Voice Onset Time vs. Articulatory Modeling for Stop Consonants

Martin Rothenberg

Published in the journal Logopedics Phoniatrics Vocology, Vol. 34, 171-180 (2009). (This is a special issue for the Jan Gauffin Memorial Symposium.)

ABSTRACT

Voice Onset Time (VOT) was developed as a parameter for the pattern-playback speech synthesizer developed approximately 50 years ago, in order to generate the acoustic effects of voiced-unvoiced differences in English stop consonants. However, problems arose when the VOT parameter was used to define stops in actual spoken language, to replace aerodynamic and physiological parameters. A representative physiological model from the same time period that avoided these problems is sketched. In this model, the manner-of-articulation of a stop is determined by the duration, timing, and extent of laryngeal, articulatory and respiratory gestures. It is concluded that the term Voice Onset Time should be used only as a parameter in speech synthesis, as originally intended, and not for the analysis of actual speech.

Introduction

To understand the appearance of Voice Onset Time, or VOT, in phonetics research, one must go back to the development of the spectrogram, or time-varying power spectrum, and the so-called pattern-playback system for speech synthesis. The spectrogram was developed at Bell Labs in the late 1940s, as a display of the acoustic parameters of speech that is more intuitive than is the raw audio waveform. However, modern computer techniques for spectral analysis and synthesis were not available in the early 1950s, so a mechanical device for converting spectral information to an audio waveform, like the pattern-playback machine developed at the Haskins Laboratories, was quite useful (1).

In the pattern-playback machine, shown diagrammatically in Figure 1, a transparent belt transmitted light modulated sinusoidally at various frequencies by a tone wheel, and summed by a light collector to form an acoustic signal. If a spectrogram of actual speech was put on the belt, the speech was heard, albeit with some distortion produced by the system.

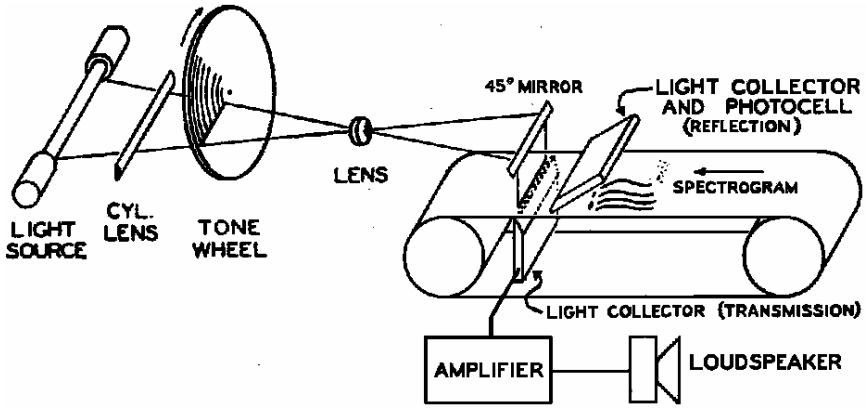

Figure 1. Sketch showing the operating principle of the Pattern-Playback device. (from Liberman, et al., 1952 (1))

An important aspect of the pattern-playback for perception research was that a highly simplified abstract representation of a spectrogram could be used, and perception of the resulting speech-like sounds studied. According to early researchers, the value of the pattern-playback was in the ability it gave the researcher to manipulate these abstract spectral patterns in order to study the perceptual importance of various acoustical features of the speech waveform, as, for example, the vowel formants.

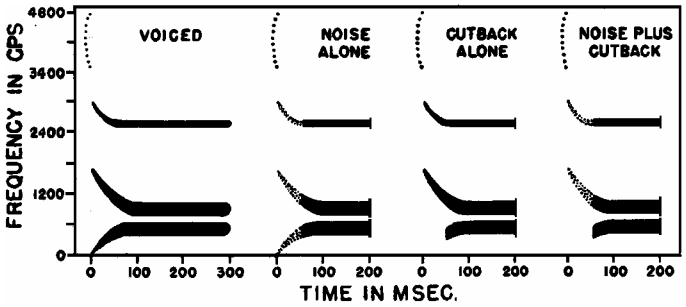

Figure 2. Patterns for a pattern-playback speech synthesizer from an early experiment (from Liberman, et al., 1958 (2))

Figure 2 shows a typical pattern used in an early experiment (2). Vowel formants were painted on the belt as bars. When consonants were to be abstracted, formant transitions were painted to simulate the spectral change as the articulators moved to and from the consonant articulation. Patterns of dots were used to make a more random sound to simulate unvoiced sound, and variations in the glottal source spectrum (not directly manipulable with the pattern-playback type of speech synthesis) were simulated by a taper in the width of the formant-bars in frequency regions that were acoustically weak, and even a cutoff of a formant bar altogether.

To put in context the type of research that could be performed with the pattern-playback system, consider the following diagram showing the links between articulatory (or physiological) models and perception.

Articulatory Model > Aerodynamic Parameters > Acoustic Parameters > Perception

While many linguists were doing research on articulatory modeling for stop consonants (see, for example, Smalley (3) and Ladefoged (4)) a number of researchers at Haskins Laboratories, including the authors in references 1 and 2, were starting at the other end of the chain – investigating the links between certain acoustic parameters and perception, using the pattern-playback system.

