PLOS ONE

RESEARCH ARTICLE

Development and evaluation of a 3-D virtual pronunciation tutor for children with autism spectrum disorders

Fei Chen[1,2], **Lan Wang**[1]*, **Gang Peng**[1,2]*, **Nan Yan**[1], **Xiaojie Pan**[3]

1 CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Shenzhen, China, **2** Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region, **3** Shenzhen Aixin Zhihui Rehabilitation Centre for Children with Special Needs, Shenzhen, China

* gpeng@polyu.edu.hk (GP); lan.wang@siat.ac.cn (LW)

OPEN ACCESS

Citation: Chen F, Wang L, Peng G, Yan N, Pan X (2019) Development and evaluation of a 3-D virtual pronunciation tutor for children with autism spectrum disorders. PLoS ONE 14(1): e0210858. https://doi.org/10.1371/journal.pone.0210858

Editor: Simone Sulpizio, Vita-Salute San Raffaele University, ITALY

Received: February 27, 2018

Accepted: January 3, 2019

Published: January 28, 2019

Data Availability Statement: All data files are available from https://osf.io/dz87h/files/.

Funding: This work was partly supported by grants from National Natural Science Foundation of China (NSFC: U1736202, 61771461, 11474300) (http://www.nsfc.gov.cn/), and Shenzhen Fundamental Research Program (JCYJ20160429184226930, JCYJ20170413161611534) (http://www.szsti.gov.cn/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The deficit in speech sound production in some children with autism spectrum disorder (ASD) adds to their communication barriers. The 3-D virtual environments have been implemented to improve their communication abilities. However, there were no previous studies on the use of a 3-D virtual pronunciation tutor designed specifically to train pronunciation for children with ASD. To fill this research gap, the current study developed and evaluated a 3-D virtual tutor which served as a multimodal and real-data-driven speech production tutor to present both places and manners of Mandarin articulation. Using an eye-tracking technique (RED 5 Eye Tracker), Experiment 1 objectively investigated children's gauged attention distribution online while learning with our computer-assisted 3-D virtual tutor in comparison to a real human face (HF) tutor. Eye-tracking results indicated most participants showed more interests in the visual speech cues of the 3-D tutor, and paid some degree of absolute attention to the additional visual speech information of both articulatory movements and airflow changes. To further compare treatment outcomes, training performance was evaluated in Experiment 2 with the ASD learners divided into two groups, with one group learning from the HF tutor and the other from the 3-D tutor (HF group vs. 3-D group). Both groups showed improvement with the help of computer-based training in the post-intervention test based on the calculation of a 5-point Likert scale. However, the 3-D group showed much higher gains in producing Mandarin stop and affricate consonants, and apical vowels. We conclude that our 3-D virtual imitation intervention system provides an effective approach of audiovisual pronunciation training for children with ASD.

Introduction

Given the high incidence of autism spectrum disorder (ASD), at close to 1% worldwide [1], there is an urgent need to improve our understanding of ASD and to refine our treatment strategies. Besides the social communication deficits and repetitive sensory-motor behaviors,

Competing interests: The authors have declared that no competing interests exist.

speech and language disorders tend to be a hallmark of ASD. Among the communicative characteristics of children in the second and third year of life who are identified with ASD, delayed onset and development of the spoken language tends to be one of the key signs and symptoms. Approximately 10–25% of all children with ASD fail to develop speech to communicate with others [2]. Due to the known deficits that characterize ASD in the theory of mind [3], researchers have identified that pragmatic skill is the most seriously impaired in terms of language deficits in ASD [4,5]. In contrast, less attention has been paid to the articulatory and phonological deficits among this population.

Articulation skills were sometimes reported to be relatively intact, or even better, when compared to other speech and language behaviors in individuals with ASD [6–8]. However, in these studies, articulation tests were only judged as correct or incorrect, giving no information on the nature of the pronunciation errors. Other studies, using more detailed phonological analyses, have detected more severe speech sound impairment in individuals with ASD [9–11]. Recently, one study has reported that approximately 41% of school-aged participants with ASD in their research produced some pronunciation errors [12]. To conclude, although there remains some disagreement in the literature on the status of phonological development in an ASD cohort, relevant research overall indicates that at least a subgroup of children with ASD may exhibit speech sound production difficulties. Moreover, several studies [13–16] have found a strong relationship between the severity of pronunciation difficulty and the severity of overall language impairment, suggesting that some children with ASD who present with more severe global language impairment may also exhibit more severe speech production difficulties.

Appropriate communication depends on the use of a vast array of language skills (e.g., phonology, lexicon, grammar, semantics, and pragmatics), with phonological decoding and encoding acting as the initial and final operations of the input-processing-output language system. Articulatory and phonological difficulties can affect speech intelligibility to some extent, and they represent an additional social and communication barrier for people with ASD. One study [12] highlighted the contribution of speech sound errors to communication barriers in ASD, even when only a few pronunciation errors occurred. Furthermore, some studies [11,12] have emphasized that there was no correlation between chronological age and number of speech errors, suggesting that speech distortions did not appear to resolve themselves over time in individuals with ASD. Thus, unlike typically developing (TD) children who refine their speech production skill through accumulated language experience with age [17], pronunciation difficulties for some children with ASD may continue into adulthood if there is no effective intervention. Consequently, speech production difficulties should be treated early in children with ASD, and the inclusion of the phonological component in treatment should be applied [15,18].

However, there have been few empirical reports so far on treatment strategies specifically targeted at enhancing speech production skill among the ASD population. There is an overall dearth of knowledge concerning articulatory interventions among children with ASD. A review article [19] found a very limited number of articles concerning ASD and phonological interventions. For example, one study [20] emphasized stimulating learning motivation in children with ASD. Later, other studies on the treatment of speech sound production focused mainly on improving teaching strategies of therapist [2,21,22]. These studies on phonological interventions were conducted among a relatively small sample size, with single case studies [21,22] or in three ASD children [2]. Since there were no controls for individual differences and other variables, it is difficult to make generalizations from these studies. Recently, a novel intonation-based intervention called Auditory-Motor Mapping Training has been developed to facilitate speech output in non-verbal children with ASD [23]. After therapy, all the six non-

verbal children showed significant improvements in their ability to articulate several word approximations. Although this training approach showed great success in promoting the production of syllable approximations in ASD children who are non-verbal, it did not focus on improving fine-grained production of speech sound elements.

Recently, several novel approaches have been proposed to help ameliorate communication deficits in children with ASD. Previous studies on the role of the mirror neuron system (MNS) and the observed links between the MNS and imitation suggested that mirror system based imitation learning could be used as an efficacious treatment strategy for children with ASD [24,25]. Moreover, another highly promising technique was to present the 3-D virtual environments to the teaching-learning process for children with ASD [26]. Virtual reality (VR), i.e. a simulation of the real world based on computer graphics, has recently emerged in many domains of rehabilitation in children with ASD, and this technique can be useful since it allows teachers or therapists to present a repeatable and diversifiable environment during learning in a very similar context to the real ones in the absence of potential risks [27]. With the progress of speech technology, various 3-D virtual tutors have been utilized as new ways to act as one of the non-immersive desktop VRs and directly show learners with pronunciation animations for imitation learning. Several reasons explain why the use of auditory and visual information from a 3-D tutor is so successful, and why it holds so much promise for pronunciation tutoring [28]. Both internal and external articulatory movements have been demonstrated in 3-D virtual tutors to successfully guide the pronunciation training for hearing-loss children [29–31] and second language learners [32,33]. Recently, a computer-animated talking head has been employed to train and develop vocabulary and grammar knowledge for children with ASD [34,35]. However, to the best of our knowledge, until now, there have been no studies on the use of a 3-D virtual pronunciation tutor designed specifically to enhance speech production skills among the ASD population.

The underlying sources of speech delays/disorders in children with ASD have important implications for the rationale of using a 3-D virtual pronunciation tutor for speech therapy in ASD. Many reasons may account for the speech sound production difficulties experienced by children with ASD. As indicated by the 'social theory' [3], some studies pointed out that failure to attend to the ambient social language environment during daily communication negatively impacts the ability to acquire spoken language in children with ASD [11,36,37]. Probably, the utilization of a 3-D virtual tutor presented on a computer screen has the potential to reduce the social learning burden for ASD learners by avoiding communication with a real human teacher. Furthermore, it is important to note that our 3-D virtual tutor, compared to real human face tutor, contains additional visual information with a profile view (such as internal articulator and airflow animations). Given that speech perception and production are presumed to be closely correlated constructs, impaired or atypical speech sound perception in the auditory modality [38,39] may be another factor leading to speech production deficits in children with ASD. Also, the traditional pronunciation training approach mainly with the help of auditory inputs may not be efficacious enough to enhance speech production in ASD learners. Hopefully, the additional visual speech cues in our 3-D virtual tutor, if noticed by ASD learners, may offer supplementary production guidance for intended imitation learning from the visual modality.

The overall goal of this study is to utilize recent advances in speech technology to develop and evaluate a 3-D virtual tutor for pronunciation training in children with ASD. We first introduced some of our previous studies on the development of a 3-D virtual tutor showing Mandarin articulatory and aspiratory animations [29,40–42]. This speech production tutor has been developed to produce articulatory movements in accordance with airflow motions when uttering Mandarin syllables. We then evaluated this 3-D virtual pronunciation tutor

with an eye-tracking study (Experiment 1) and a pronunciation training study (Experiment 2), to ascertain whether the 3-D virtual tutor is effective in improving the level of the ASD learners' interest and enhancing the accuracy of their pronunciation of Mandarin consonants and vowels.

For Experiment 1, the objective evaluation of our 3-D pronunciation tutor is the first focus of this study since a virtual tutor system involved in any application must first undergo evaluation. The uncanny valley effect [43] suggested that the learners' familiarity with language training systems with the same underlying speech model could differ significantly. The transparent face and the cooperative motions of various internal articulators and airflow in our 3-D tutor are not commonly seen in the children's daily lives. It is unclear whether children with ASD would pay attention to these uncommon visual contents. The eye-tracking approach allows for objective and quantitative observation of attention, and through the analysis of fixation patterns, can indicate which information from a tutor is available to the learner [44]. Because it is non-invasive and does not require advanced motor responses or language, eye tracking is particularly suitable for conducting studies in young children with ASD. In Experiment 1, by using an eye-tracking methodology, our 3-D virtual pronunciation tutor was evaluated in comparison to real human face videos. The pronunciation tutors were shown through a computer screen under two conditions: real human face tutor (HF tutor) and 3-D pronunciation tutor (3-D tutor), each with a front view first and a profile view afterwards. By analyzing eye-tracking related measures, three research questions (RQs) were put forward and investigated:

RQ1: When presented with visual speech cues, which tutor is more attractive to children with ASD, HF tutor or 3-D tutor?

RQ2: For the 3-D tutor with a profile view, did children with ASD pay attention to the uncommon visual contents containing internal articulator and airflow motions?

RQ3: Compared with age-matched TD children, did children with ASD show a similar attention pattern while watching the pronunciation tutors?

