Investigations on Mandarin Aspiratory Animations Using an Airflow Model

Fei Chen, Lan Wang, *Member, IEEE*, Hui Chen, *Member, IEEE*, and Gang Peng

***Abstract*—Various three-dimensional (3-D) talking heads have been developed lately for language learning, with both external and internal articulatory movements being visualized to guide learning. Mandarin pronunciation animation is challenging due to its confusable stops and affricates with similar places of articulation. Until now, less attention has been paid to the biosignal information of aspiratory airflow, which is essential in distinguishing Mandarin consonants. This study fills a research gap by presenting the quantitative analyses of airflow, and then designing an airflow model for a 3-D pronunciation system. The airflow information was collected by Phonatory Aerodynamic System, so that confusable consonants in Mandarin could be discerned by mean airflow rate, peak airflow rate, airflow duration, and peak time. Based on the airflow parameters, an airflow model using the physical equation of fluid flow was proposed and solved, which was then combined and synchronized with the existing 3-D articulatory model. Therefore, the new multimodal system was implemented to synchronously exhibit the airflow motions and articulatory movements of uttering Mandarin syllables. Both an audio-visual perception test and a pronunciation training study were conducted to assess the effectiveness of our system. Perceptual results indicated that identification accuracy was improved for both native and nonnative groups with the help of airflow motions, while native perceivers exhibited higher accuracy due to long-term language experience. Moreover, our system helped Japanese learners of Mandarin enhance their production skills of Mandarin aspirated consonants, reflected by higher gain values of voice onset time after training.**

***Index Terms*—Airflow, 3-D articulatory animations, confusable consonants, evaluation, second language learning.**

Manuscript received December 13, 2016; revised May 18, 2017 and August 9, 2017; accepted September 15, 2017. Date of current version November 27, 2017. This work was supported by the National Natural Science Foundation of China under Grants 91420301, 11474300, 61232013, and 61661146002, in part by the Shenzhen Fundamental Research Program (JCYJ20160429184226930), and in part by the National Key R&D Program of China under Grant 2016YFB1001405. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Thomas Hueber. *(Corresponding authors: Lan Wang; Gang Peng.)*

F. Chen and L. Wang are with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China (e-mail: chenfeianthony@163.com; lan.wang@siat.ac.cn).

H. Chen is with the Institute of Software, Chinese Academy of Sciences, Beijing 100864, China (e-mail: chenhui@iscas.ac.cn).

G. Peng is with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China, and also with the Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong (e-mail: gpengjack@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2755400

I. Introduction

With the development of speech technology, language learners have been paying more attention to Computer Assisted Pronunciation Training system (CAPT system). The activation of ‘audio-visual’ mirror neurons might explain human capacity to learn by imitation from multiple modalities [1]. As the carrier of audio-visual speech, various 3-D talking heads have been developed in CAPT systems by instructing learners to imitate from multimodal animations. It has been proved to reduce cognitive load through multimodal presentation [2], [3] and offers learners with one-to-one instruction.

To present more authentic articulatory animations, various physiological-data based talking heads have been devised and applied to enhance the language learning as a new mode (see Fig. 1). Among which, talking heads in (a) and (c) were developed for Mandarin visual speech [4], [5], (b) and (d) for English [6], [7], (e) for French [8], (f) for Swedish [9]. Based on MRI or CT images, physiology-based talking head models have been developed to represent human face, lips, tongue, and even to represent the human velum and nasopharyngeal wall [10]. In order to drive articulators to move, one approach is to utilize the acoustic-to-articulatory inversion model to recover articulatory movements directly from the speech audio signal [11], and this technique has been further applied to a virtual talking head with a speaker-adaptive way [12]. The other common method tries to animate both external and internal articulators in a real-data-driven articulatory talking head, using articulation data recorded through facial motion capture [13], Electro-Magnetic Articulography (EMA) [7], [14], [15], or video-fluoroscopic image [16].

Importantly, investigations on Mandarin visual speech potentially make unique contributions [17], since some kinematic characteristics and acoustic features of Mandarin phonemes are different compared with other languages. Besides the suprasegmental lexical tones, Mandarin phonology system has a variety of voiceless stops and affricates with similar places of articulation, which were difficult for non-Mandarin second language (L2) learners to acquire [18]–[20]. Based on a relatively large non-native speech corpus by L2 learners of Mandarin [21], it has been found that 6 out of the top 10 pronunciation error patterns are relevant to ‘de-aspiration’ or ‘aspiration’. The unaspirated and aspirated contrast plays a dominant role in telling apart the minimal pair of Mandarin confusable consonants (e.g., d [t] vs. t [tʻ]). As they show similar places of articulation, it is difficult to discriminate them only through the articulatory movements in all the existing 3-D Mandarin talking heads [4], [5], [22]–[25].

2329-9290 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

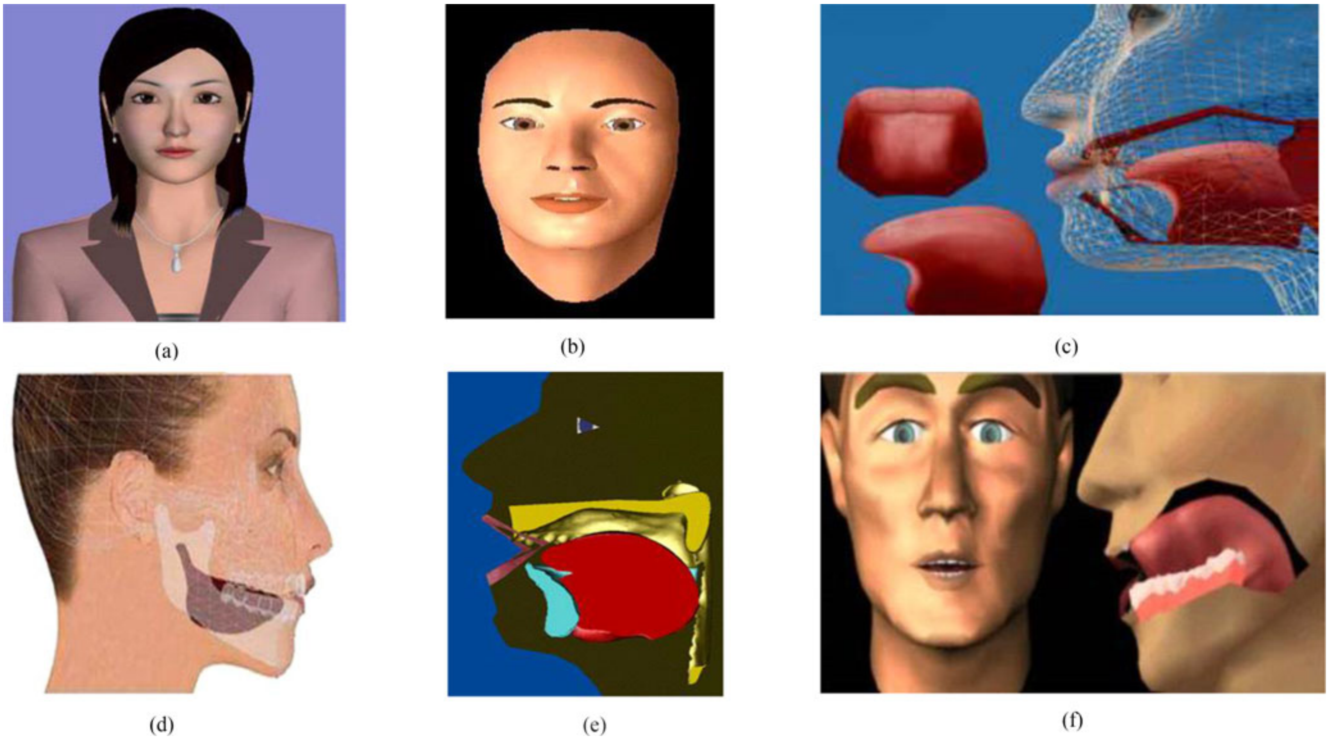

Fig. 1. Some examples of physiological-data based talking heads. Among which, talking heads in (a) and (c) were developed for Mandarin visual speech, (b) and (d) for English, (e) for French, (f) for Swedish.