To help obtain differences in the perceived manner-of-articulation of a stop – as voiced vs. unvoiced – using the pattern-playback synthesizer, a delay was sometimes used between a brief wide-band noise (the simulated consonant release) and the onset of the periodic sound representing a following vowel. This was called the Voice Onset Time or VOT. Higher values of VOT could be used to synthesize an English unvoiced stop, and smaller values were found to be generally heard as a voiced stop.

When periodic excitation was to start before the release noise burst, to simulate voicing of the closure interval, the VOT was eventually allowed to have a negative value, and a tapered bar of the (negative) VOT duration was painted near zero frequency in the simulated closure interval.

In the decade following its development, numerous hours of perceptual testing were used to determine optimum strategies for setting ad hoc synthesizer parameters, such as VOT, and much interesting perceptual research was performed by Haskins researchers using these methods. However, problems arose when researchers at Haskins Labs began using the term VOT to pertain to measurements of actual speech.

Though the concept of VOT originated as a parameter in an acoustically based speech synthesizer, and not as a tool for linguistic analysis, that changed when Lisker and Abramson began using this new terminology for phonetic research, and in 1964 published their much cited paper (3) reporting VOT numbers obtained from spectrogram and waveform measurements in a number of languages.

Many VOT measurements of actual speech for linguistic analysis do have value. For unvoiced, aspirated stops, measurements of VOT from an acoustic waveform or spectrogram were useful, since the term VOT could be interpreted as a measure of the duration of the *aspiration interval*. In other words, it appeared that

for these stops, the Haskins Labs researchers were attempting to measure the aspiration interval duration, long acknowledged to be a significant variable in the study of speech, but using a new name for it.

However, the following problems arose when the concept of VOT was applied to actual human language.

Problem 1 - Disagreements arose as to where voice was to be judged to begin for an unvoiced stop.
Problem 2 - Patterns of voicing for voiced stops could not be specified by a single number representing a negative VOT, and it was not clear why one should want to do this.
Problem 3 - Aspirated stops in certain languages could not fit into the VOT model (as in Icelandic and Hindi).

An Alternative Model

To better put these problems in context, I outline a phonetic model for stop consonants that is an alternative to the acoustical model inherent in VOT measurements.

Alongside the VOT acoustic theory of stop consonant manner-of-articulation, and actually long preceding it in the work of Stetson and others (6), was a competing theory based on articulatory or physiological parameters, which specified a stop consonant primarily in terms of the presence or absence of a laryngeal abduction or adduction gesture – usually abductory – and the timing of such a gesture relative to the articulatory closing gesture. A version of this theory was described in detail in my 1966 doctoral dissertation The Breath-Stream Dynamics of Simple-Released Plosive Production, published as a monograph of the same name in 1968 (7). For brevity, I will refer to this publication as "Breath-Stream Dynamics". Breath-Stream Dynamics was an attempt to describe a universal articulatory, or physiological, model for the production of stop (or plosive) consonants. The model included the dynamic restrictions (on speed of movement and accuracy of coordination) inherent in the vocal mechanism. More specifically, the model in Breath-Stream Dynamics was restricted to released stops having only one point of articulatory closure, and that are followed by a vowel, which mirrors the domain of applicability of the VOT measure as used in speech synthesis experiments.

Breath-Stream Dynamics focused on stops having an articulatory release, and divided them into intervocalic stops and prevocalic stops. It was argued that, because of the physiological constraints in the speech mechanism, an intervocalic plosive in any language (or at least almost any language) could be characterized by putting it into one of the seven categories illustrated in Figure 3. The chart in Figure 3 is adapted from a similar chart in Breath-Stream Dynamics. The determining feature was the presence or absence of a glottal opening or abductory gesture, and, if there was one, by the relative timing of this glottal opening gesture and the articulatory closing gesture.

