_*^Q \backslash W_*^K \backslash W_*^V$ is the learnable parameter of the fully-connected (FC) layer, N is the number of attention heads,



(a) One-layer Transformer (b) Multi-Head Attention (c) Dot-Product Attention

Figure 2 The overview of our proposed model for fast screening. The inputs of the model are the evaluating response and the excellent response which is from the same procedure. First, the response is encoded by a two-stream encoder to extract the complementary feature from pronunciation and speech content for multi-aspect evaluation. Then, the extracted features of the evaluating response and the excellent response are fed into a proposed excellent-comparison architecture. This excellent-comparison architecture is used to evaluate the response by comparison with the excellent response, by which the network is made more aware of the criterion of the procedure. Finally, the feature (after excellent comparison) is fed into a classifier to acquire the rating.

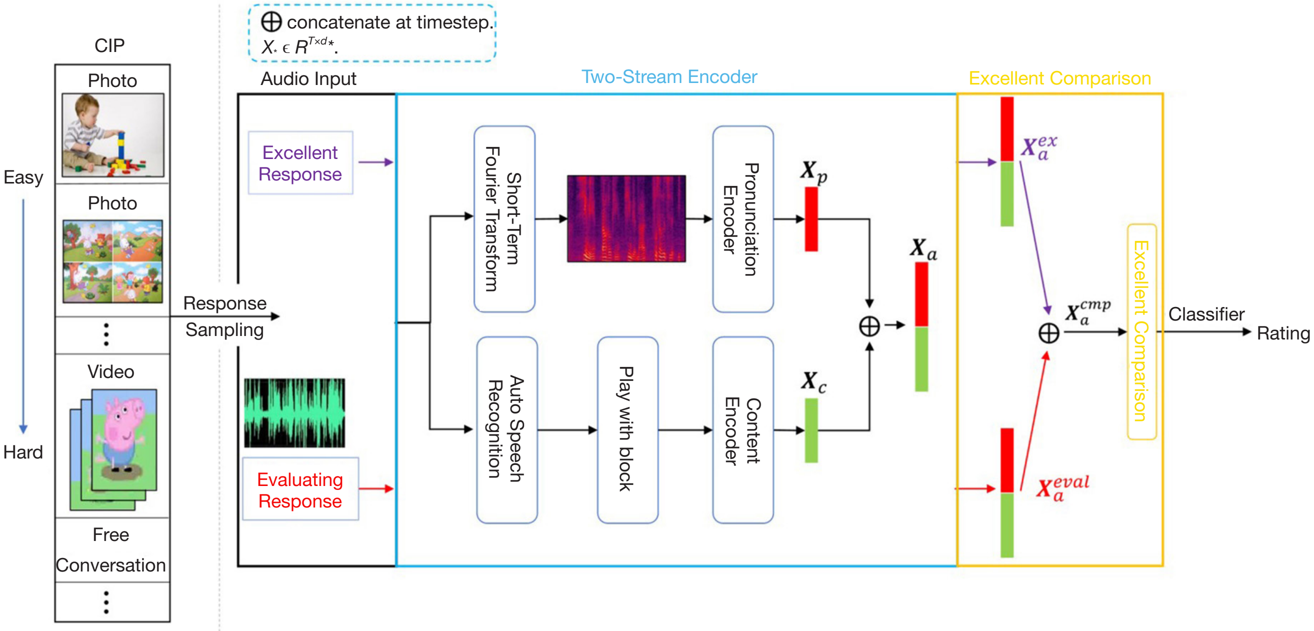


Figure 3 A brief overview of transformer encoder (47), where "FC" denotes fully-connected layer and "MatMul" denotes matrix multiply. To see more technical details, please refer to the work (47).

$* \in \{1,\cdots,N\}$ denotes the *-th head, and $DPA_*(\cdot,\cdot,\cdot)$ denotes the dot-product attention. More illustration of Eq. [4] are shown in *Figure 3B,C*.

