REVISITING PLACE OF ARTICULATION MEASURES FOR STOP CONSONANTS: IMPLICATIONS FOR MODELS OF CONSONANT PRODUCTION

Kenneth N. Stevens[*], Sharon Y. Manuel[*†,] and Melanie Matthies[*‡]
[*]*Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge MA,*
[†]*Northeastern University, Boston MA,* [‡]*Boston University, Boston MA*

ABSTRACT

This paper reports on revised analysis procedures that are being developed to identify place of articulation for stop consonants. These procedures are being used to examine the sources of variability that exist in the acoustic characteristics of these consonants for different phonetic contexts, speakers, and speaking styles. There are three motivations for this work: (1) to improve models of articulatory-to-acoustic mapping; (2) to develop procedures and data that can be used in knowledge-based automatic speech recognition and in speech synthesis; and (3) to provide a baseline of the normal range of variation of these consonants, against which clinicians can evaluate disordered production. The data show that combining a variety of acoustic properties from the bursts and transitions for stop consonants leads to robust classification of place of articulation, particularly if the gender of the speaker and if the feature [back] in the adjacent vowel is known.

1. INTRODUCTION

There is a long history of research that seeks to determine the acoustic cues for place of articulation for stop consonants. This research includes early perceptual studies [1], examination of short-time spectra and their time variation at consonantal release [3, 8, 6], and detailed quantification of locus equations [7]. Stimulated by the original work of Fant [2], researchers have developed quantitative models of the production of these consonants [5]. These models attempt to account for the acoustic events that occur when consonant releases or closures are formed with the lips, the tongue blade, or the tongue body. The models also provide some direction as to how the positioning and movement of these primary articulators and of other related articulators are influenced by the context in which the consonants occur. The coarticulation aspects of the models are less quantitative than the ones that specify events arising from the primary closure and release, and one must rely on descriptive data to quantify these aspects. In general, however, the models have directed the selection of analysis procedures so that they extract, from the signal, attributes that relate directly to the articulator that is making the closure or release, and therefore identifies the place of articulation.

2. MODELS OF STOP AND NASAL CONSONANT PRODUCTION

Since they are produced with different vocal-tract shapes, labials, alveolars, and velars are expected to show different time variation of the formant frequencies in the vowel in the vicinity of the consonant closure or release. As is well known, the adjacent vowel (or consonant) influences the vocal tract shape for the consonant and therefore modifies the formant frequencies in the immediate vicinity of the consonant. Locus equations provide one way of describing these context effects [7]. As the consonant release or closure is formed, there are two components to the movement of the articulators and the formants, particularly for labials and alveolars: one is a rapid movement of the lips or the tongue blade in a time period of 10-20 milliseconds, and the other is a slower movement of the tongue body and mandible toward the following vowel in the case of a release, and from the preceding vowel for a closure [4]. For a velar, the tongue body and mandible movement is responsible for the entire transition. A consequence of these observations is that the time course of the formant movements, particularly the durations of the movements, is dependent on consonant place and on the adjacent vowel. There is a trend for the first-formant transitions to be faster or more abrupt for labials and to be slowest for velars, with alveolars in between.

At the release of stop consonants, a brief transient and burst of frication noise is generated. The noise component is due to turbulence in the rapid airflow immediately following the release. Because of the different rates of release of the different articulators, this noise burst tends to be shortest for labials, longest for velars, and intermediate for alveolars. Both the transient and the frication noise source are filtered by the vocal tract, and this filtering is influenced primarily by the cavity anterior to the constriction. There is no cavity for labials, and a cavity of length 1-2 cm for alveolars. For velars, there is a cavity of length about 3-7 cm, depending on whether the following vowel is a front vowel or a back vowel. Thus, the spectrum of the radiated sound has either no major spectral prominence for a labial, a spectral prominence that is usually in the frequency range 3500 to 7000 Hz for an alveolar, and a lower frequency prominence for a velar.

In summary, a number of acoustic measurements can be made to distinguish different places of articulation for stop consonants. Some of these measurements are more reliable, or show less variability, than others. However, we take the view that all of this acoustic information can be used by a listener to identify place of articulation. Depending on the speaker, the phonetic environment, and the properties of the speaker-listener channel, these different cues may assume different degrees of importance.

3. THE DATABASE AND THE MEASUREMENTS

In the present study we report some acoustic measurements for consonants contained in a database of sentences produced by four speakers (two female, two male). A variety of measurements were made, guided by the above review of models of stop consonant production. We report here examples of the data. The data will be reported in two forms: (1) average data for particular consonants for the female and male speakers, and (2) data showing the variability across speakers and across phonetic environments.