It is hypothesized that, due to their attention deficits [45], children with ASD might show a much more scattered attention pattern compared with TD children. Moreover, since children with ASD tended to show a processing deficit towards real human faces [46], they might exhibit more interests in the visual speech cues of 3-D virtual tutor, and pay some degree of visual attention to the uncommon visual contents (internal articulator and airflow animations) in our 3-D tutor with a profile view.

For Experiment 2, in order to further evaluate the efficacy of our 3-D virtual tutor as a pronunciation training tool for the acquisition of Mandarin speech, a pronunciation training study was thus conducted to compare pronunciation performance between two subgroups of children with ASD, learning from HF tutors or 3-D pronunciation tutors respectively (HF group vs. 3-D group). As our 3-D virtual tutor contains real-data-driven visual speech cues which offer realistic pronunciation guidance from the visual modality, it is hypothesized that the 3-D tutor might be more effective in facilitating the imitation learning of Mandarin consonants and vowels in ASD learners.

Materials and methods

Development of a 3-D virtual pronunciation tutor

The visual speech cues in a 3-D virtual tutor may offer supplementary production guidance from the visual modality. Our 3-D virtual tutor was presented as an entire-head virtual tutor on a computer screen, showing models of the face, lips, tongue, jaw, and nasopharyngeal wall

(based on MRI data). Moreover, the internal articulator (i.e., tongue) and airflow changes in our 3-D tutor with a profile view were animated based on physiological signals during pronunciation, in order to generate a more realistic pronunciation tutor. In this way, the 'audio-visual' MNS [47] may be activated to facilitate intended imitation learning from the internal articulatory model and airflow model. Furthermore, since our 3-D virtual pronunciation tutor targeted especially at enhancing speech sounds (vowels and consonants) production in Mandarin-speaking children with ASD, the phonology system in Mandarin should also be considered. Both the internal articulatory model and airflow model were concurrently implemented in our 3-D virtual tutor with a profile view to offer realistic visual speech cues (i.e., viseme, the visual equivalent of a phoneme or unit of sound in spoken language) to ASD learners. During the training program, the additional visual information in our 3-D virtual imitation intervention system was exhibited to ASD learners for imitation learning of pronunciation.

First, to realize the articulatory model, the external and internal articulatory movements of each syllable were collected and recorded by an Electro-Magnetic Articulography (EMA AG500). The 3-D articulation data were recorded from a Mandarin-speaking female language teacher. For instruction purposes, the natural audio speech is recorded from the same native female language teacher, rather than using synthesized speech. The articulatory trajectories of 13 feature points included: Five facial feature points for calibration (nose, left head, right head, right jaw, and left jaw), four external feature points (right lip corner, and left lip corner, upper lip, lower lip) and another four internal feature points (tongue tip, middle tongue, tongue root, and middle jaw), which were normalized and smoothed. Then the static 3-D head model with tongue, uvula, and jaw was constructed based on anatomy that has great irregularity and a large number of vertices. Thus, a flexible graphic algorithm named Dirichlet free form deformation was adopted to compute the control parameters with Sibson coordinates for articulatory modeling work [29]. The EMA-based displacements were then added to the articulatory model, in order to drive multiple articulators to move smoothly and simultaneously over time. The articulatory model presented by this 3-D virtual tutor could depict realistic Mandarin articulation, giving an indication of the way of consonants and vowels being pronounced.

Second, many minimal pairs of Mandarin stops and affricates can be easily confused and are differentiated by the distinctive phonetic feature of unaspirated vs. aspirated contrast (e.g., b [p] vs. p [p']). These minimal pairs share the same place of articulation and similar trajectory of articulatory movement. To realize the visual information concerning the manner of articulation, exhaled airflow, one of the important bio-signals during speech production, was then collected and visualized to improve the identification rate of Mandarin consonants. Our recent study [41] collected airflow information with a *Phonatory Aerodynamic System* (PAS) and proposed an airflow model using a physical equation of the fluid flow. At the syllable-level animations, the new airflow model was then simultaneously combined with the 3-D articulatory model presented in [29], and a new multimodal animation system was constructed.

The overall implementation procedure of our 3-D pronunciation tutor is shown in Fig 1. This multimodal speech production tutor can present dynamic animations with both front and profile views. As shown in Fig 2, the 3-D virtual pronunciation tutor can show lip movements with a front view, and additionally, exhibit internal articulator (tongue) movements and airflow motions with a transparent profile view.

Evaluation: Eye-tracking study (Experiment 1)

Participants. The participants included ten low-functioning children with ASD (eight boys) as an experimental group and 13 TD children (nine boys) as a control group. The

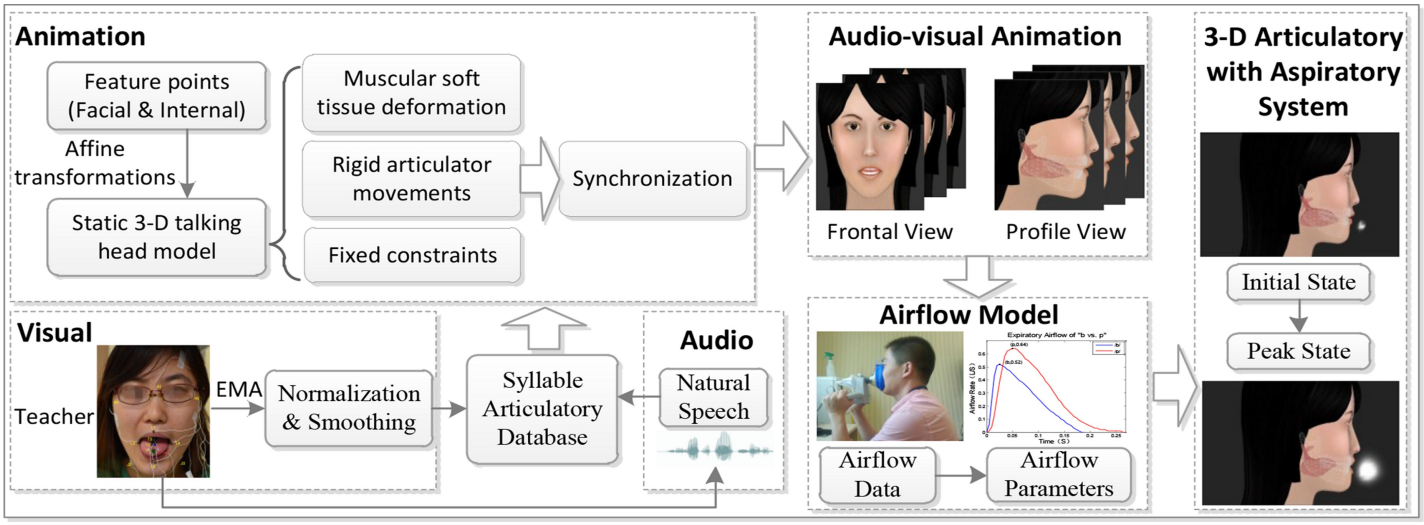


Fig 1. Implementation procedure for the 3-D virtual pronunciation tutor.

https://doi.org/10.1371/journal.pone.0210858.g001

average chronological ages of the two groups (see Table 1) were similar (t (21) = 1.07, p = 0.30, Cohen's d = 0.42). All 23 child participants were native Mandarin speakers and had normal or corrected-to-normal visual acuity and normal hearing. Approval of the research was granted by Behavioral Research Ethics Committee of the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, and a written consent form was obtained from each child's parent. The individual in this manuscript has given written informed consent (as outlined in PLOS consent form) to publish these case details.

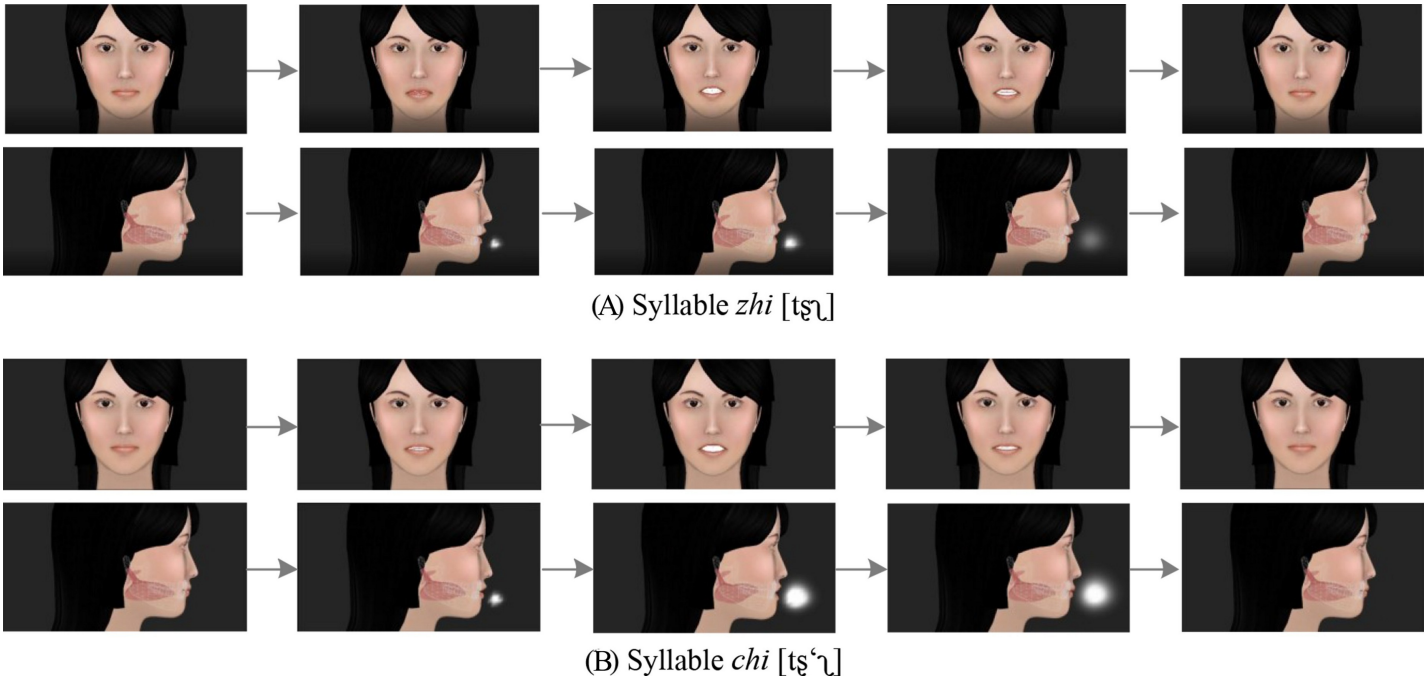


Fig 2. The 3-D articulatory with aspiratory animations for the sequence of two mandarin syllables. (A) zhi [tʂʅ], (B) chi [tʂʻʅ].

https://doi.org/10.1371/journal.pone.0210858.g002

Table 1. Chronological ages and developmental ages in child participants in Experiment 1.

