GLOTTOGRAPHIC AND AERODYNAMIC ANALYSIS ON CONSONANT ASPIRATION AND ONSET F0 IN MANDARIN CHINESE

Yujie Chi, Kiyoshi Honda, and Jianguo Wei

College of Intelligence and Computing, Tianjin University, Tianjin, China
yujiechi@qq.com, khonda@sannet.ne.jp, jianguo@tju.edu.cn

ABSTRACT

Stop consonants in Mandarin Chinese are all voiceless at word-initial positions only showing aspirated and unaspirated distinctions. Between the two phonation types, voice onset time (VOT) shows a clear contrast in duration, whereas voice onset fundamental frequency (onset F0) does not, as seen in previous studies. This study reports our first use of improved instrumentation techniques to record speech sound, oral airflow and glottal activity to examine consonant aspiration and onset F0. Glottal abduction, adduction and vibration cycles are monitored by a new external photo-glottographic system (ePGG) refined for better signal quality to detect accurate voice onset. Experimental data on Mandarin stops obtained from two male subjects suggests that consonant aspiration results in a large variation of VOT and oral airflow at voice onset. The onset F0 shows individual variation of falling and rising contours, and it is higher in aspirated stops than in unaspirated ones.

Index Terms— consonant aspiration, onset F0, external PGG, pneumotachography, Mandarin Chinese

1. INTRODUCTION

Production of stop consonants follows a strict temporal sequence of laryngeal adjustment and articulator movement. The voice onset time (VOT) is widely used to distinguish voiced, voiceless unaspirated and voiceless aspirated obstruents [1-3]. Voice onset fundamental frequency (onset F0) has been known as an effective cue for perception of voiced and voiceless consonants even when VOT is ambiguous [4, 5]. That is, voiced consonants have a lower onset F0 and a flat or rising contour, and voiceless ones have high onset F0 and a falling contour.

The causal mechanism of onset F0 has been explained by active and passive factors. The active factor is cricothyroid activity augmented for rapid cessation of vocal fold vibration in both aspirated and unaspirated voiceless stops [6]. Such a perturbation on F0 is reported to last more than 100 ms during the following vowel, but it is shorter in tonal languages [7]. The passive factor for onset F0 is the effect of glottal airflow at voice onset. Consonant aspiration has been reported to have a transient effect on onset F0 [7]. From an aerodynamic view, aspirated stops involve a higher airflow rate with lower subglottal pressure at voice onset, which may account for the lower onset F0 following voiceless aspirated stops than voiceless unaspirated stops. However, previous studies show inconsistent results across languages or speakers. Among studies on Mandarin Chinese, Xu shows an F0 lowering effect of aspiration [8]; Shimizu observes a raising effect [9]; and Shi suggests no clear pattern for the distinction [10].

Aspirated stops are characterized by the wide open glottis during closure, rapid airflow after the release, and longer VOT in contrast to unaspirated stops. Because of technical limitations, however, most related studies conventionally use speech signals to detect voice onset and extract onset F0, allowing a certain signal ambiguity. Photoglottography is suitable for such a study and also excels at voice onset detection [11]. For this purpose, the external lighting and sensing photoglottography (ePGG) was developed as a non-invasive technique having no limitation of vowels to analyze [12-14].

In this study, a new ePGG system was constructed for the better performance and combined with a new fiber-mask pneumotachography [15] (see Figure 1). Both are modified from the original systems [14] to observe glottal variation together with changes in airflow rate. The target measures are set at the release of consonants, voice onset, VOT and onset F0 in Mandarin Chinese.