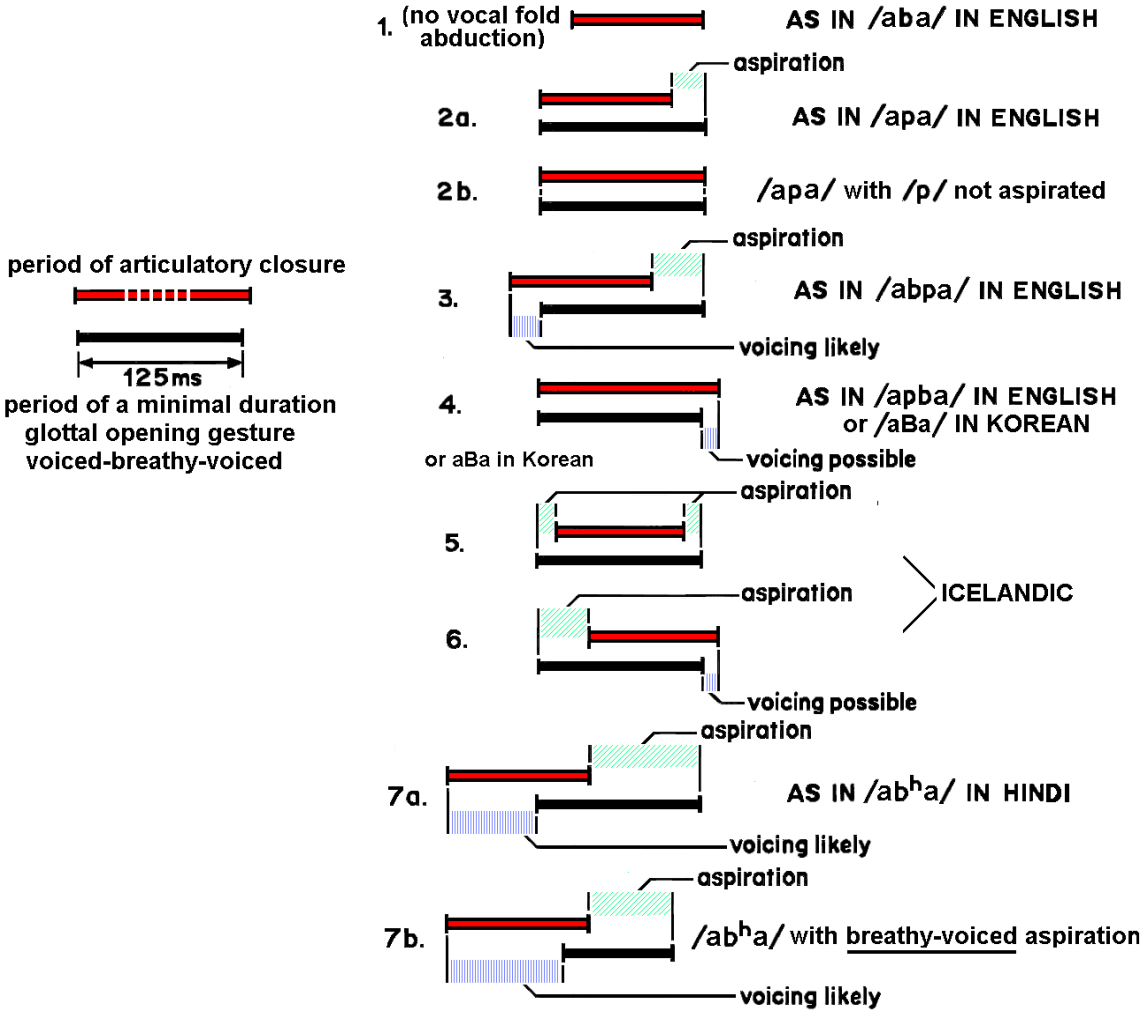

Figure 3. Possible patterns of interarticulatory timing for intervocalic stops. (adapted from Rothenberg, 1968 (7)) A question mark was placed after the word Icelandic in the original figure because definitive data were not available in 1968 as to which of the two timing patterns was preferred in the preaspirated stops of Icelandic. Löfqvist and Yoshida (8) show some subsequent data clarifying this issue.

A similar model was presented in Breath-Stream Dynamics for prevocalic stops, as a sentence-initial stop, or one following an unvoiced consonant. In a prevocalic stop, the vocal folds perform a movement from an open adjustment to a voiced adjustment. Since the laryngeal gesture is only one-directional in a prevocalic stop, the dynamic constraints are reduced from the intervocalic case, and the potential for precise interarticulatory coordination is greater. For this reason, it was argued that it is important to differentiate between prevocalic and intervocalic stops in a phonetic analysis. This differentiation is not conveniently made in VOT analysis, since VOT analysis is limited to only the adduction phase of the laryngeal gesture. But let us return to intervocalic stops.

In slow speech it was found that interarticulatory timing could be somewhat arbitrary, and depend on stress, dialect, control of breath in long sentences, etc. Therefore, to set a standard that could be compared between languages, it was proposed in Breath-Stream Dynamics that for a determination of stop-category (manner-of-articulation), the laryngeal gestures should be of minimal duration, as constrained by the dynamic limitations of the laryngeal musculature. In other words, measurements should be taken at a rate of speaking that is somewhat fast but still accurate. For intervocalic stops, a minimal duration movement from *voiced* to *breathy* to *voiced* was estimated in Breath-Stream Dynamics from various measurements to have a duration of roughly 125 ms. A gesture that only reached a condition of *breathy-voice* could have a duration of as little as half that figure.

The Three Problems Areas for VOT Use in Phonetic Analysis

First problem area

The first problem with VOT as a measure in natural language is the difficulty in defining VOT for even a simple aspirated voiceless stop, as illustrated in Figure 4. The figure shows the airflow pattern for a somewhat slowly spoken repetition of /apa/, as recorded at the mouth using a circumferentially vented pneumotachograph mask (lower trace), and also after a low pass filter that performed a very rough inverse filter function (upper trace).

Note: All airflow recordings presented in this paper were obtained using a circumferentially vented (CV) MA-1 mask and associated electronics from Glottal Enterprises, and filtered, displayed, and in some cases differentiated to obtain a representation of the acoustic waveform, using the Waveview software from the same company. In presenting images of the screen, as in Figure 4, extraneous screen elements have been removed, and explanatory labels are added in some cases.