Group	ASD (n = 10)		TD (Control) (n = 13)	
	Mean	SD	Mean	SD
Chronological Ages (Range, in years)	6.63	0.85	6.33	0.46
	(5.23–7.91)		(5.54–6.93)	
Developmental Ages (Range, in years)	3.15	0.71	N.A.	
	(2.24–4.83)		N.A.	

https://doi.org/10.1371/journal.pone.0210858.t001

The clinical diagnosis of autism was established according to the DSM-V criteria for ASD [48], and further confirmed using the Chinese versions of the Gilliam Autism Rating Scale–Second Edition (GARS-2) [49] or the Childhood Autism Rating Scale (CARS) [50] by pediatricians and child psychiatrists with expertise in diagnosing ASD. The ASD participants in Experiment 1 showed different levels of language delay and speech production deficit based on speech-language pathology and parental reports, while none of them were non-verbal. Moreover, the *Psychoeducational Profile-Third Edition* (*PEP-3*) evaluation [51] showed that the average developmental age of expressive language and receptive language in ASD subjects was around 3.15 yr (SD = 0.71).

Apparatus. Eye movement data for each participant were collected with a *RED 5* Eye Tracker (SMI Technology, Germany) non-intrusively, which was integrated into a panel of 22-inch TFT monitor with a resolution of 1280×1024 pixels. The sampling rate was set to be 60 Hz and an accuracy of 0.4˚. The freedom of head movement was 40 cm × 20 cm at 70 cm distance. Eye-tracking data were recorded online with the software of *Experiment Center 2.0* and analyzed offline with *SMI BeGaze* analysis software.

Testing materials. Training materials were comprised of 16 commonly used Mandarin syllables superimposed with high-level tone in Mandarin. They were combined with eight basic Mandarin monophthongs and 12 easily confused stops and affricates in terms of Pinyin (International Phonetic Alphabet (IPA) in the square brackets): bo ([po]), po ([pʻo]), bu ([pu]), pu ([pʻu]), de ([tɤ]), te ([tʻɤ]), ga ([ka]), ka ([kʻa]), ju ([tɕy]), qu ([tɕʻy]), ji ([tɕi]), qi ([tɕʻi]), zi ([tsɿ]), ci ([tsʻɿ]), zhi ([tʂ]), and chi ([tʂʻ]). All these 16 syllables were played under two presentation conditions (HF and 3-D tutors), totaling 32 videos, each containing a front view first and then a corresponding profile view. Videos in HF condition were time-aligned with those in 3-D condition with a duration of six seconds (three seconds for each view). The audio for each syllable was recorded from the same female speaker for both HF and 3-D tutors, and the volume was fixed at 75 dB SPL.

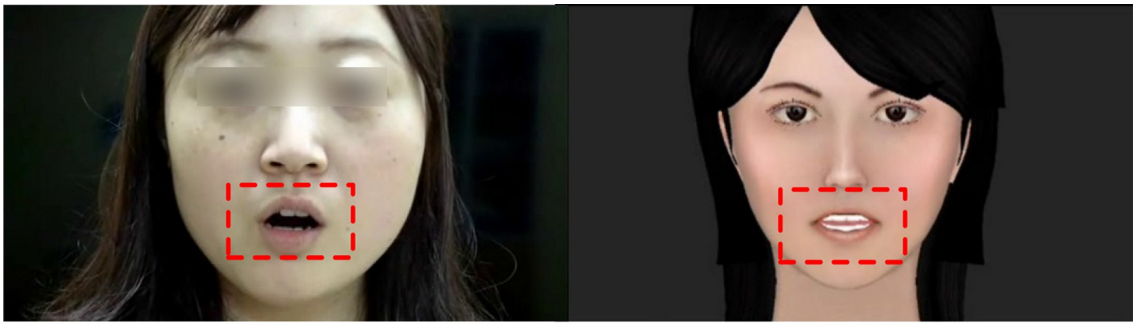
Procedure. The procedure of Experiment 1 was shown in Fig 3. Firstly, a familiarization stage was undertaken to guarantee that all the child participants could follow the instructions. The experimenter taught them to place their chins above a fixed support frame, and to watch videos of two Mandarin syllables (ge [kɤ] and ke [kʻɤ], excluded from the testing syllables)

Fig 3. Procedure of the eye-tracking study in Experiment 1.

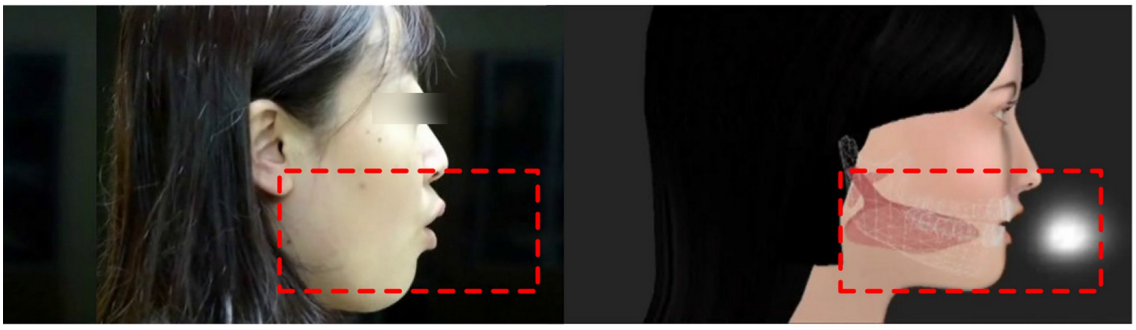
https://doi.org/10.1371/journal.pone.0210858.g003

with two presentation conditions (i.e., HF and 3-D). While watching videos, subjects were asked to concentrate on what they heard and saw and imitate the pronunciation. However, the child participants in Experiment 1 were not told about the underlying meaning of visual contents to avoid any attention bias. Then, eye movement data were calibrated with five fixation points and recalibration was required if calibration results were poor or missing. Afterward, in the formal testing stage, the 32 testing videos were randomly repeated twice (totaling 64 videos) over four testing sessions.

Parameters of eye-tracking data. With the front view (Fig 4A), the specified area containing lip movement was regarded as the area of interest (AOI). With the profile view (Fig 4B), the specified area in the 3-D video, including additional internal articulatory movements and airflow information, was regarded as the AOI. In the current study, three eye-tracking parameters were calculated during the learning process [52]. The first parameter was 'entry time' (ET), meaning the duration from the start of the trial to the first hit on the AOI. The shorter the ET, the more interest was shown for the AOI. The second parameter was 'fixation count' (FC), defined as the total number of fixations lasting more than 100 milliseconds (ms) inside the AOI, which reflects the absolute attention to the AOI during the learning process. The last parameter was 'proportion of fixation duration' (POFD), indicating the ratio of fixation duration inside the AOI to the duration of the whole video, reflecting relative attention to the AOI.



(A) Front View



(B) Profile View

Fig 4. The AOIs with a Front View (A) and a Profile View (B). They corresponded to the homolographic areas inside the dashed red rectangle, which were closely related to speech sound production (with syllable po [p'o] as an example). The AOI with a front view mainly incorporates lip movement area, and the AOI with a transparent profile view in the 3-D tutor contains the movement of external and internal articulators (including mouth, tongue, teeth, jaw, and palate) and airflow change.

https://doi.org/10.1371/journal.pone.0210858.g004

Evaluation: Pronunciation training study (Experiment 2)

Participants. To avoid any practice effect from the first eye-tracking study when children were exposed to both HF tutor and 3-D tutor, a different 28 Mandarin-speaking children diagnosed with low-functioning ASD (chronological age range: 3.33–6.90 yr) were recruited to participate in the pronunciation training study. These ASD children were diagnosed with the same diagnostic approach as that used in Experiment 1. They were recruited in the current study since they all showed severe language delay and speech sound disorders based on both clinical observation and *The Psychoeducational Profile-Third Edition* (*PEP-3*) evaluation. Moreover, exclusion criteria included a history of hearing loss and exposure to more than one language in the child's home. Not all the 28 children with ASD completed all tests and training. One child left the rehabilitation center during data collection, another three children were reluctant to cooperate with experimenters during the pre-test or post-test, and two children were dropped during the training period because they were not engaged by the computer screen and did not pay attention to the videos. These six child participants with ASD were thus excluded from any further analysis in the current study. Approval of the research was granted by Behavioral Research Ethics Committee of the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, and a written consent form was obtained from each child's parent. The individual in this manuscript has given written informed consent (as outlined in PLOS consent form) to publish these case details.

In order to control for other possible factors that might influence training performance, all children with ASD completed the *PEP-3* evaluation, designed to assess the development of communication, motor skills, and presence of maladaptive behaviors in children with ASD [51]. The *PEP-3* was demonstrated to efficiently measure skills related to learning and to capture the uneven and idiosyncratic development of skills commonly found in children with ASD [53–55]. The *PEP-3* contains current normative data from a large representative sample of 407 individuals with ASD and a comparison group of 148 TD children. Moreover, the total raw score for all test items is converted into developmental ages (based on a TD sample). The related *PEP-3* evaluation results of developmental ages (including cognitive verbal/preverbal, expressive language, receptive language, fine motor, gross motor, and visual-motor imitation) in this study are presented in Table 2.

Testing materials and instrument. In both pre-intervention and post-intervention tests, all the child subjects were asked to imitate the audios containing the same 16 Mandarin syllables used in Experiment 1. The entire imitation task was recorded by Cool Edit software (22050 Hz sampling rate, 16-bit resolution) in a quiet aural rehabilitation room, and any vocalizations that occurred simultaneously with any other sounds on the recording were abandoned. Each syllable was imitated twice, with a total of 32 speech samples from each subject.

Table 2. The Means (standard deviations) of chronological ages, pre-test scores, and developmental ages of *PEP-3* tests in Experiment 2.

	CA (in years)	Pre-test Scores (1–5)		DA of *PEP*-3 tests (in year)					
		Consonants	Vowels	EL	RL	CVP	FM	GM	VMI
HF Group	4.87 (1.15)	3.42 (0.58)	3.63 (0.45)	2.23 (0.74)	2.30 (0.91)	2.46 (0.88)	2.33 (0.78)	2.24 (0.81)	2.30 (0.68)
3-D Group	4.81 (0.87)	3.39 (0.63)	3.61 (0.56)	2.33 (0.93)	2.30 (0.83)	2.30 (0.76)	2.51 (0.82)	2.38 (0.72)	2.27 (0.61)
t-value	0.13	0.13	0.09	-0.30	0.01	0.45	-0.54	-0.42	0.08
p-value	0.90	0.90	0.93	0.77	0.99	0.66	0.60	0.68	0.94

CA, chronological age; DA, developmental age; PEP-3, psychoeducational profile-third edition; EL, expressive language; RL, receptive language; CVP, cognitive verbal/preverbal; FM, fine motor; GM, gross motor; VMI, visual-motor imitation.

https://doi.org/10.1371/journal.pone.0210858.t002

For the audio imitation method, due to the presence of a speech model, the child's true speech production abilities may have been overestimated. Nevertheless, in this special population, where speech output may be limited due to delayed language development, or a lack of desire to speak spontaneously, imitation may be the only option [18]. Moreover, one study [56] have indicated that the spontaneous and imitated production of speech in an ASD cohort did not differ in terms of sound class, and production errors tended to share both place and manner of articulation. Moreover, during pronunciation training, the computer-assisted 3-D or HF tutors were played through the software of an interactive speech training system [29].

