Proceedings of The Institute of Acoustics

ACOUSTIC VARIABILITY IN VELAR STOP CONSONANTS

Sarah Hawkins

Department of Linguistics, University of Cambridge

In the search for invariant acoustic correlates for place of articulation of stop consonants, velar stops have provided a greater challenge than bilabials, alveolars, or dentals. At the moment of release of closure, the range of frequencies of F2, F3, and F4 is usually between two and four times as great for velars in various vocalic contexts as for bilabials and dentals in the same contexts (6). One reason for this great acoustic variability is the relatively large articulatory variability permissable for velars, expecially in languages, like English, which lack other stops made with the tongue body. In English and many other languages, "velar" stops may be quite palatal in the environment of front vowels, whereas they are articulated much further back in the context of back vowels. In consequence, the length of the front cavity, whose resonances are those primarily excited at the release of closure, varies a great deal for velar stops. The data do not support a simple division of velars into those in front versus those in back vocalic environments, however. Although a fronted articulation produces a shorter front cavity with correspondingly higher resonances, there is still a relatively wide range of spectral patterns for these consonants (6).

Despite this attested variability, velar stops are often distinguishable from other places of articulation by the presence of a compact spectral prominence in the mid-frequency range of 800-4000 Hz. A compact spectral prominence is a distinct peak that is alone within that frequency range. This primary characteristic of compactness is accompanied by secondary characteristics that can distinguish velars from other places of stop articulation. Kewley-Port (8), summarizing her own and others' data, lists a longer burst (20-30 ms for velars vs. 5-10 ms for bilabials and alveolars), with a relatively long VOT, and a relatively unchanging spectrum during the burst due to the slower release of velars. Despite these dynamically-specified spectral characteristics, as a class velar stops are less well-identified than bilabials or alveolars from visual inspection of spectral displays (8).

This paper describes a preliminary study of the acoustic characteristics of syllable-initial velar stops. Two questions were asked. First, can we synthesize an acceptable velar stop simply by manipulating the spectrum of the burst alone, and if so what are the critical acoustic properties of such bursts? And second, to what extent are the properties of these velar bursts observable in naturally-spoken syllables? The focus of interest in both synthetic and natural speech is the release burst and the first few periods of the vowel.

SYNTHETIC CV SYLLABLES

The synthetic CVs described here comprise the stimuli for a perceptual experiment reported alsewhere (7), in which we asked whether spectral properties of the stop burst alone can be identified that lead to an invariant perceptual response. Our approach was to construct an acoustic continuum of bursts such that, when followed by minimal vowels, we hear velar stops at one end, and either bilabials or alveolars at the other end, depending on the continuum. This we did primarily

by changing the spectral balance - that is, by changing the amplitudes of noise-excited formants - in parallel synthesis.

Figure 1 shows short-time dft spectra of bursts at the extremes of the two continua, the velar-alveolar in the right panel, and the velar-bilabial in the left. In the velar-alveolar (gd) set of bursts, only F2 and F5 were excited. The difference between the two spectra shown lies solely in the amplitude of F2, which was some 30 dB higher in the spectrum labelled /g/, resulting in a compact mid-frequency prominence as found in velars. The lack of F2 excitation in the spectrum labelled /d/ results in the diffuse, rising spectrum typical of alveolars (4,8). For the velar-bilabial (gb) set, the changes were rather more complex: in the spectrum labelled /g/, only F2 and F5 were excited, as before, but in the spectrum labelled /b/, these formants were excited at a lower level and the amplitudes of F3 and F4 were raised to produce a falling spectrum.

