Accuracy and variability of acoustic measures of voicing onset[a)]

Alexander L. Francis,[b)] Valter Ciocca,[c)] and Jojo Man Ching Yu
Department of Speech and Hearing Sciences, University of Hong Kong

(Received 28 May 2002; revised 31 October 2002; accepted 18 November 2002)

Five commonly used methods for determining the onset of voicing of syllable-initial stop consonants were compared. The speech and glottal activity of 16 native speakers of Cantonese with normal voice quality were investigated during the production of consonant vowel (CV) syllables in Cantonese. Syllables consisted of the initial consonants /p^h/, /t^h/, /k^h/, /p/, /t/, and /k/ followed by the vowel /a/. All syllables had a high level tone, and were all real words in Cantonese. Measurements of voicing onset were made based on the onset of periodicity in the acoustic waveform, and on spectrographic measures of the onset of a voicing bar (f_0), the onset of the first formant (F1), second formant (F2), and third formant (F3). These measurements were then compared against the onset of glottal opening as determined by electroglottography. Both accuracy and variability of each measure were calculated. Results suggest that the presence of aspiration in a syllable decreased the accuracy and increased the variability of spectrogram-based measurements, but did not strongly affect measurements made from the acoustic waveform. Overall, the acoustic waveform provided the most accurate estimate of voicing onset; measurements made from the amplitude waveform were also the least variable of the five measures. These results can be explained as a consequence of differences in spectral tilt of the voicing source in breathy versus modal phonation. © *2003 Acoustical Society of America.* [DOI: 10.1121/1.1536169]

PACS numbers: 43.70.Jt [AL]

I. INTRODUCTION

The accurate determination of voicing onset from acoustic signals is important both theoretically and clinically. From a clinical perspective, the onset of voicing is often used in assessing developmental maturation of neuromotor coordination (DiSimoni, 1974; Eguchi and Hirsh, 1969; Zlatin and Koenigsknecht, 1976) and constitutes an important part of the assessment of the speech production of hearing impaired talkers (Monsen, 1976). Measurement of voicing onset is also used in therapeutic applications for stutters (Borden *et al.*, 1985). From a theoretical perspective, the voice onset time (VOT) of stop consonants often serves as a significant acoustic correlate of, and perceptual cue to, category differences between voiced and voiceless, and aspirated and unaspirated stop consonant categories (Abramson and Lisker, 1970; Klatt, 1975; Lisker, 1975, 1978; Lisker and Abramson, 1970, 1964). While the terms VOT and voicing onset are used throughout this paper in accordance with established practice, it is important to note that VOT is merely one of a large set of interrelated acoustic consequences of variation in the relative timing of glottal and oral gestures (Abramson, 1977). In principle any (or all) of these acoustic correlates of laryngeal timing could conceivably function as a perceptual cue in the right circumstances (e.g., the absence of other, more predictable or more salient cues). However, in this article we are primarily concerned with the technical requirements of accurately estimating the onset of laryngeal oscillation (''voicing'') from an acoustic signal, and do not intend to address questions of how listeners might *perceive* the onset of voicing except insofar as it may influence talkers' production patterns.

While voicing onset can be accurately determined by electroglottography (Fourcin and Abberton, 1971), there are cases in which electroglottography is not possible, for example when making field recordings at remote sites or when analyzing speech from recordings that have already been made without an accompanying record of glottal function. Thus, it is often necessary to be able to identify the onset of voicing on the basis of an acoustic analysis alone.

Currently, it appears that the most commonly used methods for measuring the onset of voicing are based on the onset of periodicity in the acoustic waveform, possibly supplemented by spectrographic analyses (Abramson, 1995) or direct measurements of airflow (e.g., Koenig, 2001). However, this issue may have been decided more on the basis of expedience than precision, and it is not clear from the existing literature whether (or why) waveform-based measures should be considered preferable to many of the other acoustic measures of voicing onset that have been proposed in the past. For example, Peterson and Lehiste (1960) identified the onset of voicing as the point at which stable striations first become visible in the frequency region of the first formant of a wide band spectrogram. In contrast, Klatt (1975) made his measurements of voicing onset at the onset of visible energy in higher formants on the grounds that voicing onset might not always be visible in the first formant region. Finally, Lisker and Abramson (1964) determined the onset of voicing according to the time of the first vertical striations visible in

[a)]Some of the material in this article was presented at the 9th meeting of the International Clinical Phonetics and Linguistic Association, May 4th, 2002, Hong Kong, China.

[b)]Present address: Audiology and Speech Sciences, Purdue University, Heavilon Hall, West Layfayette, IN 47907. Electronic mail: francisa@purdue.edu
[c)]Electronic mail: vciocca@hkusua.hku.hk

a wideband spectrogram, presumably irrespective of the frequency (or formant) at which they first appeared. In addition to these spectrographic measures, it is also possible to measure the onset of voicing as the onset of energy visible in the ''voicing bar''—the region of lowest frequency energy in a wide-band spectrogram corresponding to the fundamental frequency (f_0), typically found below the first formant (Kent and Read, 2002, p. 144). Finally, it is possible to measure voicing onset directly from the acoustic waveform itself, in terms of the onset of the first clearly periodic pattern in the acoustic signal (see, e.g., Lieberman and Blumstein, 1988, p. 216). Until now there does not seem to have been a rigorous attempt to compare the efficacy of these different techniques.

