LANGUAGE AND SPEECH, Vol. 30, Part 2, 1987

DIFFERENCE IN SECOND-FORMANT TRANSITIONS BETWEEN ASPIRATED AND UNASPIRATED STOP CONSONANTS PRECEDING [a]*

BRUNO H. REPP
and
HWEI-BING LIN**
Haskins Laboratories

Perceptual experiments with synthetic speech have shown that the category boundary on an acoustic [pa]–[ta] (/ba/–/da/) continuum (obtained by varying the onset frequencies of the second and third formants) is closer to the labial endpoint than the boundary on a [p^ha]–[t^ha] (/pa/–/ta/) continuum. Of several possible explanations, the most plausible seems to be that natural unaspirated and aspirated stops have different formant transitions. To supplement limited data on this point in the literature, we conducted an acoustic analysis of CV syllables produced by 10 male speakers of American English. The results show very clearly that the second formants of [p^ha] and [t^ha] start 100–200 Hz higher than those of [pa] and [ta] and reach comparable frequency values only at voicing onset. This difference, which is probably an acoustic consequence of subglottal coupling during aspiration, seems to be part of a listener's tacit knowledge of phonetic regularities and thus explains the perceptual boundary shift. It also needs to be taken into account in realistic speech synthesis.

Keywords: aspiration, stop consonants, formant transitions, acoustic theory of speech production

INTRODUCTION

A highly reliable finding of perceptual studies using synthetic CV syllables forming place of articulation continua is that the category boundary on an unaspirated [pa]–[ta] (i.e., English initial /ba/–/da/) continuum is closer to the labial endpoint than the corresponding boundary on an aspirated [p^ha]–[t^ha] (i.e., English /pa/–/ta/) continuum (Miller, 1977; Oden and Massaro, 1978; Repp, 1978; Alfonso and Daniloff, 1980; Massaro and Oden, 1980; Ohde and Stevens, 1983). In each of these studies, the stimuli in the two continua differed in the onset frequencies and transitions of the second and third formants (F2 and F3), whereas the difference between the two continua rested on

* This research was supported by NICHD Grant HD-01994 to Haskins Laboratories. We are grateful to Ignatius Mattingly, Richard McGowan, Susan Nittrouer, and especially Kevin Munhall for their excellent comments which helped improve this paper. A short version of the paper was presented by the second author at the 113th Meeting of the Acoustical Society of America in Indianapolis, May 1987.

** Also, The University of Connecticut

voice onset time (VOT). In the case of aspirated stops, this meant a delay in voicing onset, presence of aspiration noise, and attenuation or complete suppression of the first formant (F1) during the aspirated interval. Formant transitions and VOT thus were varied in a strictly orthogonal fashion.

No satisfactory explanation has been provided for the perceptual boundary shift, although several authors have speculated about its causes. If we include several additional possibilities that have occurred to us, no less than six different hypotheses result, which we shall discuss briefly to show that all but the one addressed by our study (No. 6) are unlikely candidates.

(1) *Feature processing interaction.* Miller (1977) attributed the boundary shift to non-independence in phonetic feature processing. (See also Haggard, 1970; Smith, 1973; Sawusch and Pisoni, 1974; Oden and Massaro, 1978). At the time, when feature detector theory was at the height of its popularity (see Remez, 1987), this hypothesis may have seemed to have some explanatory value. Basically, however, it is just a restatement of the finding, since it would be just as valid if the boundary shift went in the opposite direction. One testable prediction may be derived from this hypothesis, however. The shift in the place-of-articulation boundary should be a step function of VOT; that is, for a series of place-of-articulation continua differing by small increments in VOT, the perceptual boundary between labial and alveolar categories should change abruptly as VOT crosses the phonological voicing boundary but should remain relatively constant within voicing categories. In other words, the location of the place boundary should be a function of the perceived voicing category (the discrete response of a hypothetical "voicing detector"), not of VOT. In several experiments using appropriate stimulus arrays, Oden and Massaro (1978) and Massaro and Oden (1980) actually obtained results consistent with this prediction, although they nevertheless chose to emphasize the "relatively continuous" nature of the boundary change (Massaro and Oden, 1980, p. 1003). Repp (1978), on the other hand, obtained fairly continuous place boundary changes as a function of VOT; however, VOT varied over a smaller range in his stimuli. In view of these inconclusive data, the feature-processing interaction hypothesis cannot be dismissed, but it has little explanatory power in the context of contemporary theorizing, especially since it is indifferent to the direction of the boundary shift. The same can be said about Oden and Massaro's (1978) feature integration model, which, even though it assumes independent processing of acoustic features, represents the phonetic-feature interaction at the level of mental category prototypes. The model fits the data well, but it does not explain the direction of the effect.

(2) *Presence versus absence of F1.* A second hypothesis is that the boundary shift originates in the auditory system: Some auditory interaction may make the F2 and F3 transitions of aspirated stops appear to be lower in frequency than those of unaspirated stops, or may increase the relative perceptual salience of rising (labial) versus falling (alveolar) F2 and F3 transitions in aspirated as compared to unaspirated stops. The first formant could be involved in such an interaction. Because F1 tends to be weak during natural aspiration, and because "F1 cutback" is in fact an important cue for phonological voicelessness in initial English stop consonants (Liberman, Delattre and