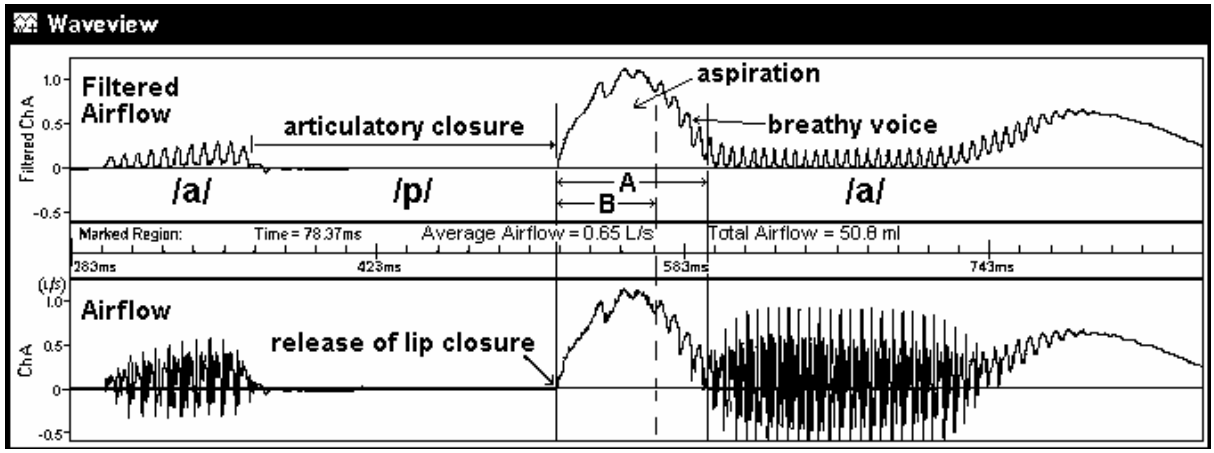

Figure 4. Airflow and approximately inverse filtered airflow for a production of /apa/ by an adult male English speaker.

In the rather typical production of the phoneme /p/ in Figure 4, what should be the value of VOT? Should it be the interval A, which includes the period of breathy-voice, or should it be it B, which excludes it, or a little less than B, to also exclude what appears to be a first post-release oscillatory movement of the vocal folds? The problem is that in actual speech, the transition from unvoiced aspiration to fully voiced speech is rarely abrupt, but is usually a gradual change in waveform and spectrum. In fact, it is likely that this spectral change during the transition (as well as the inverse change as the vocal folds separate) is the strongest perceptual factor in separating voiced from unvoiced homorganic stops.

Relevant to this measurement problem is the possibility that for instances of the same phoneme, depending on the speaker, voice effort, and numerous other factors, voicing can start at different points in the adduction phase, and progress at different rates. This means that intervals A and B in Figure 4 are not in some fixed ratio, and therefore different definitions can lead to widely different results.

This difference is illustrated in Figure 5, in which two extreme cases are shown for the measurement of VOT in an unvoiced aspirated stop. Both subjects were adult males speaking /apa/, but one subject was prelingually profoundly deaf and used bilateral cochlear implants. Though the airflow traces in both cases show rather normal and well-defined vocal fold abduction gestures and aspiration intervals, the measures of VOT as the time between the release (first cursor) and the onset of voice production (second cursor), as approximated by the author from the acoustic energy (differentiated airflow) for each signal, was widely different, because of the much longer breathy-voice transition in the top case. Thus, the use of an acoustically based VOT measure would point to a much different consonant articulation being used, while the true difference was one of voice quality.

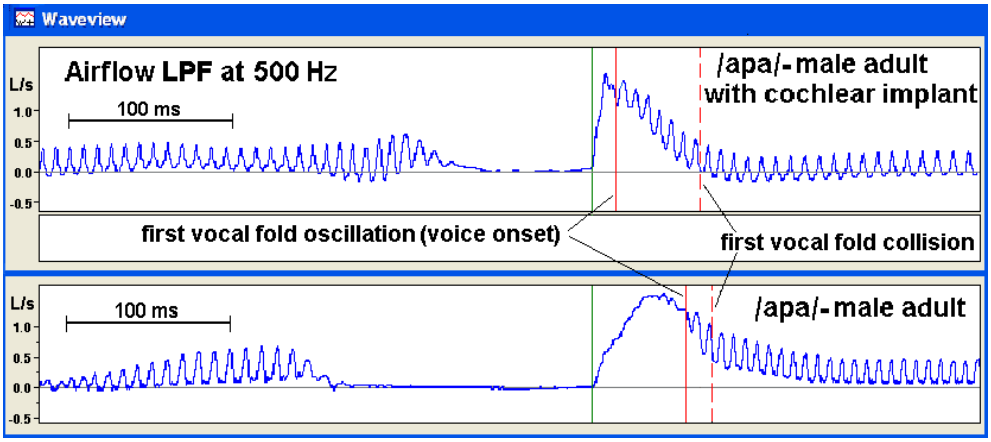

Figure 5. Two productions of /apa/ exhibiting extreme differences in a VOT measurement (intervals between the cursors) for similar durations of the aspiration interval. The two traces are aligned at their instants of articulatory release.

