J. Acoust. Soc. Am., v. 107, p. 2907

4pSC20. Modeling stop-consonant releases for synthesis

Helen M. Hanson and Kenneth N. Stevens*

Sensimetrics Corporation, Somerville MA
and
MIT Research Laboratory of Electronics, Cambridge MA

*also with the MIT Dept. of Elec. Eng. And Comp. Sci.

1. Introduction

Speech production is usually described in terms of vocal-tract configuration, articulator movement, and source characteristics. However, the properties of the vocal-tract walls may very much affect the characteristics of certain classes of segments, as described in detail in an upcoming paper by Stevens (to appear).

In the current work, we focus on the effects of vocal-tract surfaces on the releases of stop consonants. Although touched on in earlier works by Stevens (1993, 1998), in this paper we show through synthesis experiments that in order to explain acoustic data on burst durations (Klatt, 1975; Zue, 1976), it is necessary to include these effects in models of stop releases. In addition, we provide an explanation of the reported variation of VOT by place of articulation (Lisker and Abramson, 1967; Klatt, 1975; Zue, 1976) for both voiced and voiceless aspirated stop consonants.

2. Variation of VOT and burst duration by place

It has been observed that in the production of stop consonants, the voice-onset time is progressively longer for labial, alveolar, and velar stops. Lisker and Abramson (1967) found this phenomenon in stops produced in 11 languages, while Klatt (1975) and Zue (1976) observed it in American English. For the voiceless aspirated stops, Klatt (1975) and Zue (1976) segmented VOT into burst and aspiration durations. They found that it is primarily the burst that varies in duration, while the aspiration duration is fairly constant across place.

Because of the apparent universality of this property, it is believed to have a physiological basis. One cause proposed is the difference in mass of the articulators. That is, the tongue body is more massive than the tongue blade, which is more massive than the lips. In addition, the length of the constriction is longer for velars than for labials and alveolars. Consequently, the release of the constriction, and the rate of change of the pressure behind it, is believed to be slower for velars than for labials, and it should take longer for transglottal pressure to build up enough to achieve vocal-fold vibration (Klatt, 1975; Zue, 1976).

There are two problems with this explanation. The first is that while it can explain the difference in *burst* durations for voiceless aspirated stops, it does not explain the difference in VOT, because the vocal folds are not adducted until long after the consonant is released and oral pressure has fallen. Therefore, it has been suggested that on the surface, VOT variation by place is similar for voiced and voiceless stops, but the underlying process that leads to this variation is different (e.g. Klatt, 1975; Maddieson, 1997).

A second problem is that recent attempts to synthesize voiced stops using the quasi-articulatory synthesizer HLsyn suggest that the differences in rate of articulator release may not be sufficient to result in the VOT differences reported.

Figure 1 shows a schematic of the HLsyn system, in which higher-level, quasi-articulatory parameters are mapped to acoustic, Klatt-synthesizer parameters. Figure 2 illustrates the HLsyn parameters, and Fig. 3 shows the circuit model that is used to map HLsyn parameters to aerodynamic events, from which some of the acoustic (KL) parameters are derived.

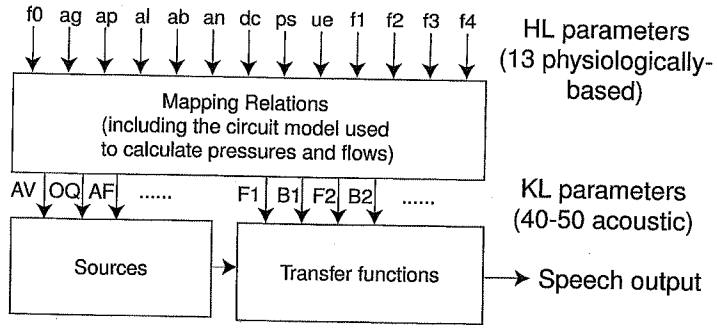

Figure 1. Schematic of the HLsyn synthesizer, in which higher-level quasi-articulatory parameters are mapped to acoustic (Klatt synthesizer) parameters.

