Is fundamental frequency a cue to aspiration in initial stops?

Alexander L. Francis,[a] Valter Ciocca, Virginia Ka Man Wong, and Jess Ka Lam Chan
Department of Speech and Hearing Sciences, University of Hong Kong, Hong Kong

(Received 11 January 2006; revised 7 August 2006; accepted 8 August 2006)

One production and one perception experiment were conducted to investigate the interaction of consonant voicing and fundamental frequency at the onset of voicing (onset f_0) in Cantonese, a tonal language. Consonantal voicing in English can affect onset f_0 up to 100 ms after voicing onset, but existing research provides inconclusive information regarding the effects of voicing on f_0 in tonal languages where f_0 variability is constrained by the demands of the lexical tone system. Previous research on consonantal effects on onset f_0 provides two contrasting theories: These effects may be automatic, resulting from physiological constraints inherent to the speech production mechanism or they may be controlled, produced as part of a process of cue enhancement for the perception of laryngeal contrasts. Results of experiment 1 showed that consonant aspiration affects onset f_0 in Cantonese only within the first 10 ms following voicing onset, comparable to results for other tonal languages. Experiment 2 showed that Cantonese listeners can use differences in onset f_0 to cue perception of the voicing contrast, but the minimum extent of f_0 perturbation necessary for this is greater than is found in Cantonese production, and comparable to that observed in acoustic studies of nontonal languages. These results suggest that consonantal effects on onset f_0 are at least partially controlled by talkers, but that their role in the perception of voicing/aspiration may be a consequence of language independent properties of audition rather than listeners' experience with the phonological contrasts of a specific language.
 [DOI: 10.1121/1.2346131]

PACS number(s): 43.70.Bk, 43.70.Fq, 43.71.Es, 43.71.An, 43.70.Kv [AL] Pages: 2884–2895

I. INTRODUCTION

In English, fundamental frequency at the vowel onset (onset f_0) is correlated with the phonological feature of voicing in initial stops, such that the onset f_0 following voiceless stops is higher than after voiced stops (Haggard, Ambler, and Callow, 1970; Hombert, 1978; House and Fairbanks, 1953; Lehiste and Peterson, 1961; Löfqvist, Baer, McGarr, and Seider Story, 1989; Ohde, 1984; Whalen, 1990). It has been argued that this pattern is *intrinsic* to voicing differences, meaning that the onset f_0 is determined by physiological and/or aerodynamic factors related to the relative timing of voicing and oral closure release in a manner that is not (entirely) under the control of the speaker (cf. Hombert, 1978; Ohala, 1978 pp. 25–29). However, the duration of consonant-related f_0 perturbations in English extends farther into the vowel than would be predicted by simple physiological or aerodynamic factors (Hombert, 1978). Because of its longer-than-expected duration, Kingston and Diehl (1994) argued that this effect is an intentional maneuver on the part of English speakers to help cue the perception of voiced consonants.

Further support for the hypothesis that English speakers may intentionally exaggerate consonant-related f_0 perturbations in order to enhance voicing perception is found in tonal languages: Gandour (1974) and Hombert (1977) both determined that voicing-related onset f_0 perturbations were shorter in Thai and Yoruba, respectively, than in nontonal languages, extending approximately 30–50 ms into the vowel. By contrast, onset f_0 perturbations are observed for 100 ms or more in nontonal languages such as English (House and Fairbanks, 1953; Lehiste and Peterson, 1961; Löfqvist, Baer, McGarr, and Seider Story, 1989; Whalen, 1990), French (Hombert, 1978) and Dutch (Löfqvist *et al.* 1989). Thus, it has been argued that tonal language speakers either actively seek to inhibit the extent of an intrinsic effect on f_0 or choose not to manipulate pitch as a voicing cue because they are already using pitch to cue lexical tone contrasts. In either case, the essential point is the same: Pitch is a strong cue for tone categorization but a relatively weak one for consonant identification (Abramson and Lisker, 1985; Haggard, Ambler, and Callow, 1970). Therefore, tone language speakers avoid allowing a consonant-associated f_0 event to affect pitch (and therefore tone) except perhaps very locally within the onset of a syllable (Hombert, 1978, p. 83). If tonal language speakers are able to control the effect of consonant voicing on onset f_0, then this effect must be at least partially controllable, and not purely intrinsic to the production of voicing.

However, one crucial causal link in this chain of reasoning is still missing: It is not clear whether tone language speakers are able to use f_0 as a cue to consonant voicing (or aspiration), let alone whether they actually do so in their native language. It has been shown that f_0 is not always a strong cue to consonant voicing even for speakers of languages that do show considerable correlation between onset f_0 and voice onset time (VOT) in production (Francis and

[a]Author to whom correspondence should be addressed. Department of Speech, Language and Hearing Sciences, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, Indiana 47907. Electronic mail: francisa@purdue.edu

Nusbaum, 2002; Haggard *et al.*, 1970; Whalen, Abramson, Lisker, and Mody, 1990). If f_0 is so critical to tone perception that speakers cannot afford to allow it to vary (much) according to voicing, then perhaps tone language listeners simply disregard this property of the signal when making consonant voicing decisions in the first place. In the present paper we revisit the interaction of consonantal perturbations of f_0 and tone and extend this research into the domain of perception. We ask whether native speakers of Cantonese, a tone language, produce consonants with limited voicing-associated f_0 perturbations like Thai and Yoruba speakers, and, if so, whether they are still able to use onset f_0 as a cue to consonant voicing.

A. Voicing and onset f_0

To some extent, the effects of English stop consonant voicing on onset f_0 are unexpected. Although English stops are often considered to contrast in terms of voicing, this distinction is primarily phonological. The phonetic distinction, especially in syllable-initial position is more properly characterized as one between voiceless unaspirated and voiceless aspirated stops (Keating, 1984; Lisker and Abramson, 1964; Zlatin, 1974). However, English voicing effects on onset f_0 follow a pattern similar to that found in languages that contrast phonetically *voiced* with phonetically *voiceless unaspirated* consonants. That is, English phonologically voiced (phonetically voiceless unaspirated) stops have the same lowering effect on onset f_0 as do French and Dutch phonologically (and phonetically) voiced stops, while both English phonologically voiceless (phonetically voiceless aspirated) and French and Dutch phonologically voiceless (phonetically voiceless unaspirated) stops have a raising effect on onset f_0 (French: Hombert, 1978; Dutch: Löfqvist *et al.* 1989). This type of patterning (among other factors) led Kingston and Diehl (1994) to argue that consonantal effects on onset f_0 are the result of controlled processes related to the *phonological* status of the consonant series (voiced vs voiceless) rather than a result of intrinsic physiological dependencies between the articulatory and/or aerodynamic properties of the production of different degrees of prevoicing or voicing delay.

Like English, syllable-initial stop consonants in Cantonese may be divided into two phonological voicing classes, typically termed *voiceless unaspirated* and *voiceless aspirated* (Bauer and Benedict, 1997; Lisker and Abramson, 1964). In this case, the phonological and phonetic terminologies agree, and this contrast is distinctive and phonemic. For example, /p/ in the word 錶/piu55/ ("watch") and /p^h/ in the word 飄/p^hiu55/ ("float") are two different phonemes as these two words carry different lexical meanings (see Sec. I B, for information on tone transcription). In terms of voice onset time (VOT), this Cantonese contrast is similar to the voicing contrast of English. For example, both English and Cantonese voiceless unaspirated stops typically have a short-lag VOT (under 30 ms) and a lower amplitude of the post-burst aspiration noise, while voiceless aspirated consonants have a long-lag VOT (greater than 30 ms) and higher amplitude aspiration (Clumeck, Barton, Macken, and Huntington, 1981; Lisker and Abramson, 1964). Based on this similarity, it seems reasonable to expect that Cantonese, like English, should be considered to manifest the distinction between [+voice] and [−voice] consonant classes in terms of two phonetic features, {voiceless unaspirated} and {voiceless aspirated} (cf. Keating, 1984).

If the effect of consonant voicing on onset f_0 in English is primarily intrinsic, resulting from physiological and/or aerodynamic consequences related to laryngeal gestures involved in voicing (delay), then we would expect Cantonese stops to exhibit the same degree of consonantal influence on onset f_0 as in English because both languages exhibit similar patterns of VOT: Voiced (voiceless unaspirated) stops should show a rising or level onset f_0, and voiceless (voiceless aspirated) stops should show a falling onset f_0. On the other hand, if the English pattern of onset f_0 is (at least partly) due to an intentional manipulation of a secondary cue for the purposes of enhancing the perception of a primary cue (e.g. VOT, cf. Kingston and Diehl, 1994) then Cantonese speakers might show little or no effect of consonant voicing on onset f_0, because f_0 manipulations are likely to be preferentially reserved for tonal cues.