Procedure and design. Fig 5 illustrates the procedure of Experiment 2. The performance comparison between pre-intervention test and post-intervention test was widely adopted to evaluate training outcomes in various pronunciation training studies [57,58]. Firstly, based on the average chronological ages, pre-test scores of consonants and vowels, and developmental ages of different *PEP-3* subtests, all the 22 children with ASD were equally divided into two subgroups (3-D group and HF group). Independent-samples *T* tests indicated that the two subgroups did not differ significantly in the above measurements (see Table 2). The 3-D group is the experimental group, while the HF group belongs to the control group. Each subgroup contained 11 subjects (one girl in each subgroup), and they were further asked to take part in a pronunciation training program which contains three sessions in total (see Fig 5). Within each session, the HF group learned the 16 syllables four times per session (three-day intervals between two consecutive sessions) from the HF tutors, while the 3-D group learned from the 3-D tutors with the same amounts of training. All the pronunciation tutors were presented as a front view first and then as a corresponding profile view, similar to that presented in Experiment 1. Each session lasted about an hour for each participant. During the training period (three repetitive sessions in total), learners were trained one by one in a quiet rehabilitation room, and they were asked to watch the videos in front of a computer screen and try to imitate pronunciation from the tutors afterwards. Besides, the ASD learners in Experiment 2 were taught the underlying meaning of various visual contents in tutors. To rule out a possible confounding effect of the daily routine training in the rehabilitation center, all the ASD learners' instructors and speech therapists agreed not to additionally teach pronunciation during the training period. After the third training session, both subgroups were asked to conduct a post-intervention test with a similar audio imitation task to that in the pre-test (see Fig 5).

Scoring and inter-rater reliability. In this study, we collected and evaluated pre-test and post-test recordings to measure pronunciation performance. All the recorded speech samples were firstly rated by an expert majoring in linguistics to pick out the better trial from two iterations of each syllable. All the chosen 16 syllable production samples were then rated by another five Mandarin-speaking experts majoring in linguistics. Since all the tested Mandarin stops and affricates were voiceless, it is not appropriate to split the consonants and vowels and rate them separately. Raters were instructed to separately score each consonant and vowel production based on the perception of the heard phonemes and to minimize the mutual influence in

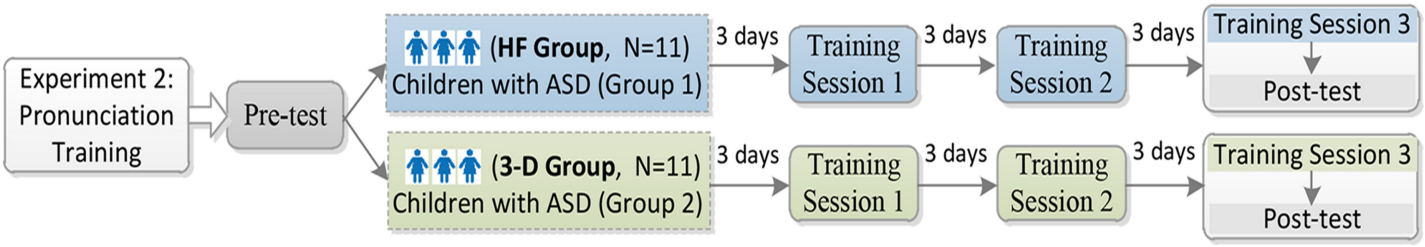


Fig 5. Procedure of the pronunciation training study in Experiment 2.

https://doi.org/10.1371/journal.pone.0210858.g005

their judgment of consonant or vowel production, even though they occurred next to each other within a carrying syllable. The 5-point Likert scale, with "1" (completely incorrect) and "5" (completely correct), was used to rate the quality of each phoneme production [58,59]. Different from previous studies with subjective judgements, the language experts in this study rated the speech tokens with a fine-grained criterion based on phonetic feature analyses. For example, a consonant b [p] was rated as "5" when the sound that was produced contained all the four phonetic features: voiceless, bilabial, unaspirated, and stop, while rated as "1" when none of the above four phonetic features were met. Moreover, to minimize transcribing bias, there was no identifiable information for each child (e.g., group, name, chronological age, gender) on the scoring sheets. To further minimize the experimental bias, raters did not know the experimental design and were blind as to which production data (pre-test or post-test) they were evaluating. Scoring for the 12 consonants and eight vowels from the 16 syllables was calculated and averaged for each ASD subject.

Inter-rater reliability was derived using SPSS (v.22.0). Kendall's Concordance Coefficient W for the inter-rater agreement was calculated for scores derived from five raters. The inter-rater reliability with Kendall's coefficients of 0.801 (consonant production in the pre-test), 0.896 (vowel production in the pre-test), 0.815 (consonant production in the post-test), 0.836 (vowel production in the post-test) were respectively reached for the scores given by the five experts, with all representing high inter-rater reliability.

Results

Experiment 1: Eye-tracking study

Entry time (ET). All the eye-tracking data were analyzed in SPSS. Firstly, a three-way repeated-measure analysis of variance (ANOVA), with Greenhouse-Geisser corrections when appropriate, was conducted on the ET with *view* (front and profile) and *presentation condition* (HF and 3-D) as two within-subject factors, and *group* (ASD and TD) as a between-subject factor. The analysis revealed significant main effects for *presentation condition* ($F(1, 21) = 7.28$, $p < 0.01$, $\eta_p^2 = 0.26$), and *group* ($F(1, 21) = 13.39$, $p < 0.001$, $\eta_p^2 = 0.39$). The main effect of *view* was not statistically significant ($F(1, 21) = 0.47$, $p = 0.50$, $\eta_p^2 = 0.02$), and no significant two- or three-way interactions were observed (all $ps > 0.05$). The results indicated that all child subjects showed a shorter ET for the 3-D tutors with both front and profile views, and the TD children exhibited a much shorter ET than the children with ASD (see the left column in Fig 6).

Fixation count (FC). Secondly, to examine the effect of *view* (front and profile), *presentation condition* (HF and 3-D), and *group* (ASD and TD) on the FC, a three-way ANOVA with repeated measures was conducted. Significant *view*×*presentation condition* interaction, $F(1, 21) = 49.37$, $p < 0.001$, $\eta_p^2 = 0.70$, and *view*×*group* interaction, $F(1, 21) = 13.48$, $p < 0.01$, $\eta_p^2 = 0.39$, as well as significant main effects of all three factors (all $ps < 0.01$), were observed. There was also a significant *view*×*presentation condition*×*group* interaction, $F(1, 21) = 9.28$, $p < 0.01$, $\eta_p^2 = 0.31$. This was followed up with a two-way analysis of factors *(presentation condition*, and *group*) with both front and profile views.

With a front view, the analysis confirmed a significant main effect for *group* on FC ($F(1, 21) = 5.69$, $p < 0.05$, $\eta_p^2 = 0.21$), while no main effect was detected for *presentation condition* on FC, such that similar absolute attention was paid to the AOI between HF and 3-D videos with a front view ($F(1, 21) = 1.24$, $p = 0.28$, $\eta_p^2 = 0.06$). There was no significant interaction between *presentation condition* and *group* ($F(1, 21) = 3.94$, $p = 0.07$, $\eta_p^2 = 0.16$) (see the middle column in Fig 6A). With a profile view, there were significant main effects for both *group* ($F(1, 21) = 22.33$, $p < 0.001$, $\eta_p^2 = 0.52$) and *presentation condition* on FC ($F(1, 21) = 74.09$, $p < 0.001$, $\eta_p^2 = 0.78$), while there was no significant interaction between *presentation condition* and *group* ($F(1, 21) =$

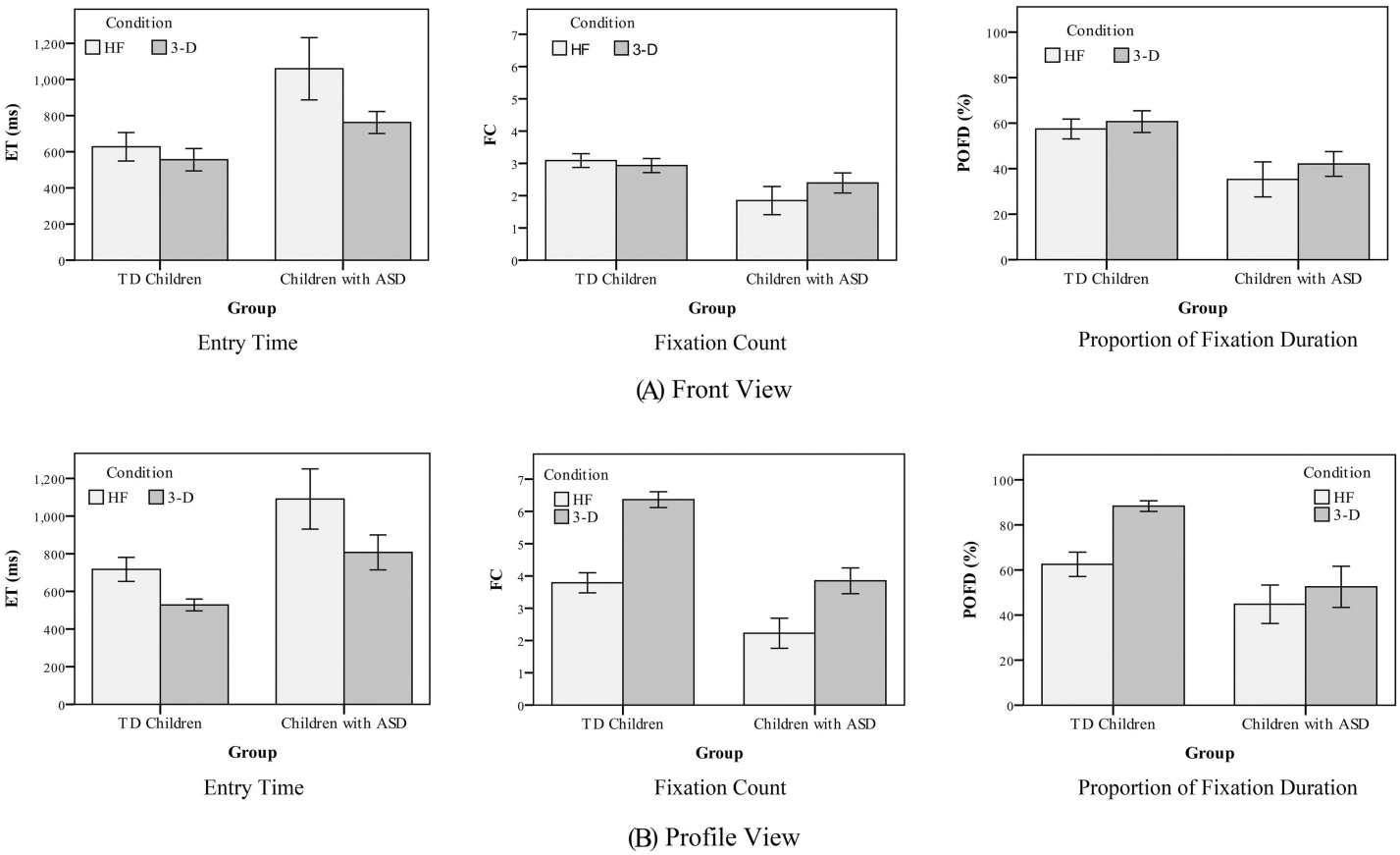


Fig 6. The results of three eye-tracking parameters. Average entry time (left), fixation count (middle), and proportion of fixation duration (right) of AOIs with a front view (A) and a profile view (B) in Experiment 1 (Error bars: +/- 1 SE).

https://doi.org/10.1371/journal.pone.0210858.g006

3.80, $p = 0.07$, $\eta_p^2 = 0.15$). The results indicated that both TD and ASD subjects paid more absolute attention to the AOI of 3-D videos with a profile view (see the middle column in Fig 6B). Furthermore, to better answer RQ2, the whole AOI in a transparent 3-D profile view was further divided into two subareas: Subarea 1 mainly contained an internal articulator showing places of articulation, and Subarea 2 mainly contained expiratory airflow showing manners of articulation. The parameter of FC was calculated to indicate the distribution of absolute attention paid to the two subareas in ASD learners. Results indicated that the average FC for Subarea 1 was approximately 2.20, and around 1.65 for Subarea 2 in children with ASD.