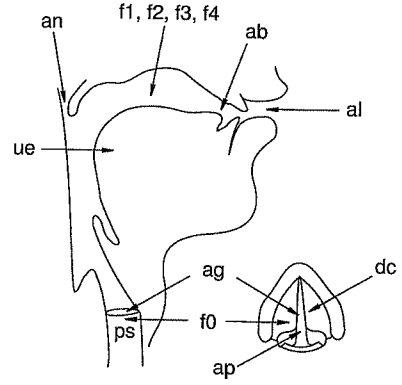

Figure 2. An illustration of the HLsyn parameters.

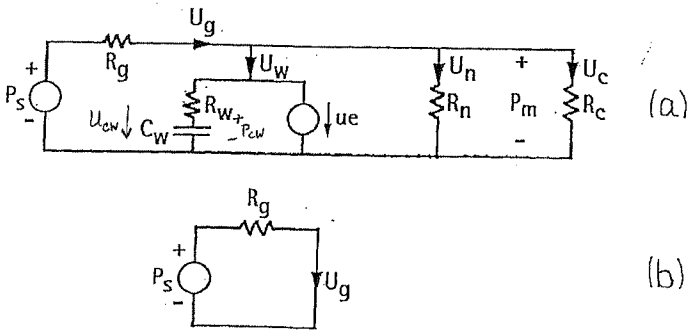

Fig. 3. (a) Low-frequency equivalent circuit used to calculate the intraoral pressure P_m in the vocal tract and various flows. (b) Simplified circuit for a non-nasalized, unconstricted vocal-tract configuration.

Figures 4-5 illustrate synthesized versions of the utterances /ə ˈba/ and /ə ˈga/. In HLsyn, the parameter **al** controls the cross-sectional area of the constriction at the lips. The cross-sectional area at the dorsum is controlled by the parameter **f1**, which is mapped to an intermediate area parameter *acd*. For our examples, these two parameters were varied so that consonantal closure takes place at 100 ms, while release occurs at 200 ms. For the labial, **al** was set such that the rate of release was 100 cm^2/sec; for the velar, **f1** was set such that the rate of release was 25 cm^2/sec. The average area of the glottis **ag** was set at 4 mm^2 for the duration of the utterances.

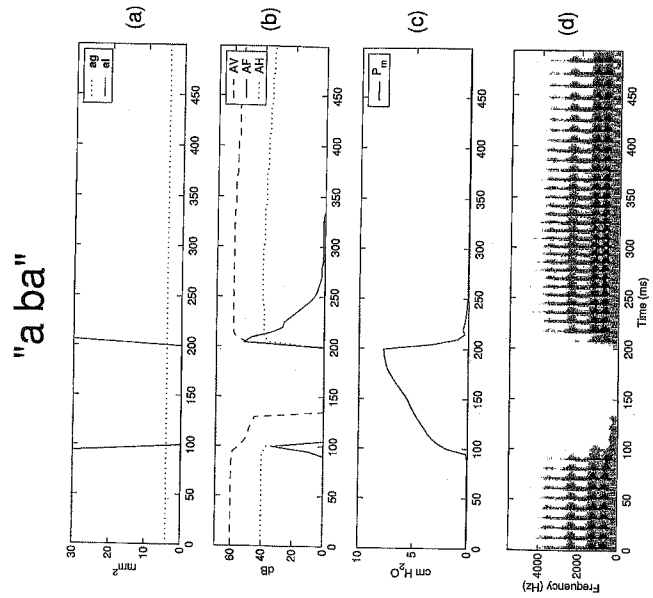

Fig. 4. Synthesized utterance "a ba". (a) At the release, the cross-sectional area at the lips, **al**, changes at a rate of 100 cm^2/sec. The average glottal area is constant at 4 mm^2 throughout the voiced utterance. (b) Amplitudes of the acoustic sources (AV = amplitude of voicing; AF = amplitude of frication; AH = amplitude of aspiration). Note that voicing starts immediately upon release. (c) The intraoral pressure, Pm. (d) Spectrogram of the utterance.