In the following part, we illustrate the technical details of the proposed deep framework.

Two stream encoder

The two-stream encoder is designed to extract a more comprehensive representation of an audio from both the pronunciation and content. Let $\mathbf{X} \in \mathbb{R}^T$ (T is the total timesteps of the audio) be the audio of the response. For

Table 4 The top-5 important CIP indicators

Indicator	Linguistic aspect				Importance
	Pro	Flu	Gra	Sem	
Semantic				√	13.2%
Initial consonant	√				11.0%
Redundant articles		√			11.0%
Content restatement		√			10.8%
Grammar			√		9.5%

CIP, comprehensive inducing procedure; Pro, pronunciation; Flu, fluency; Gra, grammar; Sem, semantic.

pronunciation stream (the top stream shown in *Figure 2*), $\mathbf{X}$ is transformed by Short-Term Fourier and is then encoded into a pronunciation feature $\mathbf{X}_p \in \mathbb{R}^{T_p \times d}$ (d is the feature dimension of each timestep) by a transformer (47). For the content stream, $\mathbf{X}$ is transformed into transcription by a pre-trained ASR such as that of Zhou *et al.* (48) and is then encoded into a content feature $\mathbf{X}_c \in \mathbb{R}^{T_c \times d}$ by another transformer. Finally, the pronunciation feature $\mathbf{X}_p$ and the content feature $\mathbf{X}_c$ are concatenated at the timestep dimension to produce the audio encoding $\mathbf{X}_a \in \mathbb{R}^{(T_p+T_c)\times d}$.

Excellent-comparison architecture

The excellent-comparison architecture is designed to let the audio feature be aware of the rating procedures and the rating criteria (see *Figure 2*). Let $\mathbf{X}_a^{\text{ex}} \in \mathbb{R}^{(T_p+T_c)\times d}$ be the audio encoding (encoded by the two-stream encoder) of the excellent response and $\mathbf{X}_a^{\text{eval}} \in \mathbb{R}^{(T_p+T_c)\times d}$ be the audio encoding (encoded by the two-stream encoder) of the evaluating response. Then, before the comparison, the $\mathbf{X}_a^{\text{ex}}$ and $\mathbf{X}_a^{\text{eval}}$ are concatenated at the timestep dimension to produce the excellent-comparison audio encoding $\mathbf{X}_a^{\text{cmp}} \in \mathbb{R}^{(2T_p+2T_c)\times d}$; i.e., $\mathbf{X}_a^{\text{cmp}} = \left[\mathbf{X}_a^{\text{ex}}; \mathbf{X}_a^{\text{eval}}\right]$. Next, $\mathbf{X}_a^{\text{cmp}}$ is input into a multi-head attention layer (47) to evaluate the responses by comparison to an excellent response. Finally, to acquire a compact feature for efficient computation, we averaged the multi-head attention output $\{MHA\left(\mathbf{X}_a^{\text{cmp}}\right)$, see Eq. [4]} at timestep dimension to produce the final feature of the audio $\mathbf{x}^{\text{feat}} \in \mathbb{R}^d$.

Rating as classification

To force the network to give an integral rating and learn the nonuniform distribution of the rating levels automatically, we considered the rating as a classified task inspired by the fact that neural net classifiers trained with class labels can automatically capture similarity among classes (49). Specifically, we let $[y_i]_{i=1}^N$ be the ground truth of the rating, where N is the total number of the responses. The loss of the model is a cross-entropy loss:

$$L_{\text{rate}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}\left(y_i = c\right)\log\left(\mathbf{w}_c^{\text{cls}}\mathbf{x}_i^{\text{feat}}\right) \quad [5]$$

where C is the total number of the rating levels, $\mathbb{I}(\cdot)$ is the indicator function, and $\mathbf{w}_c^{\text{cls}} \in \mathbb{R}^d$ is a vector to project $\mathbf{x}_i^{\text{feat}}$ into the prediction of c^{th} rating level.

