Learning challenging L2 sounds via computer assisted training: Audiovisual training with an airflow model

Fei Chen[a#*], Quansheng Xia[c#], Yan Feng[b], Lan Wang[d] and Gang Peng[b,d*]

[a]School of Foreign Languages, Hunan University, Changsha, China;

[b]Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China;

[c]College of Chinese Language and Culture, Nankai University, Tianjin, China;

[d]Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Running title: Audiovisual training with airflow model

[#]The first two authors contribute equally to this study.

[*]Corresponding authors:

Fei Chen: chenfeianthony@gmail.com

School of Foreign Languages, Hunan University, Changsha, China

Gang Peng: gpengjack@gmail.com

Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

Conflict of Interest: The authors have declared that no competing interests existed at the time of publication.

Acknowledgments: This work was partially supported by grants from National Natural Science Foundation of China (91420301, U1736203), Fundamental Research Funds for the Central Universities, Hunan University (53111801066), Young Scholars Fund of Humanities and Social Sciences of Ministry of Education (18YJC740116), and General Program of Social Science of Tianjin (TJZW15-004).

Learning challenging L2 sounds via computer assisted training: Audiovisual training with an airflow model

Abstract

The acquisition of unaspirated and aspirated consonants in Mandarin has been reported to be rather challenging for second language (L2) learners. In the current study, a 3-D airflow model was integrated into the virtual talking head for audiovisual pronunciation training in these Mandarin consonants. Using the eye-tracking technique, Experiment 1 investigated L2 learners' general acceptance and gauged attention distribution online while learning with the system. Eye-tracking results showed that the talking head was well accepted, and was successful in directing L2 learners' attention to the visual modality of the airflow model marking the contrast in aspiration. To further compare training efficacy, learning outcomes were evaluated by randomly dividing Japanese learners of Mandarin into different training groups using the 3-D tutor with and without an airflow model, respectively, in Experiment 2. The additional visual cue of the airflow model helped enhance their production of the aspirated Mandarin stops. Moreover, this computer-assisted training approach was shown to be robust as the advantage of training with an airflow model can be generalized to novel syllables with a change of tones or rimes.

Keywords: Audiovisual Pronunciation Training, Airflow Model, Talking Head, Eye Tracking, Second Language Learning

1 Introduction

While the chronological age when one begins to learn a second language (L2) seems to have the greatest influence on speech sound learning (Piper & Cansin, 1988), training can also help to enhance and hone learners' pronunciation skills to a great extent (Graeme, 2006). Haslam (2010) found that pronunciation training strategies were significantly correlated with pronunciation improvement, suggesting the importance of discovering the most successful pronunciation training strategies. Various training approaches to L2 consonant, vowel, or tone learning have been widely adopted and shown to be effective, such as 'high variability phonetic training' (e.g., Wang et al., 2003; Zhang, 2009) and 'adaptive training' (e.g., McClelland et al., 2002). Moreover, Hazan et al (2005) evaluated the efficacy of 'audiovisual training,' and demonstrated that it was more efficient compared with auditory-only training method for L2 pronunciation learning. More importantly, the effectiveness of audiovisual training has been shown to be robust as it is generalizable to new words in speech production (Hazan et al., 2005; Motohashi-Saigo & Hardison, 2009; Okuno & Hardison, 2016). For the current study, a computer-assisted virtual talking head with an airflow model was utilized and evaluated for audiovisual training in unaspirated and aspirated consonants among L2 learners of Mandarin.

Teaching Mandarin as an L2 has become an important profession and an important research area (Lam et al., 2001; Shei & Hsieh, 2012; Wang, 2016; Zhang & Roberts, 2019). However, Mandarin Chinese is regarded as one of the most challenging languages for L2 learners (Chen et al., 2015). One reason is that Mandarin phonology presents certain challenges for L2 learners. Apart from the lexical tones as a tonal language, the phonological system in Mandarin is also well-known for its varieties of voiceless stops and affricates. In total, there exist three minimal pairs of unaspirated vs. aspirated stops (bilabial stop: *b* [p] and *p* [p^h]; alveolar stop: *d* [t] and *t* [t^h]; velar stop: *g* [k] and *k* [k^h]), as well as three minimal pairs of affricates (alveolo-palatal affricate: *j* [tɕ] and *q* [$tɕ^h$]; alveolar affricate: *z* [ts] and *c* [ts^h]; retroflex affricate: *zh* [tʂ] and *ch* [$tʂ^h$]). Different pairs of voiceless stops or affricates are discriminated by place of articulation, while the two consonants within each minimal pair share the same place of articulation and are primarily differentiated by manner of articulation (i.e., aspiration contrast).

Typically developing children, whose native language is Mandarin, tend to acquire unaspirated

consonants earlier relative to aspirated consonants, and the development of speech production of all the 12 Mandarin stops and affricates is completed without much effort before age five (Zhu & Dodd, 2000; Si, 2006). However, the acquisition of unaspirated and aspirated consonants in Mandarin has been reported to be rather challenging for L2 learners (e.g., Chen et al., 2013; Lai, 2009). According to the learner corpora for learning Mandarin as an L2 (Chen et al., 2015), four out of the top ten phonetic error patterns were linked with 'deaspiration' (*c* [tsʰ]→*z* [ts], *ch* [tʂʰ]→*zh* [tʂ], *t* [tʰ]→*d* [t], *p* [pʰ]→*b* [p]) and another two were 'aspiration' (*zh* [tʂ]→*ch* [tʂʰ], *j* [tɕ]→*q* [tɕʰ]). In these parentheses, the former phonemes are the target phonemes, and the latter are the actual articulation produced by L2 learners of Mandarin. The unaspirated vs. aspirated contrast acts as one of the distinctive features in distinguishing different stops or affricates in Mandarin. However, in most languages (e.g., Japanese, French, Italian, Spanish, Arabic, Russian, German, Filipino, Urdu, and Bulgarian), different consonants are distinguished by the voicing contrast (voiceless vs. voiced), while not by the feature of aspiration. As suggested by Featural Model (Brown, 1998), non-native speech sounds tend to be judged as indistinguishable sounds if learners' L1 do not involve the required distinctive features. Consequently, learning to produce Mandarin stops and affricates accurately can require great effort if unaspirated vs. aspirated contrast does not feature in the phonology of one's L1.

1.1 Development of an Airflow Model Integrated into a Mandarin Articulatory Model

In the past few decades, technological innovation has brought rapid development in computer assisted language learning (CALL) and language teaching (Colpaert, 2012, 2016). A growing body of research tries to apply the new advances of CALL to improve learners' language skills, such as pronunciation training in learning foreign languages. Under such circumstances, there are various 3-D talking heads which are designed for Computer-Assisted Pronunciation Training (CAPT) systems to help learners in need to imitate the articulator animations. Thanks to the development of speech technology, physiological articulation data can be recorded through Electro-Magnetic Articulography (EMA), facial motion capture, and video-fluoroscopic images. By these means, articulator animations of visual speech can be exhibited accurately in different virtual talking heads. Among which, various language-specific talking heads were developed for L2 speech learning of Mandarin (Li et al., 2013; Liu et al., 2013; Wu et al., 2006; Yu & Li, 2014; Zhang et al., 2014),

English (Massaro, 1998; Wang et al., 2012), French (Badin et al., 2008), and Swedish (Bälter et al., 2005). Both external articulator (i.e., lip) and internal articulator (i.e., tongue) movements have been shown to guide the pronunciation training of L2 learners successfully (e.g., Navarra & Soto-Faraco, 2007; Peng et al., 2018; Wang et al., 2014).

However, as mentioned above, since the two Mandarin stops or affricates within a minimal pair (e.g., *b* [p] vs. *p* [p^h]) are differentiated principally by the unaspirated vs. aspirated contrast, and share the same place of articulation, it can be rather difficult to tell apart these sounds using the articulatory model alone in existing Mandarin 3-D talking heads (Li et al., 2013; Liu et al., 2013; Wu et al., 2006; Yu & Li, 2014; Zhang et al., 2014). To provide a solution to this problem, our recent study (Chen et al., 2016) collected expiratory airflow parameters (mean and peak airflow rate, peak time, and airflow duration) using the Phonatory Aerodynamic System (PAS 6600), further proposing a 3-D airflow model by modifying the fluid dynamics model (Stam, 1999) to strengthen the visual discriminability of confusable Mandarin stops and affricates within a talking head system. In the syllable-level animations, the articulatory model (based on articulation data collected with Carstens AG-501 EMA) was then concurrently integrated with the dynamic airflow model, and a new multimodal 3-D talking head was successfully constructed (see Figure 1 for the implementation procedures). Comparing the articulator motions while producing a minimal pair of retroflex consonants (*zh* [tʂ] vs. *ch* [$tʂ^h$]), we find it is very tough to visually detect the discrepancies from tongue and lip movements. Nonetheless, the airflow size, rate, and density of the aspirated consonant '*ch* [$tʂ^h$]' during the pronunciation process tend to be much faster, bigger, and heavier than those of unaspirated consonant '*zh* [tʂ]' (see Figure 2). Our airflow model (Chen et al., 2016) visualized exact and dynamic changes of consonant aspirations and can provide additional visual information for better discrimination of each minimal pair of stops and affricates in Mandarin.

