INTELLIGIBLE ENHANCEMENT OF 3D ARTICULATION ANIMATION BY INCORPORATING AIRFLOW INFORMATION

Fei Chen[1], Hui Chen[2], Lan Wang[1], Ying Zhou[1], Jiaying He[1], Nan Yan[1], Gang Peng[1,3]

[1]Key Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institutes of Advanced Technology, [2]Institute of Software, Chinese Academy of Sciences，
[3]The Chinese University of Hong Kong

chenfei@siat.ac.cn, lan.wang@siat.ac.cn

ABSTRACT

The 3D talking head has been developed fast, in which both external and internal articulators were demonstrated. For Mandarin pronunciation, the aspiration airflow is crucial to discriminate confusable Mandarin consonants. In this paper, we present a 3D talking head system for articulatory and aspiration animation with the use of EMA articulation data and airflow data simultaneously. The quantitative analyses of airflow data indicated confusable Mandarin consonants could be distinguished from each other by the mean airflow during voicing, peak expiratory airflow, and airflow duration. An airflow model was then incorporated into the 3D articulatory model to produce the airflow in accordance with articulator movements of Mandarin pronunciation. An audio-visual test was designed to evaluate the current 3D articulation and aspiration system, where minimal pairs were used to recognize the animation. The identification accuracy was significantly improved from 43.9% without airflow to 84.8% with airflow-incorporated information.

Index Terms—Airflow, PAS, 3D articulatory dynamics, confusable consonants, intelligible enhancement

1. INTRODUCTION

Speech perception based on audio-visual feedback appeared to be superior to auditory-only perception [1]. Besides the visible external information (e.g., lip, jaw), integrating visual information of the internal articulator movements (e.g., tongue, tooth, and even velum and nasopharyngeal wall) has received more and more attention [2-3]. Studies have demonstrated that information of internal articulators could improve speech comprehension [3], enhance the pronunciation training [4], and even help in speech therapy [5]. With the use of data collected from facial motion capture [6], Electro-Magnetic Articulography (EMA) [7], or video-fluoroscopic images [8], speech visualization has been recently extended to use visual synthesis methods and a transparent 3D articulatory dynamics to vividly present both external and internal articulatory animations [9-11]. Based on the related perceptual test, subjects evaluated the 3D animations with a high identification rate of minimal pairs of English phonemes [10-11].

In particular, the acoustic and kinematic features in Mandarin pronunciation are different from English [12]. For instance, aspirated consonants in English are allophones in complementary distribution with their unaspirated counterparts, but in some other languages, notably Mandarin, the difference is contrastive [13-14]. Many confusable Mandarin consonants are differentiated only by the distinctive feature of unaspirated vs. aspirated contrast (e.g., *b* [p] vs. *p* [p']). The above-mentioned confusable consonants have similar articulator movements to produce those sounds, which makes it difficult to discriminate them in the existing audio-visual pronunciation animations [12, 15-18]. Therefore, airflow information might provide cues to enhance the discrimination between confusable Mandarin consonants. Moreover，traditional airflow control during articulation plays an important role for speakers with pathological voices [19-21]. Three studies [19-21] demonstrated aerodynamic parameter of airflow rate can be useful in discriminating normal from dysphonic voices.

In this paper, airflow information incorporated into a 3D articulatory animation system was proposed. Airflow data of confusable Mandarin consonants were collected using *Phonatory Aerodynamic System* (PAS) Model 6600 (KayPENTAX Corp.). After segmentation and data processing, the features of mean airflow during voicing, peak expiratory airflow, and airflow duration have shown significant differences between confusable consonants. Based on the 3D talking head system proposed in [9-11, 22], we incorporated an airflow model into the existing 3D articulatory model. The durations, peak values and overall values of airflow data were then used to drive the airflow model, in accordance with the articulation data. Perception tests were then carried out to evaluate the 3D articulation

This work was supported by grants from National Natural Science Foundation of China (NSFC: 91420301, 61135003, 11474300, and 61401452), 863 project: 2015AA016305, Shenzhen Speech Rehabilitation Technology Laboratory, and Shenzhen Fundamental Research Program JCYJ20130401170306806.

and aspiration system, where the identification accuracies of minimal pairs were significantly improved compared to the 3D articulation only.