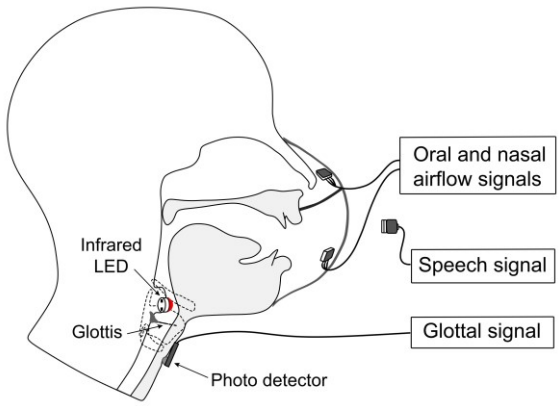


Figure 1, *Experimental setup for the combined measurement of audio, ePGG, and airflow signals.*

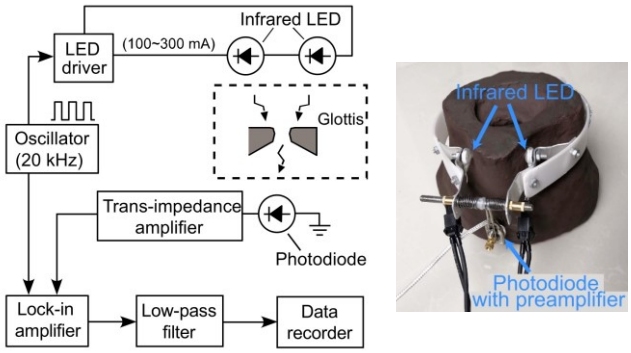


Figure 2, *Block diagram of the improved ePGG system with a picture for the component arrangement.*

2. METHOD

2.1. Improvement of external photoglottography (ePGG)

The conventional photoglottography (PGG) is a mildly invasive technique with the use of a fiberscope as a light source for glottal transillumination. The external photo-glottography (ePGG) is a non-invasive version of PGG because both light source and detector are placed outside the neck. This non-invasive technique has to detect much weaker light intensity, since the light must pass through the tissue near the larynx twice, which is critical especially for subjects with thick neck tissue or strong muscles. In addition, the larynx moves up and down during speech, and the lighting and sensing positions shift in relation to the position of the glottis. The ePGG was also sensitive to the conditions of illumination, optical distance to the airway and direction of the light beams. Those two problems remained to be solved. The most serious drawback is the insufficient gain of a photodetector circuit. The second is the fluctuation of the signal baseline due to the laryngeal movement.

To solve the first problem of a poor signal-to-noise ratio (SNR), many different high-gain and low-noise circuits were tested. Our final solution was a trans-impedance amplifier circuit using a high-performance analog IC (LTC6268) to amplify a small photo-current from a PIN photodiode (BPW34).

Another improvement was made by adopting a lock-in amplifier that is commonly used for extracting signals from high-amplitude noise. A square wave oscillator at 20 kHz for driving infrared LEDs is employed to avoid ambient noise from low-frequency environment light. The same square pulses are used as the carrier waves to operate the lock-in amplifier so as to extract clean glottal signals from the amplitude-modulated carrier waves and further increase the SNR.

During speech production, the larynx and articulators change in position, which results in the baseline fluctuation of ePGG signals. Although no good ways are found to limit the relative movement between the ePGG device and laryngeal structure, a tight strap was thought to be beneficial. Figure 2 (on the right) shows the supporting neck strap, which was used to keep the distance between the light source and photodetector and the angles of the LEDs. Two high-power infrared LEDs (1W, 850 nm) were lifted toward the neck tissue using a triangular base to optimize the direction for illumination. The distance between the two LEDs was adjusted by a double-screw bolt. The sensing unit is well shielded and supported by a bar of double torsion springs to tightly press on the skin. A small convex lens is attached on the photodiode for a better light coupling.

2.2. Fiber-mask pneumotachography

A fiber-mask pneumotachographic system was used to measure the oral airflow rate. The mask covering the mouth and nose has a septum to detect oral and nasal air pressures separately. Miniature pressure sensors (Honeywell, HSCSAAN001NDAA5) were attached directly on the mask so that the airflow rate can be computed for each channel according to the principle of pressure difference anemometry. The detail of the system is reported elsewhere [15]. In the experiment, both channels were recorded, but only the oral channel data was used for the analysis. The unit output voltage (1V) roughly corresponds to 0.45 L/s.