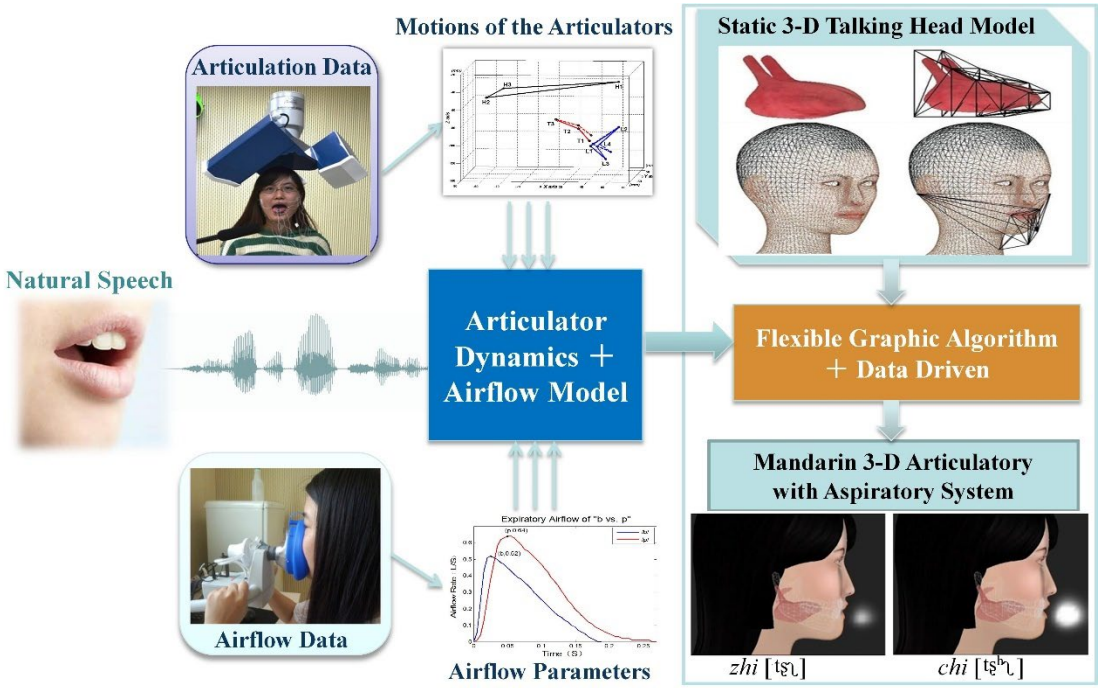


Figure 1. The implementation procedures of our Mandarin 3-D talking head.

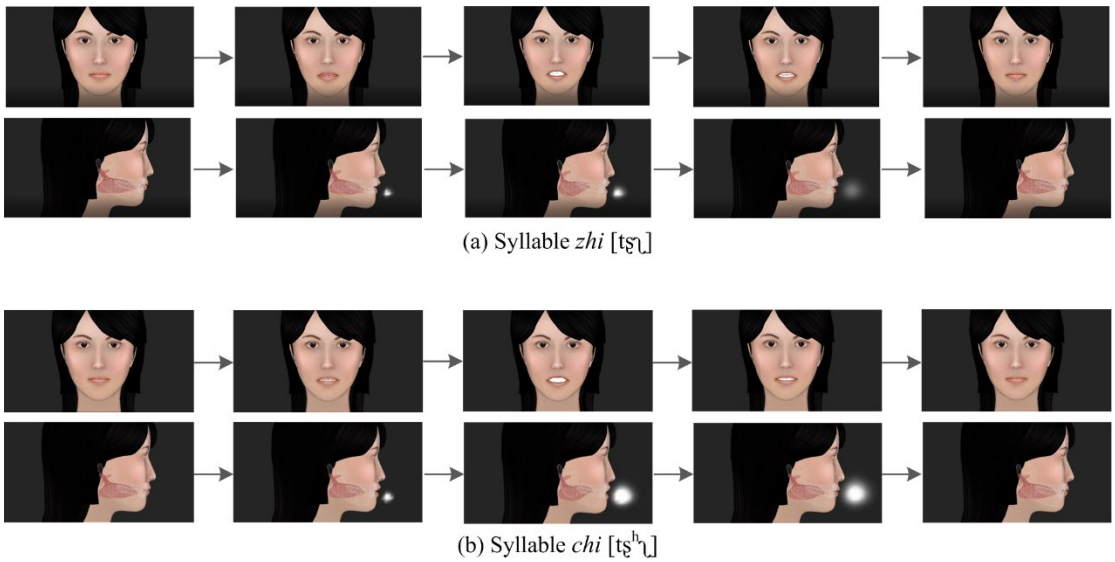


Figure 2. The dynamic 3-D articulatory with aspiratory animations in the sequence of a minimal pair of Mandarin syllables: (a) *zhi* [tʂɻ], (b) *chi* [tʂʰɻ].

1.2 Objective Assessment of 3-D Talking Heads

The objective assessment of our 3-D talking head serves as one of the key issues in the current study since the general acceptance of the talking head among L2 learners is crucial. Previous studies have always adopted a questionnaire method, with a series of subjective ratings, including realism score, quality, enjoyment, and engagement, to evaluate different types of virtual talking head offline (Liu et al., 2013; Stevens et al., 2013; Weiss et al., 2010). Eye-tracking has been utilized extensively to detect learners' psychophysiological response and to assess their attention distribution in the language learning and processing (e.g., Godfroid & Schmidtke, 2013; Tham et al., 2019; Xie et al., 2021). However, the eye-tracking methodology has rarely been applied to objectively assess L2 learners' acceptance of virtual talking heads. This approach can provide an objective and quantitative measurement of eye-movement behavior and attention distribution online in non-native L2 learners. Moreover, the eye-tracking technology has the advantage that it does not involve a secondary task, and thus enables us to measure the natural learning process without risk of altering it.

As shown in Figure 1, for our virtual talking head, an entire-head virtual tutor on a computer screen was presented, exhibiting models of the lips, face, jaw, tongue, and nasopharyngeal wall which are based on the MRI data. The 3-D articulator and airflow model were animated in accordance with physiological signals collected with EMA and PAS respectively, in order for a more realistic talking head to be generated. The 'uncanny valley effect' (Mori, 1970) suggests that the familiarity and acceptance of virtual systems may be positively related to the degree of human likeness, indicating that the acceptance level is probably higher when the virtual system is more realistic. However, unfortunately, acceptance may be severely reduced if the human likeness of a virtual system is only around 80% realistic. In such circumstances, an almost human-like virtual system seems overly 'strange' to some, and can even produce an uncomfortable uncanny feeling. On the one hand, our virtual 3-D talking head model was constructed on the basis of physiological data, and may be too human-like to escape the uncanny valley. On the other hand, the internal articulator and airflow motions, presented in a transparent profile view, tend to be unrealistic since they are uncommon in daily life. These concerns are also part of the reason why videos of a real human teacher are often adopted in many CAPT systems. Nevertheless, it is necessary to address that virtual talking heads,

if accepted by learners, offer the great advantage of being able to exhibit additional visual information (such as tongue and airflow animations in our talking head). Thus, using the eye-tracking methodology, the first focus was to evaluate L2 learners' general acceptance of our virtual 3-D talking head in comparison with videos of an actual human face (HF) tutor.

Furthermore, the transparent profile view of our 3-D talking head encompasses abundant visual cues related to sound articulation, including internal articulator, mouth corner, and expiratory airflow. When there are too many visual inputs, perception may be biased toward one visual area, particularly if the perceptual load capacities are exceeded by the attentional demands. In addition, the L2 visual cues unfamiliar to L2 learners may result in difficulty in non-native perception, and it is likely that L2 learners tend not to employ such irrespective non-native visual cues which do not mark phonetic contrasts in their own L1 (Hazan et al., 2005). The L2 learners of Mandarin may lose sensitivity and interest in the visual cues of airflow animations in our talking head if unaspirated and aspirated contrast is not a distinctive feature of their L1. Thus, the second focus of the eye-tracking study is to evaluate quantitatively whether L2 learners with various native languages distribute absolute and relative attention (as measured by eye-fixation durations) to the non-native visual information provided by the airflow model.