The rest of this paper is organized as follows: Section 2 describes the data collection and data analysis of airflow information using PAS. Section 3 presents the implement of airflow information on the existing articulatory dynamics. The experiments and results are shown in Section 4. Finally, discussion and conclusion go to Section 5.

2. AIRFLOW DATA ACQUISITION AND ANALYSES

2.1. Participants, materials and procedure

Participants were six healthy young adults (3 males) between 23 and 27 years of age (M = 25.18 years, SD = 1.08), with Mandarin as their native language. Each was free from colds or seasonal allergies on the day of testing.

To make discriminations among confusable consonants not only pronounced alone but also with a carrying syllable, the pronunciation materials were designed to contain two sessions. Session one covered 15 Mandarin consonants, which were divided into six groups sharing the same places of articulation and similar articulatory trajectory: Group 1 contains two bilabial consonants in term of *Pinyin* (IPA in the square brackets): *b* ([p]) vs. *p* ([p']); Group 2 contains two alveolar consonants: *d* ([t]) vs. *t* ([t']); Group 3 contains two dorso-velar consonants: *g* ([k]) vs. *k* ([k']); Group 4 contains three alveolo-palatal consonants: *j* ([tɕ]) vs. *q* ([tɕ']) vs. *x* ([ɕ]); Group 5 contains three front-apical consonants: *z* ([ʦ]) vs. *c* ([ʦ']) vs. *s* ([s']); Group 6 contains three post-apical consonants: *zh* ([tʂ]) vs. *ch* ([tʂ']) vs. *sh* ([ʂ]). Session two consisted of 55 valid Mandarin syllables combining all the above-mentioned 15 consonants and six monophthongs ([a], [o], [ɤ], [i], [u], [y]) in Mandarin. All the pronunciation materials were repeated three times by subjects.

Before use, the PAS flow-head and pressure transducer were calibrated according to the manufacturer's instructions. Each participant was seated comfortably in a straight-backed chair and instructed to perform each speaking task using their most comfortable vocal pitch and loudness. These instructions were provided to ensure that the speech samples obtained were as natural as possible. During data collection, participants were instructed to hold the PAS with two hands using the side handles on the device and to press the facemask firmly against the face so the nose and mouth were both covered to prevent air escape.

2.2. Data analyses

Airflow information was calculated on three aspects: mean airflow during voicing (L/S), peak expiratory airflow (L/S), and airflow duration (S). The best utterance was chosen from three repetitions. Fig. 1 shows the airflow fluctuation of "*b*" vs. "*p*", and "*ba*" vs. "*pa*" pronounced by one subject. For consonants when pronounced alone, we selected the whole articulation to analyze airflow information. However, when occurred in syllables, the consonant needs to be properly split away from the following vowel. By using WaveSurfer software, we could clearly detect segment accompanied with F0 information that actually corresponded with the vowel segment. Moreover, we would check again by listening to selected audio to confirm the correctness of segmentation.

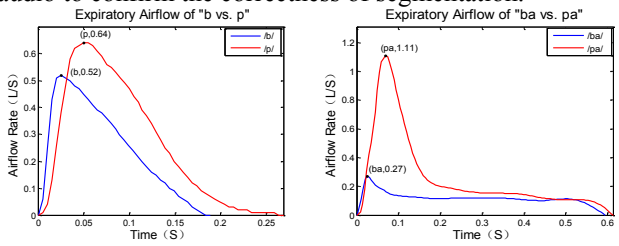

Fig.1. The expiriatory airflow of "*b*" vs. "*p*", "*ba*" vs. "*pa*"

2.2.1. The airflow of consonants when pronounced alone

The average data of mean airflow during voicing, peak expiratory airflow, and airflow duration of confusable consonants when pronounced alone were shown in Table 1.