Given the wide range of possibilities for measuring the onset of voicing from an acoustic signal, why should it matter which acoustic landmark (e.g., f_0, F1, F2, F3, or onset of waveform periodicity) is used? One important issue is the relative accuracy of measurements made at each landmark, as there are obvious differences between the latency of voicing onset that each indicates. For example, voicing striations typically appear later in higher formants, though this effect may be mitigated for voiceless stops because of the reduction in amplitude of the first formant transition following the release of such stops (Liberman *et al.*, 1958). If all measures are made using the same acoustic benchmarks (e.g., F3), accurate comparison may still be possible across syllables. For example, if it is determined that measurements of voicing onset based on the onset of striations in the third formant (F3) lag behind those made from the onset of striations in the first formant (F1), it may still be possible to reliably compare measurements made from F3 across utterances or talkers. However, such comparison depends on the assumption that measurements made at all landmarks are equally variable. If measurements made at a particular landmark vary as a function of independent factors such as consonant aspiration or talker identity, then that landmark would be considered less useful. The ideal acoustic measurement of voicing onset is one that is both accurate and relatively consistent. Its latency must closely match the physiological onset of vocal-fold vibration, and must remain unaffected by factors unrelated to the physical initiation of vocal fold vibration.

The present study compared the accuracy and variability of five acoustic measures of voicing onset. Accuracy was measured in terms of the mean asynchrony between measurements made at each landmark and the time of voicing onset determined by electroglottography. Variability was measured in terms of the variance of the mean asynchrony for each landmark.

II. METHOD

A. Subjects

Sixteen native speakers of Cantonese (eight men, eight women) with normal voice quality (as judged by unanimous agreement of three final year speech therapy students) and no reported history of speaking or hearing disability participated in this study. Their age ranged from 20 to 25 years (mean $=22.06$ years).

B. Stimuli

Stimuli consisted of six monosyllabic real words in Cantonese, all produced with a high level tone. All words had a consonant–vowel (CV) syllable structure with the vowel /a/ approximately as in the English word ''father.'' The words were /pa/ (father), /p^ha/ (on all fours), /ta/ (dozen), /t^ha/ (he), /ka/ (home), and /k^ha/ (compartment). Note that Cantonese is traditionally described as maintaining a phonological contrast between voiceless unaspirated and voiceless aspirated stop series at three places of articulation. Perceptually, aspiration cues seem to be stronger indictors of this phonological contrast than are timing (VOT) cues (Tsui and Ciocca, 2000). However, it may be argued that both timing- and aspiration-related acoustic patterns reflect the same relative timing of oral and laryngeal gestures (Abramson, 1977; Davis, 1994). Thus, regardless of what listeners are listening for (aspiration noise, VOT, or some combination), here we are primarily concerned with how best to use the acoustic signal to infer the relative timing of two articulatory events—events that together result in *both* the presence or absence of aspiration noise *and* longer or shorter VOT.

C. Procedure

Stimuli were recorded in a sound-shielded room using a low noise omnidirectional microphone (Shure Beta 87) with a Bruel and Kjær model 2812 MKII preamplifier and a Kay Elemetrics model 6094 Laryngograph connected to a TASCAM DA-30 MKII DAT tape recorder. The laryngographic signal was recorded to the left channel and the acoustic signal was recorded to the right channel. During recording, the microphone was mounted on a boom and situated approximately 10 cm in front of the talkers' lips. The laryngograph electrodes were held in place with a Velcro strap and were located slightly above and to either side of the talkers' thyroid cartilage. Each word was presented individually to the talker by means of file cards (one word, written in Chinese characters, per card). For each presentation of each word, the talker was asked to read the word aloud three times in a normal voice and at a comfortable rate with a preceding vowel /a/. For example, for the word /pa/ ''father'' the talker read [apa apa apa]. Each of the six words was presented a total of five times in randomized order. Only the middle instance of the three target syllables in each utterance was analyzed, in order to minimize any effect of the initiation or anticipation of the end of the utterance. That is, all measurements were taken from intervocalic stops at the onset of a stressed syllable. For each of the six words there were five syllables from which measurements of voicing onset were made. Items that were heard to be mispronounced during the recording session were repeated at the end of the recording list. However, talkers misspoke or misread four of the stimuli without being noticed during recording, producing syllables with an incorrect place of articulation, and in two cases also an incorrect degree of aspiration. These four tokens were not included in the final analysis.

D. Data analysis

Speech samples were low-pass filtered at 22 kHz and digitally sampled at 44.1 kHz using GW Instruments' SoundScope 16 software on a Macintosh 7200/120AV via a Digidesign Audiomedia II sound card. Acoustic and laryngographic (Lx) signals were digitized simultaneously, and stored in separate time-locked files linked by name. Initial spectrographic measurements from the acoustic signal were made with GW Instruments' SoundScope 16 using a wide-band spectrogram display set to a 300-Hz analysis bandwidth and a frame advance of 0.1 ms with a time scale of 5 ms per division. Approximately 12.5 divisions were shown, or 62.5 ms at this scale. These values were selected as a compromise to achieve good temporal resolution while still retaining accurate formant resolution for both male and female talkers. In some cases other bandwidth parameters were used to confirm measurements made from the 300-Hz display. For determining onset of periodicity in the Lx and acoustic waveform displays, varying time-scales were used. An initial estimate was made from a display window encompassing at least five or six periods of the vowel, plus the entire duration of the consonant burst release (if any) and aspiration (if any). Subsequently, the temporal resolution of the window was increased until the zero-crossing preceding the first clearly periodic component of the acoustic waveform could be accurately identified. In the case of the Lx signal the onset of periodicity did not always correspond to a zero-crossing because the low-magnitude oscillations in voltage corresponding to the glottal opening and closing gestures were superimposed on larger fluctuations in current level of an unknown origin. In such cases, the onset of periodicity was determined to be the lowest point of the wave immediately preceding the first upward-going component of the first clearly periodic cycle in the Lx waveform.