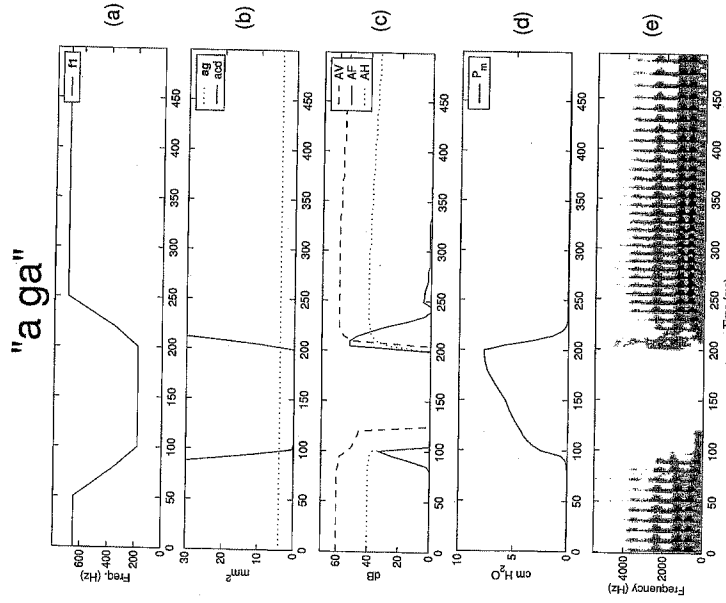

Fig. 5. Synthesized utterance "a ga". (a) The first natural frequency, f1, maps to the cross-sectional area at the dorsum. (b) At the release, the cross-sectional area at the dorsum, *acd*, changes at a rate of 25 cm^2/sec. The average glottal area is constant at 4 mm^2 throughout the voiced utterance. (c) Amplitudes of the acoustic sources (AV = amplitude of voicing; AF = amplitude of frication; AH = amplitude of aspiration). Note that voicing starts almost immediately upon release. (d) The intraoral pressure, Pm. (e) Spectrogram of the utterance.

The third panel from the bottom in both figures shows the resulting acoustic (Klatt synthesizer) parameters for the voicing and noise sources (AV = amplitude of voicing, AF = amplitude of frication, and AH = amplitude of aspiration). It can be seen that voice onset occurs at the consonantal release, simultaneously with the onset of frication noise. In this same panel, it can be seen that the amplitude of frication is not strong for a long enough time to get a proper burst. Thus, if the HLsyn model is correct, the differences in articulator rates of release are not sufficient to result in VOTs or bursts that vary by place of articulation.

The second panel from the bottom illustrates that the oral pressure falls off very rapidly following the consonantal release, and thus the delay in developing sufficient transglottal pressure is minimal. For the same reason, relatively strong frication can not be sustained for more than 5 ms or so. In order to produce appropriate VOTs and amplitude of frication, oral pressure must be maintained at a relatively high value for a longer time following release.

Finally, the last panel in each figure is a spectrogram of the synthesized utterance. It can be seen, especially for the velar, that there is not as much frication as might be expected.

In the model we used for articulator release, only movement due to active muscle forces was included. However, passive forces applied by the elevated intraoral pressure to the compliant articulator surfaces can contribute significantly to the acoustic properties of stop consonants (Stevens, to appear). We theorize in the next section that inclusion of these passive forces in a model of stop-consonant releases can account for the prolonged frication that occurs at release, and can also account for the place-related VOT differences that have been reported. The model provides a unified explanation for place-related VOT variations in both voiced and voiceless stop consonants.

3. Mechanism of release of stop consonants

Midsagittal sections showing the vocal tract shape for the three stop consonants /b/, /d/, and /g/ in English are displayed in Fig. 6.