Table 1. The average data of mean airflow, peak airflow, and airflow duration of consonants when pronounced alone

Confusable Consonants in Isolation	***Mean Airflow (L/S)***	***Peak Airflow (L/S)***	***Airflow Duration (S)***
"b" [p]	0.33	0.70	0.21
"p" [p']	0.71	1.46	0.24
"d" [t]	0.27	0.57	0.30
"t" [t']	0.34	0.81	0.30
"g"[k]	0.17	0.37	0.32
"k" [k']	0.54	1.04	0.31
"j" [tɕ]	0.17	0.36	0.40
"q"[tɕ']	0.24	0.54	0.40
"x" [ɕ]	0.27	0.43	0.58
"z" [ʦ]	0.13	0.30	0.41
"c" [ʦ']	0.28	0.69	0.45
"s" [s']	0.25	0.46	0.50
"zh" [tʂ]	0.18	0.38	0.37
"ch" [tʂ']	0.28	0.61	0.40
"sh" [ʂ]	0.30	0.53	0.55

Results of one-way ANOVA revealed that the airflow duration was significantly different among different consonants [F (14, 75) = 2.288, $p < 0.05$], and the peak airflow was also significantly different [F (14, 75) = 4.756, $p < 0.001$]. The similar results went to the mean airflow during voicing when pronounced alone [F (14, 75) = 4.367; $p < 0.001$]. Specifically, Tukey's HSD post hoc pairwise comparisons within six groups showed that mean airflow of "*p*" [p'] is significantly higher than "*b*" [p]; "*t*" [t'] > "*d*" [t]; "*k*" [k']> "*g*" [k]; "*x*" [ɕ] > "*q*" [tɕ'] > "*j*" [tɕ]; "*c*" [ʦ'] > "*s*" [s'] > "*z*" [ʦ]; "*sh*" [ʂ] > "*ch*" [tʂ'] > "*zh*" [tʂ] (all *p*s < 0.05), which could be observed more visually from Figure 2.

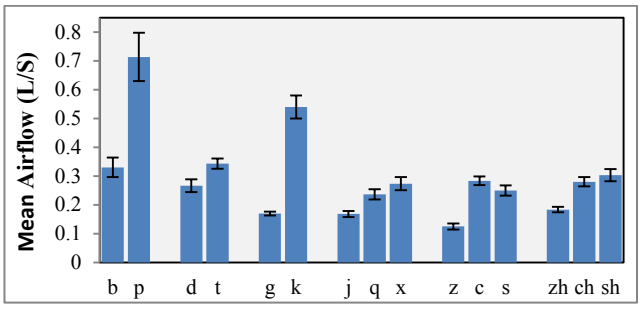

Fig.2. Mean airflow among different consonants when pronounced alone (Error Bars: ±1 SE)

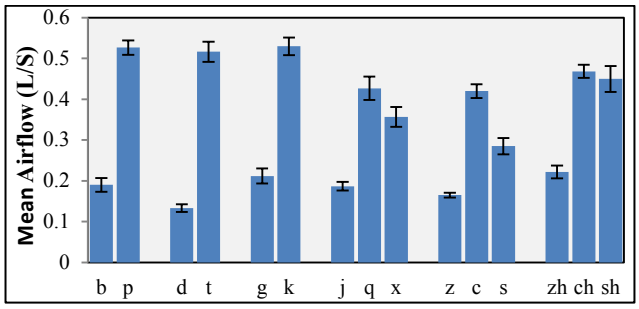

Fig.3. Mean airflow among different consonants when pronounced with a carrying syllable (Error Bars: ±1 SE)

2.2.2. The airflow of consonants pronounced with a carrying syllable

When pronounced with a carrying syllable, the average data of mean airflow during voicing, peak expiratory airflow, and airflow duration of confusable consonants were shown in Table 2. Results of one-way ANOVA revealed that airflow duration of aspirated consonants was much longer than that of unaspirated ones [F (14, 75) = 4.807; $p < 0.01$], and the peak airflow was significantly different among consonants [F (14, 75) = 16.731; $p < 0.001$].