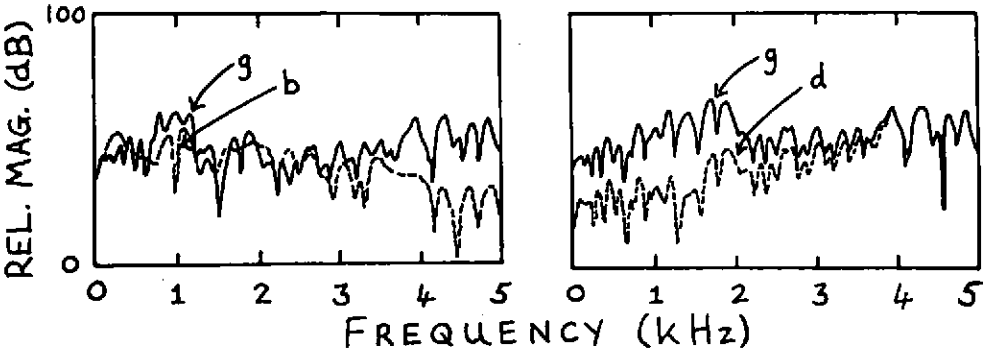

Fig. 1. Short-time dft spectra of the bursts at the extremes of the continua.

Stimuli intermediate between these extreme pairs were made by changing formant amplitudes in equal acoustic steps, resulting in two continua of ten stimuli each. Formant frequencies were the same in all stimuli in one continuum, but, to make the transition to bilabial or alveolar percepts possible, F2 and F3 were different in the two continua. In the velar-bilabial continuum, F1, F2, F3 and F4 were 500 Hz, 1000 Hz, 2000 Hz, and 3250 Hz respectively; in the velar-alveolar continuum, F2 and F3 were 1562 Hz and 2430 Hz respectively. In addition, burst duration in the velar-bilabial continuum decreased by 1 ms with each stimulus, from 15 ms for the most velar to 5 ms for the most bilabial burst.

In one experimental condition, a short (145 ms) vowel followed each burst. Formant frequencies were constant throughout the vowel, except for a 250 Hz rise in F1 over the first 20 ms, and the formants had the same frequencies as in the bursts, so there were no transitions other than in F1. Listeners identified the initial stop as /g/ or /b/, or as /g/ or /d/, as appropriate for the burst of the particular continuum. In another condition, 45 ms aspiration noise followed the burst, replacing some of the vowel; listeners identified these stimuli as /k/ or /p/, or as /k/ or /t/. In a third condition, the isolated bursts were

categorized in an ABX paradigm in which A and B were the two extreme stimuli of one continuum.

Forced-choice categorization functions for each continuum had a reasonably sharp crossover between 100% velar and 100% nonvelar responses for 9 listeners, but 4 other listeners heard only alveolars in the velar-alveolar continuum. For the isolated bursts, categorization functions were similar to the larger group's identification functions. Apparently most listeners can classify sounds by place of articulation using only the burst spectrum, perhaps relative to certain properties of the vowel spectrum.

The burst and the first two periods of the vowel of each stimulus underwent 3 types of spectral analysis: dfts on linear and Bark frequency scales (with and without preemphasis), and an analysis using the peripheral auditory system model of Bladon and Lindblom (2), henceforth the BL model. This model applies to the signal nonlinear frequency and amplitude transformations thought to occur in the human ear. The final output is a loudness-density spectrum in sones/Bark versus Bark frequency units. There are a number of auditory filter models available, each with its strengths, and some have been applied to stops with some success (5,9). Patterson's model can disambiguate spectra of velars that lack clearly classifiable conventional spectra (8). The BL model was chosen for this study mainly because it has had some success in vowel normalization between the sexes (3), and Kewley-Port (8) reported particular difficulty in classification of velar stops produced by her female talker. Although the BL model performs poorly in normalizing stop bursts (1), it has not to my knowledge been used systematically to distinguish stop consonants differing in place of articulation.