Auxiliary learning

Caused by the fact that the comprehensive evaluation has multiple aspects, only learning the rating of the response leads to poor results when the supervision from the evaluation of other aspects is absent. Hence, the auxiliary losses of other tasks were added to the primary task loss. The auxiliary tasks are designed to predict other indicators (shown in *Table 4*). The auxiliary losses are the cross-entroy loss like the primary loss L_{rate} as shown in Eq. [5]. In summary, the loss of the model is as follows:

$$L = L_{\text{rate}} + \lambda_{\text{aux}}\sum_{t_{\text{aux}}=1}^{T_{\text{aux}}} L_{\text{aux}}^{t_{\text{aux}}} \quad [6]$$

Where λ_{aux} is a hyperparameter to control the preference of learning the primary task and auxiliary tasks, T_{aux} is the total number of the auxiliary tasks, and L_{aux}^{t} is the loss of the $t_{\text{aux}}^{\text{th}}$ auxiliary task.

Results

DLD screening experiment

For screening DLDs, we followed the DLD screening protocol (see Sec. 2.3). We adopted the gradient boosting decision tree model in eXtreme Gradient Boosting (XGBoost) (50) to train a DLD screening model, where the input features are the CIP indicators or the deep learning features. The deep features are extracted from the deep model which was only trained under the supervisions of the CIP indicators and the responses' ratings of the training subjects in the DLD evaluation protocol. We set the parameters in the XGBoost model as lr =0.1, eta =0.05, depth =5, and the other parameters remained at default value in the XGBoost library.

DLD screening performance

The performance is shown in *Figure 4*. “Proenca *et al.* (6)” denotes the input of the XGBoost model is the indicators (after normalized) proposed by Proenca *et al.* (6). “Our CIP indicators” (which is professional-reliant) denotes the input is the mean CIP indicators (after normalized) of all rubrics. “Our fast screening model” denotes the input is the mean of audio features (extracted from the deep model) of all rubrics’ response. Also, we appended the additional age of the subject as the models’ input since age plays a core role in DLD screening. Clearly, our performance on DLD screening is better than the “Proenca *et al.* (6)”, since we considered comprehensive linguistic aspects when screening while the “Proenca *et al.* (6)” did not. Although only speech is available for DLD screening, our model achieved a relatively high verification accuracy, which implies that screening DLDs via SA is feasible, and our proposed CIP indicators and fast screening model are effective. Furthermore, the automated model (fast screening model) achieved comparative performance with the professional-reliant model (CIP indicators), indicating that our automated DLD screening framework is as effective as a professional evaluator.

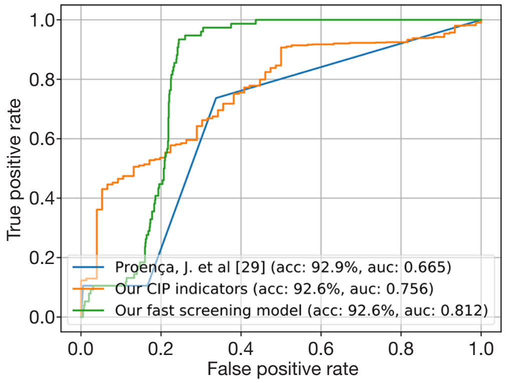


Figure 4 The ROC curve of CIP indicators and the fast model for DLD screening. The “auc” is the area under curve (the larger the better). The “acc” is the verification accuracy. The relatively high verification accuracy indicates that our DLD screening framework (via SA evaluation) was effective and outperformed the “Proenca *et al.* (6)”. The comparative performance between the CIP indicators and the fast screening model implies that our proposed fast screening model was as effective as professional evaluation. CIP, comprehensive inducing procedure; DLD, developmental language disorder; SA, speech ability.