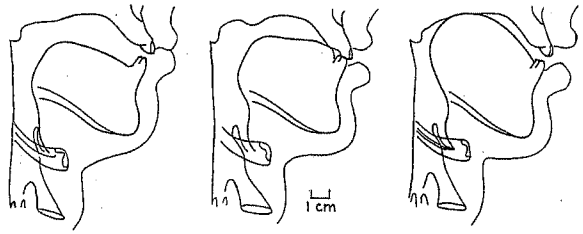

Fig. 6. Midsagittal sections for the stop consonants /b/, /d/, and /g/ in English. (after Perkell, 1969)

During the closure for an alveolar and velar consonant, the tissue of the tongue blade or tongue body is compressed and is in contact with the hard palate. For a labial consonant (left panel of Fig. 6), the two lips are pressed together. There is a buildup of pressure behind the closure. This pressure is sufficient to cause an outward displacement of the surfaces of the soft tissue. In particular, tongue or lip tissue displaces in response to this intraoral pressure.

The mechanical compliance of these surfaces is estimated to be in the range 1.0×10^{-5} to 3.0×10^{-5} cm^3/dyne (Ishizaka et al., 1975). Thus, in response to an increased intraoral pressure of, say, 6 cm H_2O (5900 dynes/cm^2) the surface displacement is in the range 0.6 mm to 1.8 mm. This compliance may be different for voiced and voiceless consonants, i.e., it is at the lower end of the range for voiceless consonants (Svirsky et al., 1997).

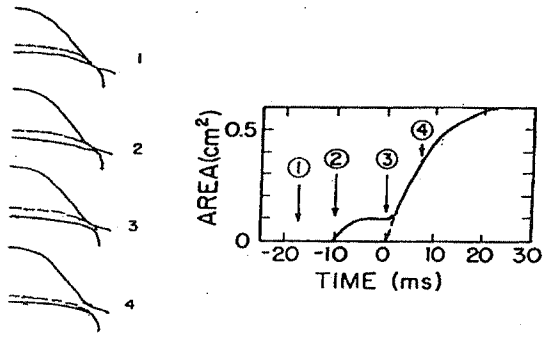

Fig. 7. (a) A series of schematicized shapes of the tongue blade as the release occurs for an alveolar stop consonant. The dashed line in each panel shows the contour of the tongue blade if there were no increased pressure behind the constriction, and the solid line gives the estimated tongue-blade shape if the effect of the pressure is included. (b) Schematization of the cross-sectional area versus time at the consonant release when the effect of the pressure is included (solid line), and when there is no increased intraoral pressure (dashed line). The labeled arrows indicate the points in time corresponding to the panels in (a). (After Stevens, to appear).

As a consequence of the displacement of the surface of the tongue or lower lip in response to the increased intraoral pressure, the time course of the consonantal release is modified relative to what it would be if there were no increase in intraoral pressure. The sequence of events is shown schematically in Fig. 7(a) for a model of an alveolar release. Before the release, the tongue surface upstream from the constriction is displaced downward about 1-2 millimeters, so that the region of contact with the palate is reduced. As the tongue body moves downward to begin the release, the length of this contact region decreases, and finally the opening occurs. This release occurs before it would if the walls were rigid. Immediately there is flow of air through the constriction, and the force on the surface due to the pressure drops to zero. The displaced surface springs back to its normal position, with a time constant R_wC_w, determined by the compliance C_w and internal resistance R_w of the tissue. (R_w is approximately 1000 dyne-s/cm^3.) This time constant is in the range 10-30 ms.

The surface displacement at the point where the closure is released is superimposed on the steady downward displacement of the tongue surface. This total displacement, then, has the form shown in Fig. 7(b). (The figure actually shows the cross-sectional area, making appropriate assumptions about the cross dimension.)

As a rough approximation we assume a fixed width of the constriction of about 5 mm. In this case, there is a plateau with a cross-sectional area of about 3-10 mm^2 and a duration of about 10 ms. This plateau intersects with the area that would have occurred had there been no increased intraoral pressure.

Some evidence that such plateaus in cross-sectional area actually occur for stop consonants (but not for nasals, for which intraoral pressure does not build up) can be found in a study of bilabials by Fujimura (1961).

The duration of the plateau in area that precedes the rapid rise in area is expected to be roughly proportional to the anterior-posterior length of the contact region, which is different for the lips (for labials), the tongue blade (for alveolars), or the surface of the tongue body (for velars). These lengths are estimated to be 0.5 to 1.0 cm for labials and alveolars and 2.0 cm for velars. Thus, the duration of the plateau at the release of labials and alveolars is expected to be 5-10 ms.

