*J. Linguistics* 22 (1986), 145–166. Printed in Great Britain

# On the naturalness of stop consonant voicing

JOHN R. WESTBURY

*Department of Orthodontics, Dental Research Center, University of North Carolina*

AND PATRICIA A. KEATING

*Phonetics Laboratory, Department of Linguistics, University of California*

(Received 1 October 1985)

## 1. Introduction

A long recognized problem for linguistic theory has been to explain why certain sounds, sound oppositions, and sound sequences are statistically preferred over others among languages of the world. The formal theory of markedness, developed by Trubetzkoy and Jakobson in the early 1930's, and extended by Chomsky and Halle (1968), represents an attempt to deal with this problem. It is at least implicit in that theory that sounds are rare when (and because) they are marked, and common when (and because) they are not. Whether sounds are marked or unmarked depends – in the latter version of the theory, particularly – upon the 'intrinsic content' of acoustic and articulatory features which define them. There was, however, no substantive attempt among early proponents of the theory to show what it was about the content of particular features and feature combinations that caused them to be marked, and others not.

In the last ten to fifteen years, the theory of markedness has been supplemented with – some might argue, supplanted by – an increasingly popular notion of linguistic naturalness, developed and discussed to varying degrees by Ohala (1974, 1983), Hooper (1976), Stampe (1979), Vennemann (1972), Schachter (1969), Schane (1972), Crothers (1978), and Lindblom (1978, 1983; cf., also Liljencrants & Lindblom, 1972). Those in concert with this notion maintain that the sounds, sound systems, and sound sequences that are most common among the world's languages are those that are most natural – natural because they are somehow easiest to articulate or perceive; because they represent physical constraints inherent to speech producing and perceiving systems; or because they otherwise represent optimal tradeoffs between competing demands of perception and articulation.

Unfortunately, 'explanations' of typological generalizations based on naturalness have, as a rule, been no more satisfying than those based on markedness. This is particularly true in the phonological literature. The claim, for example, that speakers are more readily disposed to produce some sounds

and sound sequences than others can be meaningful only if we know specifically how that is so. Thus, the notion of naturalness effectively presupposes well-developed models which specify (1) limiting properties of the production and perceiving mechanisms, thereby defining possible speech behaviours, and (2) general principles which prioritize that range of behaviours. However, among those whose work relates most directly to the notion – with the exceptions of Lindblom and Ohala – development of suitable models has been notably absent.

In this paper, we consider the general question of whether it is more 'natural' for stop consonants to be voiced or voiceless. According to the myoelastic-aerodynamic theory of phonation (van den Berg, 1958), the vocal folds will oscillate only when there exists an adequate pressure drop and airflow across them. During stops, no air exits the mouth or nose, so that this condition is not obviously met. That observation suggests that voiced stops might be more difficult to produce, and thereby less 'natural' than their voiceless counterparts (Ohala, 1983). And yet, a great many languages have voiced stops, at least in some phonetic environments. Indeed, if voiced stops are generally hard to produce, why do they exist at all? Why are they so prevalent? Why aren't they generally unstable, both synchronically and diachronically? Simple questions such as these readily lead to an examination of the actual means by which voicing might be produced and maintained during a stop. In this paper we present a systematic approach to this question based on the use of an explicit model of the articulatory mechanism to simulate the likely effects on voicing of a variety of articulatory conditions.

## 2. The model

### 2.1. *Basic approach*

To a first approximation, the vocal tract consists of two soft-walled cavities, the lungs and mouth. They are separated from each other by a constriction formed by the vocal folds, and separated from the atmosphere by constrictions at the velopharyngeal port and/or mouth opening. Over the course of an utterance, the volumes of both cavities, the dimensions of various constrictions, and the mechanical properties of the vocal tract walls and vocal folds themselves may be controlled voluntarily and independently, thereby producing the familiar low-frequency variations in air pressures and flows characteristic of speech.

Essential aspects of this approximation are represented in terms of the equivalent network (adapted from Rothenberg, 1968) shown in Figure 1, wherein voltage and current can be considered analogous to the acoustic quantities pressure and volume velocity. The elements of the network and their articulatory interpretation are as follows: the voltage source $E_s$ represents the net (inspiratory or expiratory) pressure generated by the respiratory musculature; $C_s$ and $C_o$ represent the respective acoustic compliances of air