Importance of the CIP indicators for the screening of DLDs

To explore the importance of different CIP indicators of SA evaluation for DLDs screening, we considered the indicator appearance count (after being normalized) in the tree node as the importance. The top-5 important CIP indicators are shown in *Table 4*. The table illustrates the fact that for screening DLDs, comprehensive linguistic aspects should be considered since the comparative top-5 important CIP indicators come from different linguistic aspects. Additionally, we show the importance of different linguistic aspects in *Table 5* by accumulating the importance of the CIP indicators of each aspect; this also demonstrates the necessity of comprehensive SA evaluation for DLD screening. Additionally, we also found that age plays a core role in DLD screening, with an importance of 12.0%.

SA performance of the DLDs subjects

To further illustrate the necessity of comprehensive SA evaluation, we analyzed the SA performance from six linguistic aspects of some DLD subjects (see *Figure 5*). The symptoms of DLDs subjects were various, and the variance of the performance on some linguistic aspects (e.g., semantic) was large, which implies that screening from a single (or few) linguistic aspects is infeasible.

The SA evaluation performance of the deep model

In this subsection, we illustrate that our proposed deep model achieved state-of-the-art performance on the CIP response’s ratings, which is the reason that our fast model could also perform remarkably well on DLD screening.

Table 5 The importance of different linguistic aspects for screening DLDs

Low level		Medium level		High level	
Pronunciation	Efficiency	Fluency	Grammar	Semantic	Logic
23.7%	17.3%	27.6%	9.5%	13.2%	1.5%

Note: “Efficiency” is “Expression Efficiency”. DLD, developmental language disorder.

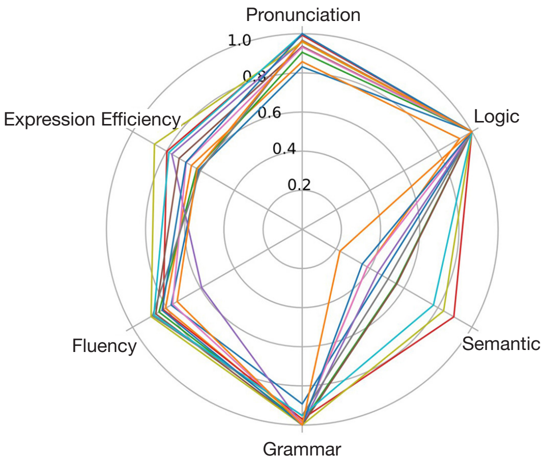


Figure 5 The SA performance of different DLD subjects. Different colors denote different subjects. DLD, developmental language disorder; SA, speech ability.

Table 6 The comparison with the state of the art

Methods	MLP (30)	L Chen (6)	Ours
Top-1 acc	70.8%	72.2%	76.1%

Note: "Top-1 acc" denotes the top-1 accuracy of all the responses rating.

Table 7 The effect of the two-stream architecture

Methods	Pronunciation stream	Content stream	Two-stream
Top-1 acc	71.2%	73.7%	74.4%

Table 8 The effect of the excellent-comparison architecture

Methods	w/o Excellent-comparison	w/ Excellent-comparison
Top-1 acc	74.6%	76.1%

For SA evaluation, we followed the SA evaluation protocol (see Sec. 2.3). Except where noted, we set λ_{aux} =0.1, and the auxiliary tasks were learning all the CIP indicators provided by the CIP dataset. A dropout layer (51) with a 0.5 dropout rate was adopted to avoid overfitting, and the Adam optimizer (52) was adopted with the learning-rate =1e-4 and batch-size =8.

Comparison with the state-of-the-art

To show that a deep model specifically designed for the CIP procedure is necessary, we compared our model with the state-of-the-art algorithms. The results are shown in *Table 6*. "*MLP*" denotes that the sophisticated features from one proposal (6) are adopted to train a multilayer perceptron (53). "*L Chen*" denotes that we reproduced the bi-directional LSTM (31,32) model with attention (29). Our method achieved the best accuracy, 76.1%, on the CIP response's rating. Compared with the conventional models ("*MLP*"), the evaluation of our deep model is better since our deep model can learn a high-level feature automatically while the conventional models cannot. Compared with the deep model ("*L Chen*"), the evaluation performance of our model was superior, and outperformed the "*L Chen*" with 3.9% at the top-1 accuracy. This performance can be attributed to our model being specifically designed for CIP procedures, addressing the multi-aspect and Var-Resp problem.