2.3. Measurement

Ten target consonant-vowel (CV) monosyllabic words contrasting two types of aspiration in Mandarin are used as test words. Aspirated and unaspirated voiceless bilabial stops, dentoalveolar stops and velar stops are combined with high and low vowels to form the words, as shown in Table 1. Velar stops are lexically missing with vowel /i/ in Mandarin.

Chinese characters having high tones (55) were selected for all the syllables according to Wu's study on Mandarin obstruents [16]. Words with affricates are included for a future analysis. Each word is repeated twelve times in a randomized order and printed in a carrier sentence /tu ʈʂʰu __ tsɻ/ ("Please speak out the word __." in English).

Speech sound was recorded simultaneously with ePGG and oral-nasal pneumotachographic signals, as shown in Figure 1. An electret condenser microphone (AT9903, Audio Technica, omnidirectional) is fixed on a head worn supporter. Totally four channels of signals are recorded with a digital data recorder (DAS40, Sefram, 14 bit, 50 kHz).

Table 1, *Consonant-vowel syllables in the stimuli*

	Bilabial	**Dento-alveolar**	**Velar**
Aspirated stops	/pʰa/ /pʰi/	/tʰa/ /tʰi/	/kʰa/
Unaspirated stops	/pa/ /pi/	/ta/ /ti/	/ka/

The recording was carried out in a soundproof room. Two male Mandarin subjects (ages of 24 and 25) participated in this experiment. After wearing the devices, the subject was asked to say /p^{h}a/ before and during intervals of the recording. If a clear peak appears in the oral airflow signal and not in the nasal airflow signal, the mask is judged to be correctly worn. At the same time, a small peak in the ePGG signal indicates the proper setting of the ePGG device.

2.4. Data analysis

The ePGG, oral airflow and speech data from the data recorder were processed with Python scripts. With assistance of the Praat software, each target word was labeled [17].

Firstly, the time delay across channels was corrected. For male subjects, the approximate distance between the glottis and oral pressure sensor was used for the delay of airflow; with an additional 3-cm distance to the microphone for the delay of speech signals. Then, ePGG and oral airflow data were low-pass filtered at 400 Hz to reduce high-frequency noise using a 3rd-order Bessel filter.

Figure 3 gives a data example of speech, glottal aperture and oral airflow in a carrier sentence. The CV syllable in the figure is the aspirated stop /p^h/ with vowel /a/. The preceding consonant is an aspirated affricate (/t$\underset{\sim}{\text{s}}^h$/). The two peaks in the ePGG signal correspond to glottal openings for the two stops with aspiration. As marked by blue dashed lines, the peak glottal opening is seen slightly after the release point. Comparing the aspirated and unaspirated stops in Figure 4, the oral airflow peak after the release indicates the initiation of aspiration. Then the glottal area and oral airflow decline to prepare for vowel /a/. The voicing may start with aspiration noise.

In this preliminary analysis, the release point and voice onset point were manually labeled on a chart combining speech, AC component of ePGG signal (ePGG-AC) and oral airflow signal in order to obtain accurate and reliable measures. The ePGG-AC signal was plotted after high-pass filtering (75Hz, 3rd-order Bessel type).

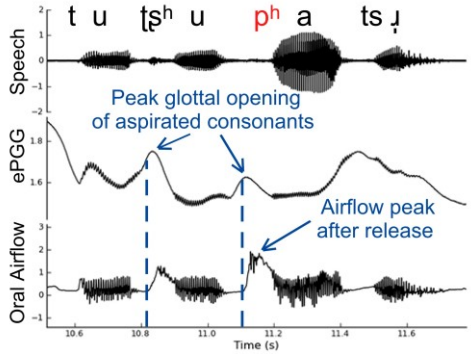
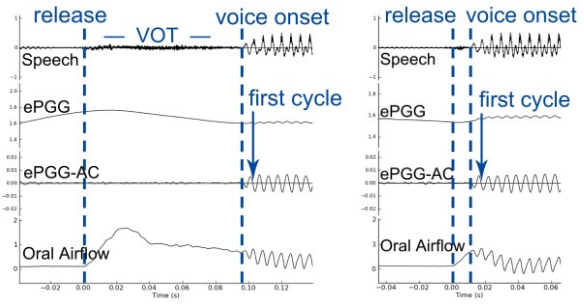