Voicing onset was measured from the acoustic signal in terms of the time from the start of the sound file to one of five acoustic landmarks (f_0, F1, F2, F3, waveform), as shown in Fig. 1. Note that we did not measure VOT (voice onset time, defined as the difference between the time of the burst release and the onset of voicing). Rather, we measured the time of voicing onset in terms of the duration from the (arbitrary) start of the file because there is no accurate landmark in the Lx signal from which to identify the time of the burst release. For the f_0 measurement, the onset of voicing was determined as the time of the onset of coherent energy visible in the lowest-frequency region of the spectrogram. For the waveform measurement, the onset of voicing was identified as the time of the zero crossing preceding the upward-going portion of the first cycle of oscillation visible in the acoustic waveform. For the F1 measurement, the onset of voicing was identified as the time of onset of the first vertical striation visible in the frequency region of the first formant. For the F2 measurement, the onset of voicing was identified as the time of onset of the first vertical striation extending upward through the frequency regions of the first and second formants without interruption. For the F3 measurement, the onset of voicing was identified as the time of onset of the first vertical striation extending upward through the frequency regions of the first, second, and third formants

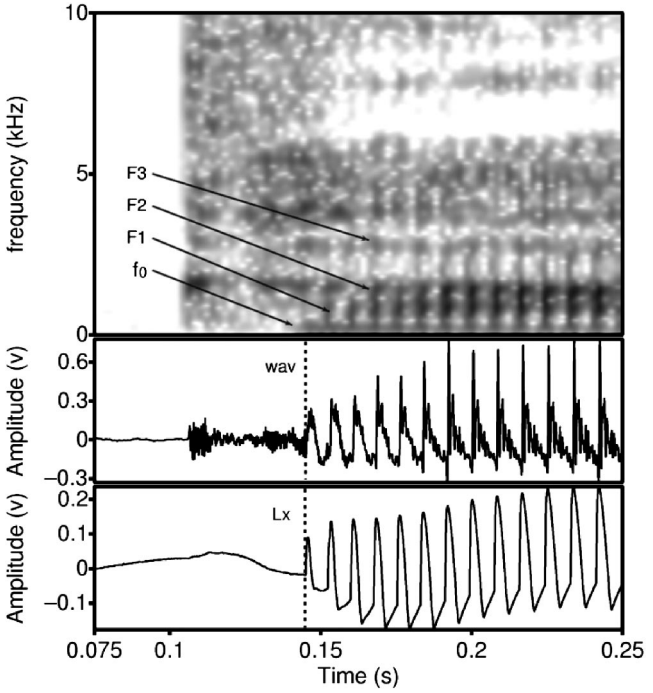

FIG. 1. Example of measurement locations in a representative syllable with an aspirated stop (male talker) in the spectrogram (top panel), acoustic waveform (middle) and Lx waveform (bottom). Measurement locations are indicated for six landmarks: fundamental frequency (f_0), first formant (F1), second formant (F2) and third formant (F3), the first upward-going zero-crossing preceding the first periodic wave component of the acoustic waveform (wav), and the lowest point preceding the first periodic wave component of the laryngographic waveform (Lx).

without interruption. These five acoustic measures of voicing onset were compared to the onset of regular vocal fold oscillation shown in the Lx waveform, defined as the lowest point immediately preceding the first cycle of regular oscillation, and calculated in terms of the measured time from the start of the sound file.

The asynchrony of each acoustic landmark was calculated as the difference between the acoustically determined time of voicing onset and that calculated from the Lx waveform, in milliseconds. Thus, an onset of voicing identified at 120 ms from the beginning of the sound file according to the F2 landmark, when compared with an onset of voicing at 110 ms from the beginning of the laryngographic waveform file according to the Lx waveform, would result in an asynchrony value of $+10$ ms. This measure of asynchrony was used as an estimate of the accuracy of measurements made at each acoustic landmark.

While accuracy is important for many purposes, in other cases it may not matter whether an acoustic measure of voicing onset is accurate (in the sense of being close in time to the actual onset of vocal fold vibration) as long as the values measured at that landmark remain relatively constant across contexts. For example, if all measures of voicing onset using a particular acoustic landmark are consistently 10 ms greater than those derived from the Lx waveform, then it should still be possible to compare measurements across different talkers, consonants, or other contextual variables since all mea-