Cooper, 1958), F1 has been attenuated as a matter of routine in the synthesis of aspirated stop consonants. There is also evidence in the literature that, in certain situations, the F1 transition, when it is present, may influence the perception of transitions in the higher formants. When a syllable is split between the ears, so that F1 goes to one ear and F2 to the other ear, the discriminability of F2 transitions is improved relative to a monaural or binaural condition (Rand, 1974; Danaher and Pickett, 1975). This improvement has been attributed to a release from peripheral "upward spread of masking" by F1. It seems reasonable that such masking would have a greater effect on F2 transitions that are close in frequency to F1 and/or have a similar (rising) trajectory; thus it might decrease the relative salience of labial transitions in unaspirated stops, so attenuation of F1 in aspirated stops would then result in a relative enhancement of these transitions, in accord with the observed perceptual boundary shift. In dichotic split-formant studies, Perl and Haggard (1974) and especially Perl (1975) did observe "a tendency for increased dichotic release from masking where initial F2 transitions tend towards the same slope as accompanying F1 transitions" (Perl, 1975, p. 36). Unfortunately, most other relevant studies failed to show such trends (Turek, Dorman, Franks and Summerfield, 1980; Schwab, 1981; Grunke and Pisoni, 1982; Hannley and Dorman, 1983; Nusbaum, Schwab and Sawusch, 1983). In addition, informal observations by the first author suggest that synthetic syllables in which phonological voicelessness is cued solely by F1 cutback without accompanying aspiration noise (cf. Liberman *et al.*, 1958) do not exhibit any place boundary shift. The upward-spread-of-masking hypothesis thus seems untenable.

(3) *Absence of release burst.* A third possible explanation takes note of the fact that most studies have employed synthetic syllables without release bursts. Alveolar release bursts, because of their different spectral energy distribution, are more intense than labial release bursts, and aspirated stops tend to have stronger bursts than unaspirated stops (Zue, 1976). Burst amplitude (with spectral properties held constant) has been shown to be a secondary place of articulation cue. Listeners report more labial stop percepts when the amplitude is low than when it is high (Ohde and Stevens, 1983; Repp, 1984). Thus, if listeners *expect* a burst, its absence may lead to a general bias toward labial stop percepts, and this bias may be larger for stimuli that normally have stronger release bursts, viz., aspirated stops. In other words, the absence of a strong burst may make a stimulus sound even more labial than does the absence of a weak burst. However, Ohde and Stevens (1983) employed aspirated and unaspirated stimuli that included synthetic bursts and still found a large place boundary shift as a function of aspiration. Therefore, the "missing burst" hypothesis seems less promising now than it did a few years ago. Besides, it is almost impossible to test rigorously because of the difficulty of synthesizing release bursts that are both realistic and matched to the formant transitions on a place of articulation continuum.

(4) *VOT as a place cue.* It is well known that alveolar stops have longer VOTs than labial stops, especially in their aspirated forms, although the difference is not very large and there is substantial overlap of the VOT distributions (see, e.g., Lisker and Abramson, 1967; Ohde, 1984). Even so, it is conceivable that the temporal aspect of VOT serves as a weak place cue in aspirated stops, such that listeners are somewhat more likely to

perceive labials when VOT is relatively short, and alveolars when VOT is relatively long. If the VOTs of the synthetic [$p^h a$]–[$t^h a$] stimuli used in earlier studies were on the short side, the place boundary shift in favor of labial responses could be accounted for. The longest VOT used by Oden and Massaro (1978) and Massaro and Oden (1980) was 40 msec; that employed by Repp (1978) was 42 msec; Miller (1977) and Ohde and Stevens (1983) used a VOT of 50 msec for their aspirated stops; and Alfonso and Daniloff (1980) used a VOT of 60 msec. The average VOT of [$p^h a$] and [$t^h a$] produced in isolation is about 70 msec, with the VOT of [$t^h a$] being some 10 msec longer than that of [$p^h a$] (Lisker and Abramson, 1967; present study). Thus all VOTs used in previous synthesis were indeed on the short (labial) side. It is noteworthy, however, that the largest place boundary shifts (about 145 Hz in terms of F2 onset frequency) were obtained by Alfonso and Daniloff (1980), who used the longest VOT for their aspirated continuum. This observation, together with the great variability of VOTs in natural speech, makes it unlikely that VOT could be responsible for the boundary shift.

(5) *Aspiration noise spectrum and/or intensity as a place cue.* Massaro and Oden (1980) proposed that the aspiration noise itself may provide a cue for labial place of articulation (see also Ohde and Stevens, 1983). At first glance, this hypothesis seems to ignore the fact that in synthetic stimuli (as in natural speech) the aperiodic source passes through the same F2 and F3 filters as the periodic source, leading to similar spectral shapes above F1. It is possible, however, that differences in the spectral slope and/or amplitude of periodic and aperiodic source spectra somehow contribute to the perceptual boundary shift, especially if they deviate from what is observed in natural speech. Unfortunately, these parameters are commonly omitted from descriptions of synthetic stimuli, and information about their magnitudes in natural speech is also hard to come by. Massaro and Oden did find that labial responses increased further when aspiration noise intensity was increased; however, since labial responses increased with VOT (up to 40 msec, the longest value used) in their study, the result may reflect the fact that stimuli with higher aspiration levels are phonetically equivalent to stimuli with longer VOTs, perhaps due to a time-intensity reciprocity in auditory perception (Repp, 1979; Darwin and Seton, 1983). Certainly there is no reason to believe that natural labial stops are characterized by more intense aspiration than alveolar stops. In summary, while the global acoustic characteristics of natural aspiration bear closer examination, it seems unlikely that they vary with place of articulation and, hence, that they could function as secondary place-of-articulation cues.