Proportion of fixation duration (POFD). Thirdly, the eye-tracking parameter of POFD was submitted to a three-way ANOVA. The analysis revealed a significant *view*×*presentation condition*×*group* interaction, $F(1, 21) = 5.14$, $p < 0.05$, $\eta_p^2 = 0.20$. There were also significant main effects of *view*, *presentation condition*, and *group* (all *p*s < 0.01), as well as a significant *view*×*presentation condition* interaction ($F(1, 21) = 6.06$, $p < 0.05$, $\eta_p^2 = 0.22$). The significant three-way interaction indicated that the effects of *view* on the POFD were modulated by both *presentation condition* and *group*. This was followed up with a two-way analysis of two independent factors (*presentation condition* and *group*) with both front and profile views.

With a front view, the specified area containing lip movement was regarded as AOI, the main effect of *group* on POFD was significant ($F(1, 21) = 7.78$, $p < 0.05$, $\eta_p^2 = 0.27$), but the main effect of *presentation condition* on POFD was not significant ($F(1, 21) = 3.27$, $p = 0.09$, $\eta_p^2 = 0.14$). No significant interaction between *presentation condition* and *group* was found (*F*

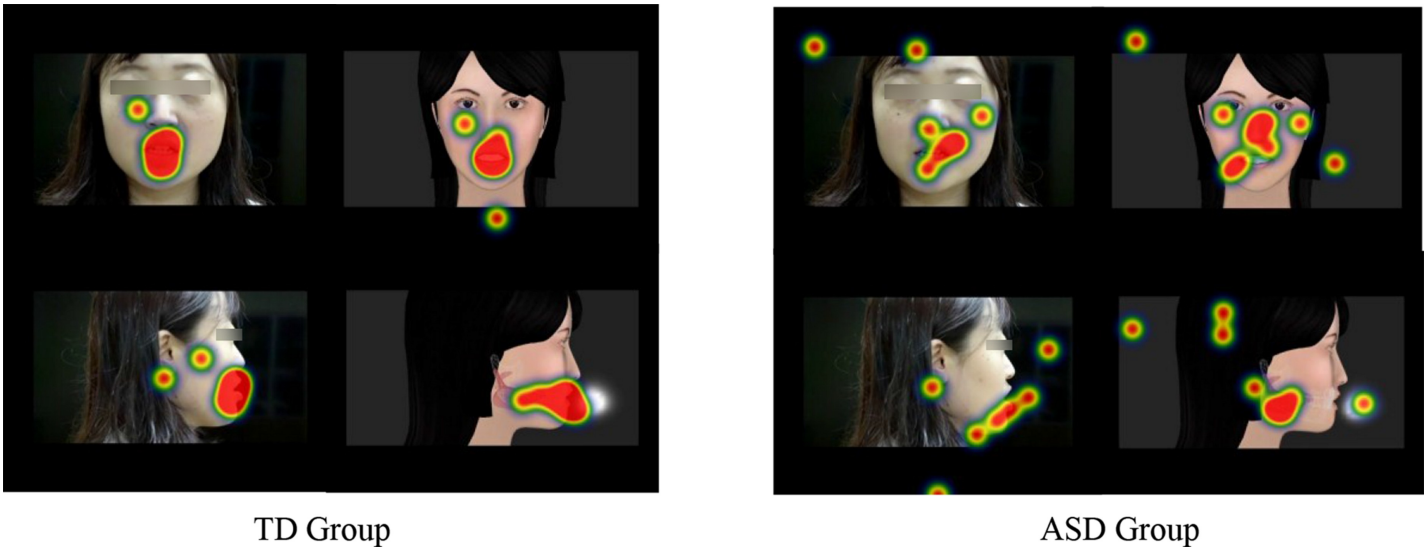


Fig 7. Heat maps of one testing syllable in Experiment 1.

https://doi.org/10.1371/journal.pone.0210858.g007

(1, 21) = 0.41, $p = 0.53$, $\eta_p^2 = 0.02$). This suggests that similar POFD (i.e., relative attention) was paid to the AOI of HF and 3-D videos with a front view (see the right column in Fig 6A). With a profile view, significant main effects of *presentation condition* (F (1, 21) = 16.82, $p = 0.001$, $\eta_p^2 = 0.45$) and *group* on the POFD (F (1, 21) = 10.69, $p < 0.01$, $\eta_p^2 = 0.34$) were found. Moreover, there was a significant two-way interaction between *presentation condition* and *group*, $F(1, 21) = 4.90$, $p < 0.05$, $\eta_p^2 = 0.19$. After this, simple main effect analyses of the *presentation condition* were performed with Bonferroni adjustment. For TD children, they showed a higher POFD towards 3-D videos with a profile view ($F(1, 21) = 22.93$, $p < 0.001$). However, for children with ASD, the *presentation condition* had no effect ($F(1, 21) = 1.58$, $p = 0.22$) (see the right column in Fig 6B).

Exploring the pattern of eye gaze in children is important to understand their attention distribution during learning. In Experiment 1, using the eye-tracking methodology, an objective evaluation comparing 3-D and HF tutors was conducted. RQ1 asked whether ASD learners showed interest in the AOIs in 3-D pronunciation tutor. Results showed that ET into the AOIs of 3-D videos was much shorter than that in the HF videos, indicating that children with ASD indeed showed more interest in the AOIs of our 3-D pronunciation tutor. The RQ2 asked whether ASD learners paid attention to additional articulation and airflow information with a 3-D profile view. The results showed that some fixation counts were indeed distributed to the motions of both internal articulators and aspirated airflow (also see the heat maps in Fig 7 for more detail). Finally, the RQ3 explored whether children with ASD showed a similar attention pattern to TD children. Although all the eye-tracking parameters indicated that, compared to TD children, ASD learners showed a much more scattered gaze behavior while watching the HF and 3-D videos, they still paid a relatively concentrated visual attention to the AOIs responsible for speech production (see Fig 7). In addition, like the TD children, ASD learners showed more interest in the AOIs in our 3-D virtual tutor.

Experiment 2: Pronunciation training study

All the pronunciation data were analyzed in SPSS. The results of means and standard deviations of pre-test and post-test scores are listed in Table 3. Paired-sample T tests showed a

Table 3. Pre-test and post-test scores in Experiment 2.

Group	Consonants					Vowels				
	Pre-test		Post-test		t-value	Pre-test		Post-test		*t*-value
	M	SD	M	SD		M	SD	M	SD	
HF Group	3.42	0.58	3.79	0.54	-4.46**	3.63	0.45	4.01	0.40	-4.07**
3-D Group	3.39	0.63	3.99	0.59	-9.44***	3.61	0.56	4.19	0.42	-10.21***

M, Mean; SD, Standard Deviation.

**$p < 0.01$

***$p < 0.001$

https://doi.org/10.1371/journal.pone.0210858.t003

significant improvement in the mean consonant scores for the HF group after training, $t(10)$ = -4.46, $p < 0.01$, Cohen's $d = 0.65$; a significant improvement in the mean consonant scores for the 3-D group, $t(10)$ = -9.44, $p < 0.001$, Cohen's $d = 0.98$; and a significant increase in mean vowel scores for the HF group, $t(10)$ = -4.07, $p < 0.01$, Cohen's $d = 0.89$; as well as a significant increase in mean vowel scores for the 3-D group, $t(10)$ = -10.21, $p < 0.001$, Cohen's $d = 1.15$.

The other key question of interest is whether the improvement in consonant or vowel scores from pre-test to post-test is greater for the 3-D group than it is for the HF group. The improvement (gain) from pre-test to post-test was computed for each participant by subtracting each person's pre-test score from his or her post-test score: *Gain score = post-test score–pre-test score*. First, a two-way 2 (*training group*: HF, 3-D) × 12 (*consonant*) ANOVA was conducted on gain scores for consonants, with the *consonant* as a within-subject factor and the *group* as a between-subject factor. The analysis confirmed that the 3-D group obtained higher gain scores for consonants than the HF group ($F(1, 20) = 5.25$, $p < 0.05$, $\eta_p^2 = 0.21$), while no main effect was found for *consonant* on gain scores ($F(11, 220) = 0.86$, $p = 0.54$, $\eta_p^2 = 0.04$). There was no significant interaction between *consonant* and *group* ($F(11, 220) = 0.72$, $p = 0.65$, $\eta_p^2 = 0.03$), (see Fig 8A and Table 4).

Second, a two-way 2 (*training group*: HF, 3-D) × 8 (*vowel*) ANOVA was performed on gain scores for vowels, with the *vowel* as a within-subject factor and the *group* as a between-subject factor. Neither main effect of *vowel* ($F(7, 140) = 1.74$, $p = 0.15$, $\eta_p^2 = 0.08$) nor *group* ($F(1, 20) = 3.17$, $p = 0.09$, $\eta_p^2 = 0.13$) were significant. However, there was a significant interaction between *vowel* and *group*, $F(7,140) = 2.50$, $p < 0.05$, $\eta_p^2 = 0.11$. After this, a simple mean effect analysis was conducted with Bonferroni adjustment. For Mandarin vowels *u* [u], *o* [o], *e* [ɤ], *a* [a], *i* [i], ü[y], *group* had no effect (*all ps > 0.05*). However, for apical vowels *-i* [ɿ] ($F(1, 20) = 9.48$, $p < 0.01$) and *-i* [ʅ] ($F(1, 20) = 16.69$; $p < 0.001$), the 3-D group showed a relatively higher gain score than the HF group (see Fig 8B and Table 4).