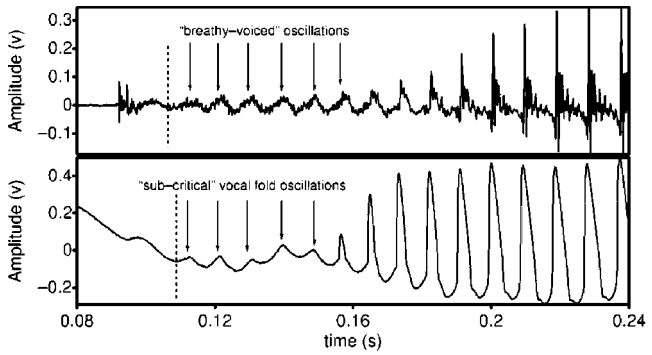

FIG. 2. Example of acoustic waveform (top) and Lx waveform (bottom) showing subcritical oscillations in vocal fold adduction visible in the Lx waveform corresponding to ''breathy-voicing'' periodicity in the acoustic signal. Measured point of onset of periodicity in the acoustic waveform is marked with a vertical cursor in the upper panel, and the vertical cursor in the lower panel indicates the measured location of the onset of oscillatory motion in the Lx signal.

sures will be biased to the same degree. However, if the asynchrony of measurements made at a particular landmark is highly variable, especially if it changes in a context dependent manner, then measurements made at that landmark cannot be considered reliable. Therefore, to determine the variability of measurements made at each landmark, we also calculated the variance of the asynchrony measured at each acoustic landmark for each syllable and each talker.

It should be noted that a large number of aspirated consonants exhibited a pattern in the Lx signal that suggested the presence of breathy voicing—subcritical oscillations of the vocal folds for around six or seven periods prior to the onset of clearly identifiable glottal opening/closing gestures, probably corresponding to the ''edge vibrations'' identified in spectrograms by Lisker and Abramson (1964, 1970). These patterns are similar to the glottal waveform of breathy vowels shown by Klatt and Klatt (1990, p. 822) (based on Stevens, 1977), and in the /aha/ productions of some of the talkers described by Löfqvist *et al.* (1995, p. 63). In the present study such patterns (illustrated in Fig. 2) were very common among female talkers, but every talker showed at least one production with such partial or breathy voicing (note that Fig. 2 shows the speech of a male talker). In all such tokens, voicing onset was estimated at the beginning of the upward-going curve of the first period of clear oscillatory motion in the Lx signal (shown with a vertical cursor in the lower panel of Fig. 2). Many utterances also showed a fluctuation in the Lx waveform coincident with the burst release in the acoustic waveform. In some cases (e.g., Fig. 1), this oscillation had a wavelength clearly much greater than that of typical voicing for the talker and was ignored (such long, solitary fluctuations may represent the influence on electrical impedance of extralaryngeal muscles involved in other aspects of speech production). In other cases, these fluctuations in the Lx waveform appeared much more like subcritical (breathy) voicing and were treated as subsequent to the onset of voicing. Finally, we note that the consonants under investigation, while syllable-initial, were not utterance-initial. As a result, it is possible that talkers continued the voicing of the preceding vowel through the stop consonant closure and into the subsequent vowel. While such articulations did occur occasionally, only productions in which there was a clear cessation of voicing preceding the burst release for the duration of at least approximately one period were included in these analyses.

III. RESULTS

A. Inter-rater reliability

In order to ensure that all measurements reported here reliably represent results that might be obtained by any experimenter, all of the waveform and Lx measures were redone by a second experimenter. In addition 20% of the tokens (one of the five repetitions of each of the consonants produced by each of the talkers) were randomly selected for remeasurement of the f_0, F1, F2, and F3 landmarks. All repeated measurements were made using the Praat 4.0 analysis software (Boersma and Weenink, 2001) with comparable spectrographic settings. Pearson's product-moment correlation analyses showed a high degree of inter-rater reliability, perhaps in part due to the strategy of using explicit, predetermined, acoustically defined measurement locations (e.g., the first visible vertical striation extending continuously through the first and second formant) facilitate inter-rater reliability. Pearson's correlation coefficients for each landmark were Lx, ($N=96$), $r=0.95$, $p<0.001$; waveform ($N=96$), $r=0.96$, $p<0.001$; f_0 ($N=96$), $r=0.97$, $p<0.001$; F1 ($N=96$), $r=0.97$, $p<0.001$; F2 ($N=96$), $r=0.96$, $p<0.001$; and F3 ($N=96$), $r=0.97$, $p<0.001$. All subsequent analyses used the measurements made by the second experimenter in those cases where measurements were repeated.

B. Accuracy

A four-way mixed factorial ANOVA (gender by aspiration by place of articulation by landmark) was calculated on the mean asynchrony values. Because of the very large differences in variance between cells in the design (for example, the mean variance for measurements of voicing onset made at the waveform landmark for [p] tokens produced by male speakers was 0.99 ms, while that for measurements made at the F3 landmark for [t^h] tokens was 428.02 ms), the Huynh–Feldt correction (Huynh and Feldt, 1970) was employed for repeated measures with more than two levels. Results showed no main effect of gender, $F(1,14)=0.059$, $p=0.81$, or of place of articulation, $F(1.48,20.72)=2.19$, $p=0.13$, but there were significant effects of landmark, $F(1.69,23.64)=57.05$, $p<0.001$, and aspiration, $F(1,14)=47.84$, $p<0.001$. There was also a significant two-way interaction between landmark and aspiration, $F(4,56)=61.54$, $p<0.00$, and a significant interaction between landmark, aspiration, and place of articulation, $F(8,112)=2.40$, $p=0.02$. None of the other interactions was significant at the $\alpha=0.05$ level. A graph of the three-way interactions (landmark by aspiration by place of articulation) is shown in Fig. 3. From this graph and posthoc analysis (Tukey HSD, $\alpha=0.05$), a number of observations can be made.

