# Models of Aspirated Stops in English

**Helen M. Hanson*** and **Kenneth N. Stevens†**

*Research Laboratory of Electronics, M.I.T., Cambridge, Massachusetts, USA 02139
hanson@speech.mit.edu

†Dept. of Electrical Engineering and Computer Science, and Research Laboratory of Electronics
M.I.T., Cambridge, Massachusetts, USA 02139
stevens@speech.mit.edu

## ABSTRACT

The releases of unvoiced aspirated stops in English are typically modeled as having three consecutive phases, which overlap somewhat in time: (1) transient, (2) frication, and (3) aspiration. Close examination of stop releases reveals that the aspiration phase is more complicated than has been assumed. In this paper we explore the possibility that frication generated during the third phase may dominate the aspiration noise. This frication may be an extension of that generated at the original supraglottal constriction, or may be additional frication generated at a tongue-body or pharyngeal constriction formed in anticipation of the following vowel. Results suggest some subjects follow the classical model, but other subjects produce a mix of frication and aspiration during the third phase. Nevertheless, listeners do not have trouble with identification. We suggest that speakers can choose between using an extended burst or formant transitions to provide enhancing cues to place of articulation.

## 1 INTRODUCTION

The releases of unvoiced aspirated stops in English are typically modeled as having three phases: (1) transient, when the pressure behind the constriction is released and the resulting abrupt increase in volume velocity excites the entire vocal tract; (2) frication, when turbulence noise generated at the supraglottal constriction excites primarily the cavity in front of the constriction; and (3) aspiration, when turbulence noise generated near the approximating vocal folds excites the entire vocal tract. These phases are expected to overlap somewhat in time, but each is expected to be marked by the dominance of one type of excitation. For example, aspiration noise may be generated at the vocal folds during the second phase, but in this phase it is dominated by frication. These phases are presented in schematic form for voiceless unaspirated stops in Fig. 1; for voiceless aspirated stops, the aspiration phase will be considerably longer.

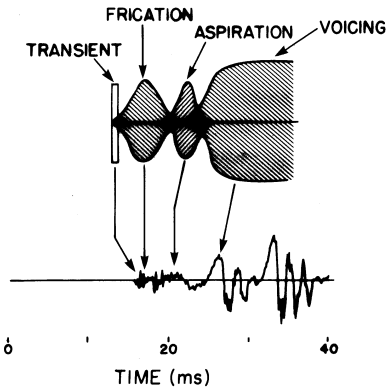


**Figure 1:** Schematic of events at the release of a voiceless unaspirated stop consonant. (From [1].)

This classical model of stop releases was cited by Fant in his seminal work on speech production [2], and, to the best of our knowledge, is the basis of earlier research on stop releases. Previous studies of stop releases are primarily temporal in nature. Lisker and Abramson [3] measured VOT (voice-onset time). Klatt [4] and Zue [5] also measured VOT, and they attempted to differentiate the frication and aspiration phases; however, they both reported that the two phases were often not unambiguous. Studies of the spectral characteristics of stops have largely focused on burst characteristics [5, 6]. Little attention has been paid to the aspiration phase.

According to the model of Fig. 1, the spectral characteristics of the burst will reflect the resonant frequencies of the cavity in front of the supraglottal constriction (for alveolar and velar places of articulation). The spectral characteristics of the "aspiration" phase might be expected to bear a similarity to the spectral characteristics of /h/. The overall amplitude might be lower, but the relative amplitudes of the peaks should be similar; that is, formant peaks should mainly be visible at mid-frequencies and higher, and their amplitudes should increase with frequency until about F3, after which their amplitudes should fall off. Similarly,


ISBN 1-876346-48-5 

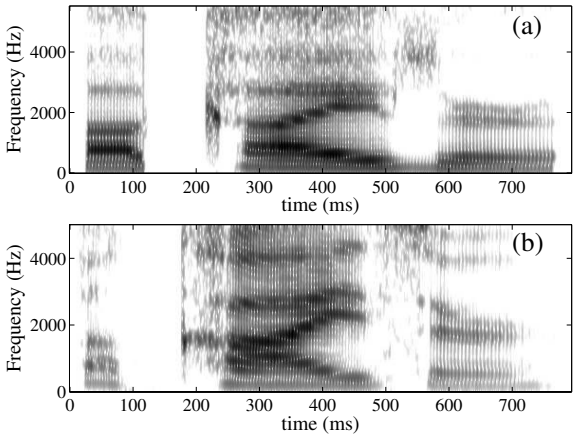


**Figure 2:** (a) Spectrogram of synthesized version of "a Kaiser", female speaker; (b) Spectrogram of natural version of "a Kaiser", female speaker.

the resonant frequencies during the aspiration phase should reflect the formants of the following vowel, as the tongue body moves into place. Figure 2(a) is a spectrogram of the phrase "a Kaiser" synthesized using HLsyn, which is based on a Klatt synthesizer. Acoustic events following the /k/ release are adjusted consistent with the typical model (Fig. 1). The frication and aspiration phases of the /k/ release are easy to distinguish, with the frication being much stronger in the F2 region, and the aspiration energy being rather evenly distributed among F2–F5.