1.3 Pronunciation Training with an Airflow Model

Researchers and educators in L2 speech acquisition have called for empirical evaluation of the progress that students have made after tapping novel methods for learning (Chun, 1998; Hincks, 2003; Morley, 1991). Another key issue of using a virtual 3-D talking head is whether such an audiovisual CAPT system can enhance learners' pronunciation skills in reality. Investigations into Mandarin-specific phonetic characteristics and applying Mandarin visual speech in audiovisual training present potentially unique and interesting contributions (Chen & Massaro, 2011). Recent developments in synthetic Mandarin visual speech have been shown to provide an effective way to facilitate Mandarin pronunciation through demonstration of internal and external articulator animations by a talking head (Liu et al., 2013; Peng et al., 2018). The second main purpose of this study is to define whether our additional airflow model embedded in a talking head can also effectively improve L2 learners' pronunciation of unaspirated and aspirated contrast in Mandarin.

To control the influence of learners' L1 background, the training group in this experiment comprised

only Japanese learners of Mandarin. In Japanese phonology, different consonants are majorly differentiated by voiceless vs. voiced contrast in terms of the manner of articulation, while 'aspiration' does not act as a distinctive feature. The so-called Japanese aspirated consonants are essentially allophones which are distributed complementarily with their unaspirated phonemes. The aspirated portion of these Japanese aspirated allophones in the word-initial position sounds much shorter compared with the corresponding aspirated phonemes in Mandarin (Vance, 1997). As calculated by a straightforward acoustic measure, the average voice onset time (VOT, the temporal relation between the onset of the release of a stop element and the onset of glottal pulsing), the voiceless aspirated consonants show a considerably longer voicing lag in Mandarin and are identified as a strongly aspirated type even among the languages in the globe (Abramson & Whalen, 2017; Cho & Ladefoged, 1999). Influenced by native Japanese phonology, Japanese learners of Mandarin present a high rate of deaspiration when producing Mandarin's strongly aspirated consonants. They tend to produce weakly aspirated forms (Zhang, 2009). Thus, the reduced aspiration of Japanese learners can be insufficient to perceptually mark a Mandarin aspirated stop or affricate as 'aspirated', which may lead to a communication barrier. A training package utilizing additional visual input from our airflow model was thus given to Japanese learners of Mandarin to evaluate the outcome of audiovisual training with a 3-D talking head.

As pointed out in the literature reviews, some common problems existed in CALL studies, such as, lack of clear description of experimental design, a nearly exclusive concern on languages in Western Europe (Felix, 2005; Hubbard, 2005; Stockwell, 2007; Zhao, 2003), and inadequate evidence that demonstrates the technologies' effectiveness for foreign language teaching and learning (Golonka et al., 2012). To overcome these limitations, a two-step experimental design was made in the current study. In the first step, an airflow model was constructed, then combined and synchronized with an articulatory model in a Mandarin virtual 3-D talking head. To evaluate the efficacy of this talking head as a training tool of pronunciation for the development and acquisition of unaspirated and aspirated consonants in Mandarin, in the second step, we conducted the eye-tracking study as Experiment 1 and then the pronunciation training study as Experiment 2 respectively. We sought to investigate the following four research questions:

1. During the learning process, do L2 learners of Mandarin show more interest in our virtual 3-D tutor than in videos of a real HF tutor?

2. When learning from our 3-D tutor with the transparent profile view, do L2 learners of Mandarin distribute some of their visual attention to non-native visual cues from the airflow model?

3. Is our 3-D tutor with an airflow model more effective in enhancing the production accuracy of Mandarin stops and affricates?

4. Do the advantages of pronunciation training with the airflow model generalize to novel syllables with a change of tones or rimes?

2 Experiment 1: Eye-Tracking Study[1]

Experiment 1 was implemented to answer the first two research questions. By adopting the eye-tracking methodology, objective and quantitative measures were designed and provided to assess learners' acceptance and attention preference for various visual cues in our multimodal 3-D talking head. We tried to determine whether this virtual 3-D talking head is effective in training of pronunciation for non-native learners from various native language backgrounds.

2.1 Methods

2.1.1 Participants

Fifteen non-Chinese learners (nine males; mean age = 27.93 year, SD = 4.61) were recruited to take part in this eye-tracking study. They came from a broad range of native language backgrounds, but all their L1 consonant systems lack the distinctive feature of unaspirated vs. aspirated contrast (see Table 1). All the participants were overseas students studying in the first year in China. They were all beginners, having been learning Mandarin as an L2 in a classroom learning setting for less than three months, and had no experience of learning languages with a computer-animated 3-D virtual talking head. All participants had normal hearing and normal or corrected-to-normal visual acuity. Approval of the study was granted by the Behavioral Research Ethics Committee of <institution redacted for peer review> , and the consent form was obtained from each participant.

Table 1

Characteristics of non-Chinese participants in the eye-tracking study

No.	Gender	Age	Country	First Language	Dominant Phonetic

[1] Parts of the first eye-tracking data have been presented in the 10th International Symposium on Chinese Spoken Language Processing (ISCSLP 2016).

		(years)			Feature of L1
1	M	36	Bulgaria	Bulgarian	Unaspirated
2	M	33	Japan	Japanese	Unaspirated
3	M	30	Japan	Japanese	Unaspirated
4	F	24	Japan	Japanese	Unaspirated
5	F	25	Japan	Japanese	Unaspirated
6	M	29	Pakistan	Urdu	Unaspirated
7	M	26	Pakistan	Urdu	Unaspirated
8	M	26	Pakistan	Urdu	Unaspirated
9	M	30	Pakistan	Urdu	Unaspirated
10	M	31	Pakistan	Urdu	Unaspirated
11	M	32	Egypt	Arabic	Unaspirated
12	F	32	Philippines	Filipino	Unaspirated
13	F	23	Philippines	Filipino	Unaspirated
14	F	20	Russia	Russian	Unaspirated
15	F	22	Germany	German	Unaspirated

2.1.2 Apparatus for eye tracking

Eye-movement data were collected using the RED 5 Eye Tracker non-intrusively in the CAPT context. RED 5 was integrated into a 22-inch TFT monitor with a resolution of 1,280 × 1,024 pixels. An accuracy standard of 0.4° and a sampling rate of 60 Hz were used. The freedom of head movement allowed 40 cm × 20 cm at around 70 cm distance. Eye-tracking data were obtained with Experiment Center 2.0 online. This eye-tracking system also provided data extensions that could be processed further by the SMI BeGaze analysis software.

2.1.3 Materials

The 16 frequently used monosyllables in Mandarin serve as the learning materials: *bu* [pu], *pu* [p^hu], *bo* [po], *po* [p^ho], *ga* [ka], *ka* [k^ha], *de* [tɤ], *te* [t^hɤ], *ji* [tɕi], *qi* [tɕ^{h}i], *ju* [tɕy], *qu* [tɕ^{h}y], *zi* [tsɿ], *ci* [ts^{h}ɿ], *zhi* [tʂɻ], *chi* [tʂ^{h}ɻ]. These monosyllables were all composed of unaspirated/aspirated Mandarin consonants and eight basic monophthongs, and were superimposed with Tone 1 (the high-level tone). These 16 monosyllables were presented in two different conditions: HF and 3-D, resulting in 32 videos in total. All the HF or 3-D videos were displayed in a front view in the first half of each video and then in a corresponding profile transparent view. Each HF video was time-aligned with the corresponding 3-D video. The same female teacher for both HF and 3-D tutors

uttered all syllables, with the volume being fixed at 75 dB SPL.

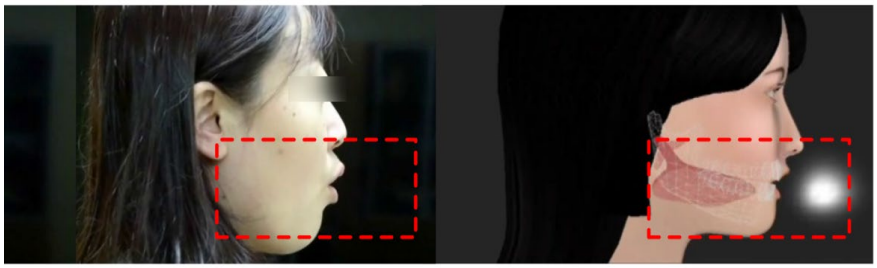
2.1.4 Procedure

Firstly, the non-native learners of Mandarin were given time to familiarize themselves with the CAPT setting. They were asked to fix their chins upon a support frame to minimize head movement, and to focus on the computer screen presenting videos of two Mandarin monosyllables (*ge* [kɤ] and *ke* [k^hɤ], which were not included in the testing syllables) under both HF and 3-D presentation conditions. The learners were asked to keep still while they watched the computer-animated videos, to pay attention to what they saw and heard, and to imitate the corresponding pronunciation. Prior to testing, eye-movement data were calibrated with five fixation points, and recalibration was conducted when calibration results were poor or missing. Next, during the formal test, the 32 videos were played twice (64 videos in total) and in random order to L2 learners.

2.1.5 Eye-Movement Data Analysis

In the front view, the area containing lip movement was marked as the area of interest (AOI). In the profile view, the area containing internal articulator, month corner, and expiratory airflow in the 3-D tutor was chosen as the AOI closely related to pronunciation. The same areas were marked in both HF and 3-D videos (see Figure 3 for more detail). The AOIs were marked manually offline in the BeGaze software. In the analysis of eye-movement data, three eye-tracking parameters for AOI were analyzed and computed: 'entry time' (ET), 'fixation count' (FC), and 'proportion of fixation duration' (POFD). Table 2 presents these three eye-tracking measures and their corresponding cognitive processes.