B. Cantonese tones and onset f_0

There are six contrastive tones in Hong Kong Cantonese and each tone has a specific contour (direction of f_0 movement: level, rising, or falling) and register (f_0 height: high, mid, or low) (Bauer and Benedict, 1997; Fok Chan, 1974). While it is not possible to speak of a defined standard for Cantonese, the tonal system in Hong Kong Cantonese differs only slightly from systems observed in speakers from other Cantonese-speaking areas where instrumental phonetic research has been conducted (primarily Macau and mainland Guangdong province). The main difference is the absence of a high falling (51) tone in Hong Kong, where it has been replaced by the high level (55) tone. It appears possible that this is the result of a sound change in progress in Cantonese more generally. Although some speakers from other regions have still been observed to produce the high falling tone consistently, the Hong Kong system is commonly considered the norm (see Bauer and Benedict, 1997 for acoustic data and further discussion). Each Cantonese tone can be represented using the Chao (1947) system with numbers from 1 to 5 expressing the talker's pitch range from lowest to highest. Two numbers are used, such that the first number indicates the pitch level at the start of the syllable while the second number indicates the syllable's ending pitch level. According to this system the tones of Cantonese are as follows. high level (HL): 55; high rising (HR): 25; mid level (ML): 33; low rising (LR): 23; low level (LL): 22; low falling (LF): 21. Given the importance of tone contour in Cantonese, and the large number (four) of tones starting at the same pitch level (2), it is possible that the influence of consonants on onset f_0 may be even more constrained than in other tone languages (e.g., Yoruba, with three phonologically level tones, see Hombert, 1977) because listeners may require more precise information about fine details of a syllable's pitch contour in order to make accurate tonal judgments.

C. Aspiration and onset f_0

Cantonese stops are more properly characterized as differing according to aspiration rather than voicing (see also Bauer and Benedict, 1997). The acoustic properties of the following vowel (f_0, frequency of formant transitions, amount of aspiration noise in the vowel) appear to be more salient perceptual cues than either VOT or (the presence or absence of) aspiration noise for the perception of aspiration in Cantonese initial stops. Tsui and Ciocca (2000) manipulated the duration of the VOT interval of naturally produced aspirated and unaspirated CV syllables to create long VOT conditions with or without aspiration noise between the burst release and the onset of voicing. They found that long VOT stimuli (created by introducing a silent interval between the burst and the onset of voicing of unaspirated stops) were perceived as "unaspirated" by native listeners, showing that VOT per se is a weak cue to the perception of (phonological) aspiration. Poon (2000) systematically manipulated aspiration noise (the level of the breathy sounds between the burst release and the onset of voicing), vowel and VOT cues of similar CV syllables in a fully crossed design, and found that stimuli that contained the vowel portion of syllables originally produced with voiceless aspirated stops were always perceived as "aspirated," independent of the level of the aspiration noise in the gap between the burst release and the onset of voicing and VOT duration. In addition, stimuli that contained a vowel taken from an originally unaspirated syllable were only perceived as "aspirated" if (both) VOT was long and aspiration noise was present between the burst and the onset of voicing, indicating that neither of the two cues (VOT and aspiration noise between the burst and the onset of voicing) could on its own override the effect of vowel type (possibly representing some sort of breathiness property or presence of aspiration noise *during* the first pitch periods of the vowel such as that described for Wu dialects by Jianfen and Maddieson, 1992).

As discussed by Löfqvist *et al.* (1989) and Ohde (1984), gestures associated with consonant voicing contrasts (specifically, increased activity of the cricothyroid muscle for voiceless consonants), not aspiration per se, appear to be the primary physiological basis for consonantal effects on onset f_0. Studies of languages that contrast aspiration, e.g., Hindi (Kagaya and Hirose, 1975) and Korean (Kagaya, 1974), support this hypothesis: Onset f_0 is lower following voiceless aspirated stops as compared with voiceless unaspirated stops, but, at least in Hindi, both types of voiceless stops have a higher onset f_0 than either voiced one. If it is the case that Cantonese stops contrast according to aspiration, not voicing, than we might expect both series of stops to induce some degree of raising of onset f_0.

Even in tonal languages aspirated stops seem to cause less of an increase in onset f_0 than do unaspirated stops. For example, Shi (1998) reported that aspirated stops in Wu and Gan Chinese showed a lower onset f_0 when compared to that of corresponding unaspirated stops, but did not mention the duration of the consonant's effect. Xu and Xu (2003) found that, in Mandarin Chinese, aspirated alveolar stops showed a significantly lower onset f_0 (measured at the first glottal pulse) than did corresponding unaspirated stops, but it is not clear how far into the vowel this effect extended. Finally, Gandour (1974) showed that, in Thai, onset f_0 following voiceless aspirated stops is lower than following voiceless unaspirated stops, and the duration of consonantal effect ranges from 10 to 50 ms. Moreover, in Gandour's (1974) study both voiceless unaspirated and voiceless aspirated stops showed a higher onset f_0 than corresponding voiced stops.

The only study found on the interaction of aspiration and f_0 in Cantonese presented evidence that aspirated stops are associated with a higher onset f_0 than unaspirated stops (Zee, 1980), suggesting that Cantonese does indeed pattern with English with respect to the influence of voicing on onset f_0. However, this groundbreaking study was somewhat limited in scope and investigated only the effects of bilabial consonants on the onset f_0 of syllables with High Level tones. Moreover, Zee (1980) computed only a single value for onset f_0 for each syllable, calculated as the mean value of the first 30 ms after voicing onset. Thus, it is not possible to determine from Zee's (1980) data whether the overall duration of the effect of aspiration was longer or shorter than 30 ms. Such averaging might also have obscured the most significant f_0 changes which were shown by Hombert (1977) to occur immediately after voicing onset.

If Cantonese stops do contrast according to aspiration rather than voicing, then we would expect that both voiceless unaspirated and voiceless aspirated stops should exhibit a relatively high onset f_0; that is, f_0 should fall or at least not rise into the vowel; but unaspirated stops should show a higher onset f_0 than aspirated ones.

D. Goals

The first goal of the present study was to investigate the effect of voicing (or aspiration) differences in Cantonese initial stops on the f_0 of the following vowels, and the interaction of this effect with tone height and contour, and consonant place of articulation. For this purpose, the f_0 of vowels preceded by voiceless aspirated and unaspirated stops was measured over the first 100 ms after voicing onset. Based on the results of previous research, it was hypothesized that there would be a significant difference between the onset f_0 following aspirated and unaspirated stops. Two possibilities of directions of f_0 changes at vowel onset were suggested: (1) a rising versus a falling f_0 pattern would be found to distinguish unaspirated from aspirated stops if the consonantal contrast in Cantonese initial stops is similar to that in English and Yoruba as suggested by Hombert (1977), or (2) the f_0 patterns would be falling for both aspirated and unaspirated stops if the consonantal contrast in Cantonese initial stops is more of an aspiration contrast as in languages including Wu, Gan, Mandarin, Korean, and Thai (Gandour, 1974; Kagaya, 1974; Shi, 1998; Xu and Xu, 2003). In all cases it was expected that any effect of aspiration on f_0 in Cantonese would be shorter than the 100 ms (or so) observed in nontonal languages such as English, French, and Dutch. The goal of the second experiment was to determine whether Cantonese listeners are able to identify consonants in initial

position as either aspirated or unaspirated based solely on differences in onset f_0 and, if so, to estimate the necessary duration of the change in onsetf_0 for such a contrast to be made.

II. EXPERIMENT 1: PRODUCTION

A. Methods

1. Subjects

Sixteen native speakers of Cantonese (eight women, eight men) reporting no hearing or speaking disability participated in this study. Participants' ages ranged from 20 to 23 years (mean 21.75). Eight participants (four women, four men) were Hong Kong University students in the Department of Speech and Hearing Sciences with some training in acoustic phonetics. The other eight participants (four women, four men) were students from other departments with no acoustic phonetic training.

2. Stimuli

Ten monosyllabic real words with unaspirated or aspirated initial stops were employed as stimuli. They were: /pa55/ "scar," /p^ha55/ "on all fours," /ta55/ "dozen," /t^ha55/ "he," /ka55/ "home," /k^ha55/ "compartment," /pa25/ "target," /p^ha25/ "steak," /pa21/ "father" and /p^ha21/ "climb." All words had consonant vowel (CV) syllable structure with the vowel /a/ (approximately as in "father" in English) to control for possible effects of intrinsic vowel f_0 (cf. Diehl and Kluender, 1989; Peterson and Barney, 1952). Only real words were used to avoid the difficulty of eliciting nonsense syllables from speakers of a language with lexicographic orthography.