The dashed cursors in the figure show estimates of where the vocal folds first clearly come into contact in their oscillatory cycles. Acoustically, this is the point at which formant energy first appears strengthened. If these instants are used for the measure of "voice onset", as some researchers advocate, then the durations measured are close to what would be the durations of the intervals of aspiration. These instants also make a good place to mark the termination of the vocal fold adductory (closing) movement in an aspirated stop.

Second problem area

The second problem with VOT measurement in natural speech is related to measuring and interpreting negative VOT values. Negative values are even more problematic than positive values for two reasons. First, we note that to get closure voicing, one needs vocal folds adjusted for voice production and also a transglottal pressure sufficient to activate them. Thus, voicing during a closure interval can occur in various patterns, depending on the mechanism, if any, that maintains a sufficient transglottal pressure during the voicing interval. The various possible mechanisms for maintaining a transglottal pressure during a closure interval were described in Breath-Stream Dynamics, Chapter 7, and are as follows.

1. Passive supraglottal expansion
2. Active supraglottal expansion
2'. Active supraglottal expansion caused by the oral opening gesture.
3. Incomplete velopharyngeal closure.
4. An increasing subglottal pressure during the period of articulatory closure, as can occur after a breath pause, depending on the timing of the respiratory gesture needed for speech initiation.

Each mechanism in this list results in a different pattern of voicing, and it is easy to show that some of these patterns are not amenable to a VOT measure. For example, mechanism 1 generally results in voicing of the beginning of the closure interval, while mechanism 2 and 2' generally result in voicing of the end of the closure interval.

However, there is a more fundamental problem with considering the value of a negative VOT as characterizing a linguistic category. A particular instance of a stop may be categorized as voiced in some languages even in the absence of a mechanism for maintaining voicing. The linguistically determining factor would be that the vocal folds were adjusted for voice during the articulatory closure, even if there was not sufficient transglottal pressure to activate them. In such cases, one may ask, how is the stop heard as 'voiced', if there is little or no voicing during the articulatory closure? The answer is that it can be recognized as voiced by the absence of a breathy-voice transition immediately before and/or after the closure interval. This is illustrated in Figure 6, in an airflow record for the sentence "Baba baked apples.", as spoken by a male adult English speaker.

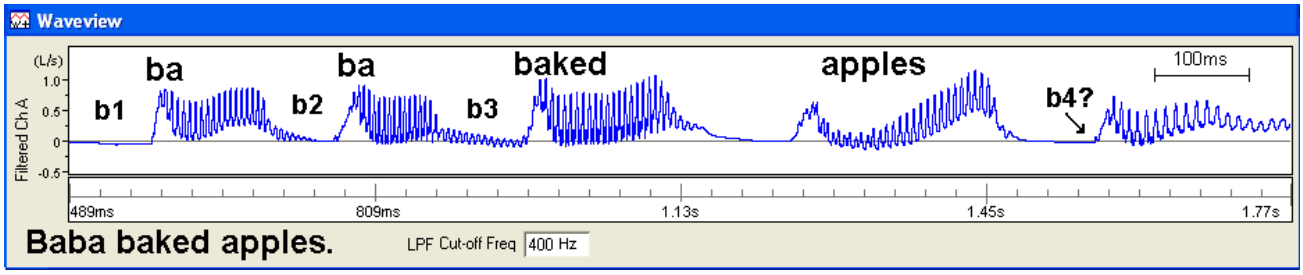

Figure 6. Airflow for the sentence "Baba baked apples." spoken by an adult male English speaker, approximately inverse filtered by low pass filtering at 400 Hz.

Illustrating the general problem with using negative values of VOT to characterize voiced stops, it may be noted that neither b1, b2, nor b3 could be readily identified with a value of negative VOT. Stop b1 in the sentence, in particular, is linguistically a /b/ and heard as a /b/, but shows no voicing before the release, since the transglottal pressure was not sufficient to initiate voicing until after the release.

The stop in "apple", labeled b4 with a question mark, also shows no voicing. The question mark indicates that though it is produced in English with a geminated /pb/ articulation – number 4 in the list of interarticulatory timing patterns in Figure 3 – the phonology of English labels the consonant as an unaspirated /p/. Other similar examples of the occurrence of stops in English that are phonologically categorized as voiced but shows no closure voicing can be seen in the Figure 9 below.

The inadequacy of VOT as characterizing voiced stops in actual language is also shown in Figure 7, which presents an example of an overemphasis on the presence of voicing (vocal fold vibratory behavior) during the closure of a nominally voiced stop.