Some previous studies have worked on optimizing visual speech through offering certain supplementary visual features [6], [26], [27]. For example, Cued Speech [26] was designed to supplement speech reading (i.e., lip reading) by including a hand gesture together with the mouth pattern. However, the mappings between gestures and their corresponding phonemes are always arbitrary. By obtaining a few robust acoustic features (including nasality, voicing, and frication) which can be mapped into supplementary visual cues, Massaro [6], [27] has successfully developed a near-real-time visual system to help disambiguate the spoken message and thus to enhance intelligibility. The supplementary visual characteristics were presented with various colored bars next to a talking head. The intensity of these colored cues varied according to the degree to which corresponding acoustic feature was present as the continuous speech signal unfolded. This methodology has been proved to be effective for listeners to perceive and detect different non-visible acoustic features of consonants to some extent.

Since the intended imitation learning from a talking head in CAPT system requires a more straightforward and vivid exhibition of key acoustic features, the current study endeavors to develop a biophysical model to show airflow visualization in accordance with articulation animations through a multimodal 3-D talking head. Although speech production requires exhalant airflow to be phonated through the vocal cavities, transient airflow changes are actually invisible during the process of pronunciation, and are immeasurable without the help of specific measuring instrument. Since few studies have focused on investigating the visualization of airflow model in a 3-D talking head, the current work aims to provide a clearer solution for this challenge, in order to improve the discriminability among confusable Mandarin voiceless consonants in a talking head, and further to apply this airflow system to teach Mandarin pronunciation among L2 learners.

The first focus of the present study is to record and analyze the physiological airflow data while producing Mandarin consonants, where relevant parameters should be extracted to exhibit aspiration differences among these confusable stops and affricates accurately. For individuals with pathological voices, the abnormal airflow control is one of the cardinal symptoms during articulation [28], [29], demonstrating that aerodynamic parameters of airflow rate could be used to discriminate normal from dysphonic voices. Recently, some works have already indicated the validity of collecting the aerodynamic characteristics of the voice by using *Phonatory Aerodynamic System* (PAS) Model 6600 [30], [31], which investigated the aerodynamic features of the voice within typical adult speakers. In these studies, multi-speaker airflow data were collected by extending the use of PAS Model 6600, since PAS system utilizes a convenient set of protocols based on typical aerodynamic tasks to minimize variability in the application, and provide airflow measure of speech and voice production to support evidence-based practice. From airflow recordings by PAS, key parameters can be extracted such as mean airflow rate, peak airflow rate, airflow duration, and peak time. It is expected that these parameters could be shown to capture differences between confusable Mandarin consonants, which are essential for further airflow modeling.

For airflow visualization, a suitable airflow model should be designed and incorporated into the existing articulatory model to produce the airflow motions in accordance with articulatory movements of Mandarin pronunciation. To make it authentic, the data-driven aspiratory animation and articulatory movements must be synchronized in a 3-D talking head system, during the production process of a syllable. The key problem is then to compute the control parameters with the use of PAS-extracted mean and peak airflow rate, and the peak time and duration are also vital to simulate the airflow model synchronized with the articulatory model. Although working on fluid control in graphics has been an active area in computer graphics [32]–[34], few approaches tried to make use of real physiological airflow data to stimulate and drive an airflow model. One study [35] has found that the flow direction and the area of mouth/nose opening did not change much during breathing, while the exhaled flow rate can be calculated and represented as a sinusoidal function with time. Furthermore, the real-time fluid dynamics [36], [37] on the basis of physical equations of fluid flow, likely fit better with our purpose. Since both forces and sources are used as important control variables to the dynamics, this physics-based model may enhance the visual effect by providing realistic fluid-like effects. In this way, a new multimodal visualization system is proposed, where 3-D articulatory movements and airflow motions would be animated simultaneously to instruct Mandarin pronunciation training as a second language.

The evaluation of our multimodal 3-D talking head for CAPT system is another point of this study, since a talking head system involved in any applications must undergo an evaluation [38]. The present study concentrated on the impact of adding a bio-data-driven airflow model in CAPT system for learning Mandarin consonants as an L2. The supplementary airflow information might be of great importance for audiovisual perceptual training among L2 learners of Mandarin. In this paper, both an audio-visual perception test and a pronunciation training study would be utilized to assess the proposed multimodal system, in order to figure out whether the supplementary airflow animations are effective to improve perception of Mandarin consonants for both native and non-native speakers and to enhance production of Mandarin stops and affricates among L2 learners of Mandarin.

The acquisition of Mandarin unaspirated vs. aspirated contrast seems pretty easy for typically-developing native speakers at an early age. However, learning speech sounds of an L2 can be effortful, if these speech sounds do not exist in the L1 or show a different phonological status between the two languages [39], [40]. For instance, in Japanese, the consonants are mainly distinguished by voiceless and voiced contrast, and ‘aspiration’ is not a distinctive feature in Japanese consonants. Some aspirated voiceless consonants in Japanese are actually allophones in complementary distribution with their unaspirated phonemes. The voice onset time (VOT, an objective and widely used acoustic measurement of stops and affricates [41]) of these Japanese aspirated allophones tends to be much shorter than the corresponding Mandarin aspirated phonemes [42], [43]. The average VOT of most Mandarin aspirated consonants is often longer than 100 ms, which can be classified as strongly aspirated type even among the world's languages [44]. For Japanese learners of Mandarin, as influenced by their native language, there is a high de-aspiration rate. They often produce the strongly aspirated Mandarin consonants as the weakly aspirated ones. Moreover, the de-aspiration phenomenon in producing Mandarin aspirated consonants by Japanese L2 learners lasts long and even becomes fossilized after years of learning [43]. The accurate display of airflow rate and duration between aspirated and unaspirated consonants may provide important visual cues to vividly teach Japanese learners how fast and how long the aspirated airflow should be produced within a syllable. In order to further testify the efficacy of our airflow model, a training package was given especially to Japanese learners of Mandarin.

In the following, Section II presents the airflow data collection of Mandarin consonants using PAS, and the related airflow data analyses. Section III realizes an airflow model that uses the airflow parameters for fluid control, and then incorporates the airflow model into an articulatory model. In Section IV, an audio-visual perception test and a pronunciation training study are shown, respectively. The discussions are shown finally in Section V.

II. Airflow Data Acquisition and Analyses

A. Participants, Materials, and Procedure

Starting from airflow data collection, a multi-speaker database was constructed. Eighteen healthy young adults (nine male) were recruited $(\mathrm{M} = 25.34\,\text{years}, \mathrm{SD} = 1.73)$, with Mandarin as their native language. All subjects had no history of significant voice problems, trauma/surgery to the oral cavity or larynx based on self-report questionnaires. During data collection, all participants were free from colds or seasonal allergies.

The pronunciation corpus was designed to contain 42 Mandarin syllables with high-level tone, combining 12 confusable Mandarin stops or affricates with 8 basic Mandarin monophthongs (i.e., [a] , [i] , [u], [ɤ], [o], [y], [ɿ], [ʅ]). Each subject was asked to repeat all the 42 syllables three times, with a totaling of 126 samples. The 12 confusable stops and affricates were divided into six minimal pairs, each sharing similar place of articulation and articulatory trajectory. There are three pairs of Mandarin voiceless stops (in terms of *Pinyin* and the corresponding IPA in square brackets): labial (Group 1: *b* [p] and *p* [p‘]), alveolar (Group 2: *d* [t] and *t* [t‘]), velar (Group 3: *g* [k] and *k* [k‘]), and three pairs of voiceless affricates: alveolar (Group 4: *z* [ʦ] and *c* [ʦ‘]), retroflex (Group 5: *zh* [tʂ] and *ch* [tʂ‘]), alveolo-palatal (Group 6: *j* [ʨ] and *q* [ʨ‘]). The two consonants within each minimal pair are mainly distinguished by the manner of articulation (i.e., the distinctive feature of unaspirated vs. aspirated contrast).