These stimuli can be divided conceptually into two overlapping sets. One set contained /pa55/ and /p^ha55/, /ta55/ and /t^ha55/, and /ka55/ and /k^ha55/. All of the tokens in this set had a high level (55) tone, thus eliminating the possibility of a tone-by-consonant interaction. This set of stimuli was used to investigate the role of place of articulation (bilabial, alveolar, and velar) in determining the effect of consonant aspiration on onset f_0. The second set consisted of /pa55/, /p^ha55/, /pa25/, /p^ha25/, /pa21/, and /p^ha21/. This set of stimuli was designed to investigate the interaction of tone with the effect of aspiration on onset f_0. Due to the constraint of using only real words, it was not possible to generate a complete set of both aspirated and unaspirated CV syllables at a single place of articulation differing according to all six possible Cantonese lexical tones. However, in the present set of stimuli all three tone contours of Cantonese (rising, falling, and level) and both the highest and lowest registers were represented.

3. Procedure

The Chinese character representing each stimulus word was written on a file card with one character per card. Recordings were made with participants seated comfortably in a single-walled sound-shielded booth. Prior to recording, participants read a set of written instructions, and were then familiarized with all of the stimuli used in the experiment. For familiarization, each stimulus card was presented, and participants were asked to pronounce the target in a carrier phrase with a speaking rate similar to conversational speech. The semantically neutral carrier phrase, /ŋ›ø23 wui23 t›k22 _ _ pɛi35 lɛi23 t^hɛŋ55/ "I will read _ _ for you to hear," was used for both familiarization and recording to encourage a natural production of each stimulus. The results of this study should be expected to generalize to both isolated productions and fluent speech since the absolute f_0 level and contour of syllables in isolation and in carrier phrase context (Ohde, 1984), and in fluent reading (Umeda, 1981) were found to be comparable. Although absolute f_0 values may be affected by coarticulation (Xu, 1994), these effects were held constant since the syllable /t›k22/ always preceded the target syllable. This preceding syllable has a low level (22) tone, so coarticulatory effects (if present) should cause a uniform lowering of f_0 at the onset of the target syllable across all target syllables (aspirated and unaspirated).

For recording, each of the ten target stimuli was written on five separate file cards for a total of 50 cards (5 repetitions of each of 10 words). All cards were presented to participants individually in randomized order, subject to the constraint that two cards with the same token could not be presented in immediate succession.

Participants were asked to read aloud each word as it was revealed by the experimenter (the third author). Approximately 1–2 production errors were observed for each talker over the course of the 50 target stimuli. After hearing a speech error, the experimenter would immediately produce the target word correctly, and ask the participants to repeat the utterance with that stimulus embedded again. There was no obvious difference between speakers' spontaneously correct productions and those following such prompting. Recording took a total of 15–20 min per talker.

All speech samples were recorded using a Shure Beta 87 microphone and a TASCAM DA-30MKII DAT tape recorder. The microphone was mounted on a boom stand and placed approximately 15 cm in front of the talker's mouth in the horizontal plane. The recorded stimuli were low-passed filtered at 22 kHz and sampled at 44.1 kHz using a Power Macintosh 7200/120AV computer equipped with a DigiDesign Audiomedia II sound card and stored in AIFF format for further analysis.

4. Analysis

Acoustic waveforms, spectrograms, and f_0 plots were obtained using Macquirer 4.9.9 (Macquirer, 1999). For analysis, spectrograms were produced with the Macquirer default wide-band filter setting (172 Hz for male talkers and 344 Hz for female talkers). Fundamental frequency was computed using the default method based on a 35 ms cepstral analysis window. In cases in which the Macquirer algorithm for calculating f_0 was unable to resolve a value at the designated time, f_0 was calculated manually from the acoustic waveform as the reciprocal of the respective glottal period at that time. Plots showing frequency ranges from 50 to 260 Hz for male talkers and 130 to 300 Hz for female talkers were used for analysis. During the measurement of f_0, each recorded stimulus was displayed as an acoustic waveform and a wide-band spectrogram. The onset of voicing was

identified as the first vertical striation (glottal pulse) extending through the first, second and third formants in the spectrogram. This definition of voicing onset is somewhat less reliable, in terms of time, than the onset of periodicity measured from the waveform alone (Francis, Ciocca, and Yu, 2003), but it is a highly conservative marker chosen to maximize the accuracy of automatic pitch-tracking. It should be noted, however, that our reliance on automatic pitch tracking (in most cases), may have resulted in some divergence of our estimate of the duration of f_0 perturbations from that reported by researchers who compute f_0 from direct measurement of single periods (e.g.,Löfqvist *et al.*, 1989; Xu and Xu, 2003). In the case of the present method, a window of 35 ms centered at the point of measurement would result in a margin of error of about 10–15 ms in terms of estimate of the duration of f_0 perturbations. In terms of accuracy of f_0 measurements, our results may be slightly less accurate for female talkers than for males (because we are averaging over more cycles), but interestingly the f_0 patterns look very much the same (see Sec. IC1, below), especially for unaspirated stops. So it is likely that any effect of averaging on precision of f_0 estimate was relatively small.

Following Hombert (1978) the most prominent f_0 changes related to aspiration were expected to be found during the first 100 ms after voicing onset and, because Cantonese is a tonal language, most probably within the first 30–50 ms. Therefore, f_0 was measured at the onset of voicing (time 0) and subsequently at 10, 20, 40, 70, and 100 ms into the vowel. These specific values were selected arbitrarily, but with the goal of maximizing the probability of pinpointing the duration of voicing-related effects on f_0 without requiring an inordinate number of measures.

To permit more accurate comparison across male and female talker groups, prior to statistical analysis frequency values in Hz were converted to cents using the equation $C = 1200 \log_2(f/127.09)$ (Traunmüller, 2005). All analyses of variance involving repeated measures with more than two levels were computed using a standard univariate model with Greenhouse-Geisser and Huynh-Feldt adjustments to degrees of freedom to compensate for the effects of violations of compound symmetry and sphericity (Greenhouse and Geisser, 1959; Huynh and Feldt, 1970). When the degrees of freedom as adjusted by the two methods differed, the more conservative values are reported.

B. Results

Before analyzing f_0, the average overall durations of vowels with each of the three tones in this experiment (HL, HR, and LF) were measured to determine whether vowel duration might interact with tone. Average vowel durations were: HL, 274 ms; HR, 286 ms; and LF, 268 ms, suggesting that f_0 is likely to be a much more important property than duration for identifying these tones. Separate analyses were carried out for the two sets of stimuli: (i) initial stops differing in aspiration and place of articulation, and (ii) initial stops differing in aspiration and in syllables with different tone.

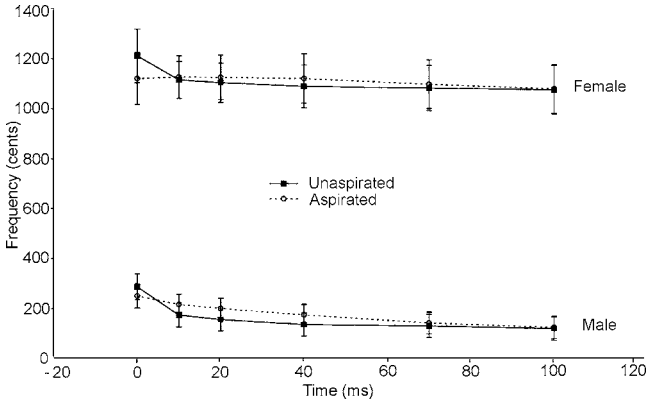

FIG. 1. Mean f_0 for aspirated and unaspirated syllables with high level tone. Error bars indicate standard error of the mean.

1. Aspiration and place of articulation

The first analysis was carried out on stimuli with the same tone (HL) but different places of articulation (bilabial, alveolar, and velar) to determine if place of articulation affects onset f_0. A mixed factorial ANOVA with two levels of aspiration, two levels of gender, three levels of place of articulation (POA), and six levels of time interval showed a significant effect of gender, $F(1,14)=119.59$, $p<0.001$, and of time, $F(1.75,24.46)=50.67$, $p<0.001$, but not of aspiration, $F(1,14)=1.22$, $p=0.29$, or of POA, $F(1.82,25.46)$, $p=0.69$. The only significant interactions were between time and gender, $F(1.75,24.46)=3.60$, $p=0.05$, and aspiration and time, $F(1.47,20.56)$, $p=0.001$. Figure 1 shows the overall average f_0 of tokens with HL tones collapsed over place of articulation during the first 100 ms after voicing onset for the two aspiration categories (aspirated stops [p^h], [t^h],[k^h] versus unaspirated stops [p], [t], [k]) for both genders.