Table 2. The average data of mean airflow, peak airflow, and airflow duration of consonants with a carrying syllable

Confusable Consonants Within Syllable	*Mean Airflow (L/S)*	*Peak Airflow (L/S)*	*Airflow Duration (S)*
"b" [p]	0.19	0.35	0.07
"p" [p']	0.53	1.02	0.12
"d" [t]	0.13	0.31	0.05
"t" [t']	0.52	1.00	0.12
"g"[k]	0.21	0.26	0.06
"k" [k']	0.53	0.95	0.11
"j" [tɕ]	0.19	0.30	0.10
"q"[tɕ']	0.43	0.72	0.16
"x" [ɕ]	0.36	0.49	0.23
"z" [ts]	0.17	0.33	0.13
"c" [ts']	0.42	1.00	0.20
"s" [s']	0.29	0.48	0.24
"zh" [tʂ]	0.22	0.43	0.11
"ch" [tʂ']	0.47	1.07	0.17
"sh" [ʂ]	0.45	0.60	0.22

Moreover, the mean airflow during voicing was significantly different among consonants with a carrying syllable [F (14, 75) = 9.126; $p < 0.001$]. Specifically, Tukey's HSD post hoc pairwise comparisons within the six groups indicated mean airflow with a carrying syllable of "*p*" [p'] was significantly higher than "*b*" [p]; "*t*" [t'] > "*d*" [t]; "*k*" [k']> "*g*" [k]; "*q*" [tɕ'] > "*x*" [ɕ]> "*j*" [tɕ]; "*c*" [ts'] > "*s*" [s'] > "*z*" [ts]; "*ch*" [tʂ'] > "*sh*" [ʂ] > "*zh*" [tʂ] (all *p*s < 0.05), which could be observed more visually from Figure 3.

3. THE IMPLEMENTATION OF AIRFLOW MODEL

3.1. The algorithm of airflow model

To visualize the state of aspiration at a given instant of time, the airflow's velocity of pronunciation is modeled with the famous Navier-Stokes equations [23, 24]:

$$\frac{\partial \mathrm{u}}{\partial \mathrm{t}} = -(\mathrm{u}\cdot\nabla)\mathrm{u} - \frac{1}{\rho}\nabla \mathrm{p} + \nu\nabla^2\mathrm{u} + \mathrm{f} \qquad (1)$$
$$\nabla\cdot\mathrm{u} = 0$$

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

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

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

To solve the above equations, a fast and stable semi-Lagrangian fluids approach proposed by Stam [23, 24] is applied. Through applying projection operator P, which projects any vector field onto its divergence free part, on both side of Eq. 1, it has:

$$\frac{\partial \mathrm{u}}{\partial \mathrm{t}} = \mathrm{P}(-(\mathrm{u}\cdot\nabla)\mathrm{u} + \nu\nabla^2\mathrm{u} + \mathrm{f}) \qquad (3)$$

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

According to Eq. 3, given the velocity field at a time $\mathrm{u}(\mathrm{t})$, then the velocity field at the next time step $\mathrm{u}(\mathrm{t}+\Delta\mathrm{t})$ over the time span $\Delta\mathrm{t}$ is generated through applying four steps of adding external force, self-advection, viscous diffusion and projection steps. A similar scheme is then used to move densities in Eq. 2, through adding source, advection and diffusion.

To simulate the consonant airflow with the above method, the initial velocity $\mathrm{u}(0)$ is zero, the external force is then added according the aerodynamic data of consonant pronunciation, having:

$$f_{ps} = ma_{ps} = m\left(\frac{vel_p}{t_p - t_s}\right)$$
$$f_{ep} = ma_{ep} = m\left(\frac{-vel_p}{t_e - t_p}\right)$$
$$m = \rho_g Q$$

where subscript 's', 'p', and 't' refer to the states at start, peak and end time of pronunciation airflow, 'ps' corresponds to the onset stage from start time of pronunciation to the peak time, and 'ep' corresponds to the offset stage from peak to the end instant of pronunciation. Given that the start and end velocities of airflow are zero, vel_p is the peak flow rate, Q is the whole volume of the airflow. During onset period of pronunciation 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_{ps}, and $\Delta t = (t_p - t_s)/N$, where N is the total time steps of this period. While in offset period of pronunciation from peak to end, the source of density is not added and the external force Δf is set as f_{ep}. In our simulation, the constant density of airflow ρ_g is 1.3g/L.