(6) *Different formant transitions in unaspirated and aspirated stops.* The sixth and final hypothesis is that the formant transitions are different in aspirated and unaspirated stops, so that listeners apply different criteria for place decisions along a formant transition continuum depending on whether aspiration is present or absent. Despite a long tradition of synthesizing unaspirated and aspirated stops with identical formant transitions for use in perceptual experiments (which may derive, in part, from the "locus" theory of Delattre, Liberman, and Cooper, 1955), there is in fact some limited support for this hypothesis in the acoustic phonetics literature. Fant (1973) reports that /p/ (i.e., [p^h]) tends to have higher F2 onsets than /b/ (i.e., [p]) before back vowels such

as /a/. However, his very limited data derive from a single speaker of Swedish, and some of the formant frequencies reported seem unusually low. Similar data for English collected by Lehiste and Peterson (1961) and replotted by Fant (1973) are suggestive at best. More convincing are Gay's (1978) spectrographic measurements of F2 onset frequencies in syllables produced by three male American speakers. F2 onset in /pap/ and /pup/ was about 180 Hz higher than in /bap/ and /bup/; however, it was about 125 Hz lower in /pip/ than in /bip/.

Gay mentions three possible causes of the difference in formant transitions preceding back vowels: (a) The *coarticulatory hypothesis:* Fant (1973) speculated that /b/ is coarticulated more strongly with a following back vowel (i.e., the tongue is more nearly in position for the vowel before the release of the stop closure) than is /p/, while no such difference exists between /d/ and /t/. (b) The *release timing hypothesis:* As the articulators begin to move towards the vowel, the release of aspirated stops may occur earlier in time than that of unaspirated stops, so that energy begins while the articulators are still farther away from the vowel target (Öhman, 1965; see Fant, 1973, p. 118). The acoustic consequences are similar to those predicted by the coarticulatory hypothesis, but it should be possible to overlay the formant trajectories of aspirated and unaspirated stops after correcting for the time shift (cf. Fant, 1973). (c) The *subglottal coupling hypothesis:* The higher F2 onsets for aspirated stops may arise from the open glottis during aspiration. This acoustic explanation appears very plausible in view of research by Lehiste (1964, cited in Lehiste, 1970) and Kallail and Emanuel (1984a, b) on whispered vowels, in which especially F1 but also F2 and F3 tend to be higher than in phonated vowels, with the possible exception of high front vowels. Indeed, glottal opening is likely to be wider at the beginning of aspiration than during whisper (Catford, 1977). Fant, Ishizaka, Lindqvist and Sundberg (1972) have modelled these effects of subglottal coupling, which may include additional subglottal formants in the aspiration spectrum, especially right after the release.

A clear demonstration of higher formant frequencies (especially of F2) in aspirated than in unaspirated stop consonants preceding [ɑ] would be of value for three reasons. First, the relevant data in the literature are incomplete and not easy to find; in particular, there have been no comparisons of the complete formant transitions in unaspirated and aspirated stops for both labial and alveolar places of articulation. Second, such data would provide an important guideline for realistic speech synthesis. Third, they would provide a sufficient explanation of the perceptual boundary shift and provide yet another illustration that listeners engaging in linguistic classification rely on tacit knowledge of a wealth of phonetic detail (see Repp, 1987).

Only the syllables [pɑ], [tɑ], [p^hɑ], [t^hɑ], were considered in this study, because they were the endpoints of the continua used in previous perceptual studies. Nevertheless, it was possible even in this limited context to address the three hypotheses about the origin of differences in formant frequencies between unaspirated and aspirated stops, if any were found. (a) If Fant's coarticulatory hypothesis is correct, the difference should be more pronounced for labial than for alveolar stops, since the tongue body is less free to anticipate the shape of the following vowel during alveolar closure. Also, the time course of the labial F2 transition should be independent of VOT in aspirated tokens;

that is, it should be a function of the movements of the upper articulators only. (b) If Öhman's release timing hypothesis is correct, the results should be similar, but in addition it should be possible to superimpose the average formant tracks of unaspirated and aspirated tokens by shifting them in time relative to each other. Thus, a finding of rising F2 transitions for [pɑ] but falling F2 transitions for [p^hɑ] would be incompatible with the release timing hypothesis, but not necessarily with Fant's coarticulatory hypothesis. (c) If the subglottal coupling hypothesis is correct, the F2 difference between aspirated and unaspirated stops should be present for both labial and alveolar stops and should disappear with voicing onset in aspirated tokens. Of course, these hypotheses are not mutually exclusive, and more than one explanation may be supported by the data.

In addition to providing measurements of F2 trajectories to address these principal hypotheses, the present study also yielded data on F1 and F3 frequencies, and on the spectral tilt and relative amplitude of aspiration – information that is difficult to locate in the literature but is useful for speech synthesis.

METHODS

Ten male speakers of American English produced the syllables [pɑ], [tɑ], [p^hɑ], [t^hɑ], five times in random order, reading from a list of randomized syllables spelled BA, DA, PA, TA. They were recorded in a sound-insulated booth using a Sennheiser microphone and an Otari MX5050 tape recorder located in an adjacent booth. The mouth-to-microphone distance was about 20 inches. All 200 utterances were low-pass filtered at 4.9 kHz and digitized at a sampling rate of 10 kHz with high-frequency pre-emphasis. Each file was edited to eliminate silence or (rare) voicing preceding the release. A 14-coefficient LPC analysis was then conducted using a 20-msec Hamming window advancing in 10-msec steps, and formant frequencies were estimated using the root solving method (ILS package, Version 4.0, distributed by Signal Technology, Inc.).