The database from which the utterances were drawn consists of a list of 100 read sentences, with an average length of seven words. Syllable-initial consonants in various vowel environments were selected for measurement. Syllable-initial consonants were defined to be either word-initial consonants or, if they were word-internal, they were prestressed or they were the final consonant in a cluster. The word-initial consonants were often in reduced syllables. Thus initial stop consonants in sequences like /pr/ or final stop consonants in sequences like /lp/ were not studied. About 15 tokens of each initial consonant type were measured, each produced by four talkers, as noted above.

For all consonants, the time of consonant release was identified. Measurements of formant frequencies were made by hand, pitch-synchronously. (An alternative would be to use averaging of power spectra over 15-odd ms to avoid identifying pitch periods.) A 6.4-millisecond Hamming window was centered over the first part of the glottal cycle, and the peaks of the discrete Fourier transform were measured. The first three formant frequencies were measured at two and sometimes three points near the closure or release, depending on the length of the vowel. The first glottal cycle where a measurement was made was the first complete cycle following the release, as judged from the waveform and the spectrum. The other two measurement points were on glottal cycles closest to 20 ms and 40 ms from the first measured cycle. (We report here only the measurements at the 0 and 20 ms points.)

The spectrum of the combined transient and burst was measured using an averaging technique. A power spectrum was obtained with a 6.4-ms window, and 15 such spectra sampled at 1-ms intervals were averaged. When the time from burst release to voicing onset was sufficiently long (greater than about 10 ms), then the 15 spectra were sampled beginning 7.5 ms before the release and extending to 7.5 ms after the release. If the time from burst to voicing onset was less than 10 ms, the rightmost spectrum in the 15 that were averaged was located so that the 6.4 ms window did not encroach on voicing onset. For comparison with the burst, a similar averaging method was used to obtain the spectrum of the following vowel over a 15-ms interval centered 20 ms after onset of voicing. The use of this averaging technique for the burst was selected because it produced a relatively smooth spectrum. The approximately 300-Hz bandwidth for the 6.4-ms window is roughly the auditory critical bandwidth in much of the frequency range where important attributes of the burst spectra are measured.

In this paper we will report a subset of the data. Three descriptors were extracted from the formant measurements near the release. These were: (1) F1o, the frequency of F1 in the first glottal period in the vowel; (2) F2o, the same measurement for F2; and (3) F2, the amount and direction of movement of F2 from the point where F2o was measured to the nearest glottal period that was 20 ms following that point. Acoustic theory shows that F1o is expected to be highest for labials, which have a fast F1 movement, and lowest for velars, for which the F1 movement is slowest. F2o tends to be lower for labials than for alveolars and velars. For back vowel environments, F2o is higher for alveolars than for velars, and the reverse is true for front vowel contexts. The measure F2 is expected to be positive for labials and is more negative for velars than for alveolars. Three descriptors were used to characterize the spectrum shape of the burst and its relation to the following vowel. These descriptors were (4) Ahi-A23 the peak spectrum amplitude of the burst in the frequency range above 3500 Hz (3000 Hz for male talkers) relative to the average of spectral peaks in the F2 and F3 range (in dB); (5) Av-Ahi, the spectrum amplitude of the F1 prominence of the vowel relative to Ahi; and (6) Av-Amax23, the difference between Av and the maximum spectrum amplitude in the F2 and F3 range. Descriptor (4) is a measure of spectrum tilt, and is expected to be higher for alveolars than for labials and velars. Measure (5) gives the high- frequency burst amplitude relative to the vowel amplitude, and is expected to be least for alveolars and greatest for labials. Measure (6) is expected to be least for velars and greatest for labials, because bursts for velars tend to have a mid-frequency spectral prominence.

4. SOME RESULTS

Average results are summarized in section 4.1 for the bursts and transitions for voiced and voiceless stops in initial position. A more detailed analysis of the combined burst and transition data for initial voiced stops is given in section 4.2.

	p	b	t	d	k	g
Av-Ahi	41	45	25	31	31	33
Ahi-A23	-9	-10	+2	-2	-8	-6
Av-maxA23	28	31	25	27	18	22

Table 1. Average data from measurements on bursts. Values in dB. See text for definition of parameters.

4.1 Bursts for initial stops

Average values of the measures for stop consonant bursts are summarized in Table 1. The data for Ahi-A23, which is a measure of the spectral difference between high and mid frequencies, show that the labial burst spectrum dips down by about 10 dB, whereas the high-frequency spectrum amplitude is about equal to the mid-frequency spectrum amplitude for alveolars. The tilt for the velars is in between. These results for labials and alveolars are in general agreement with expectations from models, but there are not good quantitative models for the average spectrum for velars at high frequencies [5]. The measure Av-Ahi shows that the labial burst at high frequencies is about 15 dB weaker than the alveolar burst, again in accordance with expectations from calculations with models.