Before airflow data collection, the PAS flow-head and pressure transducer were carefully calibrated according to PAS protocol instructions. All participants were asked to pronounce all the testing syllables at comfortable pitch height and loudness to make sure that the speech samples were produced naturally. During airflow data collection, participants were instructed to

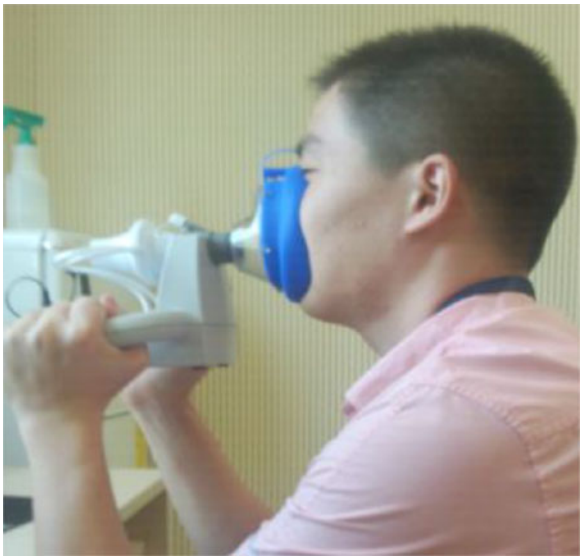

Fig. 2. The demonstration of airflow collection using PAS.

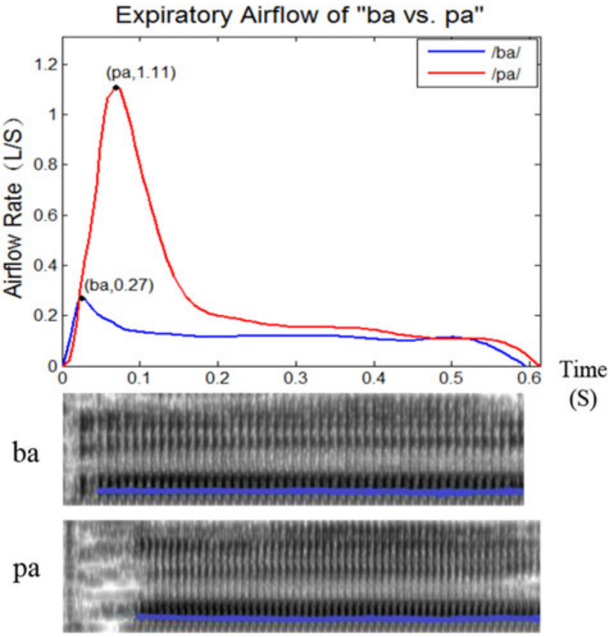

Fig. 3. The aspiratory airflow rate of “ba” vs. “pa”, and the corresponding spectrograms, with the blue line representing the F0 trajectory.

hold the PAS side handles and to firmly press the facemask over the face, so the areas of nose and mouth were both covered to stop airflow from escaping (see Fig. 2).

TABLE I
THE OBTAINED DATA OF MEAN AIRFLOW RATE, PEAK AIRFLOW RATE, AIRFLOW DURATION, AND PEAK TIME

Stops and Affricates within Syllable	Mean Airflow Rate vel_m (L/S)	Peak Airflow Rate vel_p (L/S)	Airflow Duration $t_e - t_s$(ms)	Peak Time t_p (ms)
“***b***” [p]	0.11	0.25	41.66	8.38
“***p***” [p‘]	0.46	0.81	94.44	22.94
“***d***” [t]	0.11	0.24	30.00	6.16
“***t***” [t‘]	0.56	0.90	83.33	20.04
“***g***” [k]	0.15	0.19	37.07	7.52
“***k***” [k‘]	0.50	0.84	95.18	22.77
“***j***” [ʨ]	0.13	0.24	73.88	14.83
“***q***” [ʨ‘]	0.34	0.57	160.00	38.66
“***z***” [ʦ]	0.11	0.26	83.61	16.83
“***c***” [ʦ‘]	0.39	0.82	160.69	38.62
“***zh***” [tʂ]	0.14	0.31	64.58	13.05
“***ch***” [tʂ‘]	0.49	0.92	133.61	32.31

B. *Airflow Data Analyses*

The airflow data were collected continuously within a Mandarin syllable with a sampling rate of 22 025 Hz. With the help of WaveSurfer software (version 1.8.5), the segment of voiceless consonant could be properly split away from the following monophthong. The segment of monophthong could be detected from the fundamental frequency (F0) superimposed on a vowel (see Fig. 3). What is more, we would double check through listening to the audio for the correctness validation of our segmentation. After manual segmentation, relevant airflow information of consonants was calculated in terms of four parameters: mean airflow rate (L/S), peak airflow rate (L/S), airflow duration (ms), and peak time (ms) (see Table I). The formula of average peak airflow rate for a consonant is:

$$average\ vel_p = \frac{1}{N \cdot K} \sum_{n=1}^{N} \sum_{k=1}^{K} vel_{p_{n\,k}} \qquad (1)$$

where N is the number of speakers, K is the times of recordings per speaker, and vel_p is the peak airflow rate. Fig. 3 shows the fluctuation of airflow rate and the corresponding spectrograms of a minimal pair pronounced by one subject.

Before airflow data analysis, the distribution normality was evaluated by using Shapiro-Wilk tests of normality. Moreover, the outliers (defined as over three standardized residuals away from predicted scores) were picked out with regression analysis, and were removed from further analysis. Table I shows the average airflow data of mean airflow rate (L/S), peak airflow rate (L/S), airflow duration (ms), and peak time (ms). Statistical analysis of one-way analysis of variance (ANOVA) revealed that mean airflow rate, peak airflow rate, airflow duration, and peak time of were all significantly different among different Mandarin consonants (all $ps < 0.001$). Afterwards, Tukey’s HSD post hoc pairwise comparisons of the minimal pair showed that mean airflow rate (Fig. 4(a)) and peak airflow rate (Fig. 4(b)) for “*p*” [p‘] was much faster than that for “*b*” [p]; “*t*” [t‘] > “*d*” [t]; “*k*” [k‘] > “*g*” [k]; “*c*” [ʦ‘] > “*z*” [ʦ]; “*ch*” [tʂ‘] > “*zh*” [tʂ]; “*q*” [ʨ‘] > “*j*” [ʨ] (all $ps < 0.05$). Moreover, pairwise comparisons of the minimal pair showed that airflow duration (Fig. 4(c)) was relatively longer and the peak time (Fig. 4(d)) was delayed for “*p*” [p‘] compared with “*b*” [p]; “*t*” [t‘] > “*d*” [t]; “*k*” [k‘] > “*g*” [k]; “*c*” [ʦ‘] > “*z*” [ʦ]; “*ch*” [tʂ‘] > “*zh*” [tʂ]; “*q*” [ʨ‘] > “*j*” [ʨ] (all $ps < 0.05$). Observed from Table I, it is promising that the PAS parameters can provide quantitative and differentiable information to identify aspiratory components of Mandarin consonants.