In Experiment 2, a treatment outcome study was conducted to compare pronunciation training performance between two groups of children with ASD, learning from HF tutors or 3-D tutors respectively. Both groups showed significant improvements in consonant and vowel production from pre-test to post-test, after intensive computer-assisted pronunciation training. Moreover, ASD children learning from our 3-D pronunciation tutor tended to produce Mandarin stops and affricates and two Mandarin apical vowels better than those learning from the HF tutor. It is, therefore, reasonable to conclude that children with ASD can benefit more from our 3-D virtual pronunciation tutor.

Discussion

The main goal of our investigation was to compare the treatment efficacy of a HF tutor and our 3-D virtual pronunciation tutor for pronunciation intervention in low-functioning

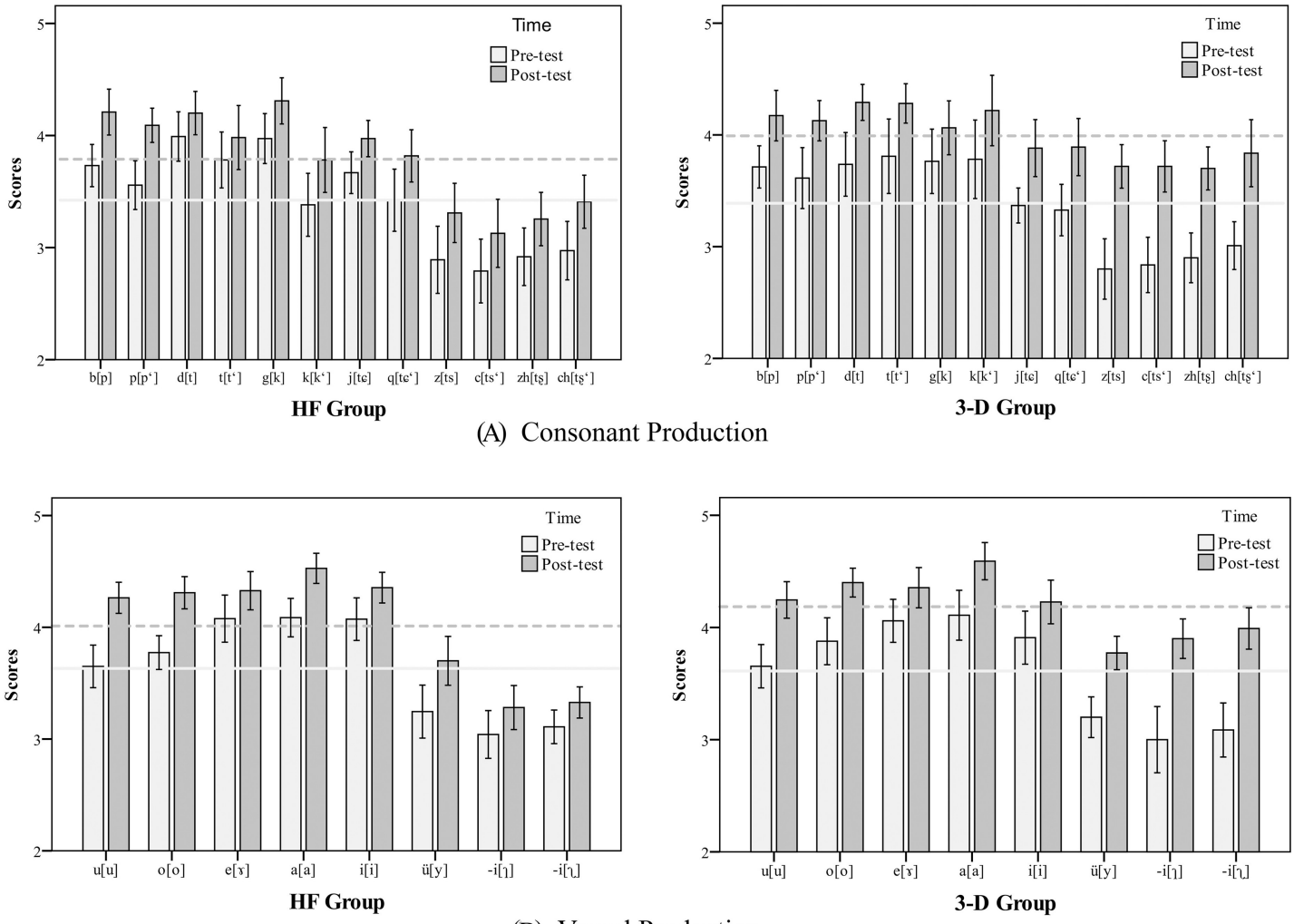


Fig 8. The Pre-test and Post-test Scores for Different Mandarin Consonants (A) and Vowels (B) in HF Group (left) and 3-D Group (right) in Experiment 2. The light solid lines indicate the average pre-test scores for different consonants or vowels, and the dark dashed lines refer to the average post-test scores (Error bars: +/- 1 SE).

https://doi.org/10.1371/journal.pone.0210858.g008

children with ASD. These two types of tutors (HF tutor and 3-D tutor) were widely used in various computer-assisted pronunciation training systems [33,58,60]. The results of the first eye-tracking study implied that, during the learning process, ASD learners showed more

Table 4. Gain Scores of different mandarin consonants and vowels for HF and 3-D groups in Experiment 2.

Group	Consonants											
	***b*[p]**	***p*[pʻ]**	***d*[t]**	***t*[tʻ]**	***g*[k]**	***k*[kʻ]**	***j*[tɕ]**	***q*[tɕʻ]**	***z*[ts]**	***c*[tsʻ]**	***zh*[tʂ]**	***ch*[tʂʻ]**
HF	0.48	0.53	0.21	0.20	0.34	0.40	0.30	0.40	0.42	0.34	0.34	0.44
3-D	0.46	0.51	0.55	0.47	0.30	0.44	0.51	0.56	0.92	0.88	0.80	0.83

Group	Vowels							
	u[u]	*o*[o]	*e*[ɤ]	a[a]	*i*[i]	*ü*[y]	-*i*[ʅ]	-*i*[ʮ]
HF	0.61	0.54	0.25	0.44	0.28	0.45	0.24	0.22
3-D	0.59	0.52	0.30	0.48	0.32	0.57	0.90	0.90

https://doi.org/10.1371/journal.pone.0210858.t004

interests in the AOIs of 3-D virtual tutor and paid some degree of absolute attention to the additional visual speech cues of articulatory and airflow models in our 3-D virtual tutor. The learners' visual attention to visual speech cues was important for language learning as the 'noticing hypothesis' [61] indicates that noticing itself does not result in acquisition, but is an essential first step in acquiring a speech item. In experiment 2, a treatment outcome study was further conducted and showed that, compared with those learning from HF tutors, the 3-D group showed a much higher increase in scores while uttering Mandarin stops, affricates, and two Mandarin apical vowels (-i[ɿ], -i[ʅ]). Based on these findings, we could conclude that relative to the HF tutor, children with ASD benefited more from noticing the visual speech cues in our multimodal 3-D pronunciation tutor. Furthermore, the user experience was evaluated by conducting a short oral interview with parents or caregivers after pronunciation training, which included three aspects: enjoyment, motivation, and acceptability. Although the enjoyment level was medium with some degree of concerns about the uncommon internal articulators in daily life, most of the parents or caregivers showed a strong motivation and a high acceptability towards our computer-assisted 3-D virtual pronunciation tutor, and even wanted to make a copy of the 3-D videos. To conclude, the above quantitative and qualitative results showed the benefit and usability of the implementation of 3-D virtual tutor for the pronunciation training in children with ASD.

The underlying learning mechanism of our current approach was built on 'imitation learning' [24,25], which relies on a straightforward and realistic presentation of key acoustic features from our 3-D virtual tutor. The mirror system based therapy has been proved to be effective in several robot-mediated training systems for learning by imitation [62–64], showing that robot-mediated imitation learning for children with ASD was effective and produced relatively better performances than a human therapist. In the current study, our multimodal 3-D pronunciation tutor with a transparent profile view can realistically exhibit synthetic visual speech with bio-data-driven external and internal articulatory movements (i.e., the place of articulation), and expiratory airflow information to discriminate Mandarin aspirated from unaspirated stops and affricates (i.e., the manner of articulation). Furthermore, the visual cues to the contrast between two Mandarin apical vowels were sufficiently salient and easily discriminated by observing the apical motion from front (-i[ɿ]) to back (-i[ʅ]) in our 3-D tutor. Mandarin-speaking children with ASD in our investigation benefited from noticing these visual speech cues through repeated exposure, and their intended learning by imitation from additional visual speech cues in our 3-D virtual tutor could partly explain the better training outcomes in children with ASD. Learning by imitation requires little priori linguistic knowledge, making this methodology suitable even for young children and individuals with severe cognitive disorders. These findings implicate a new approach in pronunciation training for ASD learners, by pedagogically illustrating visual speech cues in a 3-D virtual tutor.

The possible theoretical significance of this study is shown as follows. Firstly, our findings may provide complementary evidence that children with ASD, although somewhat restricted in their ability to use visual information from a tutor, can integrate visual and vocal information and further improve their skill of speech production. These findings are consistent with previous studies [65,66] which indicated that while children with ASD were less accurate in recognizing stimuli in a unimodal condition, they showed a normal integration of visual and auditory speech stimuli. The additional visual information related to speech sound production in our 3-D tutor is likely to be of great importance to audiovisual pronunciation training for children with ASD. Secondly, another interesting theoretical speculation [67,68] indicated that, in gaze behavior patterns looking at the eyes and mouth during face perception, children with ASD look more towards the mouth region because they tend to orient toward audiovisual synchrony. Children with ASD may focus on the mouth initially because of its physically

contingent properties, seeing the world in terms of its physical features rather than its social-affective context. This theoretical speculation suggests an alternative learning path for language acquisition: language skills are being acquired with the help of physical features (the relationship between motion and sound) rather than social-affective features (speech sounds as social cues). Similarly, in our 3-D virtual tutor, when the external and internal articulators and airflow began to change, the speech sounds started to play; when the speech sounds stopped, the animations also ceased. Children with ASD in our study tended to concentrate attention on the motion of articulators and airflow changes, probably due to their synchrony with speech sound, and this process may account partly for the greater improvement in pronunciation scores with the help of a 3-D virtual tutor.

With respect to the practical significance, our computer-assisted 3-D pronunciation tutor can be utilized remotely at home or in the community, potentially decreasing the number of in-person intervention hours that ASD children would need to have with speech-language pathologists (SLPs). Since there is a short of SLPs and therapists especially in developing countries [69], our cost-effective 3-D virtual tutor could be used as one of the substitutes which could potentially reduce the ASD family's financial burden. Generating synthetic visual speech through the 3-D virtual tutor can provide a novel mode of pronunciation training and repeatedly provide ASD learners with one-to-one instruction anytime and beyond the traditional classroom environment in a rehabilitation center, which can make a significant difference from traditional language learning methods [2,19–23].