Additional posthoc (Tukey HSD) comparisons were performed using means pooled over place of articulation to compare temporally adjacent frequency values (e.g., 0 ms vs. 10 ms) both within and between consonant classes (aspirated vs unaspirated). All results reported as significant are at the $p<0.05$ level. For female talkers, f_0 immediately following unaspirated stops (time=0 ms) was significantly higher (256 Hz) than all other points in both the aspirated and unaspirated series (ranging from 237 Hz to 244 Hz) and this was the only significant pairwise comparison either within or between series. For the male talkers the situation was somewhat more complex. In the unaspirated series, frequency at the first point (time=0 ms) was significantly greater (152 Hz) than all other points in this series, and none of these other unaspirated series' points were significantly different from one another (ranging from 138 Hz to 142 Hz). In the aspirated series no adjacent points were significantly different from one another, but there was a gradual decrease in f_0 such that, for example, the first point (0 ms,148 Hz) was significantly different from the last three (142, 139, and 138 Hz, respectively), and the second (10 ms, 142 Hz) and third (20 ms, 140 Hz) were significantly different from the last two. Comparing the aspirated and unaspirated f_0 curves showed no significant difference between the two series at any measurement point, although a trend similar to that of

the female talkers can be observed: Only at 0 ms was the frequency of the unaspirated series greater than in the aspirated one (151 vs 148 Hz).

Based on the overall similarity of the male and female f_0 contours, the ANOVA was recalculated excluding the between-subjects factor of gender. Results for the combined group were similar to, but less equivocal than, those found in the analysis including gender. There was a significant effect of time, $F(1.58, 23.65) = 43.17$, $p < 0.001$, and a significant interaction between aspiration and time, $F(1.46, 21.83) = 12.34$, $p = 0.001$. No other main effects or interactions were significant. Post hoc (Tukey HSD) analysis showed that the aspirated and unaspirated series differed only at the 0 ms point, with the unaspirated series showing a significantly higher f_0 than the aspirated one. Both curves showed a slight decline in frequency, although this effect is more gradual but also more pronounced in the aspirated series. In the unaspirated series, the first point (0 ms) was significantly higher than all other points, and the second point (10 ms) was significantly higher in frequency than the final two points (70 ms and 100 ms). There were no significant differences found for any pairwise comparison between the final four points (20 ms, 40 ms, 70 ms, and 100 ms) of the unaspirated series, suggesting that the duration of the consonantal effect was no greater than 20 ms. In the aspirated series the first point (0 ms) was significantly higher in frequency than the last three points (40 ms, 70 ms, and 100 ms), while the second (10 ms) and third (20 ms) points were significantly higher in frequency than the final two points (70 ms and 100 ms). The fourth point (40 ms) was significantly higher in frequency than the final point (100 ms), but no other pairwise comparisons among points in the aspirated series showed a significant difference, suggesting that the duration of the consonantal effect was around 40 ms.

In summary, for both male and female talkers, f_0 declined suddenly from 0 to 10 ms in the unaspirated series, but remained more or less level in the aspirated one (declining gradually for the male but not female talkers). For both male and female talkers, frequency patterns across the two series were similar, and generally corresponded to the combined (across gender) analysis in which little or no difference was found between the two series at 10 ms post voicing onset or beyond, but frequency at 0 ms was significantly higher in the unaspirated series than in the aspirated one. Overall, this suggests a pattern of decrease in f_0 following both unaspirated and aspirated stops in Cantonese, but the decline is more abrupt following unaspirated as compared with aspirated consonants.

2. Aspiration and tone

The second analysis was carried out on the measurements of the stimuli with the same place of articulation (bilabial) but different tones (55, 25, and 21). This analysis was designed to investigate whether the differential effects of aspirated and unaspirated consonants on the f_0 of following vowels is realized differently on syllables with different lexical tones. Results are shown in Fig. 2.

A four-way mixed factorial repeated measures ANOVA with factors gender, aspiration, tone (3 levels) and time was

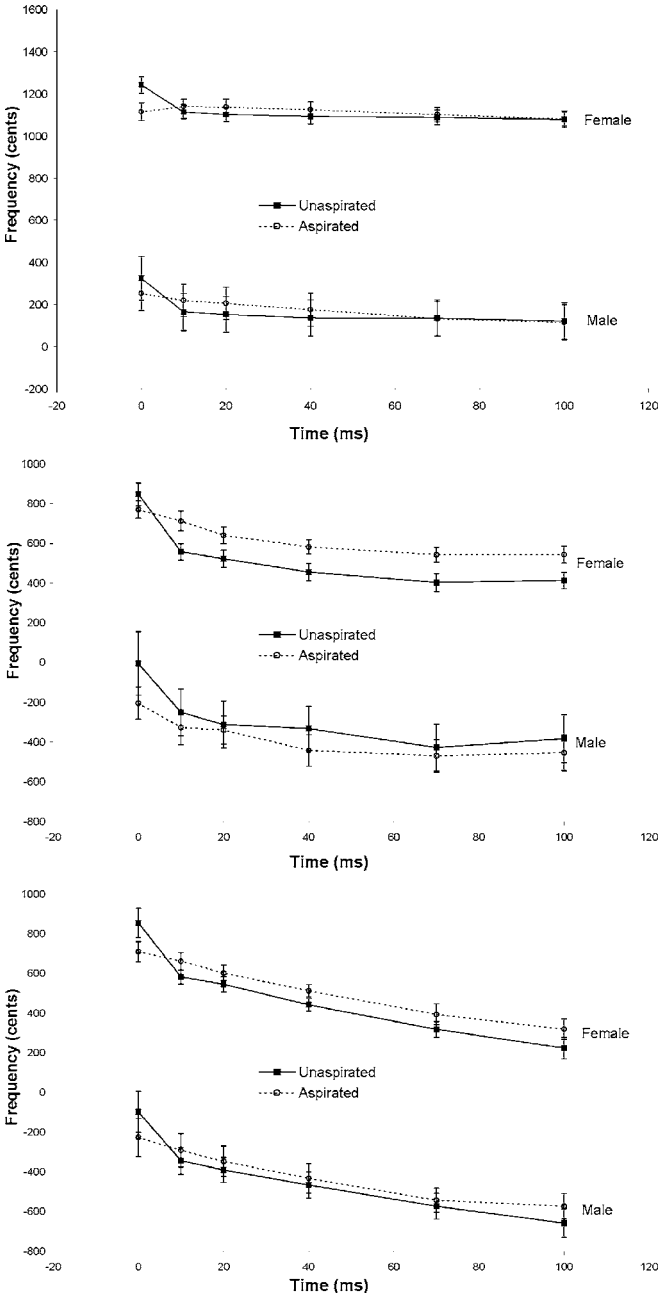

FIG. 2. Mean f_0 for syllables beginning with aspirated and unaspirated bilabial stops and having either high level (2a), high rising (2b), or low level tones (2c). Error bars indicate standard error of the mean. Note that each subgraph shows a Y axis of 1800 cents, though the starting and ending frequencies differ for different tones.

conducted. Results showed a significant effect of gender, $F(1, 14) = 142.39$, $p < 0.001$, tone, $F(1.29, 17.00) = 181.62$, $p < 0.001$, and time, $F(2.07, 28.94) = 152.75$, $p < 0.001$, and significant interactions between tone and time, $F(2.63, 36.78) = 28.52$, $p < 0.001$, and between aspiration and time $F(2.47, 34.54) = 20.77$, $p < 0.001$. No other main effects or interactions were significant.

The observation that gender did not interact with any other variable suggests that both male and female talkers showed the same effects of tone, time and aspiration on onset f_0. Therefore, post hoc (Tukey HSD) analysis were carried out on means pooled over gender. Results showed that, at contemporaneous points across all three tones, the only sig-

nificant differences due to aspiration were found at 0 ms, where the unaspirated tokens were higher in frequency than the aspirated ones. The only other case in which the aspirated and unaspirated series differed in frequency at the same time was in the 21 (low falling) tone at 100 ms post voicing onset, where there was a significant difference in f_0 between unaspirated (male 87 Hz, female 145 Hz) and aspirated (male 92 Hz, female 153 Hz) consonants.