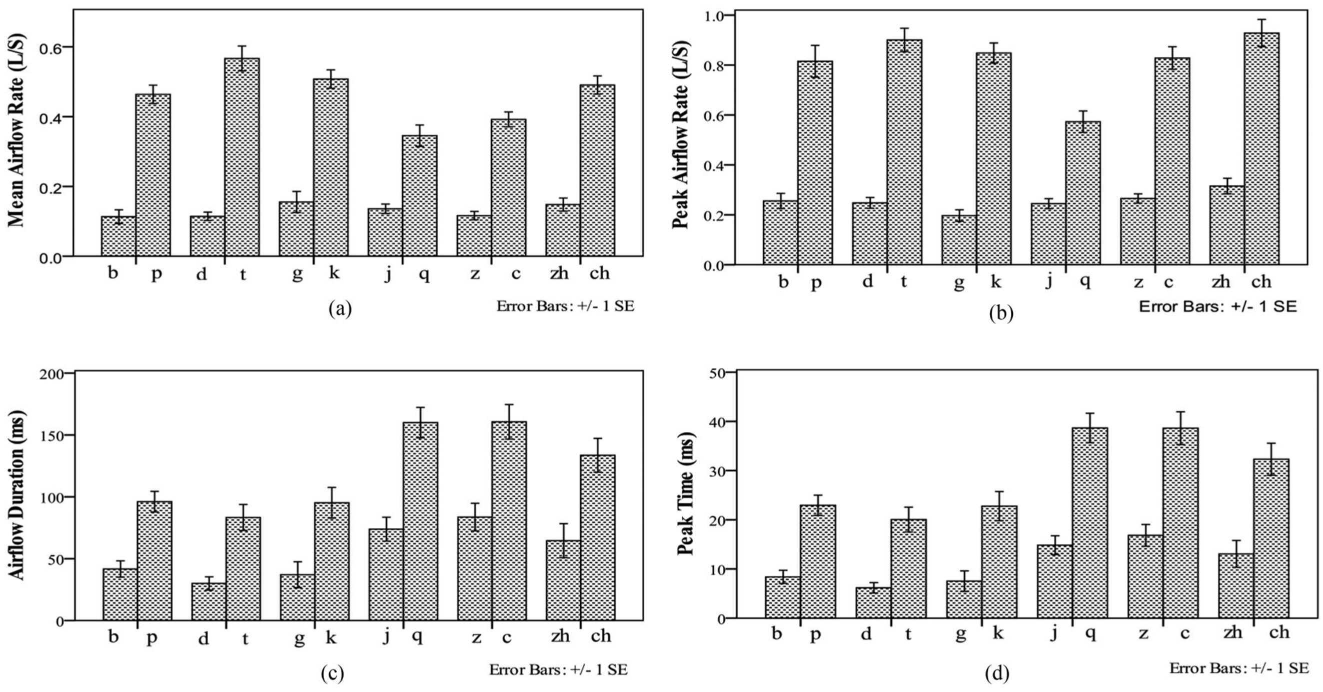

Fig. 4. The parameters of mean airflow rate (a), peak airflow rate (b), airflow duration (c), and peak time (d) of Mandarin stops and affricates in a carrying syllable $(\mathrm{Error\ Bars:}\ \pm 1\,\mathrm{SE})$.

III. The Implementation of Airflow Model Into a 3-D Talking Head

A. *Mandarin 3-D Talking Head with Internal Articulator Dynamics*

Both English talking head and Mandarin talking head have been proposed in our previous studies [7], [22], [23], [45]–[47]. For Mandarin 3-D external and internal articulator dynamics, articulatory data specifying in Mandarin production were collected with Carstens AG-501 EMA [22], [23]. The articulatory trajectories of 13 feature points were recorded, in which five facial feature points (nose, left head, right head, right jaw, left jaw) were used for calibration. In particular, four lip feature points (right lip corner, left lip corner, upper lip, lower lip,) and another four internal feature points (tongue tip, middle tongue, tongue root, middle jaw) were normalized and smoothed.

The static 3-D head model with tongue, uvula and jaw was constructed based on the templates from MRI images and anatomy, which is a generic triangular mesh model designed manually with great irregularity and a large number of vertices [7] [48]. Then the EMA data collected from any speaker could be registered to the static head model via affine transformation through related feature points manually identified in the static 3-D head model (see [7] for more details). Thus, a flexible graphic algorithm named Dirichlet free form deformation [49], [50] was then adopted for articulatory modeling work. The control parameters of multiple articulators of tongue, upper lip and lower lip were computed with Sibson coordinates of each anatomy. The EMA-based displacements were then input into the articulatory model, so as to drive multiple articulators to move smoothly and simultaneously over time.

The articulatory animations presented by this 3-D talking head could depict accurate and subtle Mandarin pronunciation at syllable level, where internal articulatory motions could be observed through a transparent profile view (see Fig. 5). This articulatory system has been successfully applied for hearing-loss children to learn Mandarin pronunciation [22].

B. *The Airflow Model*

The airflow model in this study is designed by modifying a fluid dynamics model [36], [37] which is suited for the airflow parameters extracted from the PAS collections. The algorithms are on the basis of the physical equations of fluid flow, named Navier-Stokes equations. Although it is difficult to tackle this problem when the physical accuracy is strict, an alternative solution is proposed for visual quality in [36]. To make the state of aspiration visualized at some point of time, the velocity of airflow during pronunciation is modeled with the following equation:

$$\begin{aligned}\frac{\partial u}{\partial t} &= -\left(u \cdot \nabla\right) u - \frac{1}{\rho}\nabla p + \nu \nabla^2 u + f \\ \nabla \cdot u &= 0 \end{aligned} \tag{2}$$

where u means the velocity of the airflow, ν is the kinematic viscosity, ρ is its density, and f refers to an external force. The symbol ∇ is Laplacia, having $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ in three-dimensions, and $\nabla^2 = \nabla \cdot \nabla$. The density of airflow is then simulated by the following equation:

$$\frac{\partial \rho}{\partial t} = -\left(u \cdot \nabla\right) \rho + \kappa \nabla^2 \rho + S \tag{3}$$

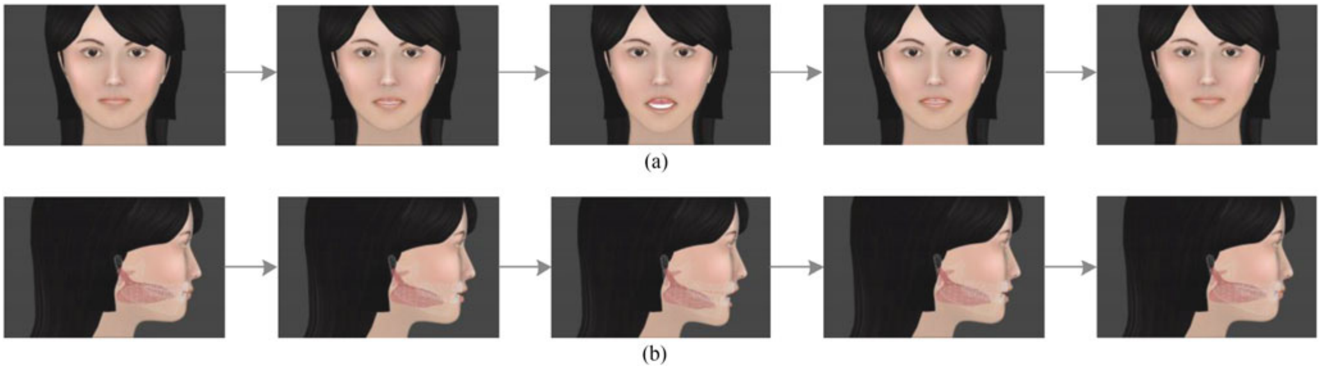

Fig. 5. The 3-D articulatory animations in sequence of Mandarin syllable ba [pa]: (a) a front view and (b) a transparent profile view.

where κ means the diffuse constant, and S is the source of density.

To solve the above equations, we have applied a fast and stable semi-Lagrangian fluids method proposed by [36], [37]. By applying a projection operator P, which projects any vector field onto its divergence free part, on both side of (2), it has:

$$\frac{\partial u}{\partial t} = P\left(-\left(u \cdot \nabla\right) u + \nu \nabla^2 u + f\right) \tag{4}$$

where having the fact $Pu = u$ and $P\nabla p = 0$. The advection, diffusion and external force items are represented by the right three parts of the equation.