First, the different places of articulation pattern together at all levels of landmark and aspiration except that the /t^h/ tokens showed significantly greater offset ($p<0.05$) from

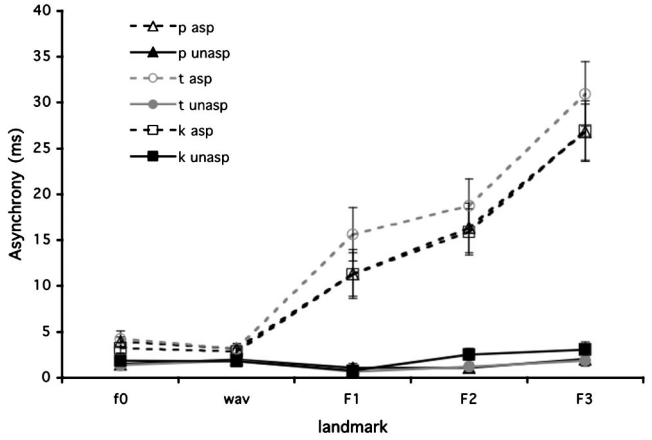

FIG. 3. Mean asynchrony between physiological (Lx) measures of voicing onset and measures of voicing onset made from the spectrographic display of the onset of energy at the fundamental frequency (f_0), first formant (F1), second formant (F2), third formant (F3), and from the first upward-going zero-crossing preceding the first periodic component of the acoustic waveform (wav). Measurements were made from syllables with aspirated stop consonants (open symbols and dotted lines) and unaspirated stop consonants (filled symbols and solid lines). Error bars indicate standard errors of the means.

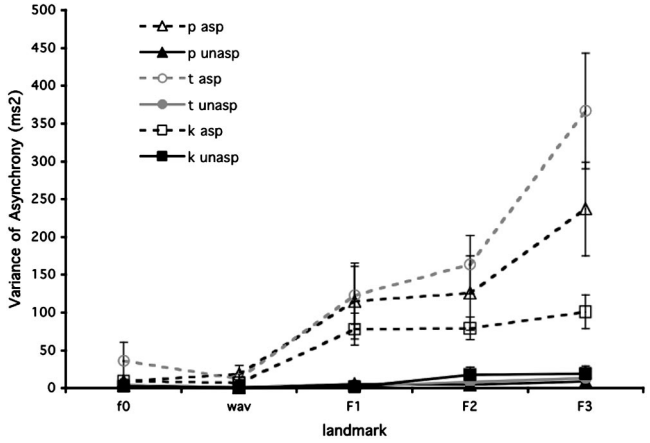

FIG. 4. Variance in mean asynchrony between physiological (Lx) measures of voicing onset and measures of voicing onset made from the spectrographic display of the onset of energy at the fundamental frequency (f_0), first formant (F1), second formant (F2), third formant (F3), and from the first upward-going zero-crossing preceding the first periodic component of the acoustic waveform (wav). Measurements were made from syllables with aspirated stop consonants (open symbols and dotted lines) and unaspirated stop consonants (filled symbols and solid lines). Error bars indicate standard errors of the means.

the /p^h/ and /k^h/ tokens at the F1 and F3 landmarks (only). These differences aside, there are no significant differences between places of articulation at the same levels of aspiration and landmark, and generalizing across place of articulation will not affect the overall validity of further conclusions. Collapsing across place of articulation, there was no significant difference between measurements of asynchrony of aspirated and unaspirated tokens made either using the f_0 or waveform landmarks ($p>0.05$). However, the asynchrony of aspirated tokens was significantly greater than that of unaspirated tokens when measured at any of the F1, F2, and F3 landmarks. These differences were due to an increased asynchrony for the aspirated tokens at the F1, F2, and F3 landmarks; in the unaspirated series there was no significant difference in voicing onset asynchrony between any of the landmarks ($p>0.05$ for all pairwise comparisons). For the aspirated series, the asynchrony of F1 measurements was significantly greater than that of measures made at f_0 or the waveform. Aspirated F1 asynchrony was also significantly smaller (more accurate) than F3, though there was no significant difference in asynchrony between the F1 and F2 measurements for aspirated consonants.

Note that the measurement criteria used for determining the location of the F2 and F3 landmarks meant that the asynchrony at these locations had to be equal to or greater than that measured from the F1 landmark because the F3 landmark could occur, by definition, no earlier than the F2, which in turn could be no earlier than the F1 landmark. However, as shown in Fig. 1, in the aspirated series the F2 and F3 landmarks could (and often did) appear at the same time, and indeed in the unaspirated series the three formant-based landmarks typically appeared at the same time. It is likely that this constraint on measurement lead to some degree of variability in measurements made from the formant-based landmarks. This is because it was often difficult to determine precisely whether the vertical striation of a particular glottal pulse extended without interruption through successive formants, even when the pulse was clearly evident within the frequency ranges of each formant. Still, the imposition of consistent display parameters may have reduced inter-rater variability due to this same uncertainty by encouraging both raters to use displays that were uncertain to a similar degree.