3.2. The 3D articulation and aspiration system

The 3D articulatory animation system presented in [22] used the articulation data collected from EMA device. By incorporating airflow model into the 3D articulation model, a new multimodal system was built up. The peak time, peak airflow, and airflow volume of the airflow data were represented by the density and velocity in the airflow model. Since the airflow data was different speaker from EMA data, the airflow duration was then warped to that of EMA articulation data. Using transparent view, the difference of airflow between minimal pairs can be seen (Fig. 4), which is hardly discriminated only through movements of articulators.

"*bo*" "*po*"

Fig.4. The 3D articulation with dynamic aspiration animations of the minimal pair "*bo*" and "*po*", at the peak states.

4. EXPERIMENTS AND RESULTS

To learn Mandarin pronunciation, confusable consonants are commonly mispronounced which are differentiated only by distinctive feature of unaspirated vs. aspirated contrast [25-26]. Hence, a set of minimal pairs was chosen to evaluate the 3D articulation with aspiration animation. A total of 12 syllables containing the confusable consonants were divided into six pairs. The audio-visual perception test [9] attempts to assess whether the animations of the airflow-incorporated 3D articulatory dynamics can be recognized without audio. A total of 11 subjects with Mandarin as their native language were recruited to participate in two perception tests. In one test, the audio streams of one minimal pair were played firstly, and then the mute 3D animations without airflow information were shown in which two syllables of one minimal pair appeared in a random order. The subjects were asked to identify which animation corresponded to the syllable. The same procedures went to the other test, while the order of the two tests was counterbalanced among subjects. Perceptual results were shown in Table 3. The average identification accuracy showed a rising trend, growing from 43.9% without airflow information to 84.8% with airflow-incorporated information. The identification accuracy of the minimal pair "*ji*" and "*qi*" even reached the ceiling level with the help of aspiration animations.

Table 3. The identification accuracy of minimal pairs

Confusable Consonants	Without Airflow Accuracy (%)	With Airflow Accuracy (%)
b([p]) vs. *p*([p'])	45.5	90.9
d([t]) vs. *t*([t'])	36.4	72.7
g([k]) vs. *k*([k'])	36.4	90.9
j([tɕ]) vs. *q*([tɕ'])	54.5	100
z([ʦ]) vs. *c*([ʦ'])	54.5	81.8
zh([tʂ]) vs. *ch*([tʂ'])	36.4	72.7
Mean	**43.9**	**84.8**

5. DISCUSSION AND CONCLUSION

In Mandarin, the aspiration airflow rate plays a decisive role in discriminating confusable consonants with the same place and similar manner of articulation [13-14]. Using PAS, our study quantitatively calculated the mean airflow during voicing, peak expiratory airflow and airflow duration of confusable Mandarin consonants. Results indicated confusable Mandarin consonants could be distinguished from each other by the airflow-related parameters. However, existing 3D Mandarin articulatory tutors mainly focus on the movements of articulators. Airflow information, which is important in distinguishing some confusable consonants, is not considered in earlier studies [12, 15-18].

Our current multimodal system contains dynamic airflow information from data collected by PAS. We aim at illustrating aerodynamic airflow differences of confusable Mandarin consonants through realistic presentation of the airflow information in our existing EMA-data-driving articulatory animation system. In the audio-visual test, the syllable identification accuracy of only 43.9% without airflow information was improved to over 84.8% with the airflow-incorporated 3D articulatory animation system which can illustrate the differences between the confusable consonants. By demonstrating visual airflow information, the current 3D articulation and aspiration system provides a promising pronunciation training method not only for Mandarin L2 learners, but also for speakers with certain pathological voices and hearing-loss children.

6. REFERENCES

[1] P. Badin, A. Ben Youssef, G. Bailly, F. Elisei and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," *Actes de SLATE*, pp. 1-10, 2010.

[2] K. Grauwinkel, B. Dewitt, S. Fagel, "Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech," *in Proceedings of Interspeech*, pp. 706–709, 2007.

[3] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication*, vol. 52, pp. 493-503, 2010.