As is shown in (4), given the velocity field at a time $u(t)$, then the velocity field at the next time step $u(t + \Delta t)$ over the time span Δt is generated, by applying four steps of adding external force, self-advection, viscous diffusion and projection steps. Afterward, a similar scheme is utilized to move densities in (3), by adding source, advection and diffusion. To simulate the consonant airflow using the aforementioned method, the initial velocity $u(0)$ is zero, and the external forces are then presented according to the aerodynamic parameters computed from PAS collections. With the use of the peak airflow rate vel_p, mean airflow rate vel_m, and the start and ending time, we can derive the following equations for the external forces:

$$\begin{aligned} f_{sp} &= ma_{sp} = m\left(\frac{vel_p}{t_p - t_s}\right) \\ f_{pe} &= ma_{pe} = m\left(\frac{-vel_p}{t_e - t_p}\right) \\ m &= \rho_g Q, \end{aligned} \tag{5}$$

where subscripts 't_s', 't_p', and 't_e' indicate the time points at start, peak and end time of airflow respectively. Moreover, the subscription 'sp' means the onset stage from start time of aspiration to the peak, and 'pe' refers to the offset stage from peak to the end instant of aspiration.

Given that both the start and end velocities of airflow are zero, m is the mass of airflow volume, and Q is the whole volume of the airflow that is the product of mean airflow rate vel_m and the duration. During the onset stage of aspiration from start to peak, the source of density ΔS is added as constant subsection of volume Q/N, the external force Δf is the same as f_{sp}, and $\Delta t = (t_p - t_s)/N$, where N means the total time steps within this period. While during the offset stage of aspiration from peak to end, the source of density is not added, and the external force Δf is set as f_{pe}. The constant density of airflow ρ_g is assigned $1.3\, g/L$ in our simulation. Thus, an airflow model is established, and the motion of the airflow is computed by the forces in (5).

C. The Implementation of a 3-D Articulatory with Aspiratory System

Modern Mandarin has a relatively simple syllable structure, consisting of an optional initial and a final on the segmental level, where the initial (i.e., onset) consists of a single consonant. At the syllable-level animation, the above airflow model was then combined with the 3-D articulatory model presented in [22], and a multimodal articulatory with aspiratory system was then constructed. The peak time, airflow rate, and airflow volume were exhibited through the velocity and density in the airflow animations. Because the airflow data were collected from speakers different from those of EMA articulatory data, the duration of syllable pronunciation would be slightly different. Thus, a dynamic time warping is used to match the airflow duration to that of EMA articulation. At the same time scale, the airflow model was synchronized with the 3-D articulatory model for each Mandarin syllable animation. As a result, the 3-D articulatory with aspiratory system was implemented and driven by the EMA articulatory data and PAS airflow data as well.

The simulations were conducted with carrying syllables containing all the 12 stops and affricates, where the airflow motions were generated when producing a consonant (i.e., initial) of a Mandarin syllable. As shown in Fig. 6, the pronunciation animations of minimal pairs, "zi" vs. "ci", "zhi" vs. "chi" have been displayed. Through the transparent views of a 3-D talking head, the motions of the airflow as well as the movements of the articulators can be clearly observed. Hereby, the size and density of the fluid change with the volume and velocity of aspiration over time.

Comparing the articulatory animations at the peak time of producing a minimal pair "zi" and "ci", it is hard to depict the differences from the lips and tongue positions of two confusable consonants. However, the airflow size and density of the consonant "c" are much bigger and heavier than those of "z" (see Fig. 6 (a)), which are in accordance with the quantitative

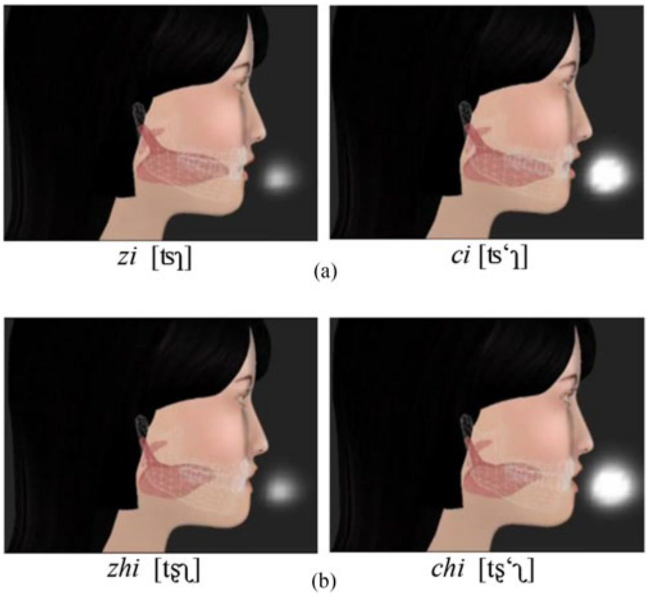

Fig. 6. The articulatory with aspiratory animations showing the minimal pairs of consonants at the peak state: (a) “zi” and “ci”, (b) “zhi” and “chi”.

analysis in Fig. 4 (b). Similarly, it is also seen the significant difference of the airflow size and density between “ch” and “zh” in Fig. 6 (b). Moreover, the tongue position of consonant “zh” can be discriminated from that of “z”, and the airflow size of “zh” is also slightly bigger than that of “z”. Similar results can be found by comparing the consonant “ch” with “c”, where the airflow size and density of producing consonant “ch” at the peak time are the greatest among all consonants of this study.

In terms of the airflow parameters extracted by PAS, the proposed airflow model can visualize accurate changes of consonant aspirations. Therefore, the 3-D articulatory with aspiratory animation can provide extra information for probably better perception of Mandarin consonants. The CAPT system with the proposed 3-D talking head has the capability of differentiating between those confusable Mandarin consonants, which may facilitate L2 learners to better control his/her aspiration during pronunciation and then to improve learning performance.

IV. Evaluation of the 3-D Articulatory with Aspiratory System

A. Audio-Visual Perception Test

For L2 learners of Mandarin, confusable Mandarin stops and affricates are often mispronounced as they are differentiated mainly by unaspirated vs. aspirated contrast. Thus, minimal pairs of Mandarin consonants were selected to evaluate our 3-D articulation with or without aspiration animations. A total of 12 Mandarin syllables with high-level tone, containing six confusable stops (bo [po], po [pʻo], de [tɤ], te [tʻɤ], ga [ka], ka [kʻa]) and six confusable affricates (ji [ʨi], qi [ʨʻi], zi [ʦɿ], ci [ʦʻɿ], zhi [tʂʅ], chi [tʂʻʅ]), were divided into six minimal pairs (see Table II). The audio-visual perception test [46] was adopted and aimed to evaluate, when the audio information was eliminated, whether the visual animations of our airflow-incorporated CAPT system could enhance the identification accuracy of confusable consonants for native speakers and non-native learners of Mandarin. Two groups of subjects were then recruited to conduct the perceptual test: Group 1 contained 21 Mandarin-speaking individuals (12 males, mean age = 24.21 years); Group 2 contained 21 Japanese learners of Mandarin (8 males, mean age = 20.63 years). Exclusion criteria for all the 42 subjects included history of hearing or visual loss, and majoring in linguistics in college. All these native and non-native subjects had no metalinguistic knowledge of Mandarin phonology, and gave responses on the basis of their own language experience. Moreover, the 21 Japanese subjects in Group 2 have studied Mandarin in Beijing Language and Culture University for more than ten months and have all passed the HSK intermediate level (HSK is an official Chinese language proficiency test).

TABLE II
The Identification Accuracies (%) of Minimal Pairs in Native and Japanese Subjects

Minimal Pairs	Group1: Native Speakers		Group2: Japanese learners of Mandarin	
	Without Airflow	With Airflow	Without Airflow	With Airflow
b([p]) vs. p([pʻ])	42.86	76.19	42.86	57.14
d([t]) vs. t([tʻ])	57.14	85.71	52.38	61.90
g[k]) vs. k([kʻ])	47.62	76.19	42.86	71.43
j([ʨ]) vs. q([ʨʻ])	47.62	80.95	42.86	66.67
z([ʦ]) vs. c([ʦʻ])	52.38	76.19	47.62	66.67
zh([tʂ]) vs. ch([tʂʻ])	47.62	80.95	57.14	71.43
Mean	**49.21**	**79.37**	**47.62**	**65.87**

All subjects were asked to conduct two audio-visual perceptual subtests. In one subtest, the two audio streams of one minimal pair were randomly played at first. After that, mute 3-D talking heads without airflow animations were shown in which two animations of minimal pair appeared in random. Participants were then asked to recognize and identify which talking head corresponded to the potential audio syllable. In the other subtest, a similar experimental procedure was adopted while the mute visual animations in 3-D talking heads were accompanied with additional airflow presentation. Subjects were told that the presentation of the airflow was related to the changes of expiratory aspiration during pronunciation. The order of the two perceptual subtests was counterbalanced within each subject group.