Effect of the two-stream architecture

We evaluated the single-stream architecture without the excellent-comparison architecture, and the results at top-1 accuracy are shown in *Table* 7. "*Two-stream*" achieves the best performance, which implies that the pronunciation stream and the content stream are complementary, and more information on the responses can be extracted by the two-stream architecture.

Effect of the excellent-comparison architecture

As shown in *Table 8*, we evaluated our model without the excellent-comparison architecture ("*w/o Excellent-Comparison*") and with the excellent-comparison architecture ("*w/ Excellent-Comparison*"). Under the setting of "*w/o Excellent-Comparison*," a multi-head attention was also adopted to fuse the pronunciation features and content features. "*w/ Excellent-Comparison*" achieved a higher top-1 accuracy with a 1.5% improvement compared with the "*w/o Excellent-Comparison*", which probably implies that, with the cue of the excellent response, the network evaluates the response better by comparing the evaluation response with the excellent response to alleviate the large variances of CIP.

Effect of the classification

To show the effectiveness of the classification loss for the CIP rating task, we also adopted a mean square error loss. The results are shown in *Table 9*. "*Mse regression*" denotes that the losses of the rating task and the auxiliary tasks are the mean square error. Since the network was forced to give an integral rating, the network achieved a better classification performance.

Table 9 The effect of the classification

Methods	Mse regression	Classification
Top-1 acc	73.0%	76.1%

Table 10 The effect of the auxiliary learning

λ_{aux}	0	0.01	0.1	1	10
Top-1 acc	73.2%	74.5%	76.1%	73.0%	73.4%

Effect of the auxiliary learning

To show the effect of λ_{aux} in the auxiliary learning, we evaluated the model with different λ_{aux}. As shown in *Table 10*, the top-1 accuracy of λ_{aux} =0.01 or 0.1 is higher than the top-1 accuracy of λ_{aux} =0, which implies that the auxiliary learning of other evaluating tasks helps the model to be aware of more evaluating aspects. However, when λ_{aux} becomes larger (λ_{aux} =1 and 10), since the network prefers to learn the auxiliary tasks than the primary rating task, the performance is hampered.

Conclusions

In this paper, the fast screening of children's DLDs via comprehensive SA evaluation was discussed. A novel CIP, a benchmark CIP dataset, and a novel deep framework specifically designed for CIP evaluation were proposed to address the novel problem. The extensive experiments showed that (I) the proposed CIP indicators and the proposed fast screening model are effective for DLD screening; (II) for screening DLDs via SA evaluation, comprehensive linguistics aspects should be considered, especially the features of pronunciation, fluency, and expression efficiency. Finally, (III) two-stream architecture and the excellent-comparison architecture are beneficial to the automated CIP responses' rating.

Acknowledgments

Funding: This work was supported by the Guangzhou City scientific research Project "Aphasia Rehabilitation Training and Neurolinguistics Study based on Chinese Language" [201904010208]; and Sun Yat-sen University Key Incubation Project for Young Teachers "Multi-model database construction for Children's developmental language disorders study based on big data deep learning" [19wkzd26]. Part of the data were collected at the affiliated kindergarten of Sun Yat-sen University and Zhantianyou Primary School in Guangzhou.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editors (Haotian Lin and Limin Yu) for the series "Medical Artificial Intelligent Research" published in *Annals of Translational Medicine*. The article was sent for external peer review organized by the Guest Editors and the editorial office.