The 'hesitation' in the time course of the cross-sectional area of the constriction at the release of a stop consonant provides a time interval in which a frication noise burst is generated preceding the full onset of glottal vibration. It is noted that the duration of the burst is not strongly dependent on the rate of increase of cross-sectional area as the consonant is released, but rather depends upon the length of the contact region during the stop-consonant closure.

4. Synthesis examples

Voiced stops. We now revisit the synthesis examples presented in Section 2. Revised versions of these examples, for which a short plateau in the cross-sectional area of the constriction has been introduced, are illustrated in Figs. 8-10. In addition to the labial and velar examples, we include an alveolar stop. These examples are seen to be superior to the earlier ones in that the bursts are longer, as are the VOTs. Because the durations of the plateaus in the articulator trajectories vary by place, the burst and VOT durations also vary by place.

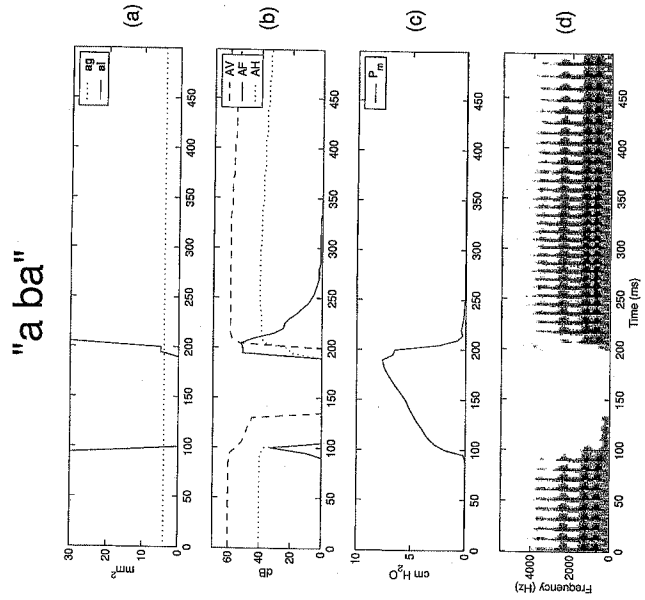

Fig. 8. Synthesized utterance "a ba". (a) At the release, the cross-sectional area at the lips, **al**, is maintained at 5 mm^2 for 5 ms, after which the rate of change becomes 100 cm^2/sec. The average glottal area is constant at 4 mm^2 throughout the voiced utterance. (b) Amplitudes of the acoustic sources. Note that voicing starts about 10 ms after release. (c) The intraoral pressure, Pm. (d) Spectrogram of the utterance.