Perceptual results for the identification accuracy were shown in Table II. A two-way 2 (group: native group, Japanese group) × 2 (presentation condition: without airflow, with airflow) ANOVA was conducted on the identification accuracy, with presentation condition as a within-subject factor and the group as a between-subject factor. The native group showed a higher identification accuracy than the Japanese group $(F\,(1,\,40) = 20.46,\ p < 0.05)$, and identification accuracy with the help of airflow information was much higher than that without airflow $(F\,(1,\,40) = 33.74,\ p < 0.001)$. There was no significant interaction between presentation condition and group, $F(1, 40) = 2.04$, $p = 0.16$.

For native speakers, the average identification accuracy was improved from 49.21% (close to chance level) without airflow to 79.37% with airflow motions. Due to the long-term language experience, most of the native subjects can correctly make use of airflow information to match the mute 3-D articulator dynamics with the minimal pairs of confusable consonants. All the native subjects gave feedback that the airflow information was essential to discriminate the confusable syllables. Consequently, our airflow model can exhibit the real airflow changes to a great extent. For the intermediate-level Japanese learners of Mandarin, they also showed an improvement of identification accuracy in the audio-visual perception test, from 47.62% without airflow to 65.87% with airflow-incorporated information. However, the accuracy gain with the help of airflow information was lower for Japanese subjects compared with native ones, indicating that these intermediate-level Japanese learners of Mandarin were less certain about the airflow distinction between Mandarin aspirated and unaspirated consonants. This result also underscores the importance of exhibiting supplementary visual information of airflow in a CAPT system to help Japanese learners of Mandarin with their pronunciation training of Mandarin stops and affricates.

B. Pronunciation Training Study

The supplementary airflow animation might offer an effective audiovisual training approach for L2 learners of Mandarin. In order to further testify the efficacy of our airflow system, a training package was given to Japanese learners of Mandarin. In this experiment, a pronunciation training study was conducted to compare learning performance between two groups of Japanese learners of Mandarin, through learning 3-D articulation with or without aspiration animations respectively. Twenty young Japanese learners were chosen from the subjects in the first perception test. Of which, Group 1 includes 10 Japanese learners of Mandarin (4 males, mean age = 21.02 years), and Group 2 also contains 10 Japanese learners (4 males, mean age = 20.10 years). Moreover, another 10 Mandarin-speaking native subjects were recruited as a control group (5 males, mean age = 23.72 years).

Since all the intermediate-level learners of Mandarin from Japan have learned Mandarin for more than 10 months in Beijing, they were able to elicit spontaneous production of Mandarin syllables from *Pinyin*. In the pretest, all subjects were asked to produce the 12 Mandarin syllables containing six minimal pairs of consonants with a natural speaking rate. The entire pronunciation task was recorded by Cool Edit software (22050 Hz sampling rate, 16-bit resolution) in a quiet room, and any vocalization that occurred simultaneously with any other sound on the recording was abandoned.

VOT value of each produced sample was calculated and analyzed (see Table III). Independent-samples T test indicated that the VOT values of unaspirated consonants in the pretest were not different between native speakers and Japanese learners of Mandarin ($t = 1.05$, $p = 0.30$), while the VOT values of aspirated consonants produced by native speakers were much longer than that spoken by Japanese learners of Mandarin ($t = -4.50$, $p < 0.001$) in the pretest. The pretest results were in accordance with the previous literature [43], indicating that Japanese learners of Mandarin tended to produce Mandarin aspirated consonants with a much shorter VOT. Due to the ‘de-aspiration’ of Mandarin aspirated consonants, these aspirated consonants spoken by Japanese learners sounded unnatural and easily confused with their unaspirated counterparts.

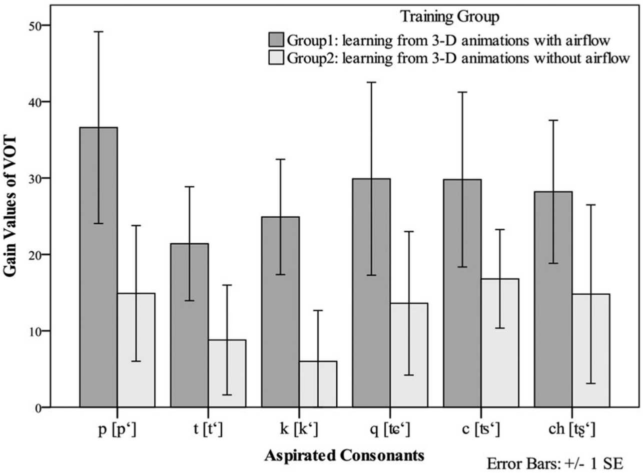

Fig. 7. The gain values of VOT for six aspirated Mandarin stops and affricates between two training groups of Japanese learners.

After the pretest, two groups of Japanese learners were then asked to take part in a training program. Ten subjects from Group 1 learned from our airflow-incorporated animations containing 12 testing syllables four times each day, while the other 10 subjects from Group 2 learned from the 3-D animations without airflow. During training, all subjects were asked to watch the videos and to follow what they heard and saw. After the fifth training day, both training groups were then asked to conduct a posttest, with a similar pronunciation task as that in the pretest.

The improvement (i.e., gain) from pretest to posttest was calculated by subtracting each participant's pretest VOT value from his/her posttest VOT value: *Gain value of VOT = posttest VOT value – pretest VOT value*. The gain values of VOT for six aspirated Mandarin stops and affricates were analyzed between two training groups of Japanese learners (see Fig. 7). Two outliers of gain values were removed before the statistical analyses. A two-way $2\,(\text{training group}) \times 6\,(\text{consonant})$ ANOVA was conducted on the gain values of VOT, with the six types of aspirated consonants as a within-subject factor and two training groups as a between-subject factor. The analysis showed that the main effect of group was significant ($F(1, 18) = 4.38$, $p < 0.05$), indicating that Japanese subjects learning from airflow-incorporated animations showed relatively higher gain values of VOT compared with those learning from animations without airflow, while no main effect for different consonants on gain values of VOT ($F(5, 90) = 0.51$, $p = 0.71$). There was no significant interaction between consonant and

TABLE III
AVERAGE VOT OF CONSONANTS PRODUCED BY NATIVE SPEAKERS AND JAPANESE LEARNERS (GROUP 1 AND GROUP 2)

Group		VOT of Aspirated Consonants (ms)						VOT of Unaspirated Consonants (ms)					
		Stops			*Affricates*			*Stops*			*Affricates*		
		p[p‘]	t[t‘]	k[k‘]	q[ʨ‘]	c[ʦ‘]	ch[tʂ‘]	b[p]	d[t]	g[k]	j[ʨ]	z[ʦ]	zh[tʂ]
Native Speakers		126.3	114.5	112.9	207.4	203.4	178.1	21.1	17.7	22.1	90.4	81.3	70.2
Group 1	Pretest	59.5	75.9	85.6	141.0	146.8	145.3	32.6	27.1	31.6	69.5	88.1	76.0
	Posttest	96.1	97.3	110.5	170.9	176.6	173.5	30.7	28.8	31.5	75.3	87.8	84.2
	t-value	-5.01^{***}						-1.27					
Group 2	Pretest	67.0	82.5	92.9	158.3	155.5	154.0	24.8	25.9	31.5	84.8	85.1	80.3
	Posttest	81.9	91.3	98.9	171.9	172.3	168.8	23.2	24.7	28.7	87.6	73.8	83.8
	t-value	-2.39^{*}						0.69					

$^{*}p < 0.05$, $^{***}p < 0.001$.

group $(F(5, 90) = 0.09,\ p = 0.98)$. Results showed that our real-data-driven articulatory with aspiratory system successfully helped intermediate-level Japanese learners of Mandarin improve their production accuracy of Mandarin aspirated consonants, reflected by much higher gain values of VOT for aspirated consonants after training. Moreover, to investigate the correlation between pretest scores and gain values of VOT for aspirated consonants, the Pearson correlation analysis was conducted for both learning groups. Results indicated that there was no correlation between the pretest scores and gain values for all the six aspirated consonants $(r = -0.10,\ p = 0.27)$.