Figure 3, *Data example showing speech, ePGG, and oral airflow waveforms.*



a. *aspirated stop* b. *unaspirated stop*

Figure 4, *Labeling of consonant release and voice onset on the filtered waveforms.*

The voice onset time (VOT) of the target syllables was measured as the time interval between the stop release and voice onset. As shown in Figure 4, the voice onset point is located at initiation of the first vibration cycle in the ePGG signal. To be consistent in measurement, the voice onset is marked at a zero-crossing point before a negative deflection, which roughly corresponds to a peak of sound pressure. The onset F0 contour was measured cycle by cycle from the voice onset point to the 100-ms point using a combination of basic zero-crossing and peak-to-peak methods. This procedure insures accurate F0 of the first cycle at the cost of a high sensitivity to noise. The onset F0 value in the analysis is an average of the initial 10 ms of the contour. The onset airflow rate was sampled at the voice onset point in the DC oral airflow signal.

3. RESULTS

Some instructive results were observed by a pairwise comparison among VOT, onset F0 value and onset oral airflow rate for the aspirated and unaspirated stops. In Figure 5, scatter diagrams were drawn in six panels from three repetitions by two male subjects (GG and ZZ). The top four panels share the same horizontal axis for VOT. It is clear to identify the unaspirated stops with shorter VOT values within 25 ms and the aspirated stops with longer VOT values beyond 70 ms. The VOT for vowel /i/ is longer than that for vowel /a/ possibly due to the higher intraoral pressure. In the figure, the onset oral airflow rate and onset F0 do not show a clear separation between the aspirated and unaspirated stops. Those values for the aspirated stops are varied even within a subject. Aspirated stops are highlighted in the bottom two panels. It is interesting to note that for subject GG, the onset oral airflow rate shows a negative relationship with onset F0, but not for subject ZZ.

The onset F0 contours also exhibit the similar results. For the unaspirated stops, both subjects demonstrate high falling curves. But for the aspirated stops, ZZ's data regularly gives high falling contours, while GG's data does not. An example of F0 contours from subject ZZ is shown in Figure 6. The left top panel is F0 contours calculated from

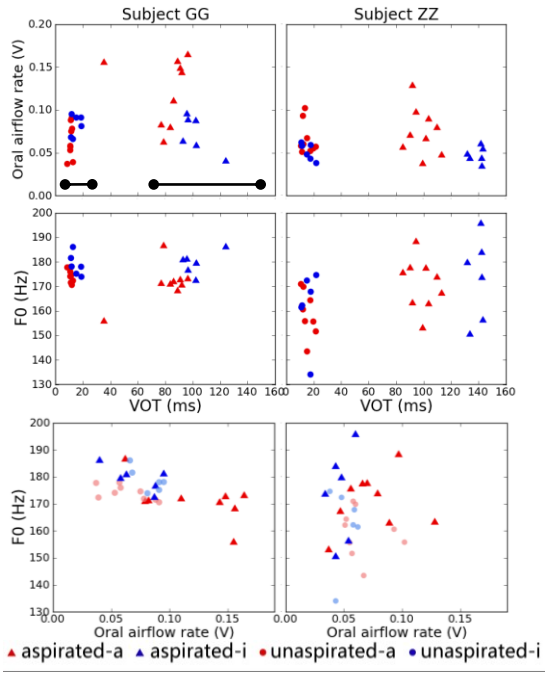


Figure 5, *Scattered diagrams showing the relations among VOT, onset F0 value, and onset oral airflow rate.*