C. Variability

A four-way mixed factorial ANOVA (gender by aspiration by place of articulation by landmark) on the variance of the asynchrony measurements made at each landmark was also calculated. Results showed a main effect of landmark, $F(2.07,29.01)=25.56$ and aspiration, $F(1,14)=28.24$, $p<0.001$, but no effect of place of articulation, $F(2.00,28.00)=2.64$, $p=0.09$, or talker gender, $F(1,14)=0.06$, $p=0.80$. There were also significant two-way interactions between landmark and aspiration, $F(4,56)=22.42$, $p<0.001$, and landmark and place of articulation, $F(8,112)=3.79$, $p=0.001$, and a significant three-way interaction between landmark, aspiration and place of articulation, $F(8,112)=3.81$, $p=0.001$. None of the other interactions was statistically significant at the $\alpha=0.05$ level. Because talker gender did not play a role in any significant interaction, and was itself not a significant factor in the analysis, it was excluded from further analyses. A graph of the three-way interaction (landmark by aspiration by place of articulation) is shown in Fig. 4. From this graph and post-hoc (Tukey HSD, $\alpha=0.05$) analysis, a number of observations can be made.

TABLE I. Pearson's product moment correlations between mean asynchrony and variance of asynchrony. Values significant at the $\alpha = 0.05$ level are shown in bold.

	Aspirated	Unaspirated
f_0	0.44	0.26
	$p = 0.085$	$p = 0.319$
Wav	**−0.57**	**−0.79**
	$p = 0.021$	$p = 0.001$
F1	**0.82**	0.04
	$p = 0.001$	$p = 0.880$
F2	**0.76**	**0.64**
	$p = 0.001$	$p = 0.007$
F3	**0.60**	**0.75**
	$p = 0.015$	$p = 0.001$

Similar to the accuracy measurements, there was no significant difference between different places of articulation at any of the five landmarks for the unaspirated consonant series, and there was no significant effect of landmark in the unaspirated series. By contrast, in the aspirated series there were some significant differences in variance at the same landmark depending on place of articulation, though only for the F3 landmark. For example, the /k^h/ tokens showed significantly less variance at the F3 landmark than the /p^h/ or /t^h/ tokens, and the /p^h/ tokens were in turn less variable than the /t^h/ ones. However, there were no significant differences due to place of articulation in the aspirated series at any other landmark. Because the overall trend of the effect of landmark is the same for all three places of articulation, the effects of landmark were investigated by pooling across place of articulation. There was no significant difference between the variability of the asynchrony of aspirated versus unaspirated consonants when measured at either the f_0 or waveform landmarks. However, at the F1, F2, and F3 landmarks the variance of measurements made from unaspirated consonants was significantly smaller than that of measurements made from aspirated consonants. Finally, for the aspirated consonants the variance of the F1 measurements was not significantly different from the F2 measure, but both were significantly smaller than F3.

D. Relationship between asynchrony and variability

It may be noted that the distribution of variances shown in Fig. 4 appears quite similar to that of the mean scores shown in Fig. 3. Indeed, an analysis of the product moment correlations between the means and variances of each of the landmarks shows strong correlations between most of the asynchrony measurement means and their corresponding variances, as shown in Table I. These findings are consistent with the observation that variance typically increases with increasing means, at least in measurements of durations associated with speech production (e.g., Ohala, 1975; Smith *et al.*, 1983; Smith, 1992, 1994).

Because of the correlation between means and variances in this case, measurements of variance alone cannot be used to determine whether measurements at certain landmarks are *intrinsically* more variable. Although these results show that measurements made at F3 do have clearly higher variance, we cannot tell whether this is simply due to the general tendency of variability to increase with large means, or whether it is truly the case that F3 measurements are always more variable than F2 measurements. This issue is further complicated by the observation, described above, that mean measurements made at higher frequency landmarks will always be equal to or greater than those made at lower frequency landmarks. Note that the use of a measure that normalizes variances according to their corresponding means, such as the coefficient of variation (CoV, defined as the standard deviation divided by the mean), is not ideal for the present data set because many of the means involved are equal to or very close to zero. This is problematic for two reasons. First, as means approach zero the coefficient of variation approaches infinity, but there are too many such means to simply ignore these cells in the design. Similarly, because the means vary around zero (both above and below), the CoV can misrepresent equivalent differences in variance. Very small differences in means near zero result in highly divergent CoV values, while the same relative difference in means farther away from zero may result in quite similar CoV values, even for the same variance. For example, given two cells with the same standard deviation of 1.5, but means of -0.1 versus $+0.1$, the CoV for the first will be -15, and for the second it will be 15. However, for means of 0.1 and 0.3, each with a standard deviation of 1.5, the respective CoV values are 15 and 5 (one third as far apart). Thus, while we can say that measurements made at higher frequencies seem to have higher variability as well, we cannot say whether this is simply an inherent property of measurements with larger means, or whether formant-based measurements of voicing onset are necessarily more variable overall.

IV. DISCUSSION

The results of the present study suggest that the presence of aspiration after the release burst of stop consonants strongly affects the accuracy and variability of estimates of voicing onset made from acoustic measures. While the unaspirated stop series showed no significant differences in accuracy and variability of measurements of voicing onset across the five measurement landmarks, aspirated stops showed a consistently larger asynchrony and variance for measures made from the spectrogram at higher frequencies. These results imply that the most accurate acoustic measurements of voicing onset across categories can be made directly from the waveform using the onset of periodicity as an indicator of the onset of voicing or from the spectrogram using the onset of energy in the voicing bar.