Visual inspection of the standard linear and Bark frequency spectra of the stimuli shows the expected changes in spectral shape of the burst as we move through each continuum, but the changes are complex and hard to describe. By far the most interesting spectra are those from the BL model. These show only one principle qualitative change as we move through the continuum, and this change takes place rather abruptly at about the stimulus corresponding to the 50% crossover in the identification functions of the listening experiment. Figure 2 shows the burst spectra of the 10 stimuli of the velar-bilabial continuum. People commenting on the shapes tend to agree that the first five form one group and stimuli 6-10 form another group. In the listening experiment, the 50% crossover for all three conditions was about 5.6. The burst spectra of the velar-alveolar continuum are shown in Figure 3. People tend to place stimuli 1-3 in one group, and stimuli 6-10 in another, with 4 and 5 forming an intermediate group that resembles 6-10 more than 1-3. For this continuum, 50% crossovers ranged from 3.77 to 5.13 depending on the experimental condition.

The two groups of bursts in the velar-bilabial continuum (Fig. 2) differ in the relative prominence of the low- and mid-frequency spectral peaks. In the first five bursts, these two peaks are similar in sone level and bandwidth, or (in 2 and 3) the mid-frequency peak dominates. In contrast, the low-frequency spectral prominence dominates the spectra of stimuli 6-10. In the velar-alveolar bursts (Fig. 3), the difference between the two groups lies largely in degree of spectral tilt towards the high frequencies. The first three stimuli each have spectral peaks and troughs across the entire spectrum, but there is no impression of a general spectral tilt, whereas stimuli 4-10 all have the upwardly tilting spectrum characteristic of an alveolar burst (4,8).

ACOUSTIC VARIABILITY IN VELAR STOP CONSONANTS

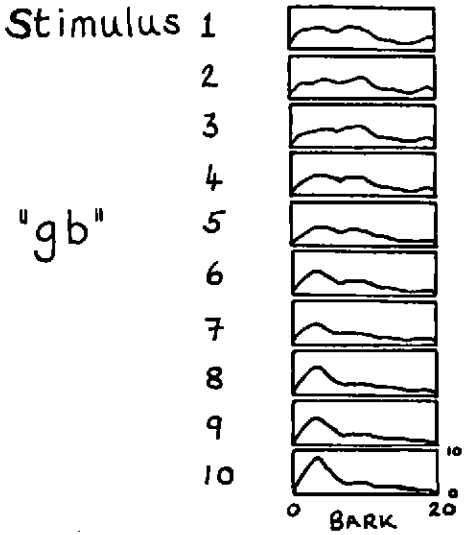

Fig. 2. "BL" spectra of gb bursts

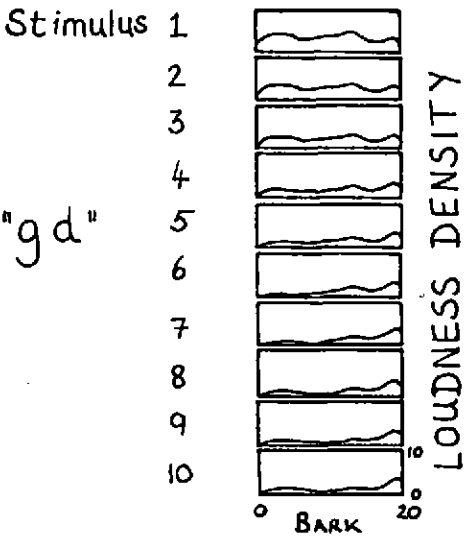

Fig. 3. "BL" spectra of gd bursts

The output of the BL auditory filter model is encouragingly congruent with the results of the listening tests. What does this congruence mean for the perceptual correlates of compact consonants? Although each continuum can be divided into two using criteria of visual similarity, the five "velar" spectra in the gb continuum are not strikingly similar to the three "velar" spectra in the gd continuum. Each does have a mid-frequency prominence and no spectral tilt, but the mid-frequency prominence is not always the most (visually) salient part of spectra, especially in the gd continuum. Note that a minority of listeners heard only alveolars and no velars in the gd continuum, however. Current experiments are examining ways to strengthen the velar percept, primarily by increasing the amplitude of the mid-frequency prominence and decreasing amplitudes at high frequencies, thereby changing spectral shape, spectral tilt, and amplitude relationships between the burst and the following vowel. Detailed analyses of the experimental results (7) encourage the speculation that when the burst is immediately followed by the vowel, as with voiced stops, the vowel spectrum is used as a reference or anchor point, and the burst is interpreted with reference to the vowel, especially by those listeners who heard no velars in the original gd continuum. If the vowel is absent or is separated from the burst by aspiration, then the overall spectral shape of the burst may become more important.