Results of the post hoc (Tukey HSD) analysis of differences between f_0 measurement points for the high level (55), high rising (25), and low falling (21) tones showed that [results for the high level (55) tone here are for the bilabial tokens only] for the high level (55) tone following unaspirated stops, f_0 was significantly greater at the first (0 ms) measurement point than at all other points, but no other pairwise comparisons showed a significant difference, suggesting that the consonantal effect extends no more than 10 ms into the vowel. For the aspirated series, the first (0 ms) and second (10 ms) points were significantly higher in frequency than the last (100 ms) point, but no other comparisons reached significance suggesting that the effect of the consonant extends no more than 20 ms into the vowel. For the 25 tone, in both series (aspirated and unaspirated) steps 5 (70 ms) and 6 (100 ms) were significantly lower in frequency than the first three steps (0 ms, 10 ms, and 20 ms), but not different from one another or the 40 ms step, suggesting that the consonantal effect ends within 40 ms after voicing onset. For the 21 tone most comparisons reached significance, and in the unaspirated condition even then last step (100 ms) was significantly lower in frequency than the preceding one (70 ms). This result suggests that the consonantal effect may extend up to or beyond 100 ms into the vowel of syllables with a low falling (21) tone (though see discussion, below, regarding the interaction of tones and consonant effects).

C. Discussion

The results of this first experiment are consistent with the hypothesis that consonantal effects on onset f_0 in Cantonese result from a combination of intrinsic physiological/aerodynamic constraints and of the intentional manipulation of f_0 for the enhancement of voicing cues.

1. Direction of change

All of the syllables investigated in this study showed a generally level or falling f_0 contour over the first 100 ms, consistent with the phonetically voiceless status of both aspirated and unaspirated stops in Cantonese. That is, Cantonese appears to follow the pattern observed in Korean (Kagaya, 1974), Hindi (Kagaya and Hirose, 1975), Thai (Gandour, 1974), and other dialects of Chinese (Wu, Gan: Shi, 1998; Mandarin: Xu and Xu, 2003) rather than that exhibited by English (House and Fairbanks, 1953; Lehiste and Peterson, 1961; Löfqvist *et al.* 1989): Both long- and short-lag consonants induce a falling f_0 contour at the onset of voicing, though this drop is stronger after unaspirated consonants than after aspirated ones.

These results support two distinct (but probably complementary) physical explanations of the effect of consonant voicing on onset f_0, a muscular/physiological one and an aerodynamic one. According to Halle and Stevens' (1971) "muscular hypothesis," the vocal folds are slack during voiced consonant closures but stiff during voiceless stop closures and subsequent aspiration through the onset of voicing, leading to a higher onset f_0 following voiceless consonants than following voiced stops (see also Löfqvist *et al.*, 1989). Since Cantonese unaspirated stops are associated with a higher onset f_0 than are aspirated ones, Cantonese unaspirated stops are likely to be produced with tenser vocal folds than aspirated consonants, though both may be produced with an overall greater degree of tenseness than, for example, English (phonologically) voiced stops. This conclusion is supported by a number of physiological studies discussed by Shi (1998), suggesting that vocal cord tension at the onset of voicing is higher for voiceless unaspirated stops than for voiceless aspirated stops in Hindi (Kagaya and Hirose, 1975), Thai (Ewan, 1976, cited in Hombert, 1978), and Korean (Kagaya, 1974).

The present findings are also consistent with an aerodynamic account such as that presented by Ladefoged (1971). He argued that the main difference between the production of voiceless aspirated and unaspirated stops is that, at the oral release phase, the glottis is widely opened for aspirated stops, but more or less closed for unaspirated stops. Thus, prior to the onset of voicing more air is released from the lungs through the glottis following aspirated stops than is released following unaspirated stops. Unaspirated stops, which also have a short voice onset time, should exhibit a smaller decrease in subglottal pressure, and therefore a higher initial rate of vocal fold vibration, than aspirated stops prior to the onset of voicing (cf. discussion by Shi, 1998) concerning similar aerodynamic effects in the production of Mandarin initial stop consonants. In the present study, unaspirated stops exhibited a consistently higher onset f_0 than aspirated ones, suggesting that their production does indeed lead to a higher trans-glottal pressure differential at the onset of voicing compared to that of aspirated stops.

Interestingly, while the present results are consistent with a broad range of physiological and aerodynamic accounts of consonantal effects on onset f_0, they appear to conflict with Zee's (1980) findings that Cantonese aspirated stops showed a higher onset f_0 than unaspirated stops. One possible reason for this disagreement is that Zee averaged f_0 over the first 30 ms of each syllable. This averaging process would have made it difficult to identify f_0 patterns similar to those presented here in which the most significant effects were found in the first 10 ms after voicing onset. Indeed, when considering only syllables with a high level tone and beginning with bilabial stops as produced by male talkers in the present study (the conditions most comparable to those of Zee, 1980), averaging the measurements taken at 0, 10, 20, and 40 ms shows an average onset f_0 of 143 Hz for the unaspirated series, and 145 Hz for the aspirated series. While the difference between these values is much smaller than those given by Zee (1980) (average differences of 10, 6, and 12 Hz for three male talkers, based on measurements from

ten productions each of /pei55/ ["sorrow"] and /p^{h}ei55/ ["to spread"]), the trend is similar; moreover, three individuals in the present study exhibited comparable f_0 differences (14 Hz, 5 Hz, and 4 Hz) between aspirated and unaspirated bilabials. Still, the present results strongly suggest that there is little or no physiological contribution of consonant voicing to onset f_0 patterns in Cantonese beyond the first few tens of milliseconds of the vowel.

The findings presented here for Cantonese, and those obtained by Xu and Xu (2003) for Mandarin, also have implications for broader theories of stop consonant production and for the mechanisms of producing a phonological voicing contrast. Keating (1984) proposed a model to resolve the terminological problem that derives from considering, for example, both English and Spanish stop consonants as contrasting according to voicing, despite clear differences in their phonetic realizations. According to Keating's proposal, different languages may instantiate the phonological features [+voice] and [−voice] in different ways such that, for example, a phonologically [+voice] segment in Cantonese or English is realized as a {voiceless unaspirated} stop, while in Spanish a phonologically [+voice] segment would be realized as a {voiced} stop. The results of the present experiment suggest that languages may differ even at the phonetic level, such that the English and Cantonese {voiceless unaspirated} stops have different effects on onset f_0 (lowering in English but raising, if only at the very onset of voicing, in Cantonese). Two possible explanations for this phenomenon have been proposed: Either onset f_0 effects are based on the *phonological* feature involved (such that all [+voice] segments depress onset f_0, while all [−voice] segments raise it), or the Cantonese and English stops may be realized with different patterns of laryngeal muscle movements, each accomplishing laryngeal closure/release in the same time relative to oral closure/release gestures but resulting in different effects on onset f_0.

Previous studies showing an effect of consonant *aspiration* (as opposed to voicing) on onset f_0 have been conducted in languages contrasting more than two phonological voicing classes (e.g., Thai: Gandour, 1974; Hindi: Kagaya and Hirose, 1975; and Korean: Kagaya, 1974); these studies specifically compared phonetically voiceless unaspirated to voiceless aspirated consonants. Interestingly, for at least two of these languages (Hindi and Korean), it has been argued that additional contrastive dimensions beyond that of laryngeal timing must be involved in generating the complete set of contrasts in the language (Abramson, 1977). Kingston and Diehl (1994) distinguish between two voicing-related features, [voice] and [spread glottis], and imply that, e.g. in Thai, the feature [±voice] distinguishes the phonetically voiced ([+voice]) from the phonetically voiceless unaspirated ([−voice]) stops, while the feature [±spread glottis] distinguishes the phonetically voiceless unaspirated ([−spread glottis]) from the phonetically voiceless aspirated ([+spread glottis]) stops. A similar pattern may obtain for Cantonese (and Mandarin), except that Cantonese and Mandarin simply lack the [+voice] category (that is, both the [+spread glottis] and the [−spread glottis] stops are [−voice]). This argument seems to be consistent with the findings of Löfqvist *et al.* (1989), who showed that the production of both Dutch and English *phonologically voiceless* stops is associated with increased cricothyroid muscle activity. Such an increase in muscle activity presumably increased the longitudinal tension in the vocal folds and resulted in a generally higher onset f_0, although these stops have a short-lag VOT in Dutch (10–15 ms) but a long-lag (60–80 ms) in English. In order to successfully instantiate the long vs short-lag VOT contrast, other laryngeal maneuvers (presumably associated with the feature [±spread glottis]) must be invoked. According to this hypothesis, increased cricothyroid activity is associated with suppression of voicing regardless of the duration of that suppression, and is present in both the voiceless aspirated and voiceless unaspirated consonants in Cantonese, but only in the voiceless aspirated stops in English. The results of the present study suggest that the presence of such cricothyroid activity is automatic, but that the gestures controlling its duration (beyond the first few tens of ms) are not.