V. DISCUSSIONS

For learning Mandarin as an L2, the aspiratory airflow information is important in telling apart minimal pair of Mandarin stops and affricates with similar places of articulation. Making use of PAS, we have collected bio-signal data of airflow from multi-speaker native pronunciations, and calculated quantitatively the mean airflow rate, peak airflow rate, airflow duration, and peak time of Mandarin stops and affricates. Results showed that minimal pairs of confusable consonants could be discriminated through the related airflow parameters. The existing 3-D articulatory systems, however, mainly focused on the animations of internal and external articulators. This study fills a research gap by presenting a multimodal system containing both articulatory and airflow models for pronunciation animations, using EMA articulatory data and dynamic airflow parameters simultaneously. The proposed 3-D multimodal pronunciation system can additionally illustrate the aerodynamic airflow changes of Mandarin consonants, and offers a comprehensive exhibition of the entire process of speech production.

To evaluate our articulatory with aspiratory system, an audio-visual perception test and a pronunciation training study have been conducted to testify whether the supplementary airflow animations are effective to improve the perception of Mandarin confusable consonants for both native and non-native speakers, and further to refine pronunciation of Mandarin consonants among L2 learners of Mandarin. In the audio-visual perception test, both native and non-native subjects showed enhanced identification skill with the help of supplementary airflow animation. However, native perceivers exhibited higher identification accuracy than the intermediate-level Japanese learners of Mandarin when the airflow information was shown. Although native subjects (excluding subjects majoring in linguistics) had no metalinguistic knowledge about the phonetic distinction of consonants, the long-term exposure to Mandarin has helped them implicitly classify the tested consonants into two subtypes, and thus greatly improve the identification accuracy when matching the tested sounds with supplementary visual animations of airflow. For the non-native Japanese subjects, however, they were more uncertain about the airflow distinction between Mandarin aspirated and unaspirated consonants compared with native ones. Moreover, the poorer perceptual results can partly explain the smaller differences of VOT values in producing Mandarin aspirated vs. unaspirated consonants by the Japanese subjects compared to native speakers (see Table III).

As influenced by native language, the de-aspiration of Mandarin aspirated consonants tended to one type of fossilized mispronunciations for Japanese learners of Mandarin [43]. The intermediate-level Japanese learners of Mandarin in the pretest tended to produce Mandarin aspirated consonants with a much shorter VOT, making these consonants sound unnatural and easily confused with their unaspirated counterparts. To further prove the efficacy of the proposed airflow model in CAPT system, a comparative pronunciation training study was conducted. The learning results showed that exhibiting additional airflow information led to better production of aspirated consonants, reflected by higher gain values of VOT after training. Consequently, utilizing supplementary visual information of airflow can efficiently help Japanese learners of Mandarin with their pronunciation training of Mandarin aspirated consonants.

To conclude, by effectively demonstrating supplementary visual information of airflow changes in a talking head, the current bio-data-driven 3-D articulatory with aspiratory system provides a novel pronunciation training approach for L2 learners of Mandarin. It is argued that the tactile modality (by putting hand close to the mouth) might also facilitate L2 learners of Mandarin to sense the expiratory airflow differences between aspirated vs. unaspirated Mandarin consonants. However, it is also known that the tactile modality has much poorer temporal and spatial resolution compared with the visual modality [27]. Our airflow model showed great advantages by offering language learners a more straightforward and accurate visual exhibition of airflow changes during syllable pronunciation.

In the future study, we will also test the changes of airflow rate from pretest to posttest for evaluation, to have a clearer view of the impact of the airflow animations on the production. Moreover, we will record the acoustic materials and have native speakers and/or phonetical experts score the pronunciation to assess whether these changes are perceptually significant. Furthermore, speech visualization that includes Mandarin fricatives, nasal $(m,\ n,\ ng)$ and lateral consonants should be investigated and exhibited in a CAPT system. Moreover, as Engwall [51] proposed that hearing-loss children showed very limited exposure to acoustic speech, and these hearing-loss children can acquire speech production through observing visual cues of phonetic features. Visualization of the current multimodal system might help hearing-loss children conceptualize articulator movements and airflow changes during speech production. Furthermore, speakers with pathological voices often showed atypical airflow control during speech production [28], [29]. The current 3-D articulatory with aspiratory animation system offers a promising training option for individuals with pathological voices as well.

References

[1] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annu. Rev. Neurosci.*, vol. 27, pp. 169–192, 2004.

[2] J. Sweller, J. J. G. V. Merrienboer, and F. G. W. C. Paas, "Cognitive architecture and instructional design," *Educ. Psychol. Rev.*, vol. 10, no. 3, pp. 251–296, 1998.

[3] R. E. Mayer, *Multimedia Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[4] Z. Y. Wu, S. Zhang, L. H. Cai, and H. M. Meng, "Real–time synthesis of Chinese visual speech and facial expressions using MPEG–4 FAP features in a three-dimensional avatar," in *Proc. Int. Speech Commun. Assoc. Interspeech*, 2006, pp. 1802–1805.

[5] J. Yu and A. J. Li, "3D visual pronunciation of Mandarin Chinese for language learning," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 2036–2040.

[6] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA, USA: MIT Press, 1998.

[7] L. Wang, H. Chen, S. Li, and H. M. Meng, "Phoneme–level articulatory animation in pronunciation training," *Speech Commun.*, vol. 54, pp. 845–856, 2012.

[8] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, "An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker's articulatory data," in *Proc. Int. Conf. Articulated Motion Deformable Objects*, 2008, pp. 132–143.

[9] O. Bälter, O. Engwall, A. M. Öster, and H. Kjellström, "Wizard-of-Oz test of ARTUR—A computer-based speech training system with articulation correction," in *Proc. 7th Int. ACM SIGACCESS Conf. Comput. Accessibility*, 2005, pp. 36–43.

[10] A. Serrurier and P. Badin, "A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2335–2355, 2008.

[11] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 175–185, Mar. 2004.

[12] T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly, "Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2246–2259, Dec. 2015.

[13] J. Ma, R. Cole, W. Pellom, and B. Ward, "Accurate automatic visible speech synthesis of arbitrary 3D models based on concatenation of diviseme motion capture data," *Comput. Animat. Virt. W.*, vol. 15, no. 5, pp. 485–500, 2004.

[14] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Commun.*, vol. 41, no. 2, pp. 303–329, 2003.

[15] S. Fagel and C. Clemens, "An articulation model for audio-visual speech synthesis—Determination, adjustment, evaluation," *Speech Commun.*, vol. 44, no. 1, pp. 141–154, 2004.

[16] N. Murray, K. I. Kirk, and L. Schum, "Making typically obscured articulatory activity available to speech readers by means of videofluoroscopy," *NCVS Status Prog. Rep.*, vol. 4, pp. 41–63, 1993.

[17] T. H. Chen and D. W. Massaro, "Evaluation of synthetic and natural Mandarin visual speech: Initial consonants, single vowels, and syllables," *Speech Commun.*, vol. 53, no. 7, pp. 955–972, 2011.

[18] N. F. Chen, V. Shivakumar, H. Mahesh, M. Bin, and H. Z. Li, "Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages," in *Proc. Int. Speech Commun. Assoc. Interspeech*, 2013, pp. 803–806.