Early research using acoustic measures of voicing onset typically relied on spectrographic measurements, perhaps because researchers could base their measurements on aspects of the signal known to be important in consonant perception. For example, Klatt (1975) suggested that measurements made according to the onset of energy in multiple (higher) formants may be more accurate than those made from the first formant alone because spectrographic displays (at that time) relied on display mechanisms (thermal printing) that were unable to adequately represent the subtle differences in energy over the small, low frequency ranges that indicate the onset of voicing in the voicing bar or first formant. However,

the development of digital spectrographic analyses may render moot such questions of visual precision, and the ability to employ high sampling rates (e.g., 44.1 kHz) makes it possible to identify patterns in the acoustic waveform with extremely high temporal precision as well. Certainly, the results presented here suggest that (*i*) measurements based on the waveform or voicing bar are generally more accurate and less variable than formant-based measurements, and (*ii*) F1-based measurements of voicing onset made from a digitally generated spectrogram are more accurate than F2- and F3-based measures.

The phenomenon of breathy voicing onsets observed in the productions of many aspirated consonants is also worthy of further discussion. Every talker in the present survey produced at least one token with a breathy voiced period between the burst release and the onset of modal voicing, and for some talkers, especially women, this pattern of glottal activity was the norm for aspirated stops. These results fit with the observations reported by Klatt and Klatt (1990), that female voices were typically perceived to be breathier than males. The Lx waveform displayed in Fig. 2 suggests that, as the vocal folds move from a fully abducted position to the nearly adducted configuration associated with modal voicing, they can pass through a phase of partial adduction. The Lx waveform does not directly measure the amount of vocal fold closure, but rather only indicates relative area of contact at the glottis (cf. Baken and Orlikoff, 2000, pp. 416–417). Therefore, from the Lx signal alone it cannot be determined whether this pattern of breathy voicing is accomplished by partially adducting the entire length of the vocal folds or by fully adducting a portion of the length of the vocal folds and maintaining an opening along the remaining length of the folds (cf. Hanson, 1997; Klatt and Klatt, 1990, p. 822). In either case, the resulting glottal configuration would support a voicing-like, periodic modulation of supra-glottal air pressure similar to the pattern indicated in Fig. 2, suggesting that this talker did achieve a degree of adduction sufficient to support periodic oscillation while maintaining sufficient abduction to allow a portion of the folds to produce breathy aspiration.

The prevalence of such breathy voicing following aspirated stop consonants suggests an articulatory account for certain acoustic observations made by Fischer-Jørgensen and Hutters (1981). As was found in the present experiment, Fischer-Jørgensen and Hutters (1981) found that, in open vowels, evidence for the onset of periodicity was often observed much earlier in lower-frequency regions of the spectrogram than in higher regions around the frequencies of the second and third formants. Acoustically, breathy phonation has been shown to exhibit greater spectral tilt than modal voicing, meaning that the amplitude of successive harmonics drops off more sharply as frequency increases in breathy phonation than in modal phonation, and higher harmonics are often replaced by aperiodic aspiration noise (Stevens, 1977). The fact that female talkers are more likely to be perceived as sounding breathy (Klatt and Klatt, 1990) has similarly been attributed to the greater rate of decrease in amplitude in successively higher harmonics in women's voices as compared with men's voices (cf. Holmberg *et al.*, 1988; Monson and Engebretson, 1977).

In the case of intervocalic stop consonants involving a cessation of voicing, a subsequent resumption of breathy voicing would mean that evidence for the onset of periodicity would not be found at higher frequencies until later in the utterance, when phonation has reverted to a modal pattern and the spectral tilt of the voicing source has once again become shallow enough to energize higher-frequency resonances of the vocal tract. Because of the complexity of the interrelationship between the articulatory gestures that control the transition from breathy to modal phonation, it seems plausible that this transition should be highly variable, both within and across talkers [see Löfqvist *et al.* (1995) for evidence of such variability]. In the present experiment, variability in estimates of voicing onset was indeed higher for measurements made at higher frequencies, especially following aspirated consonants. Because of the potential for a delayed appearance of voicing information at higher frequencies, and the concomitant increase in variability of measurements made at higher formants, the present results suggest that future investigations of voicing onset using acoustic measurements alone should be based on measurements made from the waveform or the voicing bar.

ACKNOWLEDGMENTS

This article is based on a dissertation submitted by the third author in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong. The authors would like to thank Arthur S. Abramson, Laura L. Koenig, and Anders Löfqvist for helpful comments on an earlier draft of this manuscript, and Raymond Wu, Donald Chan, and Kim-Ping Tsa for their technical assistance.

Abramson, A. S. (**1977**). "Laryngeal timing in consonant distinctions," Phonetica **34**, 295–303.

Abramson, A. S. (**1995**). "Laryngeal timing in Karen obstruents," in *Producing Speech: Contemporary Issues, for Katherine Safford Harris*, edited by F. Bell-Berti and L. J. Raphael (American Institute of Physics, New York), pp. 155–165.

Abramson, A. S., and Lisker, L. (**1970**). "Discriminability along the voicing continuum: Cross-language tests," in *Proceedings of the 6th International Congress of Phonetic Sciences*, pp. 569–573.

Baken, R. J., and Orlikoff, R. F. (**2000**). *Clinical Measurement of Speech and Voice*, 2nd ed. (Singular, San Diego, CA), pp. 416–417.