The resulting arrays of formant frequencies as a function of time were cleaned up by hand to eliminate occasional spurious peaks, to make sure that all frequencies were aligned with the appropriate formants, and to deal with the problem of missing values. One speaker was excluded from further analysis because of insufficient F2 data for labial stops. For the other speakers, missing formant frequencies were filled in by interpolating between preceding and following values or, if they occurred at the onset, by extending the first existing value backward in time. Missing frequencies were especially common in the initial time frames; this was not surprising, since release bursts often do not have a clear formant structure. Thirty-eight percent of the F2 data were missing in frame 1, 19% in frame 2, and from 12% to 3% in frames 3–10. Eighty-six percent of all missing values were in aspirated tokens; of these, 62% were in [p^hɑ] tokens and 38% in [t^hɑ] tokens. For F3, the percentages of filled-in values were 28% in frame 1 and between 7 and 15% in frames 2–10. While interpolation of missing F2 and F3 values in later frames should not have distorted the analysis results in any way, the filling in of missing initial values by level extension of later values (a conservative procedure) may have resulted in an underestimation of existing differences in formant frequencies between unaspirated

and aspirated tokens at onset. F1, of course, was generally absent during aspiration and was also spurious in unaspirated tokens for two speakers. To compare F1 in unaspirated labials and alveolars, the F1 data of the eight speakers with fairly complete values were analyzed after filling in missing values (36% in frame 1, 2–6% in frames 2–10).

Voice onset times of aspirated tokens were measured in a waveform display by locating the onset of the first glottal pulse. In addition, to corroborate the LPC analysis results and to examine the spectral and amplitude characteristics of aspiration, FFT spectra of all utterances were obtained from 20-msec Hamming windows centered 10, 30, and 50 msec after the release. To reduce random level fluctuations, the spectra were averaged over the five repetitions of each syllable by each speaker. From these average spectra we picked F2 peaks by eye wherever possible, interpolating if there were two closely adjacent peaks in the relevant region. This yielded complete estimates of F2 frequencies for all 10 speakers at the three time points for [pɑ] and [tɑ]; for [t^hɑ], only two data points (7% of the data) were missing; for [p^hɑ], however, peaks could not be located in 10 instances (33% of the data). As with the LPC data, the missing values were interpolated or extrapolated from the existing ones, so as to have a complete matrix for calculation of means and for statistical analysis.

Results and Discussion

F2 transitions

Because of considerable differences in utterance durations for different speakers, only the first 110 msec of each token (i.e., 10 overlapping 20-msec analysis time frames) were considered. The cleaned-up arrays of formant values were averaged across all tokens of all speakers to obtain an overall picture of the differences in formant transitions. These average F2 transitions are plotted as the connected points in Figure 1. It is evident that both aspirated syllable types had higher F2 onsets than their unaspirated counterparts, and that this difference gradually decreased over the first 70 msec or so. Right after the release the difference was larger for labials than for alveolars, but after 30 msec it seemed independent of place of articulation. In addition, it may be noted that F2 was higher for alveolar than labial tokens well beyond the first 100 msec. Formant transitions thus may be a good deal longer than the (approximately) 50 msec often cited in the literature and employed in speech synthesis.

A repeated-measures analysis of variance was conducted on the token averages with place of articulation, aspiration, and time as factors. All main effects and interactions were significant at $p = 0.0005$ or less, except for the place-by-aspiration interaction, which was not significant. The overall magnitude of the aspiration effect was thus similar for labial and alveolar stops. The triple interaction [$F(9, 72) = 5.71$, $p < 0.0001$], however, confirms that the aspiration effect was smaller for alveolar than for labial stops immediately after the release. Separate analyses of labial and alveolar tokens showed that the unaspirated/aspirated difference was significant for both places of articulation [labial: $F(1, 8) = 32.67$, $p = 0.0004$; alveolar: $F(1, 8) = 15.03$, $p = 0.0047$]. In addition,

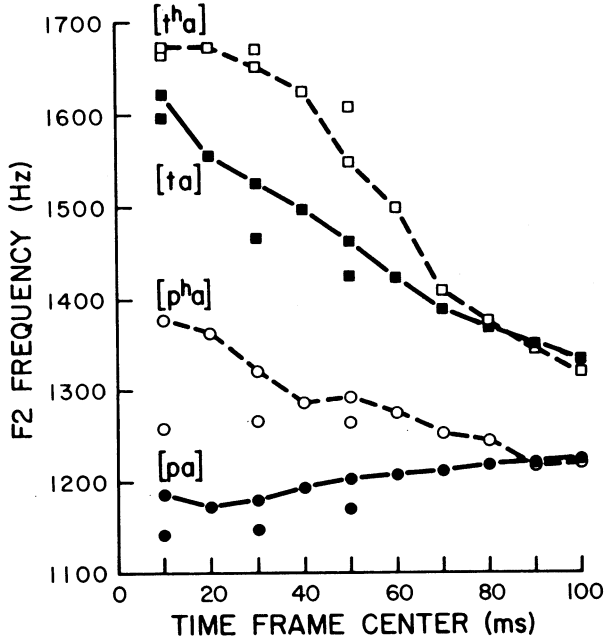

Fig. 1. The connected points show the average second formant (F2) transitions over the first 100 msec of [pa], [pʰa], [ta], and [tʰa], as determined by LPC analysis. Each transition represents the average of 45 utterances (5 tokens from each of 9 speakers). The unconnected points represent average F2 frequency estimates from FFT analysis of all 10 speakers' productions. Formant frequencies are plotted at the centers of the 20 msec time windows.

its decrease as a function of time was reflected in highly significant interactions between aspiration and time [labial: $F(9, 72) = 29.69$, $p < 0.0001$; alveolar: $F(9, 72) = 11.51$, $p < 0.0001$]. This pattern was shown by all individual speakers.