Furthermore, an increasing number of recent studies have suggested the great benefit of computerized technologies as therapeutic and educational tools for individuals with ASD [70–72]. In the current study, we also found benefits in utilizing the methodology of computer-based pronunciation training for children with ASD. In the eye-tracking study, while watching the HF or 3-D videos presented on a computer screen, ASD learners paid relatively concentrated visual attention to the AOIs which are closely related to speech production. In Experiment 2, with a short-term and intensive computer-based pronunciation training package, we found convincing evidence that production enhancement was indeed occurring with the computer-based program. Both 3-D and HF groups showed gains in consonant and vowel production from pre-test to post-test. These results are encouraging, and are in line with other positive treatment outcomes with computer-based instruction, showing it to be an effective method to train and develop vocabulary and reading knowledge in children with ASD [34,73–75]. As mentioned previously in this paper, failure to attend to the ambient language environment hinders the ability to acquire spoken language in children with ASD, and also leads to a reduced tendency to hone speech sound production from speech models produced by others in the social environment. Presenting learning materials of HF or 3-D tutors via the computers can potentially diminish the social difficulties for some children with ASD when interacting with a teacher or an SLP.

Computer-based 3-D virtual pronunciation training for ASD learners might be of great help, but some caution should be exercised before overstating this claim. As mentioned in Experiment 2, a small proportion of learners with ASD were dropped during the training period. Regardless of our coaxing and persuading, they were not attracted to the computer. This phenomenon is somewhat understandable given that the behavioral difficulties widely observed in children with ASD, such as lack of cooperation, and resistance to novel methods, often create difficult situations that are not optimal for learning [76]. However, for these small groups of ASD children, consequently, other styles of learning should be explored and supported. We need to be more ingenious in capturing ASD learners' attention. For examples, in future studies, the technique of automatic speech recognition could be integrated into the current 3-D virtual tutor, to evaluate the ASD learners' pronunciation online and give feedback in time to better enhance interaction and attract the ASD learners' attention. Moreover, as

suggested by [77], a child does not progress by acquiring units like phonemes or allophones, but rather by gradually adding lexical items to his/her repertoire. Consequently, the process of phonology acquisition in the early stages is not phoneme by phoneme, but word by word. In the future, we might make use of the treatment of specific speech sounds as a catalyst or stimulus to the word level, and combine phonology and lexicon learning together. In this way, certain speech sounds would be related to a certain word. The concurrent presentation of a 3-D virtual pronunciation tutor and a corresponding word picture with varying colors and shapes may be more engaging and motivating for children with ASD. In short, an ideal pronunciation training system should be individually tailored for each ASD student to put them in a good mood and encourage them to become more interested in working with the computer to enjoy imitating the speech sounds they hear and see.

Conclusions

A subgroup of children with ASD, especially those with more severe global language impairment, may exhibit more severe speech sound production difficulties. Clinicians and SLPs should be aware that children with more severely impaired language and behavior, may exhibit more severe speech production difficulties. Recently, two studies [37,78] have emphasized the critical need for both researchers and clinicians to address pronunciation problems and to focus on speech sound behavior in individuals with ASD. However, available interventions that aim to improve pronunciation ability in children with ASD are extremely limited.

In this study, a computer-based 3-D virtual pronunciation tutor was proposed and evaluated. The current findings indicated that individuals with ASD who are struggling with speech sound production could benefit more from our 3-D pronunciation tutor exhibited on a computer screen. By demonstrating additional visual information during speech sound production, our 3-D virtual speech production tutor provides an efficient pronunciation training method to enhance consonant and vowel production skills among the ASD cohort. We advocate multimodal learning environments to enhance speech production and other language skills among an ASD cohort.

Acknowledgments

This work was partly supported by grants from National Natural Science Foundation of China (NSFC: U1736203, 61771461, 11474300), and Shenzhen Fundamental Research Program: JCYJ20160429184226930; JCYJ20170413161611534. We sincerely thank all the child participants with ASD from *Shenzhen Aixin Zhihui Rehabilitation Centre for Children with Special Needs*, the TD children from *Taoyuan Zhuoya Primary School*, and their parents for their participation and cooperation. The authors have no conflict of interest regarding the contents of this manuscript.

Author Contributions

Conceptualization: Fei Chen, Lan Wang.

Data curation: Fei Chen, Xiaojie Pan.

Formal analysis: Xiaojie Pan.

Funding acquisition: Lan Wang, Gang Peng, Nan Yan.

Investigation: Fei Chen, Nan Yan.

Methodology: Fei Chen, Gang Peng.

Project administration: Gang Peng.

Supervision: Lan Wang, Gang Peng.

Writing – original draft: Fei Chen.

Writing – review & editing: Fei Chen, Lan Wang, Gang Peng.

References

1. Lai M-C, Lombardo M V, Baron-Cohen S. Autism. Lancet. 2014; 383: 896–910. https://doi.org/10.1016/S0140-6736(13)61539-1 PMID: 24074734
2. Koegel RL, Shirotova L, Koegel LK. Brief report: Using individualized orienting cues to facilitate first-word acquisition in non-responders with autism. J Autism Dev Disord. 2009; 39: 1587–1592. https://doi.org/10.1007/s10803-009-0765-9 PMID: 19488847
3. Baron-Cohen S. Mindblindness: An essay on autism and theory of mind. Cambridge, Mass.: MIT Press; 1995.
4. Baltaxe CAM. Pragmatic deficits in the language of autistic adolescents. J Pediatr Psychol. 1977; 2: 176–180. https://doi.org/10.1093/jpepsy/2.4.176
5. Volkmar FR, Paul R, Klin A, Cohen DJ. Handbook of autism and pervasive developmental disorders, diagnosis, development, neurobiology, and behavior. John Wiley & Sons; 2005.
6. Bartak L, Rutter M, Cox A. A comparative study of infantile autism and specific developmental receptive language disorder. Br J Psychiatry. 1975; 126: 127–145. https://doi.org/10.1192/bjp.126.2.127 PMID: 1131465
7. Kjelgaard MM, Tager-Flusberg H. An investigation of language impairment in autism: Implications for genetic subgroups. Lang Cogn Process. 2001; 16: 287–308. https://doi.org/10.1080/01690960042000058 PMID: 16703115
8. Rapin I, Dunn M. Update on the language disorders of individuals on the autistic spectrum. Brain Dev. 2003; 25: 166–172. https://doi.org/10.1016/S0387-7604(02)00191-2 PMID: 12689694
9. Boucher J. Articulation in early childhood autism. J Autism Child Schizophr. 1976; 6: 297–302. https://doi.org/10.1007/BF01537907 PMID: 1036736
10. Rapin I, Dunn MA, Allen DA, Stevens MC, Fein D. Subtypes of language disorders in school-age children with autism. Dev Neuropsychol. 2009; 34: 66–84. https://doi.org/10.1080/87565640802564648 PMID: 19142767
11. Shriberg LD, Paul R, McSweeny JL, Klin A, Cohen DJ, Volkmar FR. Speech and prosody characteristics of adolescents and adults with high-functioning autism and asperger syndrome. J Speech, Lang Hear Res. 2001; 44: 1097–1115. https://doi.org/10.1044/1092-4388(2001/087)
12. Cleland J, Gibbon FE, Peppé S, O'Hare A, Rutherford M. Phonetic and phonological errors in children with high functioning autism and Asperger syndrome. Int J Speech Lang Pathol. 2010; 12: 69–76. https://doi.org/10.3109/17549500903469980 PMID: 20380251
13. Bartolucci G, Pierce S, Streiner D, Eppel P. Phonological investigation of verbal autistic and mentally retarded subjects. J Autism Child Schizophr. 1976; 6: 303–316. https://doi.org/10.1007/BF01537908 PMID: 1036737
14. Schoen E, Paul R, Chawarska K. Phonology and vocal behavior in toddlers with autism spectrum disorders. Autism Res. 2011; 4: 177–188. https://doi.org/10.1002/aur.183 PMID: 21308998
15. Wolk L, Brennan C. Phonological investigation of speech sound errors in children with autism spectrum disorders. Speech, Lang Hear. 2013; 16: 239–246.
16. Wolk L, Giesen J. A phonological investigation of four siblings with childhood autism. J Commun Disord. 2000; 33: 371–389. https://doi.org/10.1016/S0021-9924(00)00021-6 PMID: 11081786
17. Zhu H, Dodd B. The phonological acquisition of Putonghua (Modern Standard Chinese). J Child Lang. 2000; 27: 3–42. PMID: 10740966
18. Wolk L, Edwards ML, Brennan C. Phonological difficulties in children with autism: An overview. Speech, Lang Hear. 2016; 19: 121–129. https://doi.org/10.1080/2050571X.2015.1133488
19. Goldstein H. Communication intervention for children with autism: A review of treatment efficacy. J Autism Dev Disord. 2002; 32: 373–396. https://doi.org/10.1023/A:1020589821992 PMID: 12463516
20. Koegel R, O'Dell M, Dunlap G. Producing speech use in nonverbal autistic children by reinforcing attempts. J Autism Dev Disord. 1988; 18: 525–538. https://doi.org/10.1007/BF02211871 PMID: 3215880