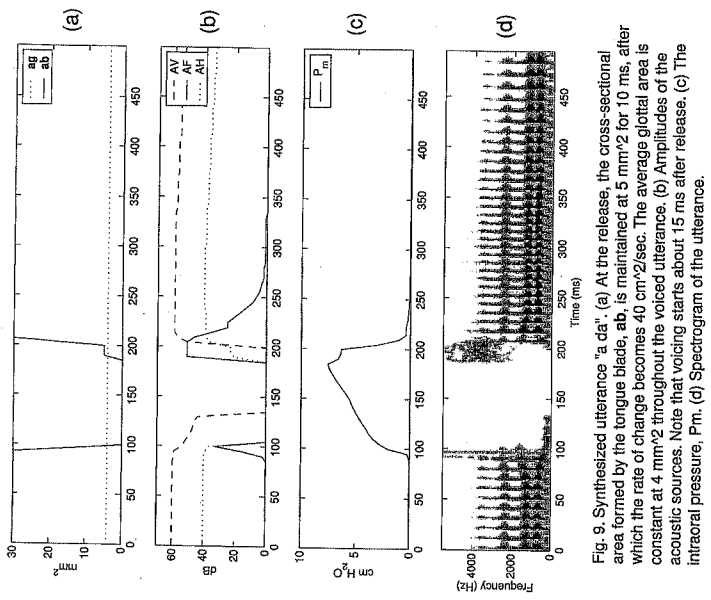

Fig. 9. Synthesized utterance "a da". (a) At the release, the cross-sectional area formed by the tongue blade, **ab**, is maintained at 5 mm^2 for 10 ms, after which the rate of change becomes 40 cm^2/sec. The average glottal area is constant at 4 mm^2 throughout the voiced utterance. (b) Amplitudes of the acoustic sources. Note that voicing starts about 15 ms after release. (c) The intraoral pressure, Pm. (d) Spectrogram of the utterance.

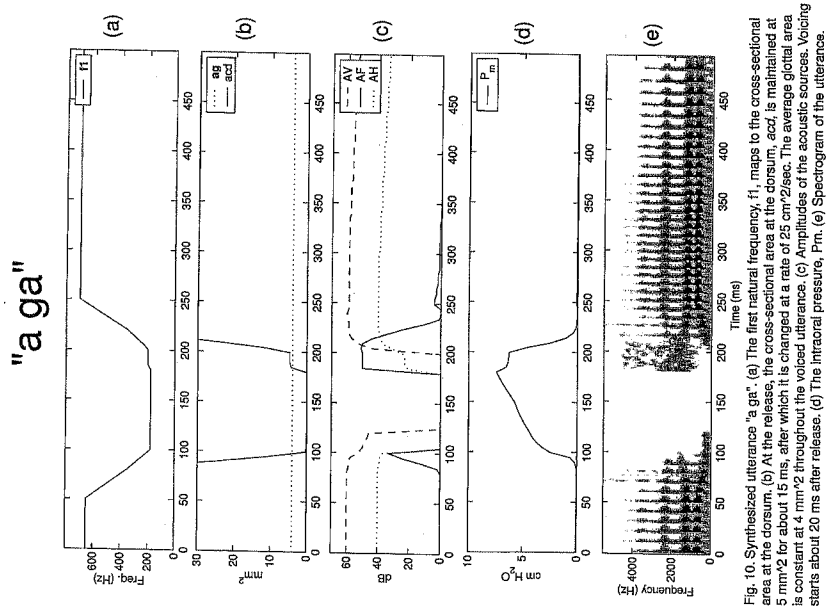

Fig. 10. Synthesized utterance "a ga". (a) The first natural frequency, f1, maps to the cross-sectional area at the dorsum. (b) At the release, the cross-sectional area at the dorsum, *acd*, is maintained at 5 mm^2 for about 15 ms, after which it is changed at a rate of 25 cm^2/sec. The average glottal area is constant at 4 mm^2 throughout the voiced utterance. (c) Amplitudes of the acoustic sources. Voicing starts about 20 ms after release. (d) The intraoral pressure, Pm. (e) Spectrogram of the utterance.

Voiceless stops. That the VOTs of voiceless stops also vary by place easily falls out from the model. Figures 11-13 below illustrate the synthesis of labial, alveolar, and velar voiceless aspirated stops. For these voiceless stops, the average glottal area begins to increase shortly before consonantal closure occurs. It gradually increases to its maximum value of 30 mm^2 just before the active release of the constriction would occur (200 ms). At 200 ms, the adduction gesture begins and the glottal area reaches its modal value of 4 mm^2 about 20 ms into the following vowel. The articulator trajectories are nearly identical to those of the voiced stops, except that the cross-sectional area of the constriction just after release is slightly larger for the voiceless stops.

As seen in the plots of the acoustic source amplitudes, VOT varies by place as expected. Thus, our model suggests that the same underlying process leads to VOT variation by place in both voiced and voiceless stop consonants.