Similar analyses were conducted on the F2 frequency estimates derived from FFT spectra; the averages are plotted as the unconnected points in Figure 1. As pointed out in the Methods section, the data for [pʰa] were somewhat unreliable, which explains the major discrepancy between the LPC and FFT frequency estimates for that syllable. For the other syllables, there was reasonable agreement between the two sets of data, although FFT estimates seemed to be systematically lower than LPC estimates for unaspirated stops. Absolute differences aside, the FFT data clearly corroborate the finding of higher F2 frequencies during aspiration. In the overall analysis of variance, all effects except the place-by-aspiration interaction were significant at $p = 0.01$ or less. Tested separately, the main effect of aspiration was significant for both labial [$F(1, 9)$

= 7.72, p = 0.0214] and alveolar stops [$F(1, 9)$ = 34.09, p = 0.0002]; for the latter there was also a significant change of the effect over time [$F(2, 18) = 8.27, p = 0.0028$].

The magnitude of the difference for labials at release is in good agreement with Gay's (1978) data, as are the absolute LPC-derived formant frequencies. The magnitude of F2 difference between phonated and whispered [a] reported by Kallail and Emanuel (1984b) is also similar. This last observation, together with the finding of similar differences for labials and alveolars, except right after the release, suggests that the explanation is to be found in the open glottis during aspiration.

Of the two alternative explanations, Öhman's release timing hypothesis seems to be inconsistent with the present data. Even granting possible distortions due to averaging over tokens representing different vocal tract sizes and speaking rates, there is no way the transitions for unaspirated and aspirated tokens could be time-shifted to coincide in Figure 1. This is especially true in the case of [pa], which has a barely rising F2 transition, and [p^ha], which has a clearly falling one. Thus, this hypothesis can be dismissed. Fant's coarticulation hypothesis predicted a smaller difference for alveolars than for labials, which was found immediately after the release but not some tens of milliseconds later. It is possible that, as the tongue is freed from the constraint of the alveolar closure, it rapidly adjusts to the following vowel shape, and more so in [ta] than in [t^ha]. (Alternatively, the presence of a frication source at the alveolar constriction may obscure any existing F2 differences during alveolar release bursts.) The coarticulatory hypothesis thus is not incompatible with the data in Figure 1, even though Fant himself commented only on labial stops.

Another prediction of Fant's hypothesis, however, is that the time course of the F2 difference should be independent of when voicing starts in aspirated tokens. The subglottal coupling hypothesis, on the other hand, predicts that the difference should end at voicing onset. The F2 trajectories for [p^ha] and [t^ha] shown in Figure 1 were obtained by averaging over aspirated tokens with VOTs ranging from 40 to 126 msec, with an average of 70 msec (66 msec for labials, 73 msec for alveolars), which resulted in considerable smearing in the time domain. An alternative way to analyze the data is to line up all aspirated tokens at voice onset rather than at the release. Figure 2 shows the average F2 frequencies in the vicinity of voice onset after lining up aspirated tokens in this way, with unaspirated tokens lined up correspondingly by yoking them to aspirated tokens of the same speaker and shifting them by the same amount along the time axis. It can be seen that the F2 difference indeed disappears at voice onset for alveolar stops, and nearly so for labial stops. In analyses of variance on the five time frames following voice onset, no significant F2 differences were obtained for either labials or alveolars. An additional analysis including rank-ordered VOT as a factor was conducted to determine whether F2 onset frequency in aspirated stops increased with VOT. The result was negative.

Had the differences in F2 trajectories extended beyond voicing onset or had they ended much sooner, the coarticulatory hypothesis might have been favored over the subglottal coupling hypothesis. On the other hand, a positive correlation between F2 onset frequency and VOT in aspirated stops would have supported the latter hypothesis. As it is, the data are still consistent with both hypotheses, though the subglottal coupling

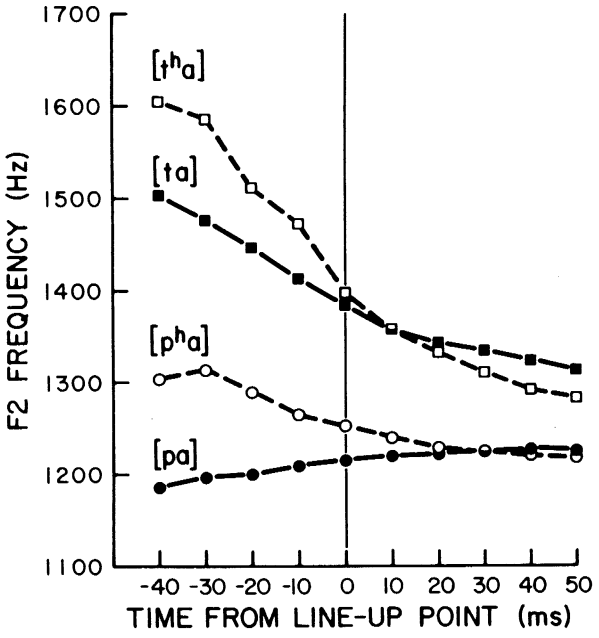

Fig. 2. Average second formant (F2) frequencies in the vicinity of voicing onset for [pʰa] and [tʰa] tokens lined up at voicing onset, and for yoked [pa] and [ta] tokens lined up at corresponding time points.

hypothesis would seem to provide a more parsimonious account. The acoustic consequences of subglottal coupling are necessary effects, while differences in the position of the upper articulators are not (as long as no direct observations of articulation show they do exist). The gradual decline in the F2 difference prior to voice onset probably reflects the gradual narrowing of the glottal opening before voicing starts (see, e.g., Kagaya, 1974; Hirose, 1977). The smaller difference between F2 of [ta] and [tʰa] right after release may be due to broadband frication noise generated while the constriction is narrow. Subglottal coupling thus provides a sufficient explanation of the observed differences in F2 trajectories.