For the purpose of improving speech synthesis, we are examining more closely the time course of the various phases of production of aspirated stops. Examination of spectrograms and spectrum-amplitude variation during stop releases reveal that the "aspiration" phase is more complicated than has been assumed. In some cases, the noise spectrum in this phase is dominated by one spectral prominence, rather than exhibiting several prominences as would be expected with a source near the glottis. Figure 2(b) is a spectrogram of a natural rendition of "a Kaiser." The second formant is rather strongly excited throughout the time prior to voice onset, while F3 is only weakly excited shortly before the vowel onset. The prominence in the F4 region is presumably a second resonance of the cavity anterior to the velar constriction.

Both Klatt [4] and Zue [5] have noted the difficulty of differentiating the frication from the aspiration phase on spectrograms. Together, these observations suggest that during the "aspiration" interval, frication noise is being generated either at the original supraglottal constriction, or at a tongue-body or pharyngeal constriction. For example, during a stop release preceding the vowel /i/ or /ɑ/, the vocal tract may be narrowed in anticipation of the vowel, and with the relatively large airflow due to the spread vocal folds, turbulence noise may be generated at this narrowing. Indeed Fant [2] (p. 185) has suggested that the interval we refer to as the aspiration phase is either aspirated or a mix of fricative and aspirated sound.

Thus, we are motivated to look in further depth at the spectral properties of stop releases, and to attempt to interpret the acoustic data in terms of the articulatory movements and the resulting airflows, pressures, and sound-generating mechanisms. The nature of the radiated sound during the production of aspirated stop consonants has been examined for the three places of articulation in English, and compared with the radiated sound for /h/. Recording methods and analysis are described in the next section, and results are presented in Section 3.

## 2 METHODS

Data were recorded from four subjects, two males and two females. The stimulus phrases are of the form /ə CV Cə ˈCVC/, where V is one of /ɑ, æ, ɪ, ʌ/, and C is one of /h, p, t, k, b, d, g/. Six tokens of each phrase were recorded for a total of 168 phrases; these 168 phrases were presented to the subjects in random order. The data were lowpass filtered, sampled at 16 kHz, and recorded directly to a computer disk. Only syllables formed by /h, p, t, k/ and vowel /ɑ/ are considered here, and only data for the first CV in each phrase is presented. Five tokens of each phrase were analyzed.

For the utterances where C is a stop, the stop release $t_r$ and voice onset $t_{v+}$ were labeled. Where C is /h/, the voice offset $t_{v-}$ and onset $t_{v+}$ were labeled.

The variation of spectral-peak amplitude was captured using average wideband spectra. These spectra were obtained in the following way: at 1-ms steps over a 6-ms interval, the speech signal was multiplied with a 3-ms Hamming window and a DFT was computed. The magnitudes of the DFTs were squared and averaged together before being converted to dB. For the mathematically inclined, we might define such an average spectrum $\overline{S}$ as follows:

$$\overline{S}_{t_0}(\omega) = 10\log_{10}\left[\frac{1}{K+1}\sum_{t=t_0-\frac{K}{2}}^{t_0+\frac{K}{2}}|S_t(\omega)|^2\right] \qquad (1)$$

where $t_0$ is the center of the averaging frame, $K$ is the frame size in ms over which we average (that is, 36 ms), and

$$S_t(\omega) = \mathcal{F}[s(\tau)w(\tau - [t - T_w/2])] \qquad (2)$$

is the DFT of the speech signal $s(t)$ multiplied by a Hamming window of length $T_w = 3$ ms. Figure 3 is an example of such an average wideband spectrum.

ISBN 1-876346-48-5  

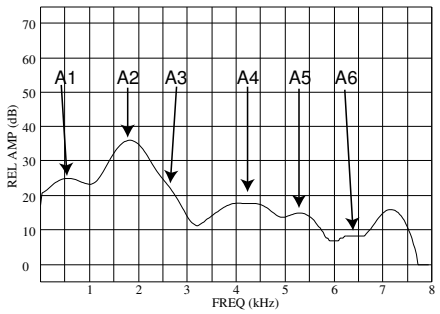


**Figure 3:** Average spectrum of a /k/ burst (male speaker), computed over a 6-ms interval. The labels A1–A6 indicate the spectral peaks associated with the first six formant frequencies.