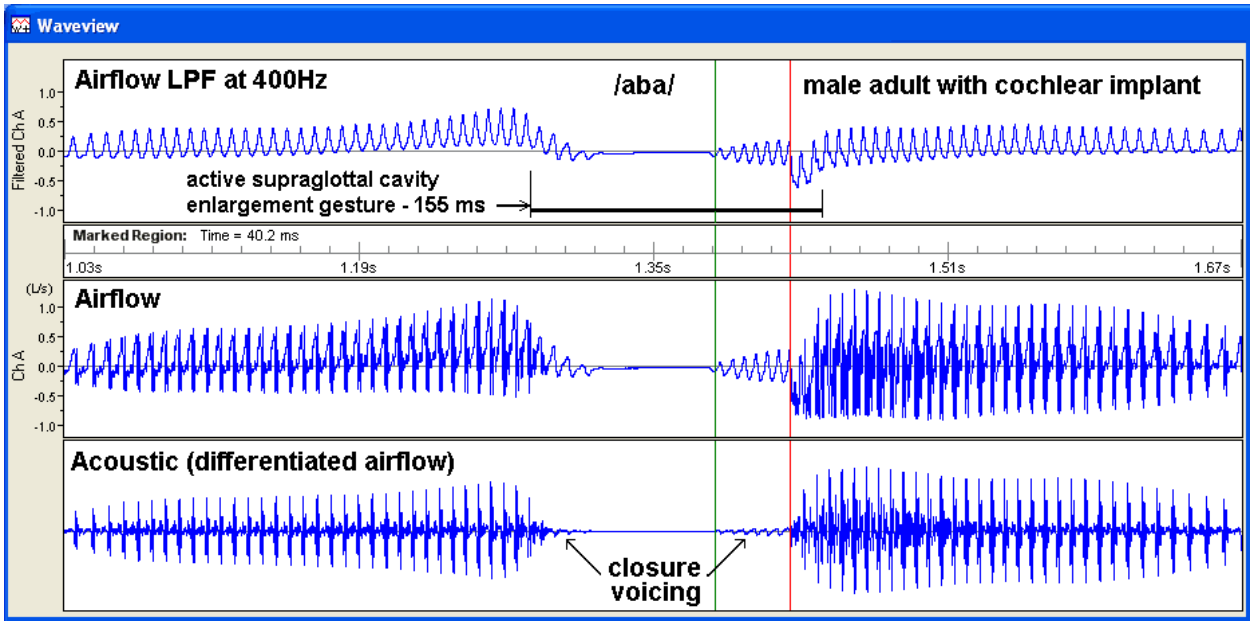

Figure 7. Airflow and acoustic waveforms for an attempted production of /aba/ by a prelingually deaf adult using a cochlear implant. Vertical cursors indicate the voice onset time that would be estimated from the acoustic waveform.

In extensive speech training, this prelingually deaf subject has been taught, perhaps incorrectly, that the English consonant /b/ must be produced with voicing during the occlusion, and not just with a voiced adjustment of the vocal folds that may or may not result in voicing. Therefore he used the mechanism of a strong active expansion of the supraglottal cavity – mechanism 2 in the list of voicing mechanisms above – to augment voicing, and consequently produced an "implosive" articulation, as described in 1935 by Hudgins

and Stetson (9). The evidence of an implosive is the negative airflow that immediately follows the release. There is also a slightly negative airflow at the onset of the closure, marking the beginning of the gesture of active supraglottal cavity enlargement.

The incorrectness of this English articulation is not apparent in the acoustic record at the bottom, and a VOT measure from the acoustic signal would merely show an acceptable negative value for voice onset (delineated by the cursors) which is in a range that could be found in a normal production of an English /b/.

Implosive variations of [b] are found in languages other than English and would be acceptable in such languages. They would also be expected with other deaf speakers receiving this type of training.

Third problem area

The third problem area in applying VOT to linguistic analysis is that it could not fit certain stop categories in a number of languages, with the examples of Icelandic and Hindi mentioned. As illustrated in patterns 5 and 6 in the articulatory coordination model of Figure 3, some Icelandic stops have pre-aspiration, which is not accounted for in the VOT model. In addition, it is well documented that Indic languages often include voiced-aspirated stops that have no "voice onset time", per se (7b in Figure 3). In contrast, parameters of the articulatory coordination model do allow a description of languages in both categories.

The Role of Perception in Phonetic Theory

The discussion to this point has emphasized the importance in phonetic models for stop consonants of the static and dynamic limitations on laryngeal and articulatory gestures, with respiratory and velar movements also being important. However, a universal phonetic model would not be complete without also considering human perceptual limitations, since two different articulatory timing patterns cannot represent different phonetic categories unless they can be heard as different.

In general, the fact that two articulatory patterns contrast in at least one natural language can be taken as proof that they are readily differentiable auditorally, and this test is usually sufficient. However there are cases in which the speech mechanism is physically capable of making a difference in articulation that is not perceptible, and therefore this difference cannot be used contrastively in a language. One such case in English is the different timing patterns possible for an unaspirated prevocalic stop that follows an unvoiced constricted or "pressure" consonant, such as /s/ at the beginning of a word, as in "spa". An example of this is the pronunciation of /a spa/ is shown in Figure 8.