NATURALLY-SPOKEN CV SYLLABLES

As a first step in relating the synthetic bursts to natural speech, I have examined the spectra of CV syllables (/g/ followed by a vowel) spoken by seven talkers. Utterances from two of these talkers have been analysed in greater detail, with two aims: first, to try to identify spectral and temporal properties of the burst or vowel onset that are common to all the syllables, and

ACOUSTIC VARIABILITY IN VELAR STOP CONSONANTS

second, to compare the success of different types of acoustic analysis in showing these critical properties. These two talkers, one female and one male, both of whom are native speakers of British English, said the syllables /gi/, /ge/, /ga/ /gu/ twice each. The syllables were digitized at 10 kHz and the following types of spectra were made of the burst and at least the first two periods of the vowel, in successive 5 ms steps and, for some syllables, in 1 ms steps: 256-point dft spectra (smoothed by padding with zeros to 512 points, and not so smoothed); 128-point, 14-pole lpc spectra, and the output of the BL model. Hamming windows were used in all cases, and the dft spectra were displayed on both linear and Bark frequency scales.

Although many of the spectra conform closely to the classical compact description, a large proportion deviate from this description, as other investigators have also found (e.g. 8). However, as shown below, most of these "deviant" utterances possess some of the classical properties of /g/ at least some of the time. One point to note is that whereas Kewley-Port (8) reported more nonconformity in spectra from her female talker than from her two male talkers, on the whole the male talker's spectra in this study departed from the classical norm more than the female's. There were also, however, considerable variations in degree of compactness between different utterances of the same syllable by a single talker. In the case of velar stops, the loss of precision experienced in dealing with high-pitched voices appears to be offset by inherent variability in production, as well as by differences between talkers.

The departure from classical compactness appears to depend partly upon phonetic context, as does the distribution of secondary cues. The classical spectral characteristics occurred more reliably in /ga/ than in the other syllables, /gi/, /ge/, and /gu/, particularly for the female talker. On the other hand, late onset of F1, identified as a secondary feature of velar stops (8), appeared more frequently in the context of these high and front vowels. F1 onsets as late as 20-30 ms after the release burst were by no means consistently present, however; the conditions determining their appearance need more study. In perception, late vowel onsets can determine perceived place of articulation in synthetic syllables lacking bursts (10), and our experiments indicate that perception of bilabials requires a short burst relative to that required for velars. However, satisfactory velar and alveolar bursts can be synthesized with identical durations, given appropriate spectral shapes. We have yet to manipulate burst duration with a constant spectral shape appropriate for velars, and this area requires further work.

Two characteristics occurred in several of the naturally-spoken syllables of this study, although neither was identified as criterial of velars by Kewley-Port (8). One is the presence of a very sharp, narrow-bandwidth formant in the mid-frequencies at the vowel onset, often appearing rather abruptly and in relative isolation from surrounding peaks, which are of much lower amplitude. Like late onset of F1, this sharply-prominent formant appears to be associated more with /g/ before high or front vowels, and possibly with weak bursts. Figure 4 shows examples of the male talker's /ge/ and the female's /gi/. The second characteristic is the occasional fluctuation in amplitude of spectral peaks and valleys during the burst, such that a compact spectrum appears intermittently. An example is shown in Figure 5, for the male talker's /gi/. The female talker's /gi/ (Fig. 4) also shows some fluctuation, although it is in fact clearer in the BL spectra than in the lpc spectra shown here. Such