Average spectra were computed at 3-ms steps over the stop releases, starting at $t_r$ and continuing to about $t = t_{v+} + 30$ ms. For /h/ we started at mid-consonant, or slightly before. For each average spectrum, we measured the amplitudes of the spectral peaks that by our best guess corresponded to formant frequencies. In Fig. 3 we have indicated the locations of the formant peaks. Note that, as might be expected for the fricative portion of a stop release, there is not always a "peak" per se. For example, in the F3 region, there is merely the hint of a shoulder on the spectrum. In these cases, we simply made a measurement on the spectrum where the peak would be expected for the following vowel. In this way we obtained formant-amplitude tracks for F1–F6. For each consonant (/p, t, k, h/), the tracks for each token were aligned at the release and averaged, yielding formant-amplitude tracks for F1–F6 representative of each consonant. An average VOT was also computed for each consonant.

Finally, we compared the amplitude tracks across consonants by aligning them (1) at the release, and (2) at the average onset of voicing. The former allows us to compare the frication across consonants, while the latter allows us to more easily compare the "aspiration" across consonants, because VOT can vary across place of articulation.

## 3 RESULTS

Because of space constraints, it is impossible to present a comprehensive picture of the data, but we can give a general idea of our results thus far. Figures 4–5 summarize the results for F2-amplitude and F5-amplitude variations. The top row in each figure shows graphs of the male data and the bottom row shows the female data. Each graph shows the average formant-amplitude track for each of the four consonants, aligned at the voice onset, that is $t = 0$ corresponds to voice onset. The heavy solid line is for /h/, which we can think of as a reference. For aspirated stop consonants, the vocal folds should be approximating during the "aspiration" phase. Thus, one might expect that if excitation during this phase was primarily aspiration, it would not be stronger than what is observed for /h/ (relative to the following vowel), especially just before voice onset.

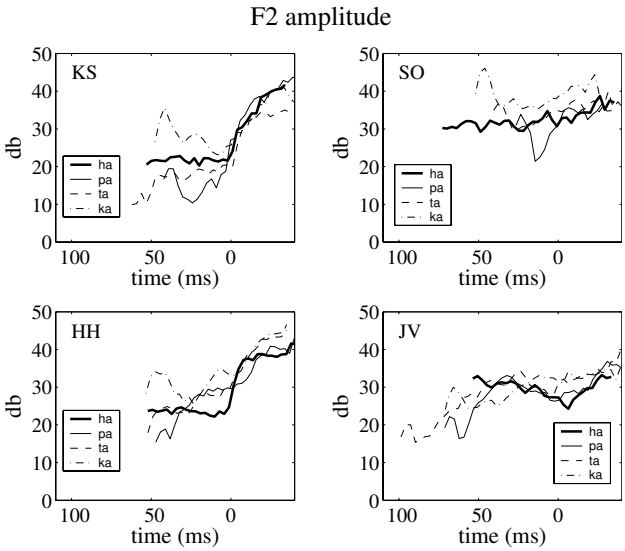


**Figure 4:** F2-amplitude tracks for two male speakers (top row) and two female speakers (bottom row).

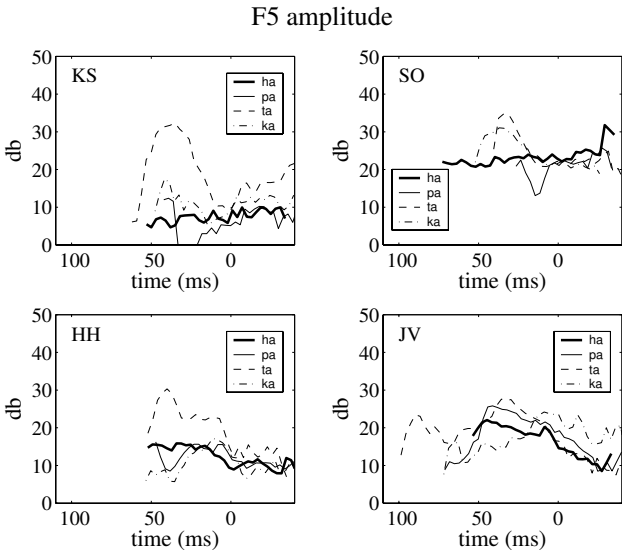


**Figure 5:** F5-amplitude tracks for two male speakers (top row) and two female speakers (bottom row).

Looking first at Fig. 4, we see that the F2 amplitude during "aspiration" for /p, t/ has a tendency to be equal to or weaker than that for /h/. For subjects SO and JV, that is also true for /k/, but for subjects KS and HH, the F2 amplitude exceeds /h/ during both the /k/ burst and aspiration, suggesting some frication near the velar constriction for these two subjects. In Fig. 5, the F5 amplitude track for /p/ again has a tendency to be equal to or weaker than that for /h/. The