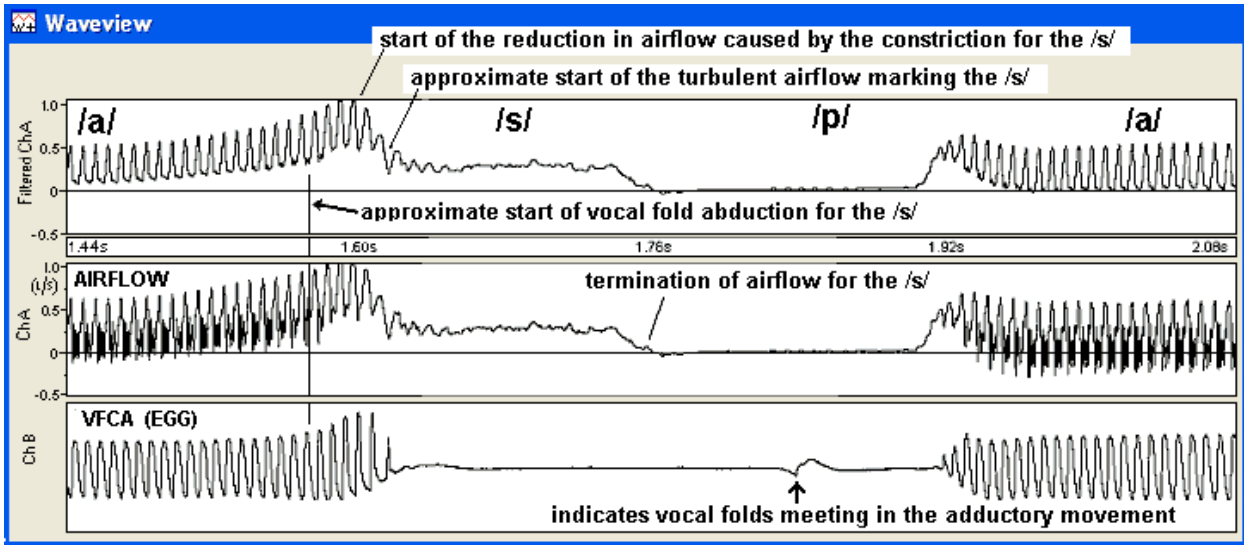

Figure 8. Simultaneous recordings of (bottom to top) an electroglottograph (vocal fold contact area or VFCA), wideband airflow, and airflow approximately inverse filtered, for a production of "a spa" by an adult male English speaker.

In this figure, an electroglottograph (Glottal Enterprises model EG-2) was used to determine noninvasively the time during the period of articulatory closure at which the vocal folds came together for a voiced adjustment. This instant is identified clearly by a perturbation in the EGG trace, as marked by the arrow in the figure. Since the oral pressure needed for the /s/ was high at the time of articulatory closure, the transglottal pressure was small during the closure, and therefore there was no voicing produced when the vocal folds came together during the closure. Voicing began only just after the release of the articulatory closure, when the supraglottal pressure goes to near zero, and the transglottal pressure rises to approximate the subglottal pressure.

It is interesting that the vocal folds approximate for voice early in the closure interval and not timed to the articulatory release, as might be expected in an unaspirated /p/. This early glottal adduction makes the production appear more like that of an unvoiced /b/. In view of this phonetic ambiguity, the question to be considered is; how can /a spa/ and /a sba/ be differentiated perceptually? The answer is that, absent some other differentiating features such as duration, pitch, etc., they cannot be differentiated. The assignment of a /p/ phoneme to the stop in “a spa” or in “a spot” cannot be made on strictly phonetic or perceptual grounds, since the stops are phonetically ambiguous as to manner-of-articulation.

But it is important to note that this phonetic ambiguity is not a problem in English, since “a spa” will never contrast with “a sba” and “a spot” will never contrast with “a sbot” and “the sport” will not contrast with “the sbort” and “the star” will not contrast with “the sdar”, etc. (If a loan word or artificially constructed noun were to be spelled “sdar” and need to be contrasted with “star”, the /d/ would be probably pronounced with a supraglottal cavity expansion (either passive or active) that is not common in English, to produce pre-release voicing that would make the differentiation possible. Likewise, in a language that had that differentiation as part of the language, there would be expected to be a similar expansion. Also possible, but less likely, would be a slight nasalization of all or part of the period of articulatory closure.

As noted above, there are examples in English of phonemically voiced stops (b, d, g) that may show no overt voicing because of the lack of transglottal pressure. One example was shown in the sentence-initial stop /b/ in Figure 6. In addition, in English phonology, when there is a word juncture separating the unvoiced pressure consonant and the unaspirated stop, the stop is labeled as voiced, even though no closure voicing may be present. Thus, phoneme sequences represented in English orthography by “us ba“ and ”a spa“ would have a different phoneme for the stop, even though the patterns of laryngeal-articulatory coordination are similar. This is illustrated in Figure 9 for the sentence “I’ll stop by for a hotdog in the background.”