[19] C. Y. Chiu, Y. F. Lia, D. Kulls, H. Mixdorff, and S. L. Chen, "A Preliminary study on corpus design for computer-assisted German and mandarin language learning," in *Proc. Orient. COCOSDA Int. Conf.*, 2009, pp. 154–159.

[20] Y. H. Lai, "Asymmetry in Mandarin affricate perception by learners of Mandarin Chinese," *Lang. Cogn. Proc.*, vol. 24, no. 7/8, pp. 1265–1285, 2009.

[21] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, and H. Li, "iCALL corpus: Mandarin Chinese spoken by non-native speakers of european descent," in *Proc. Int. Speech Commun. Assoc. Interspeech*, 2015, pp. 324–328.

[22] X. Q. Liu, N. Yan, L. Wang, and X. L. Wu, "An interactive speech training system with virtual reality articulation for Mandarin-speaking hearing impaired children," in *Proc. Int. Conf. Inf. Autom.*, 2013, pp. 191–196.

[23] D. Zhang, X. Q. Liu, N. Yan, L. Wang, Y. Zhu, and H. Chen, "A multi-channel/multi-speaker articulatory database in Mandarin for speech visualization," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, 2014, pp. 299–303.

[24] J. Yu, A. J. Li, F. Hu, and Z. F. Wang, "Data–driven 3D visual pronunciation of Chinese IPA for language learning," in *Proc. Orient. COCOSDA Int. Conf.*, 2013, pp. 93–98.

[25] H. Li, M. H. Yang, and J. H. Tao, "Speaker–independent lips and tongue visualization of vowels," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8106–8110.

[26] C. J. LaSasso, K. L. Crain, and J. Leybaert, *Cued Speech and Cued Language Development for Deaf and Hard of Hearing Children*. San Diego, CA, USA: Plural Publishing, 2010.

[27] D. W. Massaro, M. A. Carreira-Perpinan, and D. J. Merrill, "Optimizing visual feature perception for an automatic wearable speech supplement in face-to-face communication and classroom situations," in *Proc. 42nd Hawaii Int. Conf. Syst. Sci.*, 2009, pp. 1–10.

[28] R. Netsell, W. Lotz, and A. L.Shaughnessy, "Laryngeal aerodynamics associated with selected voice disorders," *Amer. J. Otolaryng.*, vol. 5, no. 6, pp. 397–403, 1984.

[29] E. M. Yiu, Y. M. Yuen, T. Whitehill, and A. Winkworth, "Reliability and applicability of aerodynamic measures in dysphonia assessment," *Clin. Linguist. Phon.*, vol. 18, pp. 463–478, 2004.

[30] S. N. Awan, C. K. Novaleski, and J. R. Yingling, "Test-retest reliability for aerodynamic measures of voice," *J. Voice*, vol. 27, no. 6, pp. 674–684, 2013.

[31] R. I. Zraick, L. Smith-Olinde, and L. L. Shotts, "Adult normative data for the Kay-PENTAX phonatory aerodynamic system model 6600," *J. Voice*, vol. 26, no. 2, pp. 164–176, 2012.

[32] N. Foster and D. Metaxas, "Controlling fluid animation," in *Proc. Comput. Graph. Int.*, 1997, pp. 178–188.

[33] N. Foster and R. Fedkiw, "Practical animation of liquids," in *Proc. Annu. Conf. Comput. Graph. Interactive Techn.*, 2001, pp. 23–30.

[34] A. McNamara, A. Treuille, Z. Popović, and J. Stam, "Fluid control using the adjoint method," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 449–456, 2004.

[35] J. K. Gupta, C. H. Lin, and Q. Chen, "Characterizing exhaled airflow from breathing and talking," *Indoor Air*, vol. 20, no. 1, pp. 31–39, 2010.

[36] J. Stam, "Stable fluids," in *Proc. Annu. Conf. Comput. Graph. Interactive Techn.*, 1999, pp. 121–128.

[37] J. Stam, "Real-time fluid dynamics for games," in *Proc. Game Developer Conf.*, 2003.

[38] B. J. Theobald, S. Fagel, G. Bailly, and F. Elisei, "Lips 2008: Visual speech synthesis challenge," in *Proc. Int. Speech Commun. Assoc., Interspeech*, 2008, pp. 2310–2313.

[39] J. E. Flege, "Second-language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, W. Strange Ed. Baltimore, MD, USA: York Press, 1995, pp. 229–273.

[40] S. G. Guion, J. E. Flege, R. Akahane-Yamada, and J. C. Pruitt, "An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants," *J. Acoust. Soc. Amer.*, vol. 107, no. 5, pp. 2711–2724, 2000.

[41] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[42] T. J. Vance, *An Introduction to Japanese Phonology*. Albany, NY, USA: State University of New York Press, 1987.

[43] L. J. Zhang, "Perceptual training and the acquisition of Chinese aspirated/unaspirated consonants by Japanese students," (in Chinese), *Lang. Teach. Linguist. Stud.*, no. 4, pp. 560–566, 2009.

[44] T. Cho and P. Ladefoged, "Variation and universals in VOT: Evidence from 18 languages," *J. Phonetics*, vol. 27, no. 2, pp. 207–229, 1999.

[45] F. Chen *et al.*, "Intelligible enhancement of 3D articulation animation by incorporating airflow information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 6130–6134.

[46] L. Wang, H. Chen, and J. J. Ouyang, "Evaluation of external and internal articulator dynamics for pronunciation learning," in *Proc. Int. Speech Commun. Assoc., Interspeech*, 2009, pp. 2247–2250.

[47] H. Chen, L. Wang, W. Liu, and P. A. Heng, "Combined X-ray and facial videos for phoneme-level articulator dynamics," *Vis. Comput.*, vol. 26, no. 6, pp. 477–486, Apr. 2010.

[48] M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, New York, NY, USA: Springer-Verlag, 1993, pp. 139–156.

[49] S. Ilic and P. Fua, "Using dirichlet free form deformation to fit deformable models to noisy 3-D data," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 704–717.

[50] L. Moccozet and N. Magnenat-Thalmann, "Dirichlet free-form deformations and their application to hand simulation," in *Proc. Comput. Animat.*, Geneva, Switzerland, 1997, pp. 93–102.

[51] O. Engwall, O. Bälter, A. M. Öster, and H. Kjellström, "Designing the user interface of the computer-based speech training system ARTUR based on early user tests," *Behav. Inf. Technol.*, vol. 25, no. 4, pp. 353–365, Feb. 2006.

Fei Chen received the M.S. degree in linguistics from Nankai University, Tianjin, China, in 2014. He is currently working as a Research Assistant with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His current research interests include psycholinguistics, experimental phonetics, language acquisition, and pathological linguistics.

Lan Wang received the M.S. degree from the Center of Information Science, Peking University, Beijing, China and the Ph.D. degree from the Machine Intelligence Laboratory, Department of Engineering, University of Cambridge, Cambridge, U.K., in 2006. She worked on the Autonomous Global Integrated Language Exploitation project funded under DARPAs Global Autonomous Language Exploitation program. She is currently a Research Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include large vocabulary continuous speech recognition, speech visualization, and audio information indexing.

Hui Chen received the B.S. and M.S. degrees in computer science from Shandong University, Jinan, China, and the Ph.D. degree in computer science from Chinese University of Hong Kong, Hong Kong. She is currently an Associate Professor in the Institute of Software Chinese Academy of Sciences, Beijing, China. Her research interests include human computer interaction, haptics, virtual reality, and computer-assisted surgery.

Gang Peng received the Ph.D. degree in language engineering from the City University of Hong Kong, Hong Kong, in 2002. He has published several research articles in various high-profile international journals. He is currently working as an Associate Professor in the Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, and Adjunct Professor of Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His central research interests include how language is represented and processed in the human brain, and how different cultures, reflected in their languages, shape perception differently. His research areas include psycholinguistics, neurolinguistics, experimental phonetics, computational linguistics, hearing disorders, and related topics.