The parameter Av-maxA23 shows a smaller value for velars than for labials and alveolars, reflecting the spectral prominence for velars in the F2 or F3 range. It is noted that there is a 2-6 dB greater high-frequency amplitude (relative to the vowel) for the burst for voiceless stops than for voiced stops, presumably reflecting the greater intraoral pressure and greater airflow at the release.

4.2. Voiced stops: Combining transitions and bursts

Analysis of the data for stop and nasal consonants shows that there is no single measure on bursts or formant transitions that reliably separates the consonants into the three groups --- labial alveolar, and labial. Figure 1 shows the average values (and standard deviation) across all speakers for all six of the parameters for each of the voiced stop consonant classes.

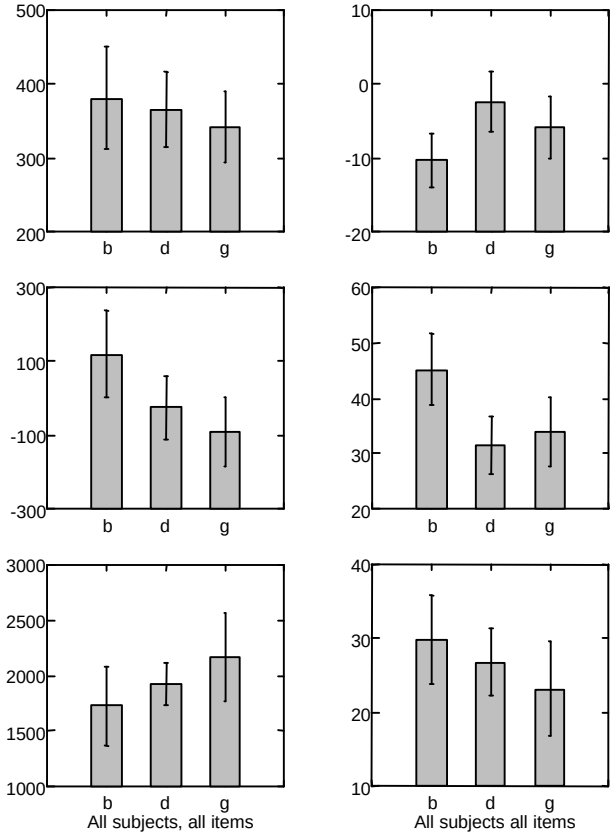

Figure 1. Average data from formant starting frequencies and F2 transition (left panels) and from the burst (right panels) for syllable-initial voiced stop consonants in sentences. Data from four speakers (two male, two female). The acoustic measures are described in the text. The vertical lines give standard deviations.

While the parameter F1o shows some differences, on average, in the expected direction, the variability is large. Less variability of this parameter would be observed if F1o were examined separately for high and low vowel contexts, i.e., if the high-low features for the vowel context were taken into account. Likewise, the measure F2 would separate the places of articulation more effectively if vowel backness were separated into two categories. The F2 onset frequency (F2o) is lowest for /b/ and highest for /g/, with /d/ in between. The variability across vowel contexts is least for /d/, and, as expected, the vowel backness feature introduces considerable variation in F2o for /b/ and /g/.

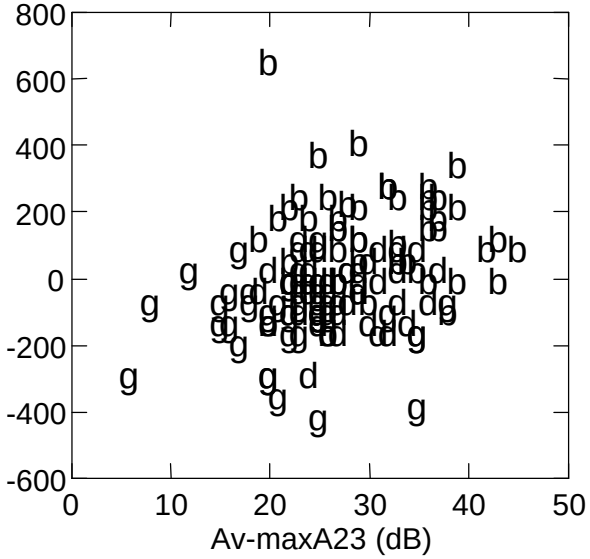

Figure 2. Scatterplot of data from individual utterances of voiced stop consonants. The measures are the same as those in Fig.1.