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-19-3097

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-19-3097). The series "Medical Artificial Intelligent Research" was commissioned by the editorial office without any funding or sponsorship. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Fleming C, Whitlock EP, Beil TL, et al. Screening for abdominal aortic aneurysm: a best-evidence systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2005;142:203-11.
2. Guo Fu W. Screening and identification of language development disorders in children. Chinese Journal of Practical Pediatrics 2016;31:748-51.
3. Schum RL. Language screening in the pediatric office setting. Pediatr Clin North Am 2007;54:425-36, v.

4. US Preventive Services Task Force. Screening for speech and language delay in preschool children: recommendation statement. Pediatrics 2006;117:497-501.
5. Bolanos D, Cole RA, Ward WH, et al. Automatic assessment of expressive oral reading. Speech Commun 2013;55:221-36.
6. Proenca J, Lopes C, Tjalve M, et al. Automatic evaluation of reading aloud performance in children. Speech Commun 2017;94:1-14.
7. Loukina A, Zechner K, Chen L, et al. Feature selection for automated speech scoring. Available online: https://www.aclweb.org/anthology/W15-0602.pdf?WT.ac=clk
8. Bernstein J, De Jong J, Pisoni D, et al. Two experiments on automatic scoring of spoken language proficiency. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.488.8087&rep=rep1&type=pdf
9. Witt SM, Young SJ. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Commun 2000;30:95-108.
10. Yoon SY, Loukina A, Lee CM, et al. Word-embedding based content features for automated oral proficiency scoring. Available online: https://www.aclweb.org/anthology/W18-4002/
11. De Wet F, Van der Walt C, Niesler T. Automatic assessment of oral language proficiency and listening comprehension. Speech Commun 2009;51:864-74.
12. Remschmidt H, Belfer M, Goodyer P. Facilitating pathways: Care, treatment and prevention in child and adolescent mental health. Heidelberg: Springer-Verlag Berlin Heidelberg, 2004.
13. Verhoeven L, Balkom H. Classification of developmental language disorders: Theoretical issues and clinical implications. East Sussex: Psychology Press, 2003.
14. Ridder H, Stege H. Early detection of developmental language disorders. In: Verhoeven L, van Balkom H. editor. Classification of Developmental Language Disorders. Mahwah: Psychology Press, 2004:349.
15. Paul R, Norbury C, Gosse C. Language disorders from infancy through adolescence: Assessment & intervention. Maryland Heights, Missouri: Mosby, 2007:324.
16. Bliss LS, Allen DV. Screening kit of language development: a preschool language screening instrument. J Commun Disord 1984;17:133-41.
17. Rescorla L. The language development survey: A screening tool for delayed language in toddlers. J Speech Hear Disord 1989;54:587-99.
18. Wiig EH, Semel E, Secord WA. Clinical Evaluation of Language Fundamentals. Fifth edition. Available online: https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Speech-%26-Language/Clinical-Evaluation-of-Language-Fundamentals-%7C-Fifth-Edition/p/100000705.html?tab=product-details
19. Downey R, Rubin D, Cheng J, et al. Performance of automated scoring for children's oral reading. Available online: https://www.aclweb.org/anthology/W11-1406.pdf
20. Sutherland D, Gillon GT. Assessment of phonological representations in children with speech impairment. Lang Speech Hear Serv Sch 2015. Available online: https://pubs.asha.org/doi/10.1044/0161-1461%282005/030%29
21. Liu Z. A brief statement on implementation outline for Putonghua proficiency test. Applied Linguistics 2004;3.
22. Zechner K, Higgins D, Xi X, et al. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Commun 2009;51:883-95.
23. Lennon P. Investigating fluency in efl: A quantitative approach. Language Learning 1990;40:387-417.
24. Cucchiarini C, Strik H, Boves L. Quantitative assessment of second language learners fluency: Comparisons between read and spontaneous speech. J Acoust Soc Am 2002;111:2862-73.
25. Zechner K, Xi X. Towards automatic scoring of a test of spoken language with heterogeneous task types. Columbus, Ohio: Association for Computational Linguistics, 2008:98-106.
26. Xie S, Evanini K, Zechner K. Exploring content features for automated speech scoring. Montréal, Canada: Montréal, 2012:103-11.
27. McKibbin RC. Serving children from the culture of poverty: practical strategies for speech-language pathologists. ASHA Lead 2002;6:4-17.
28. Bykbaev VR, Lopez-Nores M, Pazos-Arias JJ, et al. Maturation assessment system for speech and language therapy based on multilevel pam and knn. Procedia Technology 2014;16:1265-70.
29. Chen L, Tao J, Ghaffffarzadegan S, et al. End-to-end neural network based automated speech scoring, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE 2018;6234-8.
30. Hans C. Bayesian lasso regression. Biometrika 2009;96:835-45.
31. Hochreiter S, Schmidhuber. Long short-term memory. Neural Comput 1997;9:735-1780.
32. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 1997;45:2673-81.