(a) Front View



(b) Profile View

Figure 3. The AOIs in front view (a) and profile view (b) refer to the specific visual areas inside the dashed red rectangle, which are closely related to speech sound production (with syllable *po* [$p^h o$] as an example).

Table 2

Three eye-tracking measurements and their cognitive processes

Measures	Description	Cognitive processes
Entry Time (ET)	The duration from the start of the trial to the first hit of the AOI	The shorter the ET, the greater the interest in the AOI
Fixation Count (FC)	The total number of fixations lasting more than 100 ms inside the AOI	The absolute visual attention to the AOI during learning, reflecting the depth of learning
Proportion of Fixation Duration (POFD)	The ratio of fixation duration inside the AOI to the duration of the whole video screen	The relative visual attention to the AOI during learning, reflecting relative learning time

2.2 Results

To answer the first research question, the first eye-tracking parameter (ET) was computed (see Figure 4) to indicate whether L2 learners showed interest in the pronunciation-related visual cues presented by our 3-D tutor. For the statistical analysis, ET was compared via repeated-measures analysis of variance (ANOVA) with the R (R Development Core Team, 2017) package of *ez*, with *presentation condition* (3-D and HF) and *view* (profile and front) as two within-subject factors. The ANOVA analysis revealed a significant main effect for *presentation condition* ($F(1, 14) = 33.81$, $p < 0.001$, $\eta_p^2 = 0.71$), while the main effect for *view* was not significant ($F(1, 14) = 0.83$, $p = 0.38$, $\eta_p^2 = 0.06$). No significant interaction between *presentation condition* and *view* was found ($F(1, 14) = 3.30$, $p = 0.09$, $\eta_p^2 = 0.19$). The main effect of *presentation condition* and lack of interaction between *presentation condition* and *view* indicate a shorter ET into the AOI of our virtual 3-D tutor than that of the HF tutor (see Figure 4). These eye-tracking results indicate that L2 learners of Mandarin read the visual cues of lip movement in our virtual 3-D talking head more efficiently and were also interested in the dynamic changes of additional visual cues despite the fact that they were virtual and uncommon.

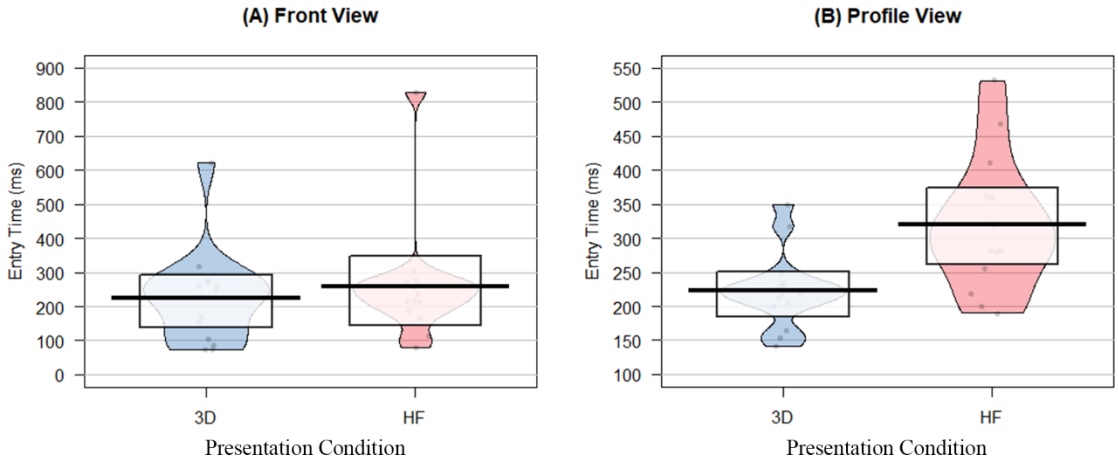


Figure 4. The entry time (ET) into AOIs under 3-D and HF presentation conditions with (A) front view and (B) Profile view.

In order to answer the second research question, the AOI was further separated into two subareas of the same size in a transparent 3-D profile view: an internal articulator exhibiting places of articulation was incorporated in Subarea 1, and the expiratory airflow showing manners of articulation was contained in Subarea 2. The parameters of FC and POFD were calculated to implicate the distribution of absolute and relative attention that had been paid to the two subareas. In profile view, the mean FC was around 2.92 for Subarea 1, and 2.65 for Subarea 2. The average FC of the two subareas did not differ in an independent samples *t*-test ($t(28) = 0.94$, $p = 0.36$). Moreover, the POFD toward Subarea 1 and Subarea 2 was around 45.22% and 50.25%, respectively, indicating similar relative attention paid to both subareas ($t(28) = -1.05$, $p = 0.31$). These results indicate that while watching a profile view of our 3-D tutor, L2 learners of Mandarin paid almost equal absolute and relative attention to non-native visual cues from the airflow model and the internal articulator. We can also see vividly and intuitively from the heat map (Figure 5) that the non-native learners' visual attention was allocated mainly to the areas overlapping with the whole AOI in the profile view of the 3-D tutor.

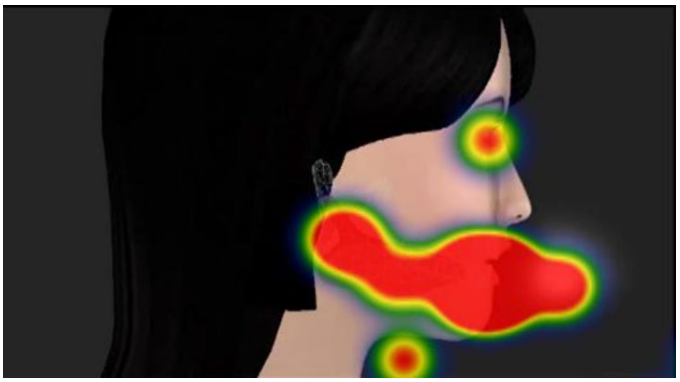


Figure 5. Heat map of one testing syllable in the 3-D tutor in profile view.

3 Experiment 2: Training Study

The results of the first eye-tracking experiment imply that our virtual 3-D talking head was successful in attracting L2 learners' attention to the airflow model marking the aspiration contrast. To further test the effectiveness of the airflow model, each Japanese learner of Mandarin was given a training package in Experiment 2. We compared the learning performance of two groups of Japanese learners of Mandarin learning with 3-D tutors with and without the airflow model respectively. Experiment 2 was thus devised to answer the third and fourth research questions: Whether the visual cues of the airflow model can produce more beneficial outcomes for the learning of both trained consonants and those in novel contexts.

3.1 Methods

3.1.1 Participants

In total, fifty Japanese adult learners of Mandarin (13 males, mean age = 20.43 years) volunteered to participate in the pretest. They had been studying at Beijing Language and Culture University or Nankai University for at least ten months in a classroom setting and had all reached intermediate level of HSK (HSK refers to *Hanyu Shuiping Kaoshi*, an official Chinese language proficiency test). Forty-two native Mandarin speakers were also recruited as a control group (9 males, mean age = 22.71 years). Exclusion criteria for all 92 participants included a history of hearing or visual impairment. During data collection, all subjects were free of colds or voice problems. Approval of the study was granted by the Behavioral Research Ethics Committee of <institution redacted for peer review>, and a written consent form was signed from each participant.

After the pretest, the Japanese learners were divided into two subgroups and invited to participate in a pronunciation training program. Group 1 contained 25 participants (6 males, mean age = 21.84

years) learning with a 3-D tutor with the airflow model. Group 2 was comprised of another 25 participants (7 males, mean age = 20.02 years) learning with a 3-D tutor without the airflow model.

3.1.2 Materials and Procedure

In the pretest, all the Chinese and Japanese subjects were asked to produce three sets of Mandarin monosyllable tokens, with each set involving six minimal pairs of unaspirated vs. aspirated contrast (see Appendix A for more details). The first set contained 12 tokens superimposed with Mandarin high-level tone, which were further utilized as the training syllables. The second and third sets of tokens contained novel syllables with different tones or rimes respectively, which were not part of the training materials. Since all the participants had learned the *Pinyin* system (an alphabetic phonological coding system widely adopted in Mainland China) in college, they could spontaneously produce all the tokens written in *Pinyin* on a sheet of paper. All the native and Japanese participants were required to produce three sets of tokens in isolation at a natural speaking rate. They were allowed to utter each token more than once, until a clear pronunciation was obtained. The speech samples were recorded using *Cool Edit* software (22,050 Hz sampling rate, 16-bit resolution) in quiet condition.