The tilt parameter Ahi-A23 distinguishes labials from alveolars, but velars overlap with both places of articulation. Of the two parameters that measure spectral aspects of the burst in relation to the vowel, Av-Ahi separates labials (with weak high-frequency amplitude) from alveolars and velars. Since the mid-frequency amplitude for the /g/ burst has a major peak, Av-maxA23 is lowest for /g/.

Not unexpectedly, the variability with vowel context and with gender is sufficiently large that any one of these parameters alone is not effective in classifying the consonants. Figure 2 shows a scatterplot of the data for two of the measures --- F2 and Av-maxA23. Since both of these measures shows some separation between the three consonant classes, one might expect that the combination would be more effective. The points in Fig.2 still exhibit considerable overlap, particularly in the middle.

Classification is even more robust if information is available concerning the gender of the speaker and the vowel context. It is also noted that some of the stop consonants in this study were word-internal and intervocalic, so that acoustic information at the time of consonant closure is available as well as the burst and formant transitions at the time of release. Thus the presence or absence of a preconsonantal vowel is also a factor in evaluating the classification of consonants based only on the release.

	Male	Female	All speakers

All vowels	87	88	85
Front vowels	96	96	92
Front initial	97	100	94

Table 2. Percent correct classification of consonant place based on discriminant analyses using 4 descriptors, for different genders and vowel environments. Front initial means word-initial consonants before front vowels.

A summary of the percent correct classification based on discriminant analysis with the four most effective descriptors is given in Table 2, with several combinations of factors. Percent correct classification for all speakers and for vowels in all contexts is 85 percent. At the other extreme, if one assumes that the gender of the speaker is known and if the consonant is word-initial, classification is correct for 97 to 100 percent of the tokens.

5. DISCUSSION AND CONCLUSIONS

(1) Examination of models of speech production can be used to guide the selection of acoustic cues for the classification of place of articulation. The places in the speech signal where these acoustic properties are to be measured must be carefully specified. Once an acoustic landmark such as a consonant release has been identified, a fine-grained analysis in the vicinity of this landmark must be made, with a precision of 2-3 ms in locating the measurement point. Although this analysis was done by hand in this study, automation of extraction of the properties should be relatively straightforward. A frame-based analysis at five-or ten-millisecond intervals, however, will not capture the required acoustic attributes.

(2) No single simple acoustic property is sufficient to identify place of articulation of consonants. A combination of properties is necessary, and these properties must be combined judiciously. The simple discriminant analysis used here is just a first step in developing appropriate methods for combining properties.

(3) The properties used in this study are just a subset of the properties that could potentially be used. For example, the third formant transition is clearly important, particularly for velars and for front vowels, the formant transitions beyond the first 20 milliseconds can also provide useful information, and burst duration is known to be place-dependent. Furthermore, if the consonant is preceded by a vowel, information about consonant place is present in that vowel, and it should be combined with information at the release. Classification scores like those in Table 2 can certainly be improved by using more care in the selection of descriptors.

(4) Nevertheless, it is apparent that adjustments need to be made for gender and for some gross features of the adjacent vowel, particularly the feature [back].

(5) Data like those in Table 1 and in Fig. 1 delineate the acoustic properties of stop consonants and their normal range of variation. The availability of acoustic data of this kind is important in the clinical evaluation of speech disorders. Knowledge of the normal range of particular acoustic parameters can help in determining whether the utterances of a speaker are outside this range, and can aid the diagnosis and quantitative assessment of the efficacy of intervention.

ACKNOWLEDGMENTS

This research was supported in part by Grants DC00075, DC01925, and DC02978 from the National Institutes of Health.

REFERENCES

[1] Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J. and Gerstman, L.J. 1952. Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, 24, 597-606.

[2] Fant, G. 1960. *Acoustic theory of speech production*. The Hague, Mouton.

[3] Halle, M., Hughes, G.W. and Radley, J.-P.A. 1957. Acoustic properties of stop consonants. *Journal of the Acoustical Society of America*, 29, 107-116.

[4] Manuel, S.Y. and K.N. Stevens 1995. Formant transitions: Teasing apart consonant and vowel contributions. In Proceedings *of ICPhS 95*, Stockholm, Vol. 4, 436-439.

[5] Stevens, K.N. 1998. *Acoustic phonetics*. Cambridge MA: MIT Press.

[6] Stevens, K.N. and Blumstein, S.E. 1978. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64, 1358-1368.

[7] Sussman, H.M., McCaffrey, H.A. and Matthews, S.A. 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90, 1309-1325.

[8] Zue, V.W. 1976. Acoustic characteristics of stop consonants: A controlled study. ScD. Thesis, Massachusetts Institute of Technology.