2. Duration of effect

In the present study the maximum duration of the effect of consonants on onset f_0 appears to be about 40 ms for the high level (55) and high rising (25) tones, with the effect being somewhat shorter in the high level than in the high rising conditions, while in the low falling condition f_0 continued to decline beyond 70 ms. This pattern is quite commensurate with the results shown by Xu and Xu (2003), and is clearly related to the overall f_0 pattern of each tone. In Cantonese, as in Mandarin, the f_0 contours of level tones typically remain mostly level, sometimes with a slight rise at the onset (Bauer and Benedict, 1997; Xu and Xu, 2003). Thus, these syllables show us the effect of the consonant without any interaction with a tonal contour. In contrast, rising tones typically show a short fall over the course of the first 1/4 to 1/3 of the syllable (Bauer and Benedict, 1997; Francis *et al.*, 2003; Moore and Jongman, 1997; Xu and Xu, 2003), and falling tones typically fall over the course of the entire syllable (Bauer and Benedict, 1997; Xu and Xu, 2003). Thus, in the high level tone, we see the "purest" effect of the consonant, with the f_0 contour more or less leveling off within about 20 ms of the onset of voicing. In the high rising (25) syllables we see the consonantal effect occurring in conjunction with the expected, initially slightly falling tonal contour, leading to the appearance of a consonantal effect out to about 40 ms post voicing onset. Finally, in the low falling (21) tone the longer consonantal effect (70 ms or more in duration) is most likely due to the normal falling f_0 contour for this tone. A close examination of the graphs shown by Xu and Xu (2003) suggests a similar pattern for Mandarin: Consonantal effects do not appear much beyond the first few tens of ms following voicing onset, even in the case of the Mandarin Falling (51) tone.

This finding is consistent with the proposal of Hombert (1978) that tone language speakers have less freedom to vary f_0 patterns as a consequence of voicing (or aspiration). The generally level or falling f_0 pattern in both the aspirated and unaspirated series is consistent with Halle and Stevens' (1971) model of voicing (both series are voiceless). However, the brevity of the effect in Cantonese as compared to

English supports Hombert's (1977) viewpoint that speakers of tone languages actively minimize the intrinsic effect of preceding consonants on f_0 of the following vowels so as to keep each tone maximally perceptually distinct. Alternatively, the aspiration effects on onset f_0 observed in tonal languages might reflect the natural extent of consonant-related influence (extending somewhere between 10 and 50 ms following voicing onset), while the patterns observed for English might result from language-specific exaggeration of these tendencies. On the basis of the present evidence it is not possible to draw any firm conclusions regarding these two conflicting hypotheses, but some speculation is possible.

Onset f_0 differences have been shown to be a sufficient cue to the voicing contrast in English initial stops when the voicing-related f_0 perturbations lasted for about 55 to 60 ms (Haggard *et al.*, 1970). Robinson and Patterson (1995) demonstrated that a minimum of six to eight cycles of a periodic sound are necessary before pitch can be accurately identified. With respect to the present case, this would translate to at least 42–56 ms for a male with an f_0 of 142 Hz, and 25–33 ms for a female with a f_0 of 230 Hz. These durations are consistent with the results presented by Haggard *et al.* (1970), but they are considerably longer than those observed in the present study. If the present results reflect the natural (intrinsic) extent of aspiration effects on onset f_0, then the English pattern (extending out to 100 ms or more) would result from an extension of that natural tendency to nearly twice what would be necessary for listeners to hear the f_0 effects. On the other hand, if the English pattern reflects the natural (unsuppressed) extent of aspiration effects on onset f_0, then the pattern observed here would indicate a much greater degree of suppression than would be necessary to eliminate perceptual evidence of the consonant effects on f_0. A reasonable conclusion then would seem to be that the English pattern reflects some degree of enhancement, while the pattern observed here reflects at least some degree of suppression. In the next experiment, we explore Cantonese listeners' perception of onset f_0 as a cue to the aspiration contrast in syllable-initial stop consonants.

III. EXPERIMENT 2: PERCEPTION

Following Kingston and Diehl (1995), VOT and onset f_0 may "cohere" perceptually (that is, they may both function as cues to the same phonetic feature contrast) either because they arise from the same physiological phenomena in production (cf. Abramson, 1977) or because they affect the auditory system in similar ways (cf. Kingston and Macmillan, 1995). In both cases, the coherence of the two cues may be considered automatic in that it derives from basic physiological properties and does not have to be learned through of linguistic experience. By contrast, it is possible that two cues may cohere due to learned co-variation. That is, over the course of development, listeners may become so accustomed to hearing the two together that the presence of either cue is alone sufficient for perceiving a given contrast, even though there may be no necessary physiological or perceptual relationship between them (cf. Ohala, 1981 and related discussion by Kingston and Macmillan, 1995). In the case of Cantonese listeners, who have experienced relatively little perceptible relationship between onset f_0 and VOT in their native language, the ability to use onset f_0 as a cue to the voicing contrast would suggest a basic perceptual (not learned) basis for the relationship between onset f_0 and other stronger cues to aspiration. To investigate this possibility, in the second experiment we investigated Cantonese listeners' ability to use onset f_0 (alone) as a cue to the perception of the Cantonese stop consonant voicing contrast.

A. Methods

1. Subjects

Eighteen native Cantonese speakers, six men and twelve women, reporting normal speech and hearing abilities volunteered for this study. Their age ranged from 22 to 31 (mean age=23.3). All had some experience listening to English in the course of their education (comparable to the experience of the participants in experiment 1), though none used English significantly outside of the university.

2. Stimuli and apparatus

All stimuli were derived from a single CV syllable created by digitally splicing naturally-produced, but resynthesized, burst and aspiration noise onto a synthetic vowel. The natural token (the word /pʰa55/ "on all fours") was selected from a set of many CV syllables produced by a college-aged male native speaker of Cantonese. An aspirated token was selected as the base rather than an unaspirated one to simplify resynthesis. For example, in our experience, it is generally easier (results in a more natural sounding resynthesis) to reduce the amplitude of aspiration noise that is already present than it is to add in aspiration noise that is not there to begin with. A syllable with a high level tone was selected because experiment 1 showed that syllables with this tone do not exhibit an effect of consonant aspiration on onset f_0 beyond the first 10 ms of the vowel at any place of articulation. Thus, Cantonese speakers likely have no prior experience with the covariation of onset f_0 (beyond the first 10 ms) and aspiration in syllables with this tone. The specific token was selected based on the cleanliness of the recording, and the relative smoothness of the acoustic signal (lack of any unexplained acoustic events that might interfere with resynthesis).

The burst and aspiration noise from this token were subsequently manipulated in a manner designed to give rise to approximately equal numbers of /pa55/ and /pʰa55/ percepts (the base syllable). Resynthesis was accomplished using the pitch synchronous overlap and add (PSOLA) algorithm (cf. Huang, Acero, and Hon, 2001, pp. 820–823; Moulines and Charpentier, 1990) as implemented in Praat 4.0.5. To make the burst and aspiration-related cues of the base syllable ambiguous, VOT was compressed to 20 ms and the peak amplitude of the burst and aspiration noise was reduced by 50%, corresponding to values that were shown in an unpublished study in our lab to result in approximately equal numbers of aspirated and unaspirated responses. Informal testing of the final syllable (including the synthetic vowel described below) suggested that the resulting syllable was heard approximately equally as a /p/ or a /pʰ/. The base syllable was

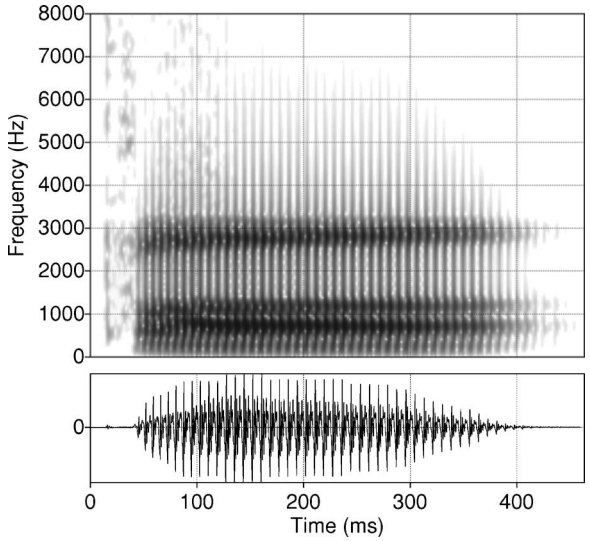

FIG. 3. Spectrogram and acoustic waveform of the syllable [p^{h}a] with 157 Hz onset f_0 and 80 ms duration of f_0 change, as used in experiment 2.