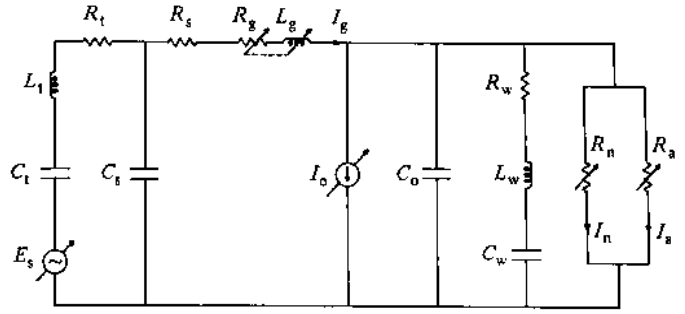


*Figure 1*
A circuit representation of the breath-stream control mechanism. Elements of the model are defined in the text.

volumes below and above the glottis; $R_t$, $L_t$, $C_t$, and $R_w$, $L_w$, $C_w$ represent the lumped viscous-mass-compliant character of walls surrounding the subglottal and supraglottal cavities; the resistances $R_g$, $R_n$, and $R_a$ represent viscous and turbulent losses generated by flows through potentially time-variable constrictions at the glottis, velopharyngeal orifice, and mouth opening, while $R_s$ represents some nominal loss generated by airflow through the trachea and bronchi; $L_g$ represents the reactive part of the glottal impedance; and finally, the current source $I_o$ represents muscularly actuated rate of change of volume of the supraglottal cavity.

It is possible to write a set of differential equations whose solutions will describe the response of this physical system to variations over time in its control elements. The forcing functions or control inputs for the system are time functions which specify values for each control element – e.g. the dimensions and thereby cross-sectional areas of various articulatory constrictions. These control inputs are derived from physiological interpretations of the familiar row-by-column feature matrix representation of an utterance. The outputs or responses of the system are also time functions which describe the consequences of any particular set of control inputs and initial conditions in terms of air pressures and flows which can be observed at various points along the vocal tract. Relevant details of our implementation of such a model, developed earlier by Rothenberg (1968), are available elsewhere (Westbury, 1983; Keating, 1984). Other models incorporating the same general approach toward understanding the breath-stream dynamics of speech, though differing in their complexity and implementation, have been described by Muller and Brown (1980), Flanagan *et al.* (1975), Ohala (1976), and Scully (1969).

### 2.2 *Assumptions made in modelling voicing*

Subsequent sections of this paper describe expectations which can be developed by using such a model to investigate when and to what extent stop

voicing is likely to occur. These expectations depend foremost upon two major assumptions. The first of these is that voicing will occur whenever the states of the glottis and vocal folds are suitable for voicing, and there exists a sufficient pressure drop between the trachea and pharynx. The model does not include any direct representation of vocal cords *per se*: the glottal opening alone is represented. In effect, we assume the aforementioned condition on the glottal state to hold whenever (constant) cross-sectional area of the glottal slit is a fair approximation of the average glottal area during a vocalic period. We also assume particular pressure drops across the glottis, which depend upon vocal fold tension, to be necessary for voicing initiation and maintenance, respectively. The model is then used to determine when voicing will occur by calculating when the pressure drops across the glottis exceed those thresholds. In all experimental cases to be presented, discussions of 'voicing' are to be understood to mean a sufficient pressure drop for voicing.

The second major assumption influencing expectations derived from the model is that the acoustic and physiological realization of an utterance depends heavily upon a well-defined interpretation of the notion ease of articulation. An utterance which consists underlyingly of a serially-ordered string of STATES, each with its own defining properties, must be input to the model as a set of control functions that vary over TIME. Each such control function may itself be segmented into a string of steady states and transitions. We define the easiest string of adjacent states to produce – and thereby, the most 'natural', from an articulatory point of view – to be the one in which the velocities of articulatory transitions, in each and all control functions, are least.

This characterization of ease of articulation is undoubtedly too simple to be of general use in articulatory models. A more general characterization might consider intrinsic properties of the various articulators themselves – e.g. their mass and compliance – as well as the 'levels' of states which must be sustained, and between which changes occur. An articulation metric which assigns cost only on the basis of the rates at which state changes occur would consider all states held indefinitely long, or all state changes of the same velocity, to be equally easy. However, the simple characterization provided above is sufficient for our interest in determining what will likely happen in certain of the easier 'utterances' – namely, utterances wherein very few states change between segments.