After the pretest, two subgroups of Japanese learners continued to attend a training program. Group 1 (n = 25) learned the 12 training syllables in the first set with a 3-D tutor with the airflow model five times per day, while Group 2 (n = 25) learned the same materials with a 3-D tutor but without the airflow model also five times per day. During training, the six minimal pairs were played randomly by computer software, while the two videos within a minimal pair were played next to each other to contrast aspiration differences. None of the Japanese learners were told about the underlying linguistic meanings of the visual content. They were only required to concentrate on the videos and to try to follow what they saw and heard. Right after the fifth day of training, all the Japanese learners were required to perform a posttest, with a pronunciation task similar to that in the pretest.

3.1.3 Data Analyses

The VOTs (in milliseconds) of all the speech samples produced in pretest and posttest were calculated and analyzed using *Praat* software. Figure 6 shows examples of how to obtain the VOT

segments of both stops and affricates. Since Mandarin stops and affricates are voiceless consonants, speech samples produced by native controls showed voicing lag, and the VOTs of consonants all had positive values. However, thirteen trials of 'voiceless unaspirated' consonants (7 in pretest and 6 in posttest) produced by 5 Japanese learners (3 learners from Group 1, and 2 leaners from Group 2) were realized as 'voiced unaspirated' counterparts (reflected by negative VOTs). Such a small proportion of samples (13 out of 3600) was excluded from further data analysis.

Given the production data showed great individual differences across subjects and trials, linear mixed-effect models (LMMs) in R (R Development Core Team, 2017) were adopted to evaluate the VOTs with 'subject' and 'trial' as two random factors, and with 'syllable duration' as the control factor. The package of lme4 (Bates et al., 2014) was used to create the LMMs. By-participant and by-trial random intercepts and slopes for all possible fixed factors were contained in the initial model. Model comparisons using the ANOVA function in lmerTest package was performed to obtain the main and interaction effects (Kuznetsova et al., 2017). The Akaike Information Criterion (AIC) was calculated to observe whether the chosen fixed factor contributed to an improved fit for the constructed model, and the model with the lowest AIC was selected to estimate the significance of the fixed factors. Post-hoc pairwise comparisons were conducted with the lsmeans package (Lenth, 2016) with Bonferroni adjustment.

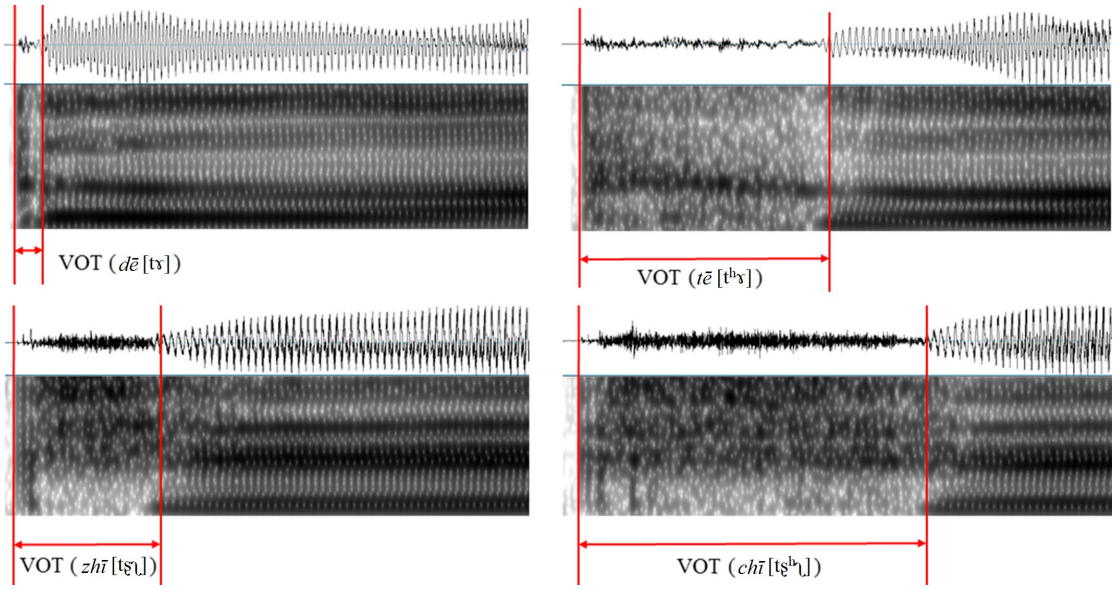


Figure 6. Praat display of VOT segments of stops and affricates in the beginning part of syllables. The VOTs are highlighted between the two vertical red lines.

3.2 Results

Table 3 compares the VOTs uttered by Japanese learners and native Mandarin speakers in the pretest before training. The visual inspection of Q-Q plots and plots of residuals showed no obvious deviations from homoskedasticity. In the LMM model, *consonant type* (2: stop and affricate), *aspiration condition* (2: unaspirated and aspirated), *language group* (2: Japanese and Mandarin), and all possible interactions were treated as fixed factors. LMM analysis revealed significant main effects of *consonant type* ($F = 20.13$, $p < 0.001$), *aspiration condition* ($F = 46.53$, $p < 0.001$), and *language group* ($F = 109.61$, $p < 0.001$) on VOT. Importantly, the interaction effects of *language group* by *consonant type* ($F = 55.11$, $p < 0.001$) and *language group* by *aspiration condition* ($F = 1364.68$, $p < 0.001$) were also significant. In terms of *language group*consonant type* interaction, post-hoc analyses demonstrated that the VOTs of stops produced by both Japanese and Mandarin were significantly shorter than those of affricates as expected (both ps < 0.001). For the *language group*aspiration condition* interaction, both Mandarin and Japanese groups produced aspirated consonants much longer than the unaspirated ones (both ps < 0.001). Moreover, the VOTs of unaspirated consonants produced by Japanese learners (mean = 50.9 ms) were similar to those produced by native Mandarin speakers (mean = 53.2 ms) ($\beta = -2.36$, $SE = 2.85$, $t = -0.83$, $p = 0.41$). However, Mandarin native speakers produced much longer VOTs of the aspirated consonants (mean = 153.8 ms) compared with the intermediate-level Japanese learners of Mandarin (mean = 98.4 ms) ($\beta = -55.43$, $SE = 2.85$, $t = -19.43$, $p < 0.001$).

Table 3

The obtained mean VOTs in ms and corresponding standard deviations in parentheses, produced by Japanese learners (n = 50) and Mandarin native speakers (n = 42) in the pretest. Note. Unasp. = Unaspirated, Asp. = Aspirated.

Consonants in Syllables		First Set: Trained Syllables		Second Set: Tone Variations		Third Set: Rime Variations	
		Japanese	Mandarin	Japanese	Mandarin	Japanese	Mandarin
Unasp. Stops	*b* [p]	22.3(9.8)	22.5(8.6)	24.1(10.1)	26.5(12.4)	17.1(4.7)	11.6(4.6)
	d [t]	24.7(7.8)	16.8(5.8)	25.6(6.8)	16.9(5.3)	24.6(6.0)	17.3(6.8)

	g [k]	29.4(7.7)	24.6(7.1)	28.9(9.0)	21.7(6.8)	37.8(10.6)	41.9(12.9)
Unasp. Affricates	*j* [tɕ]	77.9(14.4)	90.8(30.8)	82.4(18.6)	99.7(25.9)	69.2(17.5)	81.3(18.5)
	z [ʦ]	80.9(23.6)	90.9(30.0)	84.0(19.7)	95.2(27.0)	62.6(13.1)	72.0(20.6)
	zh [tʂ]	77.3(19.6)	82.1(22.6)	89.3(13.9)	85.5(26.1)	60.1(13.3)	60.6(18.8)
Asp. Stops	*p* [p^{h}]	56.5(23.4)	126.2(22.8)	57.4(21.6)	125.5(24.1)	48.4(24.6)	111.2(25.6)
	t [t^{h}]	67.3(26.1)	112.7(25.1)	63.6(24.4)	113.0(24.1)	71.2(26.8)	122.3(22.0)
	k [k^{h}]	84.9(26.7)	118.6(23.4)	92.7(26.4)	130.5(24.0)	94.7(31.4)	135.7(30.0)
Asp. Affricates	*q* [tɕ^{h}]	133.3(34.2)	198.8(37.6)	159.8(34.5)	215.0(42.3)	106.5(29.8)	167.8(35.0)
	c [ʦ^{h}]	132.4(32.7)	199.7(41.9)	120.4(29.4)	189.8(32.6)	92.9(23.8)	155.8(39.9)
	ch [tʂ^{h}]	136.9(33.1)	184.7(35.7)	154.5(33.8)	201.3(35.1)	97.0(25.6)	159.6(28.5)

It can be drawn from the pretest results that the intermediate-level Japanese learners showed the trend to produce Mandarin aspirated consonants with much shorter VOTs, but had no problem in producing unaspirated Mandarin consonants with normal VOT values. After pronunciation training (see Figure 7), the training does not change the VOTs of unaspirated consonants for both groups (all *p*s > 0.05) (Figure 7A, 7B). However, the VOTs of aspirated consonants turned to be much longer in posttest in comparison to those in pretest for both Group 1 (t (898) = -6.18, $p < 0.001$) and Group 2 (t (898) = -2.17, $p < 0.05$) (Figure 7C, 7D). Then, the gain (i.e., enhancement) from pretest to posttest in aspirated consonants was further calculated by the subtraction of each Japanese learner's VOT in pretest from the posttest VOT: Gain value of VOT = posttest VOT – pretest VOT.