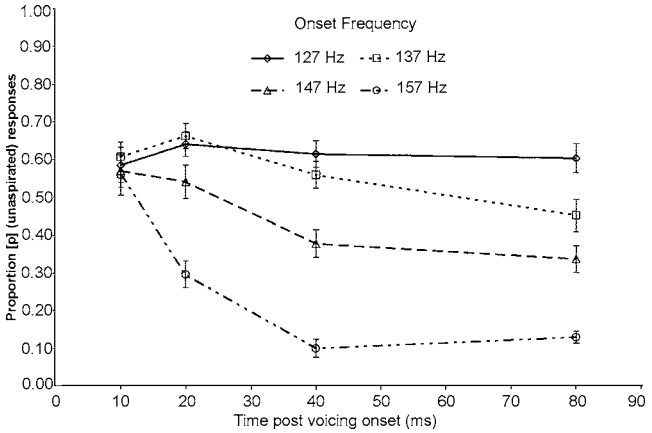

FIG. 4. Mean proportion of "aspirated" responses to stimuli with f_0 beginning at one of four onset frequencies and one of four durations of onset transition. Error bars indicate standard error of the mean.

420 ms in duration (equivalent to that of the natural syllable), and it was generated using SenSyn (Sensimetrics, Inc.). Synthesis parameters were derived from measurements made at 5 ms intervals of the f_0, amplitude envelope and the first four formant frequencies of the same syllable [p^{h}a55] "on all fours" from which the burst release was extracted. A spectrogram and acoustic waveform are shown in Fig. 3.

Sixteen different syllables were generated from this base syllable by fully crossing four levels of starting f_0 (127 Hz, 137 Hz, 147 Hz, and 157 Hz) and four levels of f_0 transition duration (10 ms, 20 ms, 40 ms, and 80 ms). For example, in the first stimulus, f_0 began at 127 Hz and fell to the original vowel f_0 (125 Hz) over the course of the first 10 ms following the onset of voicing (ending at a point corresponding to the 30 ms synthesis entry in the synthesizer parameters). The f_0 of the next token also began at 127 Hz, but took 20 ms to reach 122 Hz f_0 (corresponding to an endpoint at the 40 ms point in the synthesized sound). The choice of 157 Hz as the maximum onset f_0 value was based on the mean score of male subjects producing [p] in experiment 1, and subsequent values were selected in arbitrary 10 ms steps below that, encompassing the average range of productions of [p] and [p^h] by male speakers in experiment 1. The average f_0 of the talker selected for stimulus production in experiment 2 was comparable to the average range of male talkers in experiment 1. All properties of the burst release, aspiration noise, and all properties of the vowel other than f_0 remained the same across all tokens.

3. Procedure

The experiment was carried out in an IAC single-walled sound booth. Stimuli were presented binaurally in randomized order via Sennheiser HD-590 headphones. Stimulus presentation and response collection was controlled by a Hypercard stack running on Macintosh 7100 computer. On each trial listeners heard a single syllable and were instructed to identify it as either /p^{h}a55/ "on all fours" or /pa55/ "father" by clicking on one of two on-screen button inscribed with the appropriate Chinese characters. Characters were always presented side-by-side, in the same order for each participant. Participants completed a total of 16 blocks of 16 trials each (4 levels of onset f_0 times 4 levels of transition duration), with stimulus order randomized within blocks. The first block was treated as a practice block and not scored, although participants were not aware of this at the time of testing. Thus there were a total of 15 scored responses to each stimulus by each participant.

B. Results

Responses were collected and scored in terms of the proportion of "unaspirated" /pa55/ "father" responses for each stimulus, and were submitted to a two-way analysis of variance with repeated measures. Results (shown in Fig. 4) showed a significant effect of onset f_0, $F(3,51)=44.23$, $p<0.001$ and of duration, $F(3,51)=30.42$, $p<0.001$, and a significant interaction between the two, $F(9,153)=11.89$, $p<0.001$. Posthoc (Tukey HSD, $\alpha=0.05$) analysis showed that there was no significant difference in response proportion between any of the tokens with different onset frequencies when the duration of f_0 transition was only 10 ms in duration. For all stimuli with longer f_0 transition durations, the contour with a 157 Hz starting frequency was identified as /p^{h}a/ "on all fours" (aspirated) significantly more often than any other contour regardless of the duration of the f_0 transition. Listeners identified the 147 Hz contour as /p^{h}a/significantly more often than the 127 Hz or 137 Hz contour when the transition was at least 40 ms (though the difference between the 137 Hz and 147 Hz contours disappeared at 80 ms of transition duration). Only at 80 ms transition did listeners perceive the 137 Hz contour as significantly more aspirated (less unaspirated) than the 127 Hz contour.

C. Discussion

Cantonese listeners were able to use onset f_0 as a cue to consonant aspiration in the absence of other, more typical cues. However, there are two important qualifications to this conclusion. First, listeners in experiment 2 were unable to use onset f_0 as a cue to aspiration when onset f_0 varied only

over the 10-ms window shown in experiment 1 to differ between aspirated and unaspirated productions. Thus, Cantonese listeners do not appear to be able to perceive aspiration-related onset f_0 differences that occur in the production of their native language. Second, the manner in which onset f_0 functioned as a cue to consonant aspiration was precisely opposite to what one might have expected if Cantonese listeners' decisions were based on their prior experience with Cantonese. As shown in experiment 1, the onset f_0 of Cantonese unaspirated consonants starts *higher* than that of aspirated consonants, but in experiment 2 stimuli with higher onset f_0 values (e.g., 157 Hz) were generally identified *less* often as unaspirated stops than were those with lower onset f_0 values. Thus, Cantonese listeners' use of onset f_0 in making aspiration decisions must result from something other than their native language experience.

This pattern of results supports a model of voicing/aspiration perception in which onset f_0 can serve as a cue, not because it habitually occurs in conjunction with the cues to the primary contrast (as it does in English), but because it contributes indirectly to some more general, language-independent perceptual phenomenon (cf. Kingston and Diehl's, 1994 "low frequency property"). That is, onset f_0 differences may contribute indirectly to some more general property of the signal that Cantonese listeners are still able to interpret as cuing an aspiration contrast. One factor that may have helped in this transfer or extension of a more general cue is the listeners' experience with English. All of the participants in the present experiment were familiar with English, and used it on a daily basis in the course of their studies at the University of Hong Kong, where they were tested. An attempt was made to minimize the influence of English in the testing context, as all experiments were run in Cantonese and responses were made by clicking on Cantonese characters. Moreover, the English that listeners are exposed to in Hong Kong is typically a non-native variety; most secondary school teachers are native speakers of Cantonese, not English, and the same is increasingly true at the University level. Thus, it is not clear whether the English to which these listeners were exposed even presented a truly English-like pattern of covariation between VOT and onset f_0 cues. However, it is still possible that these listeners obtained some degree of experience with using the low frequency property to distinguish between English voiced and voiceless stops, and that this experience may have contributed to their ability to apply the same cue to the native aspiration contrast in the context of the present experiment.

IV. GENERAL DISCUSSION

The present results provide further support for the hypothesis that the effect of consonant voicing on onset f_0 in English results in part from what Kingston and Diehl (1994) call a controlled process—one that is intentionally manipulated by talkers to produce a desired perceptual effect, and in part from a intrinsic effect of the laryngeal gestures involved in voicing. The intrinsic effect may involve activity of the cricothyroid muscle, perhaps as part of a system for suppressing voicing (Löfqvist *et al.*, 1989), that results in an increased onset f_0 following voiceless consonants. The active control, in turn, appears primarily to limit the duration of this effect. In the case of voicing in nontonal languages, the role of this active control may be to prolong this effect in order to enhance of consonantal distinctiveness: Presumably, stops tend to sound more voiced when f_0 can be heard to clearly rise into the vowel, and more voiceless when they are clearly followed by a falling f_0. In Cantonese, Mandarin, Thai and other tonal languages, talkers are apparently able to restrict the production of consonant-related perturbations of onset f_0 to the first few tens of milliseconds following the onset of voicing, presumably in order to maintain the integrity of the f_0 contour as a cue to lexical tone identity. The fact that Cantonese listeners are still able to use onset f_0 as a cue for consonantal aspiration differences, despite the lack of such distinctions in their native language, is consistent with Kingston and Diehl's (1994) suggestion that (i) distinguishing voicing/aspiration contrasts may involve attention to a "low frequency property," and that (ii) this property is perceptually enhanced by a variety of independent acoustic cues, including onset f_0. The present results demonstrate that this property (or something similar) is accessible to Cantonese listeners despite their lack of consistent native language experience with onset f_0 differences beyond the first few tens of milliseconds into the vowel.