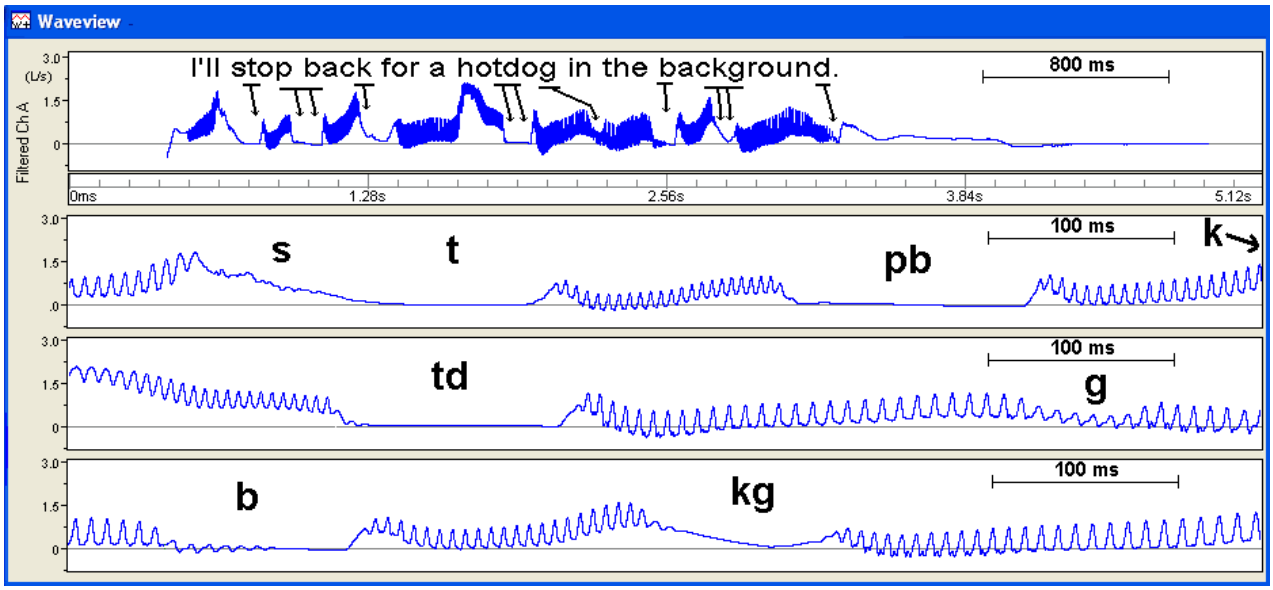

Figure 9. Wideband airflow, approximately inverse filtered by low-pass filtering at 400 Hz, for a sentence constructed to contain a number of examples of voiced stops exhibiting no closure voicing. Speaker is an adult male.

In this figure, the voiced stops (b, d, and g, in the order in which they occur) preceded by a pressure consonant (p, t, and k, in the order in which they occur) show no voicing in the closure interval, with the post-release airflow pattern for the /t/ in the /st/ of “stop” similar to the pattern for the homorganic /d/ in “hotdog”. Whether or not the adductory movement consistently occurred early in the closure interval in each case, as in the example in Figure 8, or if there were consistent differences in adduction time between phonemically voiced and unvoiced stops for this speaker, would be a subject for future research, presumably using an electroglottograph for this determination.

The intervocalic voiced stops /g/ and /b/ in the figure show typical closure voicing patterns, however, relating to the theme of this paper, it can be noted that neither can be represented by a negative value of VOT.

Some Conclusions

In summary, I would like to suggest that:

1. The term Voice Onset Time should be used only as a speech synthesis parameter, as originally intended, and not for the phonetic analysis of actual speech.

2. Measurements of aspiration interval duration during actual speech should be identified as such, and not as a ‘voice onset time’, even if the measurement is made from an acoustic record. The instant at which the vocal folds clearly come into contact during the vibratory cycle, whether measured from airflow, acoustic waveform or EGG signal, is suggested as a standard for marking the end of the aspiration interval or the vocal fold adduction phase.

3. Phonetic analysis of stop consonants should be based on a physiological or articulatory model, whether it is the one discussed here or some other alternative framework.

4. In the phonetic analysis of released stops, intervocalic and prevocalic stops should be treated differently because of their different dynamic constraints.

5. In measuring closure voicing, one should have a physiological/aerodynamic model in mind and a question to be answered within that model. If this is done, I believe that it will be found that the magnitude of negative VOT values, per se, have little meaning for real speech.

6. If at all feasible, wide-band airflow measurements should be used to judge articulatory coordination, rather than acoustic records. Airflow data are closer to physiology than are acoustic data, while also displaying the acoustic data. EGG signals provide another noninvasive technique for monitoring vocal fold abduction and adduction.

REFERENCES

1. Liberman A.M., Delattre PC, Cooper FS. The role of selected stimulus variables in the perception of the unvoiced stop consonants. American J. of Psychology 1952; 65: 497-516.

2. Liberman AM, Delattre PC, Cooper FS. Some cues for the distinction between voiced and voiceless stops in initial position. Language and Speech 1958; 1: 153-67.

3. Smalley WA, Manual of Articulatory Phonetics. Tarrytown, New York: Practical Anthropology, 1961, Rev. 1963.

4. Ladefoged P, A Phonetic Study of West African Languages. Cambridge: Cambridge University Press: 1964.

5. Lisker L, Abramson AS. A cross-language study of voicing in initial stops: acoustical measurements. Word 1964; 20: 384-422.

6. Stetson RH. Motor Phonetics, 2nd ed. Amsterdam: North-Holland Publishing Company: 1951.

7. Rothenberg M. The Breath Stream Dynamics of Simple Released Plosive Production. Vol. 6 Bibliotheca Phonetica. Basel: Karger, 1968.

8. Löfqvist A, Yoshida H, "Laryngeal activity in Icelandic obstruent production," Nordic J. Linguistics 4: 1-18, 1981.

9. Hudgins CV, Stetson RH. Voicing of consonants by depression of the larynx. Arch. Neerl. Phon. Exper. 1935; 11: 1-28.