[4] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three dimensional articulator model for speech acquisition by children with hearing loss," *in Proceedings of the 4th Internation Conference on Universal Access in Human Computer Interaction*, vol. 4554, pp. 786-794, Jul. 2007.

[5] S. Fagel, K. Madany, "A 3D virtual head as a tool for speech therapy for children," *in Proceedings of Interspeech*, pp. 2643–2646, 2008.

[6] J. Ma, R. Cole, W. Pellom, B. Ward, "Accurate automatic visible speech synthesis of arbitrary 3d models based on concatenation of diviseme motion capture data," *Computer Animation and Virtual Worlds*, pp. 485–500, 2004.

[7] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Communication*, vol. 41, pp. 303–329, 2003.

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

[9] L. Wang, H. Chen, J. J. Ouyang, "Evaluation of external and internal articulator dynamics for pronunciation learning," *in Proceedings of Interspeech*, pp. 2247-2250, 2009.

[10] H. Chen, L. Wang, W. Liu, P.A. Heng, "Combined X-ray and facial videos for phoneme-level articulator dynamics." *The Visual Computer*, vol. 26, pp. 477-486, Apr. 2010.

[11] L. Wang, H. Chen, S. Li, H. M. Meng, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication*, vol. 54, pp. 845-856, Sep. 2012,

[12] D. Zhang, X. Q. Liu, N. Yan, L. Wang, Y. Zhu, H. Chen, "A multi-channel/multi-speaker articulatory database in Mandarin for speech visualization," *in Proceedings of ISCSLP 2014*, pp. 299-303, 2014.

[13] K. Y. Chao, G. Khattab, and L. M. Chen, "Comparison of VOT Patterns in Mandarin Chinese and in English," *in Proceedings of the 4th Annual Hawaii International Conference on Arts and Humanities*, pp. 840–859, 2006.

[14] T. Cho and P. Ladefoged, "Variation and universals in VOT: evidence from 18 languages," *Journal of Phonetics*, vol. 27, pp. 207–29, 1999.

[15] 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 Proceedings of Int. Conf. Spoken Lang. Process.*, pp. 1802–1805, 2006.

[16] H. Li, M. H. Yang, J. H. Tao, "Speaker-independent lips and tongue visualization of vowels," *in Proceedings of ICASSP*, pp. 8106-8110, 2013.

[17] J. Yu, A. J. LI, Z. F. Wang, "Data-Driven 3D Visual Pronunciation of Chinese IPA for Language Learning," *in Proceedings of International Oriental COCOSDA Conference*, vol. 75, no. 1, pp. 93-98, 2013.

[18] J. Yu, A. J. LI, "3D visual pronunciation of Mandarin Chinese for language learning," *in Proceedings of IEEE International Conference on Image Processing*, pp. 2036-2040, 2014.

[19] E. M. Yiu, Y. M., Yuen, T. Whitehill, A. Winkworth, "Reliability and applicability of aerodynamic measures in dysphonia assessment," *Clinical Linguistics & Phonetics*, vol. 18, pp. 463–478, 2004.

[20] R. Netsell, W. Lotz, A. L.Shaughnessy, "Laryngeal aerodynamics associated with selected voice disorders," *American Journal of Otolaryngology,* vol. 5, pp. 397–403, 1984.

[21] S. Iwata, H. von Leden, D. Williams, "Airflow measurement during phonation," *Journal of Communicable Diseases*, vol. 5, pp. 67–79, 1972.

[22] X. Q. Liu, N. Yan, L. Wang, X. L. Wu, Manwa L. Ng, "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, 2013.

[23] J. Stam, "Real-Time Fluid Dynamics for Games," in *Proceedings of the Game Developer Conference*, 2003.

[24] J. Stam, "Stable Fluids," *in Proceedings of ACM SIGGRAPH'99 Conference Proceedings*, pp. 121-128, 1999.

[25] N. F. Chen, V, Shivakumar, M. Harikumar, B. Ma, H. Li, "Large-Scale Characterization of Mandarin Pronunciation Errors Made by native Speakers of European Languages," *in Proceedings of Interspeech,* pp. 2370-2374, 2013.

[26] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, H. Li, "iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent," *in Proceedings of Interspeech,* pp. 324-328, 2015.