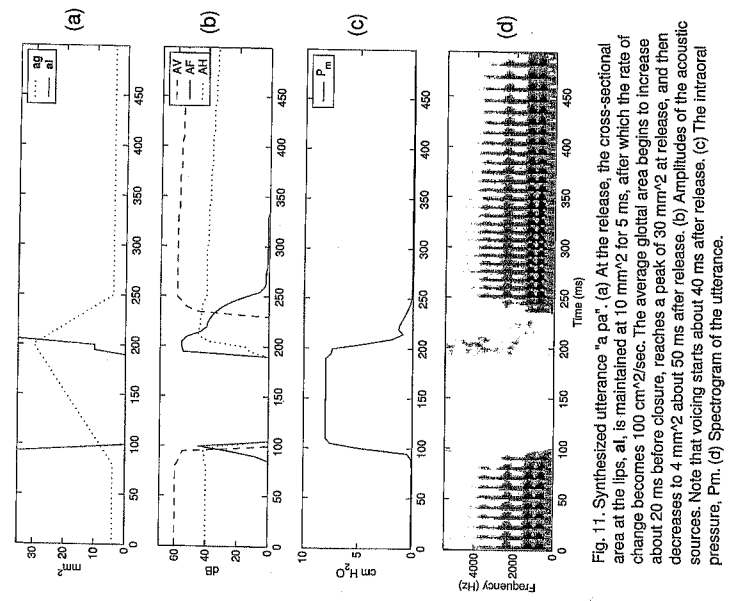

Fig. 11. Synthesized utterance "a pa". (a) At the release, the cross-sectional area at the lips, **al**, is maintained at 10 mm^2 for 5 ms, after which the rate of change becomes 100 cm^2/sec. The average glottal area begins to increase about 20 ms before closure, reaches a peak of 30 mm^2 at release, and then decreases to 4 mm^2 about 50 ms after release. (b) Amplitudes of the acoustic sources. Note that voicing starts about 40 ms after release. (c) The intraoral pressure, Pm. (d) Spectrogram of the utterance.

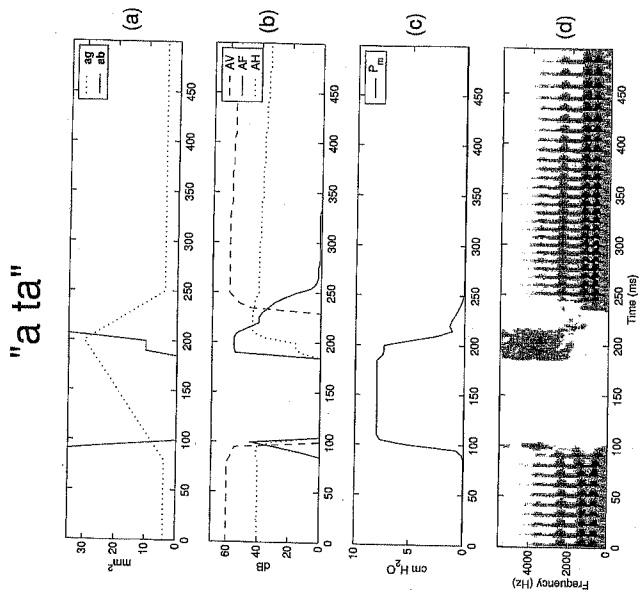

Fig. 12. Synthesized utterance "a ta". (a) At the release, the cross-sectional area of the constriction formed by the tongue blade, **ab**, is maintained at 10 mm^2 for 10 ms, after which the rate of change becomes 40 cm^2/sec. The average glottal area begins to increase about 20 ms before closure, reaches a peak of 30 mm^2 at release, and then decreases to 4 mm^2 about 50 ms after release. (b) Amplitudes of the acoustic sources. Note that voicing starts about 45 ms after release. (c) The intraoral pressure, Pm. (d) Spectrogram of the utterance.