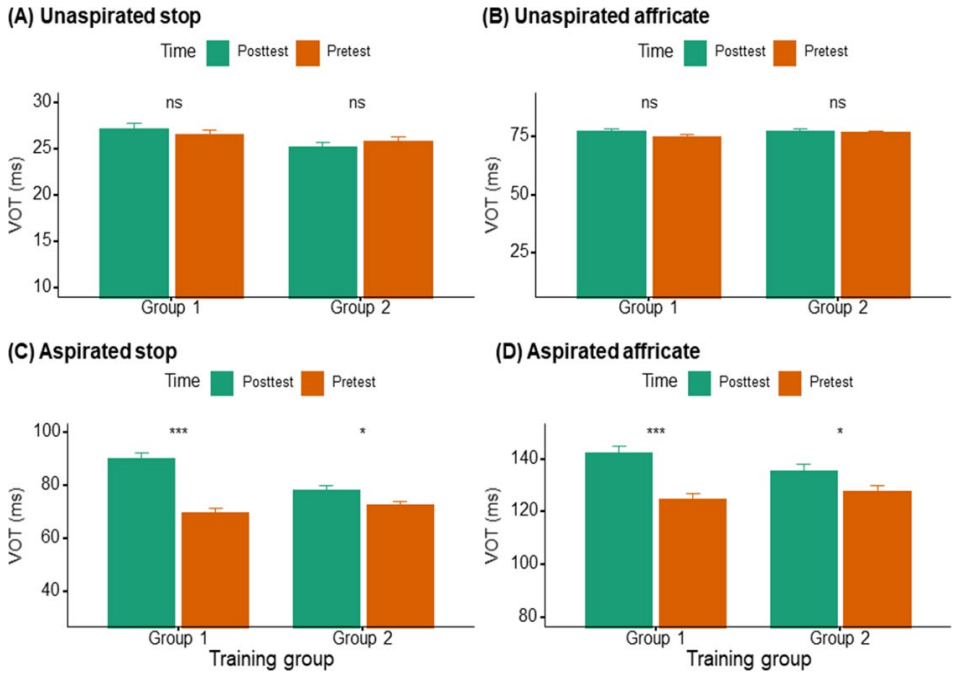


Figure 7. The mean VOTs of (A) unaspirated stop, (B) unaspirated affricate, (C) aspirated stop, and (D) aspirated affricate in pretest and posttest for two training groups: Group 1 (learning from 3-D tutors with an airflow model) and Group 2 (learning from 3-D tutors without an airflow model). Error Bars: ± 1 SE.

To further compare the training efficacy across different training groups and token sets (see Figure 8), another LMM was generated for the gain value of VOT of aspirated stops and affricates. In the LMM, *training group* (2: Group 1 vs. Group 2), *token set* (3: first set, second set, and third set), and *consonant type* (2: aspirated stop vs. aspirated affricate), and interactions among these variables were entered as fixed effects; subject and trial were entered as random effects. The LMM revealed significant main effects of *token set* ($F = 22.70$, $p < 0.001$) and *training group* ($F = 15.87$, $p < 0.001$). However, the main effect of *consonant type* and all possible two-way and three-way interactions were not significant (all *p*s > 0.05). For the main effect of *token set*, pairwise comparisons showed that, for both training groups, the gain values of VOT in the training materials (i.e., the first set) were much higher than those in novel materials with a change of tones ($\beta = -9.75$, $SE = 1.80$, $t = -5.43$, $p < 0.001$) or rimes ($\beta = -10.69$, $SE = 1.80$, $t = -5.95$, $p < 0.001$), while there was no difference between two novel token sets ($\beta = -0.94$, $SE = 1.80$, $t = -0.52$, $p = 0.86$). For the main effect of *training group*, Group 1 elicited higher gain values of VOT of aspirated stops and affricates than Group 2 ($\beta = 12.30$, $SE = 3.16$, $t = 3.90$, $p < 0.001$), indicating that our 3-D tutor with an additional

airflow model was better at helping the Japanese learners enhance the production accuracy of the aspirated Mandarin consonants embedded in both trained syllables and two types of novel syllables (see Figure 8).

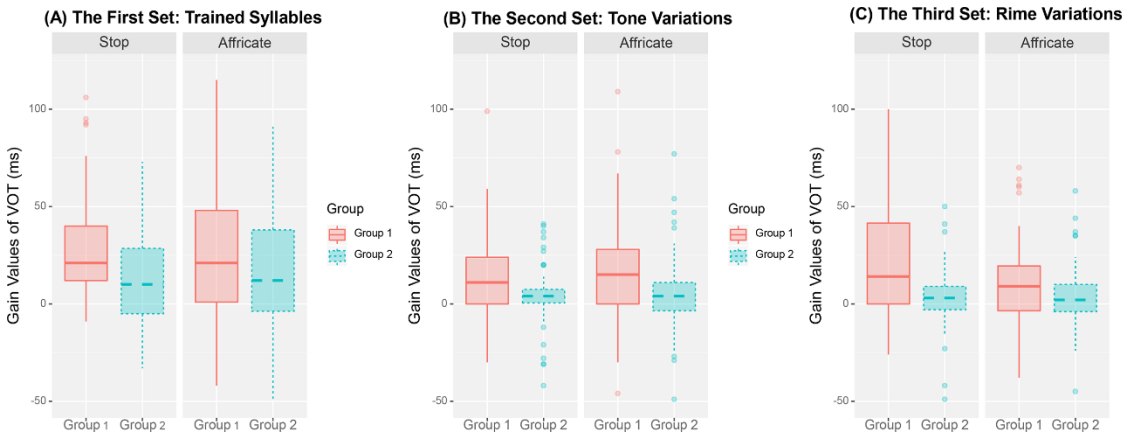


Figure 8. Box plots of gain values of VOT in aspirated stops and affricates across three token sets spoken by two training groups: Group 1 (learning from 3-D tutors with an airflow model) and Group 2 (learning from 3-D tutors without an airflow model). The bold line inside the boxes marks the median, and the upper and lower boundaries of the box correspond to its upper and lower quartiles.

4 General Discussion

This first eye-tracking study investigated L2 learners' general acceptance and attention distribution online when learning from our virtual talking head. The human likeness of the virtual head and the way in which the visual cues were presented may have affected the extent to which learners' attention was drawn on the relevant visual cues. The 'information processing model' (Segalowitz, 2003) indicates that at first, L2 learners must concentrate on every aspect of the language that they are attempting to understand or produce. Thus, the talking head system should be successful both in attracting L2 learners' attention and efficiently directing their attention to the visual areas related to speech sound production. However, most studies focus on the technical implementation of virtual 3-D talking heads and their animation, and few objective methods have been adopted to evaluate learners' acceptance of virtual 3-D talking heads online (Theobald et al., 2008).

The benefit of the eye-tracking approach lies in that it offers a stable flow of information about the

learner in real time, which can be collected to analyze and evaluate L2 learners' mental state and to penetrate where and how they are distributing their attention (Liu & Chuang, 2011). In this eye-tracking study, the proposed 3-D virtual talking head tutor was compared with a pronunciation video of a real human face. Eye-tracking results indicate that foreign learners were comfortable with our bio data-driven virtual tutor, which is fairly realistic, and that there was no 'uncanny valley effect' (Mori, 1970). Moreover, the L2 learners were more efficient in reaching the AOIs of our 3-D tutor in both front and profile views. Longer entry time into the lip movement area of the HF tutor may be due to learners' attention being distracted by the social eye region of the real human face. Furthermore, the AOIs of the 3-D tutor's profile view contain additional visual contents (e.g., internal articulator and airflow models) not presented by the HF tutor. Although these additional visual contents cannot be seen during daily communication, their physical motion feature synchronized with speech sounds successfully attracted L2 learners' interest and was well accepted by L2 adult learners.