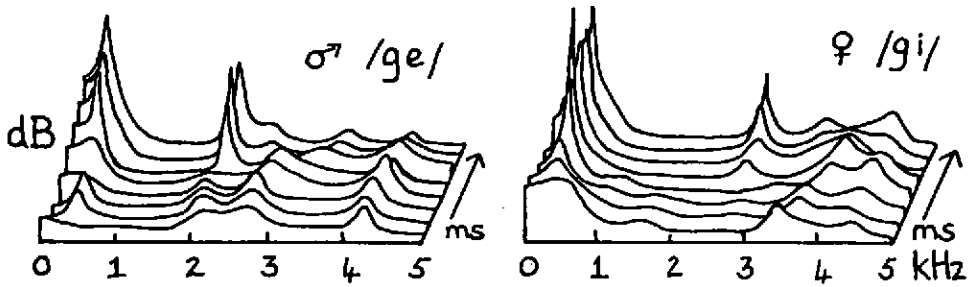

Fig. 4. LPC spectra at successive 5 ms intervals, illustrating a compact spectral prominence in the vowel onset.

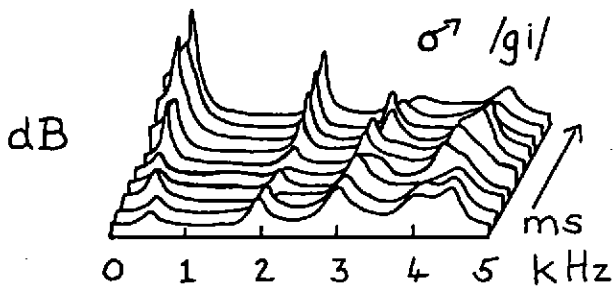

Fig. 5. LPC spectra at successive 5 ms intervals, illustrating fluctuations in the degree of compactness of the burst.

fluctuating spectra will occur with repeated bursts, which are more common in velars than other stops, and which may in themselves serve as a cue to velarity. Many of the observed fluctuations occurred within a single burst rather than between repeated bursts, however. A rarer variation on this fluctuating burst is one in which the classical compact spectral shape appears gradually as the burst proceeds, so the spectrum is more clearly compact just before the vowel onset than at the release of the closure.

These two phenomena together, a strongly compact mid-frequency peak at vowel onset, and a spectrum that fluctuates or increases in compactness during the burst, may each serve to enhance the compact percept. A fluctuating or changing spectrum may "draw attention" to itself by virtue of the rapid changes, and thereby emphasize or compensate for what in a more static spectrum may be only a weakly compact burst. Compactness in the vowel onset may override any ambiguities of the burst, consistent with what we know about the role of formant transitions. It may also contribute to the anchoring effect of the vowel on the perceptual interpretation of the burst spectrum, as hypothesized above from the results of the perceptual experiment.

Finally, we compare the success of the various types of spectral analysis in portraying the common properties of velar stops. For both male and female

voices, the BL and lpc spectra were generally preferable to the dft spectra, whether the latter had linear or Bark frequency scales. In many cases the picture was only clearer, rather than qualitatively different, and this clarity probably results mainly from greater visual simplicity, which may not be relevant to the auditory cortex. There were a number of cases, however, in which syllables which appeared very noncompact in the dft spectra appeared quite compact in the BL and lpc spectra. It may be worthwhile smoothing the dfts to enhance their visual simplicity, but it is not clear that even smoothed dfts will be preferable to lpc or auditory filter models.