ISBN 1-876346-48-5 

F5 amplitude for /k/ behaves similarly, except that for subject SO it is greater than that for /h/ during most of the VOT. The F5 amplitude for /t/ is greater than /h/ for all four subjects, although it is less obvious for subjects SO and JV.

Although we have not presented data for the other formants, the results can be summarized as follows:

- Subjects KS, SO, and HH show signs of mixed frication and aspiration in the F3 region for /t/.
- Subjects KS and HH show evidence of mixed frication and aspiration in the F4 region for /t/.
- Subject KS shows evidence of mixed frication and aspiration in the F6 regions for /t/.

Thus, it seems that frication may be generated during the "aspiration" phase of voiceless aspirated stop consonants, but usually only for velar and alveolar consonants. In addition, not all subjects seem to exhibit this mix of aspiration and frication. In particular, for the small amount of data analyzed here, subjects JV and SO seem to follow the classical model of voiceless aspirated stop production.

Because frication in the F2 region was only present for /k/, and frication in the F3–F6 regions was mainly present for /t/, it seems unlikely that the frication is being generated at a tongue body constriction formed in anticipation of the following vowel. Rather, it may be that the supraglottal constriction formed for the stop consonants /k/ and /t/ is maintained beyond the burst, and additional frication is generated there.

## 4 DISCUSSION AND SUMMARY

In this paper we have made a first attempt to reconsider the classical model of the production of aspirated stop releases. For the small amount of data considered here, we found that there is considerable variability in the acoustic attributes studied, both within a speaker and across speakers. Some speakers seem to follow the classical model, but other speakers seem to produce both frication and aspiration during the "aspiration" phase of stop releases. Despite this variability among subjects, listeners can easily identify the intended consonants. While that may seem puzzling, we offer a tentative explanation. In the classical model, the burst phase gives listeners cues to place of articulation based on the length of the cavity in front of the supraglottal transition, and the aspiration phase gives cues to place based on formant movements as the tongue body moves into place for the following vowel. Speakers for whom frication dominates in the aspiration phase are simply replacing one cue to place of articulation (formant movements) with another (length of the front cavity). One can say that speakers are free to choose between two methods of enhancing place of articulation, and that the two strategies are equally effective.

For speakers who choose to extend frication into the aspiration phase, our initial findings indicate

1. For velar stops there is an extended interval in which F2 or F3 is the major spectrum prominence following the release.
2. For alveolar stops there is a long interval of frication, appearing as a prominence in the F4–F5 region.
3. For labial stops, the initial transient and brief frication are followed by a classical aspiration interval with excitation of several formants.

In summary, the classical model of the production of voiceless aspirated stop consonants may be oversimplified. The work reported here is only a first step in trying to produce an improved model. Such a model has implications for speech synthesis, and for speech recognition. In future work we will examine the same consonants in additional vowel contexts and add additional subjects.

## ACKNOWLEDGMENTS

The work reported in this paper was supported by NIH grant no. DC00075. We thank our subjects for their cooperation.

## REFERENCES

[1] K. N. Stevens, "Models for the production and acoustics of stop consonants," *Speech Commun.*, vol. 13, pp. 367–375, 1993.

[2] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[3] Leigh Lisker and Arthur S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, pp. 384–422, 1964.

[4] Dennis H. Klatt, "Voice onset time, frication, and aspiration in word-initial consonant clusters," *J. Speech Hear. Res.*, vol. 18, pp. 686–706, 1975.

[5] Victor W. Zue, "Acoustic characteristics of stop consonants: A controlled study," Tech. Rep. 523, M.I.T. Lincoln Laboratory, Lexington, MA, 1976.

[6] Kenneth N. Stevens, Sharon Y. Manuel, and Melanie Matthies, "Revisiting place of articulation measures for stop consonants: Implications for models of consonant production," in *Proceedings of the XIVth International Congress of Phonetic Sciences, ICPhS 99*, John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville, and Ashlee C. Bailey, Eds., San Francisco, 1999, vol. 2, pp. 1117–1120.