The profile view of our virtual 3-D talking head offers simultaneous presentation of visual cues of both place contrast, reflected by the articulatory model, and aspiration contrast (i.e., manner contrast) indicated by the airflow model. Both models were considered to be necessary for Mandarin pronunciation training as they represent different aspects of visual transformations of acoustic features in the production of different Mandarin stops and affricates. The visual information of aspiration contrast in our airflow model can be regarded as a non-native visual cue for L2 learners with various L1s which show no distinctive feature of unaspirated vs. aspirated contrast. Hazan et al. (2005) indicated a perceptual deficit regarding the non-native visual cues of articulatory gestures, even when they are clearly different in a talking head. However, the current eye-tracking results show that foreign learners remained focus on the entire AOI with the airflow and internal articulators of our 3-D tutor, rather than exhibiting an attention bias toward the visual contents of internal articulators. This is probably due to the dynamic and vivid feature of the animations of airflow in our 3-D airflow model, which occupies an area away from the articulatory model and is sufficiently salient to draw L2 learners' attention. The visual attention of learners to our airflow model was crucial for language learning as the 'noticing hypothesis' (Robinson, 1995; Schmidt, 1990) suggests that the acquisition is not a result of the noticing itself; however, the noticing serves as a critical first step in acquiring a speech token. In keeping with the 'noticing hypothesis,' the novel linguistic

element represented by our airflow model, which is in the input, will not be transformed into intake unless L2 learners notice it, which gives rise to awareness.

We conducted another pronunciation training study in Experiment 2 among Japanese learners, to indicate whether visual attention to airflow animations can be further transformed into enhanced production of L2 Mandarin affricates and stops. In the pretest, the intermediate-level Japanese learners showed the tendency to produce aspirated Mandarin consonants with much shorter VOTs although they had been learning the target language for a long time. This inter-lingual error might be caused by a negative transfer from the Japanese 'aspirated allophones' which are much shorter in terms of VOT compared with Mandarin aspirated consonants, since the prior knowledge of L1 interferes with the acquisition of L2, especially when the two are phonetically and acoustically similar (Best & Tyler, 2007; Flege, 1995). Such an error pattern of 'deaspiration' tended to last long and even became fossilized, which may have prohibited the Japanese learners from further improving proficiency even with positive input. The acoustic feature of aspiration contrast of a minimal pair was presented with an airflow model for audiovisual training in one training group (Group 1). In contrast, the other training group (Group 2) learned the aspiration contrast only through audio input, since no such aspiration contrast can be detected visually in the 3-D articulatory model alone. Training results indicated that learning with the 3-D tutor with the airflow model, which utilized auditory-visual input, provided a greater boost in learning outcomes for Japanese learners in Group 1, and this training advantage generalized to novel tokens with tone or rime variations. Generalization of superior training outcomes to novel tokens can be a valuable target that carries skill improvement in speech production into new contexts.

The benefits of utilizing visual information relating to speech sound for audiovisual pronunciation training are consistent with previous studies (e.g., Kipp, 2001; Motohashi-Saigo & Hardison, 2009; Okuno & Hardison, 2016; Olson, 2014). In these studies, the visual cues (e.g., fundamental frequency, spectrograms, and waveforms) were often displayed through speech analysis software. The acoustic characteristics of speech elements were transformed into a visual demonstration, and this methodology was shown to be effective and useful for L2 listeners to perceive different acoustic features that are non-visible in phonemes. However, under these circumstances, L2 learners must acquire meta-linguistic knowledge about the acoustic characteristics of target speech sounds in advance. As indicated by the 'input hypothesis' (Krashen, 1985), learners must receive

comprehensible input in order to acquire an L2. Our talking head system thus showed advantages by directly simulating and exhibiting a biophysical airflow model for visualization during the process of speech production. The accurate display of airflow rate and duration of a minimal pair (aspirated vs. unaspirated contrast) in our airflow model may provide vivid visual cues showing L2 learners of Mandarin how quickly and how long airflow should be produced within a syllable. In this way, the 'audio-visual' mirror neurons (Rizzolatti and Craighero, 2004) were activated to accelerate intended learning of imitation from the airflow model in a talking head. To make it authentic, the real data-driven articulatory and aspiratory models must be synchronized in our 3-D talking head system. This is one of the reasons why such a complex and realistic 3-D airflow model was constructed in the present study.

This study has several limitations. Firstly, the current study though provided evidence for the superiority of audiovisual pronunciation training by using a 3-D airflow model, the duration for training was relatively short, which could partly explain the limited gain values of VOT even for the aspirated consonants within the trained syllables in Group 1 (mean: 27.1 ms). Such small gain values of VOT in the acoustic level might not necessarily lead to an enhancement in the perceptual identification. Secondly, we were still not sure whether the efficacy of our training methodology would last long since only the immediate posttest was conducted in this study. Future studies are needed to perform a longer training package, and to test whether our training advantage survives with a delayed posttest.

5 Conclusion and Implications of Pedagogy

In recent years, the research into the acquisition of non-native speech sounds using a virtual talking head has consistently shown the benefit of providing additional visual cues to facilitate pronunciation training. In the current study, both eye-tracking methodology and the pronunciation training approach were used to objectively appraise the proposed airflow model, in order to determine whether 3-D articulatory and aspiratory animations are helpful in improving the level of learners' interest and the production of aspirated vs. unaspirated Mandarin consonants. Viewed from the first eye-tracking study, it is suggested that our virtual 3-D talking head was well accepted by L2 learners, and effectively demonstrated supplementary visual information of airflow changes during speech production. Audiovisual pronunciation training with an airflow model further

contributed to greater improvement in the pronunciation of the sounds being trained and in novel syllables. To conclude, by effectively presenting visual airflow information, the current 3-D articulatory model with aspiratory animations provides an ideal pronunciation training method for L2 learners of Mandarin.

Generating synthetic and dynamic virtual talking head can offer a different mode of pronunciation training and provide L2 learners with one-to-one and face-to-face instruction repeatedly besides the classroom learning environment, thereby making a significant difference from traditional methods of language learning. The learning mechanism underlying this approach was based on 'imitation learning', which depends on the direct and accurate presentation of acoustic features from a talking head in the CAPT system. Pronunciation training with an airflow model embedded in a talking head requires little linguistic knowledge in advance, making the method suitable for young children or individuals with cognitive disorders. Moreover, visualization of the current airflow system may be effective in helping children with hearing loss to conceptualize airflow changes during speech sound production.

Finally, the traditional classroom teaching approach, in which one puts one's hand in front of one's mouth when producing unaspirated vs. aspirated Mandarin consonants, might help learners to sense airflow changes, but this approach of utilizing tactile modality shows much coarser temporal and spatial resolution. Our airflow model was based on multi-speaker airflow recordings during pronunciation, with several key bio-signal parameters (such as mean airflow rate, peak airflow rate, peak time, and airflow duration) extracted. These parameters were essential for accurate airflow modeling to capture precise visual differences between unaspirated vs. aspirated Mandarin consonants. Consequently, our airflow model offers language learners a more accurate visual exhibition of airflow changes during speech sound production to a great extent.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon request.

References

Abramson, A. S., & Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics, 63*, 75–86.

Badin, P., Elisei, F., Bailly, G., & Tarabalka, Y. (2008). An Audiovisual Talking Head for Augmented Speech Generation: Models and Animations Based on a Real Speaker's Articulatory Data. In F. J. Perales & R. B. Fisher (Eds.), *Articulated Motion and Deformable Objects* (pp. 132–143). Springer.

Bälter, O., Engwall, O., Öster, A.-M., & Kjellström, H. (2005). Wizard-of-Oz test of ARTUR: A computer-based speech training system with articulation correction. *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*, 36–43.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Prepr. arXiv1406.5823*.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities, In *Language Experience in Second Language Speech Learning: In Honor of James Flege*, edited by M. J. Munro and O.-S. Bohn (John Benjamins, Amsterdam), pp. 13–34.

Brown, C. A. (1998). The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research*, 14(2), 136-193.

Chen, F., Chen, H., Wang, L., Zhou, Y., He, J., Yan, N., & Peng, G. (2016). Intelligible enhancement of 3D articulation animation by incorporating airflow information. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6130–6134).

Chen, N. F., Shivakumar, V., Harikumar, M., Ma, B., & Li, H. (2013). Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages. In *Proceedings of Interspeech 2013* (pp. 2370–2374).

Chen, N. F., Tong, R., Wee, D., Lee, P., Ma, B., & Li, H. (2015). iCALL corpus: Mandarin Chinese spoken by non-native speakers of European descent. In *Proceedings of Interspeech 2015* (pp. 324–328).

Chen, T. H., & Massaro, D. W. (2011). Evaluation of synthetic and natural Mandarin visual speech: Initial consonants, single vowels, and syllables. *Speech Communication, 53*(7), 955–972.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal*

of Phonetics, 27(2), 207–229.

Chun, D. (1998). Signal analysis software for teaching discourse intonation. *Language Learning & Technology*, 2(1), 74-93.

Colpaert, J. (2012). The ‘‘Publish and Perish’’ syndrome. *Computer Assisted Language Learning*, 25(5), 383–391.

Colpaert, J. (2016). Big content in an educational engineering approach. *Journal of Technology and Chinese Language Teaching*, 7(1), 1–14.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research, 92*, 233–277.

Felix, U. (2005). Analysing recent CALL effectiveness research – Toward a common agenda. *Computer Assisted Language Learning*, 18(1–2), 1–32.