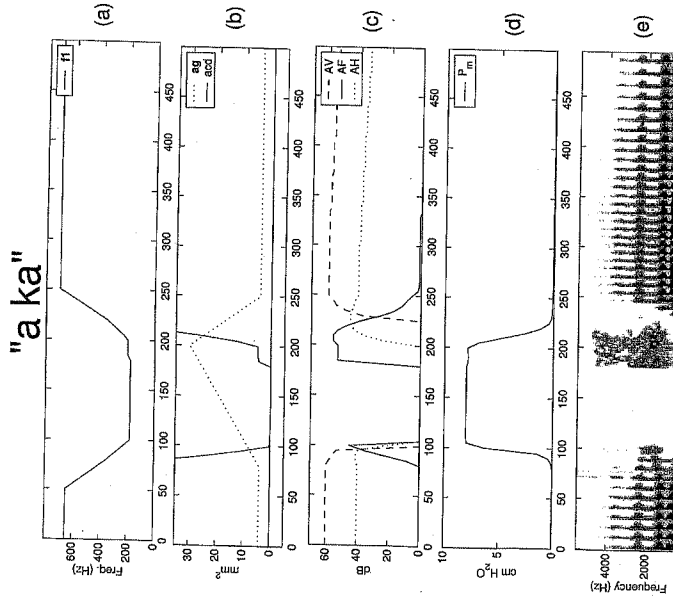

Fig. 13. Synthesized utterance "a ka". (a) The first natural frequency, **f1**, maps to the cross-sectional area at the dorsum. (b) At the release, the cross-sectional area at the dorsum, *acd*, is maintained at 10 mm^2 for about 15 ms, after which it is changed at a rate of 25 cm^2/sec. The average glottal area begins to increase about 20 ms before closure, reaches a peak of 30 mm^2 at release, and then decreases to 4 mm^2 about 45 ms after release. (c) Amplitudes of the acoustic sources. Voicing starts about 45 ms after release. (d) The intraoral pressure, Pm. (e) Spectrogram of the utterance.

5. Conclusion

In this work we have described how the properties of the vocal tract walls help to shape the acoustic characteristics of stop releases in English. Through synthesis experiments using a quasi-articulatory synthesizer, we found that inclusion of the vocal-tract wall effects on the articulator trajectory at the consonantal release is necessary to get stop bursts that are as strong in amplitude and long in duration as have been observed.

In addition, wall effects appear to be the source of the variation in VOT by place of articulation. Differences in articulator rates (due to differences in mass) are not sufficient. If it is assumed that for voiceless aspirated stops vocal-fold abduction is timed with the active release of the articulator, and not with the actual release, VOT variation by place can be explained by only one underlying process, rather than two as has been previously proposed.

References

Fujimura, O. (1961). "Bilabial stop and nasal consonants: A motion picture study and its acoustical implications," *J. Speech Hear. Res.*, 4, 233-247.

Ishizaka, K., J.C. French, and J.L. Flanagan (1975). "Direct determination of vocal tract wall impedance," *IEEE Trans. Acoust. Speech and Signal Proc.*, ASSP-23, 370-373.

Klatt, D.H. (1975). "Voice onset time, frication, and aspiration in word-initial consonant clusters," *J. Speech Hear. Res.*, 18, 686-706.

Lisker, L. and A.S. Abramson (1964). "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, 20, 527-565.

Maddieson, I. (1997). "Phonetic universals," in W.J. Hardcastle and J. Laver (eds.), *The Handbook of Phonetic Sciences*, Blackwell Pub., Malden, MA.

Perkell, J.S. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*, Research monograph No. 53, MIT Press, Cambridge MA.

Stevens, K.N. (1993). "Models for the production and acoustics of stop consonants," *Speech Commun.*, 13, 367-375.

Stevens, K.N. (1998). *Acoustic Phonetics*, The MIT Press, Cambridge, MA.

Stevens, K.N. (to appear). "The properties of the vocal-tract walls help to shape several phonetic distinctions in language," *Travaux Cercle Ling. Copenhague*.

Svirsky, M.A., K.N. Stevens, M.L. Matthies, J. Manzella, J.S. Perkell, and R. Wilhelms-Tricarico (1997). "Tongue surface displacement during obstruent stop consonants," *J. Acoust. Soc. Am.*, 102, 562-571.

Weismer, G. (1980) "Control of the voicing distinction for intervocalic stops and fricatives: Some data and theoretical considerations," *J. Phon.*, 8, 427-438.

Zue, V.W. (1976) "Acoustic characteristics of stop consonants: A controlled study," Tech. Rep. 523, MIT Lincoln Laboratory, Lexington, MA.