F1 and F3 transitions

We also examined differences in F3 transitions in the same manner; however, there were no significant F3 differences as a function of aspiration in either labials or alveolars, whether aligned at release or at voice onset. Kallail and Emanuel (1984b), too, found only a very small (presumably nonsignificant) F3 difference between voiced and whispered [a].

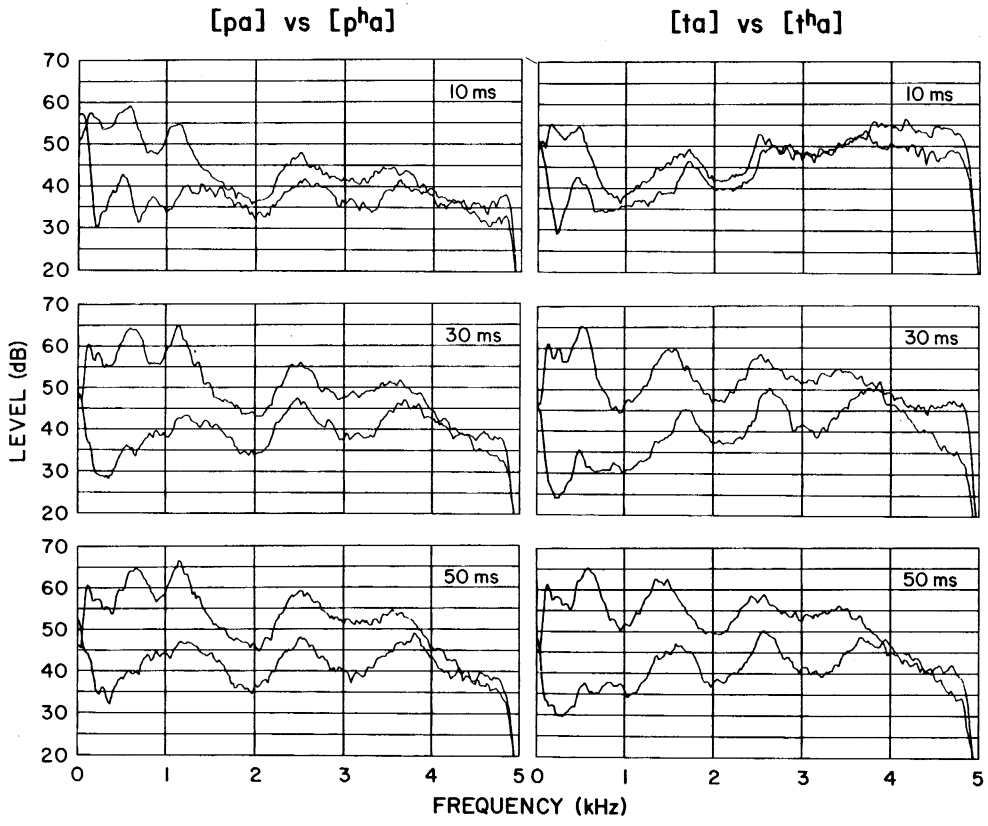

Fig. 3. Average Fourier (FFT) spectra of unaspirated and aspirated stops at three points in time, calculated using Hamming windows centered 10, 30, and 50 msec after the release. Each spectrum represents the average of 50 utterances (5 tokens from each of 10 speakers). The upper function in each panel represents the unaspirated stops, and the lower function the aspirated ones. All spectra include high-frequency pre-emphasis of approximately 6 dB/octave above 1 kHz, and less below.

F1, on the other hand, is strongly affected by a change in source, being about 250 Hz higher in whispered than in phonated male [a] (Kallail and Emanuel, 1984b), but its increased bandwidth makes frequency measurements difficult, and we did not attempt to determine F1 frequencies during aspiration. We did compare F1 transitions in unaspirated [pa] and [ta] for eight subjects (for two subjects the LPC analysis did not yield reliable F1 estimates, but the subject excluded from the F2 analysis was included here) and found a significant difference [$F(1, 7) = 21.87, p = 0.0023$] which decreased

over time [$F(9, 63) = 17.33$, $p < 0.0001$]. All subjects showed higher F1 onsets in [pɑ] than in [tɑ]; the averages were 669 and 589 Hz, respectively. After 100 msec, this 80 Hz difference had dwindled to 28 Hz.

Aspiration noise: Spectral tilt and relative amplitude

Finally, we compared spectral cross-sections of aspirated and unaspirated tokens at three points in time (10, 30, and 50 msec after the release). Figure 3 shows these spectra averaged over all tokens of all speakers. Although the formant peaks in these grand average spectra are somewhat flattened because of between-speaker variability in absolute formant frequencies, the general pattern is fairly representative of individual speakers' utterances. Three aspects deserve attention. First, the upward shift in F2 during aspiration is evident, except in the first time frame for alveolar stops, where the spectrum reflects the [s]-like frication noise that is part of the release burst (cf. Fig. 1). The F2 peak is rather broad for [$p^h ɑ$], which was also true for most individual speakers' spectra. On its lower skirt, a raised and attenuated F1 (see Kallail and Emanuel, 1984a, b) may have contributed to this prominence. On the upper skirt, additional subglottal resonances may have occurred (Fant, Ishizaka, Lindqvist and Sundberg, 1972), although we did not observe any distinct peaks in individual spectra that could be identified with such resonances.