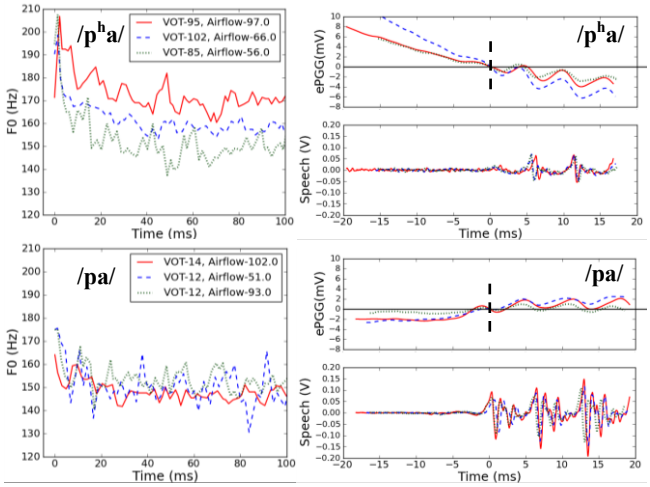


Figure 6, *F0 contour for 100 ms (left) and onset three glottal cycles of ePGG and speech waveforms (right) in three repetitions*

the aspirated syllable /pʰa/ for three repetitions, and the left bottom is from the unaspirated /pa/.

In this example, consonant aspiration appears to have a raising effect on onset F0 (205 Hz for aspirated vs. 175 Hz for unaspirated). The onset F0 values (average of the first 10 ms) show an unclear contrast between the aspirate and unaspirated stops, as seen in the middle right panel in Figure 5.

The right row of Figure 6 may explain the mechanism. The ePGG signals of the first three glottal cycles are plotted together with speech waves. All the waveforms are aligned at the voice onset point, i.e., zero point of the timeline. For the unaspirated stop /p/, the glottis is nearly closed during stop closure and does not open for aspiration after the release. Therefore, glottal vibration starts from the open glottis and initiates modal voice in the first cycle. On the contrary, for the aspirated stop /pʰ/, the glottis is wide open during the stop closure and release and then narrows gradually. During the beginning glottal cycles, vibration starts without full glottal closure. The patterns of incomplete vibration vary across places of articulation, vowels, etc. In the aspirated bilabial stop of subject ZZ, the first cycles start at higher F0 (nearly 200 Hz) than in the unaspirated stop.

4. DISCUSSIONS AND CONCLUSION

This study attempted to investigate the relationship between consonant aspiration and onset F0 in production of word-initial aspirated and unaspirated stops in Mandarin. To do so, improved photoglottography and pneumotachography were used simultaneously to record glottal aperture, oral airflow and speech signals. The external photoglottography (ePGG) was shown to be capable of tracing subtle variations of glottal events in frequency and amplitude. Experimental data on Mandarin stops obtained from two male subjects shows inconsistent onset F0 contours after the aspirated stops, partly agreeing with previous reports. Thus, the consonant aspiration results in a large variation of onset F0 across subjects, despite a smaller range of variation within a subject.

As reported by Francis [18], the aerodynamic effect of consonant aspiration on F0 is seen in the first glottal cycle within 10 ms. The similar pattern is found in many related studies [4, 9, 19]. In the rest of the period, both unaspirated and aspirated stops have high-falling F0 contours. Thus, combination of speech, ePGG and oral airflow signals describes such varied glottal activities in the first cycle. Note that the voice onset points defined in this analysis may not be consistent with what are measured from speech signals and may also be affected by noise in ePGG data.

In Mandarin, a clear contrast of VOT and consonant aspiration was observed between aspirated and unaspirated stops, whereas onset F0 does not appear to reflect such a contrast. This disparity may arise from ongoing parallel controls for the productions of tones, aspiration, and vowel onset. Clarifications for the manner of such complex controls may not be achieved only by point-based measurements, and examinations of patterns in glottal opening and airflow changes during consonant periods may provide a hint to resolve the question in a future study.