In most cases there was little to choose between the lpc and BL spectra, but, consistent with Kewley-Port's observations (8), auditory filtering can enhance the compact appearance of some spectra, notably where lpc analysis produces two or more mid-frequency peaks instead of a single one, and in /gi/, in which the "compact" spectral prominence is often high in frequency, very broad, and without great amplitude. In both these cases, auditory filtering tends to produce a single, comparatively well-defined compact spectral prominence. Reducing the number of poles in the lpc analysis may have similar effects, but has yet to be examined systematically. On the other hand, there were a significant number of syllables whose bursts appeared more compact in the lpc rather than the BL spectra. I am unable as yet to identify consistent principles governing the membership of this group, which included tokens from both talkers and all vowels examined except /i/. This group did not include the most classically compact bursts, but these, not surprisingly, appeared compact in at least the BL and lpc spectra, and sometimes in all the analyses. Expanding the rather restricted sample of this study may shed some light on this issue.

CONCLUSIONS

The feature compact appears to depend primarily on the presence of a relatively isolated mid-frequency spectral prominence. This acoustic property is not present throughout the burst of every velar stop, but a compact spectral prominence can usually be seen at some point(s) during the burst or the first few periods of the following vowel. Spectral compactness is more clearly seen with some acoustic analyses: the BL auditory filter and lpc models are probably preferable to dfts, but more work is needed to identify the most successful acoustic analyses. Velars can be synthesized in CV syllables with no transitions other than in F1, by varying only characteristics of the burst - its spectral shape, and sometimes its duration. Given appropriate duration, a compact stop is perceived when the mid-frequency spectral prominence is sufficiently well-defined in relation to the spectrum of the following vowel onset. However, just as individual talkers differ in the degree of classical compactness of their velars, so listeners appear to differ in the properties they expect in order to perceive a velar. Future work will continue the attempt to identify the critical acoustic properties of velar stops, particularly emphasising the shape of the burst spectrum in relation to the spectrum of the following vowel onset.

ACKNOWLEDGEMENTS

I am grateful to Mike Allerhand, Anthony Bladon, and David Deterding for making their programs and expertise available to me, and to Ken Stevens for collaborating on the perceptual experiment.

REFERENCES

(1) R.A.W. Bladon, 'Using auditory models for speaker normalization in speech recognition', Canadian Acoustical Association: Proceedings of the Montreal Symposium on Speech Recognition, 16-17, (1986).

(2) R.A.W. Bladon and B. Lindblom, 'Modeling the judgment of vowel quality differences', J.A.S.A., Vol. 69, no.5, 1414-1422, (1981).

(3) R.A.W. Bladon, C.G. Henton, and J.B. Pickering, 'Outline of an auditory theory of speaker normalization', in M.P.R. Van den Broecke & A. Cohen (eds.) Proceedings of the Tenth International Congress of Phonetic Sciences, Dordrecht: Foris Publications, 313-317, (1984).

(4) S.E. Blumstein and K.N. Stevens, 'Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants', J.A.S.A., Vol. 66, 1001-1017, (1979).

(5) B. Delgutte, 'Analysis of French stop consonants using a model of the peripheral auditory system', in J.S. Perkell & D.H. Klatt (eds.) Invariance and Variability in Speech Processes, Hillsdale, NJ: Lawrence Erlbaum Associates, 163-177, (1986).

(6) G. Fant, 'Stops in CV syllables', in G. Fant (ed.) Speech Sounds and Features, Cambridge, MA: MIT Press, 110-139, (1973).

(7) S. Hawkins and K.N. Stevens, 'Perceptual basis for the compact-diffuse distinction for consonants', J.A.S.A., Vol. 80, Suppl. 1, Abstract, (1986).

(8) D. Kewley-Port, 'Time-varying features as correlates of place of articulation in stop consonants', J.A.S.A., Vol. 73, no.1, 322-335, (1983).

(9) R.D. Patterson, 'Auditory filter shapes derived with noise stimuli', J.A.S.A., Vol. 59, 640-654, (1976).

(10) J.R. Sawusch and D.B. Pisoni, 'On the identification of place and voicing features in synthetic stop consonants', J.Phon., Vol. 2, 181-194, (1974).