Second, it is obvious that the spectrum during aspiration has a different tilt from that during voicing. Acoustic theory predicts a −12 dB/octave spectral slope when the source is voiced, and a −6 dB/octave slope when the source is noise from the glottis (Fant, 1960; Hillman, Oesterle and Feth, 1983). Although the spectra in Figure 3 are plotted on a linear frequency scale and include high-frequency pre-emphasis of approximately 6 dB/octave above 1 kHz, it is clear that they roughly conform to the predictions. If a correction for pre-emphasis were applied, all spectra would have a downward tilt, the voiced spectra more so than the aspirated ones, as predicted. Labial and alveolar tokens do not seem to differ in spectral tilt.

Third, the relative amplitude of aspiration should be noted. It is especially difficult to locate information in the literature on this parameter, which is often a source of frustration in synthesizing aspirated stops. As can be seen, the levels of voiced and aspirated spectra converge between 3.5 and 4 kHz but diverge increasingly at lower frequencies. The differences observed are somewhat larger than predicted on the basis of a 6 dB/octave slope difference; in fact, they are more in accord with a linear 6 dB/kHz slope difference (cf. Hillman *et al.*, 1983). On the average, the levels of voiced and aspirated F3 peaks differed by 11 dB, and those of F2 peaks by 18 dB, with very similar differences for labials and alveolars. Level differences were even larger in the F1 region, due to the reduction of F1 during aspiration. There was enormous individual variability, however, in the absolute magnitude of these differences. F3 level differences ranged from 4 to 17 dB across speakers, and F2 level differences from 7 to 27 dB, probably reflecting individual differences in source spectra.

SUMMARY AND CONCLUSIONS

We have shown that aspirated labial and alveolar stop consonants preceding [ɑ] have F2 transitions that start at significantly higher frequencies than those of unaspirated cognates. The difference gets smaller over time and disappears with voice onset, which suggests that it is due to upward shifts in vocal tract resonances caused by the open (and gradually closing) glottis during aspiration. These data replicate and extend earlier observations by others, and they provide a valuable guideline for improved speech synthesis. Fant (1973, p. 131) recommended long ago that a "minor correction for the effect of glottal opening on the F-pattern" be added in synthesis, and noted that "an open glottis increases F2 and F3 by about 50–100 Hz." Our data suggest that, in the context of [ɑ], the effect is about twice as large but restricted to F2. It is astonishing that this difference has gone relatively unnoticed for so long, and that it has been completely ignored in the long series of studies employing synthetic stop-consonant-vowel (mostly [–ɑ] or [–æ]) syllables and VOT continua over the last 20 years.

For reasons that are not well understood, the raising of F2 during aspiration seems to be absent for high front vowels such as [i] (Gay, 1978; Kallail and Emanuel, 1984a, b). It might be predicted, then, that the perceptual category boundaries on [pi]–[ti] and [p^hi]–[t^hi] continua should be similar. Unfortunately, this interesting prediction is not testable because F2 transitions do not reliably differentiate labial and alveolar stops in [i] context (see, e.g., Kewley-Port, 1982). Another prediction more amenable to test is that, unless there is differential coarticulation (Fant, 1973), the F2 transitions of whispered [pɑ] and [tɑ] (i.e., intended /bɑ/ and /dɑ/) should not differ from those of [p^hɑ] and [t^hɑ], and the category boundary on a noise-excited synthetic labial-alveolar continuum should likewise be similar to that on a [p^hɑ]–[t^hɑ] continuum.

The difference in F2 onset frequencies between aspirated and unaspirated stops preceding [ɑ] provides a sufficient explanation of the reliable perceptual shift in the labial-alveolar category boundary on a formant transition continuum as a function of VOT. The magnitude of perceptual boundary shifts reported in the literature (expressed in terms of F2 onset frequency, about 100 Hz on the average) matches the magnitude of the average acoustic difference in F2 transitions. If aspiration is introduced in a synthetic syllable without changing the F2 transition, as has been the custom, listeners *expect* the transition to be higher and therefore perceive the stimulus as relatively more labial. The effect of glottal opening on vocal tract resonances thus seems to be represented in listeners' tacit knowledge of phonetic regularities. Even though the boundary shift is essentially an artifact of primitive synthesis methods, it serves to remind us of the rich store of phonetic knowledge that listeners refer to in speech classification. Identification of speech depends as much on what listeners know about the sounds and gestures of their language as on what is in the acoustic signal (cf. Repp, 1987).

REFERENCES

ALFONSO, P. and DANILOFF, R. (1980). Allophonic backward masking of stop consonants. *Phonetica*, **37**, 355-376.

CATFORD, J.C. (1977). *Fundamental Problems in Phonetics.* Bloomington, IN: Indiana University Press.

DANAHER, E.M. and PICKETT, J.M. (1975). Some masking effects produced by low-frequency vowel formants in persons with sensorineural hearing loss. *Journal of Speech and Hearing Research*, **18**, 261-271.

DARWIN, C.J. and SETON, J. (1983). Perceptual cues to the onset of voiced excitation in aspirated initial stops. *Journal of the Acoustical Society of America*, **74**, 1126-1135.