21. Vashdi E. The influence of initial phoneme cue technique on word formation: a case study of a child with apraxia of speech and autism. Int J Child Heal Hum Dev. 2014; 7: 198–203.
22. Nordgren PM. Phonological development in a child with autism spectrum disorder: Case study of an intervention. J Interact Res Commun Disord. 2015; 6: 25–51. https://doi.org/10.1558/jircd.v6i1.25
23. Wan CY, Bazen L, Baars R, Libenson A, Zipse L, Zuk J, et al. Auditory-motor mapping training as an intervention to facilitate speech output in non-verbal children with autism: A proof of concept study. PLoS One. 2011;6. https://doi.org/10.1371/journal.pone.0025505 PMID: 21980480
24. Chen W, Zhang J, Ding J. Mirror system based therapy for autism spectrum disorders. Front Med China. 2008; 2: 344–347.
25. Iacoboni M, Dapretto M. The mirror neuron system and the consequences of its dysfunction. Nat Rev Neurosci. 2006; 7: 942–951. https://doi.org/10.1038/nrn2024 PMID: 17115076
26. Naranjo CA, Ortiz JS, Álvarez VM, Sánchez JS, Tamayo VM, Acosta FA, et al. Teaching Process for Children with Autism in Virtual Reality Environments. Proceedings of the 2017 9th International Conference on Education Technology and Computers. ACM; 2017. pp. 41–45.
27. Bellani M, Fornasari L, Chittaro L, Brambilla P. Virtual reality in autism: state of the art. Epidemiol Psychiatr Sci. 2011; 20: 235–238. https://doi.org/10.1017/S2045796011000448 PMID: 21922965
28. Massaro DW. Perceiving talking faces: from speech perception to a behavioral principle. Cambridge, Mass.: MIT Press; 1998.
29. Liu X, Yan N, Wang L, Wu X, Ng ML. An interactive speech training system with virtual reality articulation for Mandarin-speaking hearing impaired children. IEEE International Conference on Information and Automation (ICIA). 2013. pp. 191–196. https://doi.org/10.1109/ICInfA.2013.6720294
30. Massaro DW, Light J. Using visible speech to train perception and production of speech for individuals with hearing loss. J Speech, Lang Hear Res. 2004; 47: 304–320. https://doi.org/10.1044/1092-4388(2004/025)
31. Rathinavelu A, Thiagarajan H, Rajkumar A. Three dimensional articulator model for speech acquisition by children with hearing loss. International Conference on Universal Access in Human-Computer Interaction. 2007. pp. 786–794. https://doi.org/10.1007/978-3-540-73279-2_87
32. Navarra J, Soto-Faraco S. Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychol Res. 2007; 71: 4–12. https://doi.org/10.1007/s00426-005-0031-5 PMID: 16362332
33. Wang X, Hueber T, Badin P. On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French. 10th International Seminar on Speech Production (ISSP 2014). 2014. pp. 449–452.
34. Bosseler A, Massaro DW. Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. J Autism Dev Disord. 2003; 33: 653–672. https://doi.org/10.1023/B:JADD.0000006002.82367.4f PMID: 14714934
35. Massaro DW, Bosseler A. Read my lips: The importance of the face in a computer-animated tutor for vocabulary learning by children with autism. Autism. 2006; 10: 495–510. https://doi.org/10.1177/1362361306066599 PMID: 16940315
36. Paul R, Shriberg LD, McSweeny J, Cicchetti D, Klin A, Volkmar F. Brief report: Relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders. J Autism Dev Disord. 2005; 35: 861–869. https://doi.org/10.1007/s10803-005-0031-8 PMID: 16283080
37. Shriberg LD, Paul R, Black LM, van Santen JP. The hypothesis of apraxia of speech in children with autism spectrum disorder. J Autism Dev Disord. 2011; 41: 405–426. https://doi.org/10.1007/s10803-010-1117-5 PMID: 20972615
38. Haesen B, Boets B, Wagemans J. A review of behavioural and electrophysiological studies on auditory processing and speech perception in autism spectrum disorders. Res Autism Spectr Disord. Elsevier; 2011; 5: 701–714. https://doi.org/10.1016/j.rasd.2010.11.006
39. O'connor K. Auditory processing in autism spectrum disorder: a review. Neurosci Biobehav Rev. 2012; 36: 836–854. https://doi.org/10.1016/j.neubiorev.2011.11.008 PMID: 22155284
40. Chen F, Chen H, Wang L, Zhou Y, He J, Yan N, et al. Intelligible enhancement of 3D articulation animation by incorporating airflow information. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016). 2016. pp. 6130–6134. https://doi.org/10.1109/ICASSP.2016.7472855
41. Chen F, Wang L, Chen H, Peng G. Investigations on Mandarin aspiratory animations using an airflow model. IEEE/ACM Trans Audio, Speech, Lang Process. 2017; 25: 2399–2409. https://doi.org/10.1109/TASLP.2017.2755400

42. Zhang D, Liu X, Yan N, Wang L, Zhu Y, Chen H. A multi-channel/multi-speaker articulatory database in Mandarin for speech visualization. The 9th International Symposium on Chinese Spoken Language Processing. 2014. pp. 299–303. https://doi.org/10.1109/ISCSLP.2014.6936629

43. Mori M, MacDorman KF, Kageki N. The uncanny valley [from the field]. IEEE Robot Autom Mag. 2012; 19: 98–100. https://doi.org/10.1109/MRA.2012.2192811

44. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J. Eye tracking: A comprehensive guide to methods and measures. OUP Oxford; 2011.

45. Happé F, Booth R, Charlton R, Hughes C. Executive function deficits in autism spectrum disorders and attention-deficit/hyperactivity disorder: examining profiles across domains and ages. Brain Cogn. 2006; 61: 25–39. https://doi.org/10.1016/j.bandc.2006.03.004 PMID: 16682102

46. Behrmann M, Thomas C, Humphreys K. Seeing it differently: visual processing in autism. Trends Cogn Sci. 2006; 10: 258–264. https://doi.org/10.1016/j.tics.2006.05.001 PMID: 16713326

47. Rizzolatti G, Craighero L. The mirror-neuron system. Annu Rev Neurosci. 2004; 27: 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230 PMID: 15217330

48. Association AP. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. Arlington, VA: American Psychiatric Publishing; 2013.

49. Gilliam JE. Gilliam Autism Rating Scale: GARS 2. Austin, TX: PRO-ED; 2006.

50. Schopler E, Reichler RJ, Renner BR. The Childhood Autism Rating Scale (CARS). Western Psychological Services Los Angeles, CA; 2002.

51. Schopler E, Lansing MD, Reichler RJ, Marcus LM. Examiner's manual of Psychoeducational Profile. Vol. 3 Pro-Ed Inc, Austin, TX. 2005.

52. Zhou Y, Chen F, Chen H, Wang L, Yan N. Evaluation of a multimodal 3-D pronunciation tutor for learning Mandarin as a second language: An eye-tracking study. 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). 2016.

53. Chen K-L, Chiang F-M, Tseng M-H, Fu C-P, Hsieh C-L. Responsiveness of the psychoeducational profile-third edition for children with autism spectrum disorders. J Autism Dev Disord. 2011; 41: 1658–1664. https://doi.org/10.1007/s10803-011-1201-5 PMID: 21336523

54. Fu C-P, Hsieh C-L, Tseng M-H, Chen Y-L, Huang W-T, Wu P-C, et al. Inter-rater reliability and smallest real difference of the Chinese Psychoeducational Profile-third edition for children with Autism Spectrum Disorder. Res Autism Spectr Disord. 2010; 4: 89–94. https://doi.org/10.1016/j.rasd.2009.09.002

55. Fulton ML, D'Entremont B. Utility of the Psychoeducational Profile-3 for assessing cognitive and language skills of children with autism spectrum disorders. J Autism Dev Disord. 2013; 43: 2460–2471. https://doi.org/10.1007/s10803-013-1794-y PMID: 23446992

56. McCleery JP, Tully L, Slevc LR, Schreibman L. Consonant production patterns of young severely language-delayed children with autism. J Commun Disord. 2006; 39: 217–231. https://doi.org/10.1016/j.jcomdis.2005.12.002 PMID: 16480738

57. Ali M, Zamzuri A, Segaran K, Hoe TW. Effects of verbal Components in 3D talking-head on pronunciation learning among non-native speakers. J Educ Technol Soc. 2015; 18: 313–322.

58. Peng X, Chen H, Wang L, Wang H. Evaluating a 3-D virtual talking head on pronunciation learning. Int J Hum Comput Stud. 2018; 109: 26–40. https://doi.org/10.1016/j.ijhcs.2017.08.001

59. Seferoğlu G. Improving students' pronunciation through accent reduction software. Br J Educ Technol. 2005; 36: 303–316.

60. Hamdan MN, Ali AZM, Hassan A. The effects of realism level of talking-head animated character on students' pronunciation learning. International Conference on Science in Information Technology (ICSITech). 2015. pp. 58–62.

61. Robinson P. Attention, memory, and the "noticing" hypothesis. Lang Learn. 1995; 45: 283–331.

62. Zheng Z, Das S, Young EM, Swanson A, Warren Z, Sarkar N. Autonomous robot-mediated imitation learning for children with autism. IEEE International Conference on Robotics and Automation (ICRA). 2014. pp. 2707–2712.

63. Zheng Z, Young EM, Swanson A, Weitlauf A, Warren Z, Sarkar N. Robot-mediated mixed gesture imitation skill training for young children with ASD. IEEE International Conference onAdvanced Robotics (ICAR). 2015. pp. 72–77.

64. Liu X, Zhou X, Liu C, Wang J, Zhou X, Xu N, et al. An interactive training system of motor learning by imitation and speech instructions for children with autism. 9th International Conference on Human System Interactions (HSI). 2016. pp. 56–61.

65. Massaro DW, Bosseler A. Perceiving speech by ear and eye: Multimodal integration by children with autism. J Dev Learn Disord. 2003; 7: 111–144.

66. Williams JHG, Massaro DW, Peel NJ, Bosseler A, Suddendorf T. Visual–auditory integration during speech imitation in autism. Res Dev Disabil. 2004; 25: 559–575. https://doi.org/10.1016/j.ridd.2004.01.008 PMID: 15541632

67. Jones W, Carr K, Klin A. Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. Arch Gen Psychiatry. 2008; 65: 946–954. https://doi.org/10.1001/archpsyc.65.8.946 PMID: 18678799

68. Klin A, Lin DJ, Gorrindo P, Ramsay G, Jones W. Two-year-olds with autism orient to non-social contingencies rather than biological motion. Nature. 2009; 459: 257. https://doi.org/10.1038/nature07868 PMID: 19329996

69. Gillon G, Hyter Y, Fernandes FD, Ferman S, Hus Y, Petinou K, et al. International survey of speech-language pathologists' practices in working with children with autism spectrum disorder. Folia Phoniatr Logop. 2017; 69: 8–19. https://doi.org/10.1159/000479063 PMID: 29248908

70. Constantin A, Johnson H, Smith E, Lengyel D, Brosnan M. Designing computer-based rewards with and for children with Autism Spectrum Disorder and/or Intellectual Disability. Comput Human Behav. 2017; 75: 404–414. https://doi.org/10.1016/j.chb.2017.05.030

71. Miller N, Wyatt J, Casey LB, Smith JB. Using computer-assisted instruction to increase the eye gaze of children with autism. Behav Interv. 2018; 33: 3–12. https://doi.org/10.1002/bin.1507

72. Root JR, Stevenson BS, Davis LL, Geddes-Hall J, Test DW. Establishing computer-assisted instruction to teach academics to students with autism as an evidence-based practice. J Autism Dev Disord. 2017; 47: 275–284. https://doi.org/10.1007/s10803-016-2947-6 PMID: 27812773

73. Colby K. The rationale for computer-based treatment of language difficulties in nonspeaking autistic children. J Autism Child Schizophr. 1973; 3: 254–260. https://doi.org/10.1007/BF01538283 PMID: 4800391

74. Heimann M, Nelson KE, Tjus T, Gillberg C. Increasing reading and communication skills in children with autism through an interactive multimedia computer program. J Autism Dev Disord. 1995; 25: 459–480. https://doi.org/10.1007/BF02178294 PMID: 8567593

75. Moore M, Calvert S. Brief report: Vocabulary acquisition for children with autism: Teacher or computer instruction. J Autism Dev Disord. 2000; 30: 359–362. https://doi.org/10.1023/A:1005535602064 PMID: 11039862

76. Koegel LK. Interventions to facilitate communication in autism. J Autism Dev Disord. 2000; 30: 383–391. PMID: 11098873

77. Wang WSY. The three scales of diachrony. In: Kachru BB, editor. Linguistics in the Seventies: Directions and Prospects. University of Illinois: Urbana, IL; 1978. pp. 63–75.

78. Eigsti I-M, de Marchena AB, Schuh JM, Kelley E. Language acquisition in autism spectrum disorders: A developmental review. Res Autism Spectr Disord. 2011; 5: 681–691. https://doi.org/10.1016/j.rasd.2010.09.001