33. Povey D, Ghoshal A, Boulianne G, et al. The kaldi speech recognition toolkit. Available online: https://publications.idiap.ch/downloads/papers/2012/Povey_ASRU2011_2011.pdf
34. Shivakumar PG, Georgiou P. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. arXiv preprint arXiv:1805.03322, 2018.
35. Semel E, Wiig E, Secord E. Clinical evaluation of language fundamentals preschool, Second. San Antonio, TX: The Psychological Corporation, 2004.
36. Stokes SF, Wong AM, Fletcher P, et al. Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. J Speech Lang Hear Res 2006;49:219-36.
37. Rodekohr RK, Haynes WO. Differentiating dialect from disorder: A comparison of two processing tasks and a standardized language test. J Commun Disord 2001;34:255-72.
38. McCauley RJ. Familiar strangers: Criterion-referenced measures in communication disorders. Lang Speech Hear Serv Sch 1996;27:122-31.
39. Hasson N, Joffe V. The case for dynamic assessment in speech and language therapy. Child Lang Teach Ther 2007;23:9-25.
40. Gaulin CA, Campbell TF. Procedure for assessing verbal working memory in normal school-age children: Some preliminary data. Percept Mot Skills 1994;79:55-64.
41. Shili L. Phonetic thinking of Putonghua proficiency test standard. Language Studies 2001;2:26-8.
42. Issitt S. Improving scores on the IELTS speaking test. ELT Journal 2007;62:131-8.
43. Bridgeman B, Powers D, Stone E, et al. Toefl ibt speaking test scores as indicators of oral communicative language proficiency. Language Testing 2012;29:91-108.
44. Whaley LJ. Introduction to typology: The unity and diversity of language. New York, NY: Sage Publications, 1996.
45. Chambers F. What do we mean by fluency? System 1997;25:535-44.
46. Fillmore CJ, Kempler D, Wang D. Individual differences in language ability and language behavior. Cambridge, Massachusetts: Academic Press, 2014.
47. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Available online: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
48. Zhou S, Dong L, Xu S, et al. Syllable-based sequence-to sequence speech recognition with the transformer in mandarin Chinese. arXiv preprint arXiv:1804.10752, 2018.
49. Wu Z, Xiong Y, Yu S, et al. Unsupervised feature learning via non-parametric instance discrimination. Available online: https://arxiv.org/abs/1805.01978
50. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Available online: https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
51. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from over-fitting. The Journal of Machine Learning Research 2014;15:1929-58.
52. Kingma DP, Ba J. Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 2014.
53. Gardner MW, Dorling S. Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences. Atmos Environ 1998;32:2627-36.

Cite this article as: Zhang X, Qin F, Chen Z, Gao L, Qiu G, Lu S. Fast screening for children's developmental language disorders via comprehensive speech ability evaluation—using a novel deep learning framework. Ann Transl Med 2020;8(11):707. doi: 10.21037/atm-19-3097