DELATTRE, P.C., LIBERMAN, A.M. and COOPER, F.S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, **27**, 769-773.

FANT, G. (1960). *Acoustic Theory of Speech Production.* The Hague: Mouton.

FANT, G. (1973). Stops in CV-syllables. In *Speech Sounds and Features* (pp. 110-142). Cambridge, MA: MIT Press.

FANT, G., ISHIZAKA, K., LINDQVIST, J. and SUNDBERG, J. (1972). Subglottal formants. *Speech Transmission Laboratory Quarterly Progress and Status Report* (Stockholm: Speech Transmission Laboratory, Royal Institute of Technology), **1**, 1-12.

GAY, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, **63**, 223-230.

GRUNKE, M.E. and PISONI, D.B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception & Psychophysics*, **31**, 210-218.

HAGGARD, M.P. (1970). The use of voicing information. *Speech Synthesis and Perception* (University of Cambridge), **2**, 1-14.

HANNLEY, M. and DORMAN, M.F. (1983). Susceptibility to intraspeech spread of masking in listeners with sensorineural hearing loss. *Journal of the Acoustical Society of America*, **74**, 40-51.

HILLMAN, R.E., OESTERLE, E. and FETH, L.L. (1983). Characteristics of the glottal turbulent noise source. *Journal of the Acoustical Society of America*, **74**, 691-694.

HIROSE, H. (1977). Laryngeal adjustments in consonant production. *Phonetica*, **34**, 289-294.

KAGAYA, R. (1974). A fiberscopic and acoustic study of the Korean stops, affricates, and fricatives. *Journal of Phonetics*, **2**, 161-180.

KALLAIL, K.J. and EMANUEL, F.W. (1984a). Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. *Journal of Speech and Hearing Research*, **27**, 245-251.

KALLAIL, K.J. and EMANUEL, F.W. (1984b). An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects. *Journal of Phonetics*, **12**, 175-186.

KEWLEY-PORT, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, **72**, 379-389.

LEHISTE, I. (1964). *Acoustical Characteristics of Selected English Consonants.* Publication 34. Bloomington, IN: Indiana University Research Center in Anthropology, Folklore, and Linguistics.

LEHISTE, I. (1970). *Suprasegmentals.* Cambridge, MA: MIT Press.

LEHISTE, I. and PETERSON, G.E. (1961). Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America*, **33**, 268-277.

LIBERMAN, A.M., DELATTRE, P.C. and COOPER, F.S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, **1**, 153-167.

LISKER, L. and ABRAMSON, A. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, **10**, 1-28.

MASSARO, D.W. and ODEN, G.C. (1980). Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America*, **67**, 996-1013.

MILLER, J.L. (1977). Nonindependence of feature processing in initial consonants. *Journal of Speech and Hearing Research*, **20**, 519-528.

NUSBAUM, H.C., SCHWAB, E.C. and SAWUSCH, J.R. (1983). The role of 'chirp' identification in duplex perception. *Perception & Psychophysics*, **33**, 323-332.

ODEN, G.C. and MASSARO, D.W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

OHDE, R.N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America*, **75**, 224-230.

OHDE, R.N. and STEVENS, K.N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America*, **74**, 706-714.

ÖHMAN, S.E.G. (1965). On the coordination of articulatory and phonatory activity in the production of Swedish tonal accents. *Speech Transmission Laboratory Quarterly Progress and Status Report* (Stockholm: Royal Institute of Technology), **2**, 14-19.

PERL, N. (1975). Masking versus hemispheric sharing in the processing of split-formant stimuli. *Speech Perception* (Belfast: Department of Psychology, Queen's University), **2**, No. 4, 31-37.

PERL, N. and HAGGARD, M. (1974). Masking versus hemispheric sharing of processing for speech sounds. *Speech Perception* (Belfast: Department of Psychology, Queen's University), **2**, No. 3, 47-54.

RAND, T.C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**, 678-680.

REMEZ, R.E. (1987). Neural models of speech perception: A case history. In S. Harnad (ed.), *Categorical Perception* (pp. 199-225). New York: Cambridge University Press.

REPP, B.H. (1978). Interdependence of voicing and place decisions for stop consonants in initial position. *Haskins Laboratories Status Report on Speech Research*, **SR-53** (vol. 2), 117-150.

REPP, B.H. (1979). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, **27**, 173-189.

REPP, B.H. (1984). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language and Speech*, **27**, 245-254.

REPP, B.H. (1987). The role of psychophysics in understanding speech perception. In M.E.H. Schouten (ed.), *The Psychophysics of Speech Perception* (pp. 3-27). Dordrecht: Martinus Nijhoff.

SAWUSCH, J.R. and PISONI, D.B. (1974). On the identification of place and voicing features in synthetic stop consonants. *Journal of Phonetics*, **2**, 181-194.

SCHWAB, E.C. (1981). *Auditory and phonetic processing for tone analogs of speech.* Doctoral dissertation, State University of New York at Buffalo.

SMITH, P.T. (1973). Feature-testing models and their application to perception and memory for speech. *Quarterly Journal of Experimental Psychology*, **25**, 511-534.

TUREK, S.V., DORMAN, M.F., FRANKS, J.R. and SUMMERFIELD, Q. (1980). Identification of synthetic /bdg/ by hearing-impaired listeners under monotic and dichotic formant presentation. *Journal of the Acoustical Society of America*, **67**, 1031-1040.

ZUE, V.W. (1976). *Acoustic characteristics of stop consonants: A controlled study.* Technical Report No. 523. Lexington, MA: Lincoln Laboratory.