Since onset f_0 variation beyond the first few tens of ms can be controlled by speakers, longer-term changes in f_0 (e.g., in English) cannot be entirely the involuntary consequences of voicing gestures (i.e., they are not purely intrinsic, physiologically determined, by-products of such gestures). The observation that Cantonese listeners were able to use onset f_0 as a cue to a their native aspiration contrast is consistent with the hypothesis that the use of onset f_0 as a cue to voicing in English and other nontonal languages may derive from some kind of auditory enhancement, perhaps similar to the low frequency property discussed by Kingston and Diehl (1994). However, the present results can also be explained in terms of transfer from listeners' experience with the English pattern of consonantal influence on onset f_0. Future research will be necessary to investigate the use of this cue by native speakers of Cantonese (or other tonal language) who are not familiar with English (or any other language with an English-like pattern of consonantal influence on onset f_0).

ACKNOWLEDGMENTS

Material in this article derives from dissertations submitted by the third and fourth authors in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong. We would like to thank Jack Gandour for helpful comments on an earlier draft of this article.

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

Abramson, A. S., and Lisker, L. (**1985**). "Relative power of cues: F0 shift versus voice timing," in *Phonetic Linquistics*, edited by V. Fromkin (Academic, New York).

Bauer, R. S., and Benedict, P. K. (**1997**). *Modern Cantonese Phonology* (Mouton de Gruyter, Berlin).

Chao, Y. R. (**1947**). *Cantonese Primer* (Harvard University Press, Cambridge).
Clumeck, H., Barton, D., Macken, M. A., and Huntington, D. A. (**1981**). "The aspiration contrast in Cantonese word-initial stops: Data from children and adults," J. Chin. Linguist. **9**, 210–225.
Diehl, R. L., and Kluender, K. R. (**1989**). "On the objects of speech perception," Ecological Psychol. 1, 121–144.
Fok Chan, Y. Y. (**1974**). *A perceptual study of tones in Cantonese*, University of Hong Kong, Centre of Asian Studies, Hong Kong.
Francis, A. L., and Nusbaum, H. C. (**2002**). "Selective attention and the acquisition of new phonetic categories," J. Exp. Psychol. Hum. Percept. Perform. **28**, 349–366.
Francis, A. L., Ciocca, V., and, Yu, J. M. C. (**2003**). "Accuracy and variability of acoustic measures of voicing onset," J. Acoust. Soc. Am. **113**(2), 1025–1110.
Francis, A. L., Ciocca, V. C., and Ng, B. K. C. (**2003**). "On the (non)categorical perception of lexical tones," Percept. Psychophys. **65**(6), 1029–1044.
Gandour, J. (**1974**). "Consonant types and tone in Siamese," J. Phonetics **2**, 337–350.
Greenhouse, S. W., and Geisser, S. (**1959**). "On methods in the analysis of profile data," Psychometrika **24**, 95–112.
Haggard, M., Ambler, S., and Callow, M. (**1970**). "Pitch as a voicing cue," J. Acoust. Soc. Am. **47**, 613–617.
Halle, M., and Stevens, K. N. (**1971**). "A note on laryngeal features," Quarterly Progress Report, MIT Research Laboratory, Vol. **101**, pp. 198–213.
Hombert, J. M. (**1977**). "Consonant types, vowel height and tone in Yoruba," Studies in African Linguistics **8**, 173–190.
Hombert, J. M. (**1979**). "Consonant types, vowel quality, and tone," in *Tone: A Linguistic Survey*, edited by V. A. Fromkin (Academic, New York), pp. 77–111.
House, A. S., and Fairbanks, G. (**1953**). "The influence of consonant environment upon the secondary acoustical characteristics of vowels," J. Acoust. Soc. Am. **25**, 105–113.
Huang, X., Acero, A., and Hon, H.-W. (**2001**). *Spoken Language Processing*, Prentice-Hall, Upper Saddle River.
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.
Jianfen, C., and Maddieson, I. (**1992**). "An exploration of phonation types in Wu dialects of Chinese," J. Phonetics **20**, 77–92.
Kagaya, R. (**1974**). "A fiberoptic and acoustic study of the Korean stops, affricates and fricatives," J. Phonetics **2**, 161–180.
Kagaya, R., and Hirose, H. (**1975**). "Fiberoptic electromyographic and acoustic analyses of Hindi stop consonants," Annual Bulletin, Research Institute of Logopedics and Phoniatrics **9**, 27–46. Available at http://www.umin.ac.jp/memorial/rilp-tokyo/, last accessed May 19, 2006.
Keating, P. A. (**1984**). "Phonetic and phonological representation of stop consonant voicing," Language **60**(2), 286–319.
Kingston, J., and Diehl, R. L. (**1994**). "Phonetic knowledge," Language **70**, 419–494.
Kingston, J., and Diehl, R. L. (**1995**). "Intermediate properties in the perception of distinctive feature values," in *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvaniti (Cambridge University Press, Cambridge), pp. 7–27.
Kingston, J., and Macmillan, N. A. (**1995**). "Integrality of nasalization and F1 in vowels in isolation and before oral and nasal consonants: A detection-theoretic application of the Garner paradigm," J. Acoust. Soc. Am. **97**, 1261–1285.
Ladefoged, P. (**1971**). *Preliminaries to Linguistic Phonetics* (Chicago University Press, Chicago).
Lehiste, I., and Peterson, G. E. (**1961**). "Some basic considerations in the analysis of intonation," J. Acoust. Soc. Am. **33**, 419–425.
Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422.
Löfqvist, A., Baer, T., McGarr, N. S., and Seider Story, R. (**1989**). "The cricothyroid muscle in voicing control," J. Acoust. Soc. Am. **85**(3), 1314–1321.
Macquirer (Version 4.9.9). (**1999**). (Computer software), Scicon R & D Inc., Los Angeles, CA, (www.sciconrd.com).
Moore, C. B., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864–1877.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**(5), 453–467.
Ohala, J. J. (**1978**). "Production of tone", *Tone: A Linguistic Survey*, edited by V. A. Fromkin (Academic, New York), pp. 5–39.
Ohala, J. J. (**1981**). "The listener as a source of sound change," in *Papers from the Parasession on Language and Behavior*, edited by C. S. Masek, R. A. Hendrick, and M. F. Miller (Chicago Linguistic Society, Chicago), pp. 178–203.
Ohde, R. N. (**1984**). "Fundamental frequency as an acoustic correlate of stop consonant voicing," J. Acoust. Soc. Am. **75**, 224–230.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Poon, M. M. W. (**2000**). "Acoustic cues for the perception of aspiration in Cantonese initial stops," B.Sc. dissertation, University of Hong Kong, Hong Kong (unpublished).
Robinson, K., and Patterson, R. D. (**1995**). "The stimulus duration required to identify vowels, their octave, and their pitch chroma," J. Acoust. Soc. Am. **98**(4), 1858–1865.
Shi, F. (**1998**). "The influence of aspiration on tones," J. Chin. Linguist. **26**, 126–145 (in Chinese).
Traunmüller, H. (**2005**). "Auditory scales of frequency representation," http://www.ling.su.se/staff/hartmut/bark.htm. Last accessed June 5, 2006.
Tsui, I. Y. H., and Ciocca, V. (**2000**). "The 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.
Umeda, N. (**1981**). "Influence of segmental factors on fundamental frequency in fluent speech," J. Acoust. Soc. Am. **70**, 350–355.
Whalen, D. H. (**1990**). "Coarticulation is largely planned," J. Phonetics **18**, 3–35.
Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1990**). "Gradient effects of fundamental frequency on stop consonant voicing judgments," Phonetica **47**, 36–49.
Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1993**). "F0 gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **93**, 2152–2159.
Xu, Y. (**1994**). "Production and perception of coarticulated tones," J. Acoust. Soc. Am. **95**(4), 2240–2253.
Xu, C. X., and Xu, Y. (**2003**). "Effects of consonant aspiration on Mandarin tones," J. Int. Phonetic Assoc. **33**, 165–181.
Zee, E. (**1980**). "The effect of aspiration on the fundamental frequency of the following vowel in Cantonese," UCLA Working Papers in Phonetics, Vol. **49**, pp. 90–97.
Zlatin, M. A. (**1974**). "Voicing contrast: Perceptual and productive voice onset time characteristics of adults," J. Acoust. Soc. Am. **56**(3), 981–994.