Boersma, P., and Weenink, D. (**2001**). *Praat 4.0: A system for doing phonetics by computer* (computer software) (University of Amsterdam, Amsterdam, The Netherlands). Available online: http://www.praat.org

Borden, G. J., Baer, T., and Kenney, M. K. (**1985**). "Onset of voicing in stuttered and fluent utterances," J. Speech Hear. Res. **28**, 363–372.

Davis, K. (**1994**). "Stop voicing in Hindi," J. Phonetics **22**, 177–193.

DiSimoni, F. G. (**1974**). "Effect of vowel environment on the duration of consonants in the speech of three-, six-, and nine-year-old children," J. Acoust. Soc. Am. **55**, 360–361.

Eguchi, S., and Hirsh, I. J. (**1969**). "Development of speech sounds in children," Acta Otolaryngol. (Stockh) **257**, 1–51.

Fischer-Jørgensen, E., and Hutters, B. (**1981**). "Aspirated stop consonants before low vowels. A problem of delimitation, its causes and consequences," Annual Report of the Institute of Phonetics, University of Copenhagen, Vol. 15.

Fourcin, A. J., and Abberton, E. (**1971**). "First applications of a new laryngograph," Medical and Biological Illustration, **21**, 172–182; reprinted in Volta Rev. **74**, 161–176.

Hanson, H. M. (**1997**). ''Glottal characteristics of female speakers: Acoustic correlates,'' J. Acoust. Soc. Am. **101**, 466–481.
Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (**1988**). ''Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice,'' J. Acoust. Soc. Am. **84**, 511–529.
Huynh, H., and Feldt, L. S. (**1970**). ''Conditions under which mean square ratios in repeated measures designs have exact *F*-distributions,'' J. Am. Stat. Assoc. **65**, 1582–1589.
Kent, R. D., and Read, C. (**2002**). *The Acoustic Analysis of Speech*, 2nd ed. (Singular, San Diego, CA), p. 144.
Klatt, D. H. (**1975**). ''Voice onset time, frication, and aspiration in word-initial consonant clusters,'' J. Speech Hear. Res. **18**, 686–706.
Klatt, D. H., and Klatt, L. C. (**1990**). ''Analysis, synthesis, and perception of voice quality variations among female and male talkers,'' J. Acoust. Soc. Am. **87**, 820–857.
Koenig, L. L. (**2001**). ''Distributional characteristics of VOT in children's voiceless aspirated stops and interpretation of developmental trends,'' J. Speech Lang. Hear. Res. **44**, 1058–1068.
Liberman, A. M., Delattre, P., and Cooper, F. S. (**1958**). ''Some cues for the distinction between voiced and voiceless stops in initial position,'' Lang Speech **1**, 153–167.
Lieberman, P., and Blumstein, S. E. (**1988**). *Speech Physiology, Speech Perception, and Acoustic Phonetics* (Cambridge U.P., New York), p. 216.
Lisker, L. (**1975**). ''Is it VOT or a first formant detector?'' J. Acoust. Soc. Am. **57**, 1547–1551.
Lisker, L. (**1978**). ''Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction,'' Status Report on Speech Research, SR-54 (Haskins Laboratories, New Haven, CT), pp. 127–132.
Lisker, L., and Abramson, A. S. (**1964**). ''A cross-language study of voicing in initial stops,'' Word **20**, 384–422.
Lisker, L., and Abramson, A. S. (**1970**). ''Some effects of context on voice onset time in English stops,'' in *Proceedings of the 6th International Congress of Phonetic Sciences*, pp. 563–567.
Löfqvist, A., Koenig, L. L., and McGowan, R. S. (**1995**). ''Vocal tract aerodynamics in /aCa/ utterances: Measurements,'' Speech Commun. **16**, 49–66.
Monsen, R. B. (**1976**). ''Normal and reduced phonological space: The production of vowels by a deaf adolescent,'' J. Phonetics **4**, 189–198.
Monson, R. B., and Engebretson, A. M. (**1977**). ''Study of variations in the male and female glottal wave,'' J. Acoust. Soc. Am. **62**, 981–993.
Ohala, J. J. (**1975**). ''The temporal regulation of speech,'' in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. Tatham (Academic, New York), pp. 431–453.
Peterson, G. E., and Lehiste, I. (**1960**). ''Duration of syllabic nuclei in English,'' J. Acoust. Soc. Am. **32**, 693–703.
Smith, B. L. (**1992**). ''Relationships between duration and temporal variability in children's speech,'' J. Acoust. Soc. Am. **91**, 2165–2174.
Smith, B. L. (**1994**). ''Effects of experimental manipulations and intrinsic contrasts on relationships between duration and temporal variability in children's and adult's speech,'' J. Phonetics **22**, 155–175.
Smith, B. L., Sugarman, M. D., and Long, S. H. (**1983**). ''Experimental manipulation of speaking rate for studying temporal variability in children's speech,'' J. Acoust. Soc. Am. **74**, 744–749.
Stevens, K. N. (**1977**). ''Physics of larynx behavior and larynx modes,'' Phonetica **34**, 264–279.
Tsui, I. Y. H., and Ciocca, V. (**2000**). ''Perception of aspiration and place of articulation of Cantonese initial stops by normal and sensorineural hearing-impaired listeners,'' Int. J. Lang. Commun. Disord. **35**, 507–525.
Zlatin, M., and Koenigsknecht, R. (**1976**). ''Development of voicing contrast: A comparison of voice onset time in perception and production,'' J. Speech Hear. Res. **19**, 93–111.