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports and vocabulary learning scores. In J. M. Bergsleithner, S. N. Frota, J. K. Yoshioka, & R. W. Schmidt (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 183–205). Honolulu: National Foreign Language Resource Center, University of Hawaii.

Golonka, E.M., Bowles, A.R., Frank, V.M., Richardson, D.L., & Freynik, S. (2012). Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer Assisted Language Learning*, 27(1), 70-105.

Graeme, C. (2006). The short and long-term effects of pronunciation instruction. *Prospect, 21*(1), 46–66.

Haslam, N. O. (2010). *The relationship of three L2 learning factors with pronunciation proficiency: Language aptitude, strategy use, and learning context* (Unpublished master dissertation). Brigham Young University, Provo, UT.

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication, 47*(3), 360–378.

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL*, 15(1), 3-

20.

Hubbard, P. (2005). A review of subject characteristics in CALL research. *Computer Assisted Language Learning*, 18, 351–368.

Kipp. M. (2001). Anvil: A generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology* (pp. 1367–1370). Aalborg, Denmark: Eurospeech.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. London: Longman.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13).

Lai, Y. H. (2009). Asymmetry in Mandarin affricate perception by learners of Mandarin Chinese. *Language and Cognitive Processes, 24*(7–8), 1265–1285.

Lam, H. C., Ki, W. W., Law, N., Chung, A. L. S., Ko, P. Y., Ho, A. H. S., & Pun, S. W. (2001). Designing CALL for learning Chinese characters. *Journal of Computer Assisted Learning*, 17(1), 115-128.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 69(1), 1–33. https://doi.org/10.18637/jss.v069.i01

Li, H., Yang, M., & Tao, J. (2013). Speaker-independent lips and tongue visualization of vowels. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8106–8110).

Liu, H. C., & Chuang, H. H. (2011). An examination of cognitive processing of multimedia information based on viewers' eye movements. *Interactive Learning Environments, 19*(5), 503–517.

Liu, X., Yan, N., Wang, L., Wu, X., & Ng, M. L. (2013). An interactive speech training system with virtual reality articulation for Mandarin-speaking hearing impaired children. In *Proceedings of International Conference on Information and Automation* (pp. 191–196).

Massaro, D. W., & Palmer Jr, S. E. (1998). Perceiving talking faces: From speech perception to a behavioral principle. MA, USA: MIT Press.

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/–/l/ discrimination to Japanese adults: Behavioral and neural aspects. *Physiology & Behavior, 77*(4), 657–662.

Mori, M. (1970). The uncanny valley. *Energy, 7*, 33–35.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL quarterly*, 25(3), 481-520.

Motohashi-Saigo, M., & Hardison, D. M. (2009). Acquisition of L2 Japanese geminates: Training with waveform displays. *Language Learning & Technology, 13*(2), 29–47.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research, 71*(1), 4–12.

Okuno, T., & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology, 20*(2), 61–80.

Olson, D. J. (2014). Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning & Technology, 18*(3), 173–192.

Peng, X., Chen, H., Wang, L., & Wang, H. (2018). Evaluating a 3-D virtual talking head on pronunciation learning. *International Journal of Human-Computer Studies, 109*, 26–40.

Piper, T., & Cansin, D. (1988). Factors influencing the foreign accent. *The Canadian Modern Language Review, 44*(2), 334–342.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*(5), 169–192.

Robinson, P. (1995). Attention, memory, and the 'noticing' hypothesis. *Language Learning, 45*(2), 283–331.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158.

Segalowitz, N. (2003). Automaticity and second language learning. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Oxford: Blackwell.

Shei, C., & Hsieh, H. P. (2012). Linkit: a CALL system for learning Chinese characters, words, and

phrases. *Computer Assisted Language Learning*, 25(4), 319-338.

Si Y. Y. (2006). Mandarin phonological acquisition: A case study. *Contemporary Linguistics, 8*(1), 1–16.

Stam, J. (1999). Stable fluids. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 121–128).

Stevens, C. J., Gibert, G., Leung, Y., & Zhang, Z. (2013). Evaluation a synthetic talking head using a dual task: Modality effects on speech understanding and cognitive load. *International Journal of Human-Computer Studies, 71*(4), 440–454.

Stockwell, G. (2007). A review of technology choice for teaching language skills and areas in the CALL literature. *ReCALL*, 19, 105–120.

Tham, I., Chau, M. H., & Thang, S. M. (2019). Bilinguals' processing of lexical cues in L1 and L2: an eye-tracking study. *Computer Assisted Language Learning*, 1-23.

Theobald, B. J., Fagel, S., Bailly, G., & Elisei, F. (2008). LIPS2008: Visual speech synthesis challenge. In *Proceedings of Interspeech 2008* (pp. 2310–2313). Brisbane: ISCA.

Vance, T. J. (1997). *An introduction to Japanese phonology*. New York: SUNY Press.

Wang, L., Chen, H., Li, S., & Meng, H. M. (2012). Phoneme-level articulatory animation in pronunciation training. *Speech Communication*, 54(7), 845-856.

Wang, X., Hueber, T., & Badin, P. (2014). On the use of an articulatory talking head for second language pronunciation training: The case of Chinese learners of French. In *Proceedings of 10th International Seminar on Speech Production* (pp. 449–452).

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America, 113*(2), 1033–1043.

Wang, Y. H. (2016). Could a mobile‐assisted learning system support flipped classrooms for classical Chinese learning?. *Journal of Computer Assisted Learning*, 32(5), 391-415.

Wu, Z. Y., Zhang, S., Cai, L. H., Meng, H. M. (2006). Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar. *In Proceedings of INTERSPEECH-2006* (pp. 1802–1805).

Weiss, B., Kühnel, C., Wechsung, I., Fagel, S., & Möller, S. (2010). Quality of talking heads in different interaction and media contexts. *Speech Communication, 52*(6), 481–492.

Xie, H., Zhao, T., Deng, S., Peng, J., Wang, F., & Zhou, Z. (2021). Using eye movement modelling examples to guide visual attention and foster cognitive performance: A meta-analysis. *Journal of Computer Assisted Learning*, 37(4), 1194–1206.

Yu, J., & Li, A. (2014). 3D visual pronunciation of Mandarin Chinese for language learning. *2014 IEEE International Conference on Image Processing (ICIP)*, 2036–2040.

Zhang, H., & Roberts, L. (2019). The role of phonological awareness and phonetic radical awareness in acquiring Chinese literacy skills in learners of Chinese as a second language. *System*, 81, 163-178.

Zhang, D., Liu, X., Yan, N., Wang, L., Zhu, Y., & Chen, H. (2014). A multi-channel/multi-speaker articulatory database in Mandarin for speech visualization. In *Proceedings of ISCSLP 2014* (pp. 299–303).

Zhang, L. J. (2009). Perceptual training and the acquisition of Chinese aspirated/unaspirated consonants by Japanese students. *Language Teaching and Linguistic Studies, 4*, 560–566.

Zhao, Y. (2003). Recent developments in technology and language learning: A literature review and meta-analysis. *CALICO Journal*, 21(1), 7–28.

Zhu, H., & Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). *Journal of Child Language, 27*(1), 3–42.

Appendix A. Monosyllable tokens for pronunciation training study

First Set: Trained Syllables		Second Set: Tone Variations		Third Set: Rime Variations	
Unasp.	Asp.	Unasp.	Asp.	Unasp.	Asp.
bō [po1]	*pō* [p^{h}o1]	*bó* [po2]	*pò* [p^{h}o4]	*bā* [pa1]	*pā* [p^{h}a1]
dē [tɤ1]	*tē* [t^{h}ɤ1]	*dé* [tɤ2]	*tè* [t^{h}ɤ4]	*duō* [tuo1]	*tuō* [t^{h}uo1]
gā [ka1]	*kā* [k^{h}a1]	*gà* [ka4]	*kǎ* [k^{h}a3]	*gē* [kɤ1]	*kē* [k^{h}ɤ1]
jī [tɕi1]	*qī* [tɕ^{h}i1]	*jí* [tɕi2]	*qǐ* [tɕ^{h}i3]	*jiāo* [tɕiɑu1]	*qiāo* [tɕ^{h}iɑu1]
zī [tsɿ1]	*cī* [ts^{h}ɿ1]	*zǐ* [tsɿ3]	*cì* [ts^{h}ɿ4]	*zāi* [tsai1]	*cāi* [ts^{h}ai1]
zhī [tʂʅ1]	*chī* [tʂ^{h}ʅ1]	*zhǐ* [tʂʅ3]	*chí* [tʂ^{h}ʅ2]	*zhuān* [tʂuan1]	*chuān* [tʂ^{h}uan1]

Note. Unasp. = Unaspirated, Asp. = Aspirated. There are four Mandarin tones, traditionally named as Tone 1 (high-level tone), Tone 2 (rising tone), Tone 3 (dipping tone), and Tone 4 (falling tone). *Pinyin* and the corresponding IPA phonetic transcriptions in square brackets are provided.