Acknowledgement

This work was supported in part by NSFC key project of Tianjin (No. 16JCZDJC35400), 435 and in part by grants from the National Natural Science Foundation of China (General 436 Program No.61471259 and No.61573254).

5. REFERENCES

[1] 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.

[2] L. Lisker and A. S. Abramson, "Distinctive features and laryngeal control," *Language*, vol. 47, no. 4, pp. 767-785, 1971.

[3] A. S. Abramson and D. Whalen, "Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions," *Journal of Phonetics*, vol. 64, pp. 75-86, 2017.

[4] R. N. Ohde, "Fundamental frequency as an acoustic correlate of stop consonant voicing," *Journal of the Acoustical Society of America*, vol.75, no. 1, pp. 224-230, 1984.

[5] D. H. Whalen, A. S. Abramson, L. Lisker, and M. Mody, "F0 gives voicing information even with unambiguous voice onset times," *Journal of the Acoustical Society of America*, vol. 93, no. 4, pp. 2152-2159, 1993.

[6] A. Löfqvist, T. Baer, N. S. McGarr, R. S. Story, "The cricothyroid muscle in voicing control," *Journal of the Acoustical Society of America*, vol. 85, no. 3, pp. 1314-1321, 1989.

[7] J.-M. Hombert, J. J. Ohala, and W. G. Ewan, "Phonetic explanations for the development of tones," *Language*, vol. 55, no. 1, pp. 37-58, 1979.

[8] C. X. Xu and Y. Xu, "Effects of consonant aspiration on Mandarin tones," *Journal of the International Phonetic Association*, vol. 33, pp. 165-181, 2003

[9] K. Shimizu, *A Cross-language Study of Voicing Contrasts of Stop Consonants in Asian Languages*, Ph.D. Thesis, University of Edinburgh, 1990.

[10] F. Shi, "The influence of aspiration on tones," *Journal of Chinese Linguistics*, vol. 26, no. 1, pp. 126–145, 1998. (in Chinese).

[11] J. Ohala, "A new photo-electric glottograph," *UCLA Working Papers in Phonetics*, vol. 4, pp. 40-52, 1966.

[12] K. Honda and S. Maeda, "Glottal-opening and airflow pattern during production of voiceless fricatives: a new non-invasive instrumentation," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3738, 2008.

[13] K. Honda and S. Maeda, "Non-invasive photoelectro-glottography method and device," *United States Patent Application Publication,* US 2010/0256503, 2010.

[14] J. Vaissière, K. Honda, A. Amelot, S. Maeda, and L. Crevier-Buchman, "Multisensor platform for speech physiology research in a phonetics laboratory," *Journal of the Phonetic Society of Japan*, vol. 14, no. 2, pp. 65-77, 2010.

[15] Y. Chi, K. Honda, J. Wei, H. Feng, and J. Wu, "Measuring oral and nasal airflow in production of Chinese plosive," *Proceedings of Interspeech 2015*, pp. 2167-2171, 2015.

[16] Z. J. Wu, "Aspirated vs non-aspirated stops and affricates in Standard Chinese," *Proceedings of the 11th International Congress of Phonetic Sciences (ICPhS)*, pp. 209-212, 1987.

[17] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341-345, 2001.

[18] A. L. Francis, V. Ciocca, V. K. Wong, and J. K. Chan, "Is fundamental frequency a cue to aspiration in initial stops?" *Journal of the Acoustical Society of America,* vol. 120, no. 5, pp. 2884-2895, 2006.

[19] Q. Luo, K. Durvasula, and Y. H. Lin, "Inconsistent consonantal effects on F0 in Cantonese and Mandarin," *Proceedings of Tonal Aspects of Languages* 2016, pp. 52-55, 2016.