In spirit, this characterization of EASE OF ARTICULATION is not new (cf. Ladefoged, 1982, for example). The advantage of such a characterization, of course, is that it can be used within the context of a suitably explicit description of the articulatory mechanism, and of the control functions which drive it, to determine a continuum of articulatory ease which associates cost with particular sequences of speech sounds. A second important aspect of this characterization is that it emphasizes the role which phonetic context must





play in determining whether a sound is 'easy' or 'hard' to produce. Sounds are generally not produced in isolation in natural languages. Rather, any given sound is customarily bounded, at least to one side, by other sounds whose properties are certain – and are always shown – to influence its own acoustic and articulatory manifestation. It is not implausible to assume that the degree of difficulty in producing a particular acoustic or articulatory state will depend as much upon the difference between that state and temporally-adjacent ones as upon the inherent difficulty in maintaining that state. For that reason the 'ease' of stop consonant voicing can only be ascertained by considering stops in a variety of phonetic contexts.

## 3. EXPERIMENTS ON POSITION IN UTTERANCE

### 3.1. *The medial case*

Consider now the following problem: Is it more likely for a stop to be voiced or voiceless in an articulatorily 'simple' vowel+stop+vowel string, where the initial and final vowels are identical, and the stop is chosen so that its articulation between the flanking vowels involves as few changes in control parameters as possible? In order to use the breath-stream dynamics model to calculate whether and how long the conditions sufficient for voicing might exist during such a string, appropriate input functions are specified which, in effect, fix as constant throughout the string all but one of the model elements subject to voluntary control – namely, the oral constriction.

These input functions represent simple, stylized generalizations about data reported elsewhere in the literature, and incorporate the following descriptions:

Specifically, the tissues surrounding the lungs are considered stretched, so that subglottal pressure derives entirely from their elastic recoil, allowing the thoracic musculature to be quiescent. Moreover, there are no muscularly induced changes in supraglottal volume, or in the mechanical properties of tissues surrounding the lungs and mouth. Additionally, the velopharyngeal port remains fully occluded. Finally, the vocal folds are appropriately and constantly adducted and tensed for voicing. Over the entire VCV string, only cross-sectional area of the mouth opening is varied, as it must be, first to produce a constriction in the mouth and eventually, some time later, to release it.

Under conditions such as these, pressures above and below the glottis ($P_m$ and $P_s$, respectively) can be expected to change with time as shown in Figure 2. There seems to be some consensus that the vocal folds will continue vibrating as long as the pressure drop across them is greater than roughly 2000 dyne/cm$^2$ (Ladefoged, 1964; Ishizaka & Matsudaira, 1972; Lindqvist, 1972; Baer, 1975). Note from this figure that the difference between $P_s$ and $P_m$, though decreasing, is clearly greater than that amount for the first 60-odd ms of a hypothetical 80 ms closure interval. Thus, voicing would be expected

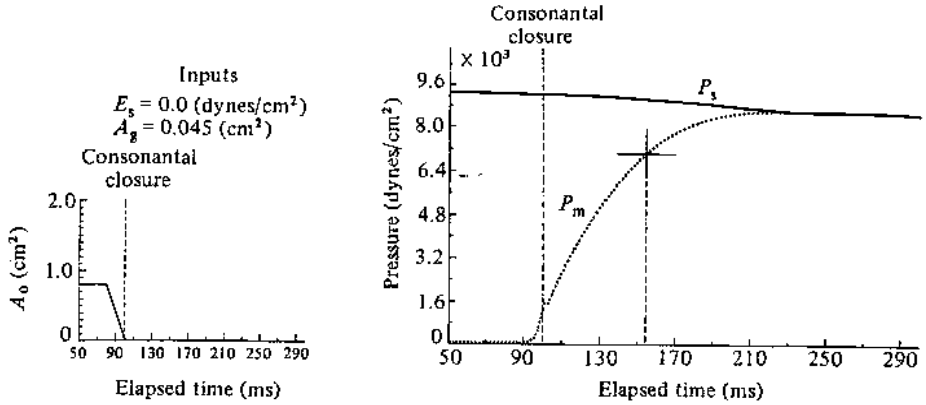


*Figure 2*

Calculated subglottal ($P_s$) and supraglottal ($P_m$) air pressure waveforms following closure for an intervocalic labial stop consonant. The state of the vocal folds, suitable for voicing during the vocalic intervals preceding and following the occlusion, is assumed to remain constant during the closure interval. Similarly, all other control elements in the model, except for that representing the oral constriction have been set to constants.

during that portion of an intervocalic stop, with offset occurring only late during its closure, within 20 ms of release for an 80 ms closure.

The relatively lengthy interval of closure voicing during such a stop would be due almost entirely to the yielding walls which surround the supraglottal cavity. In effect, their outward motion during the stop closure – in response to the increasing air pressure they contain – retards the rate at which the transglottal pressure drop decreases, and thereby lengthens the interval during closure when voicing is possible.

The precise duration of voicing will be influenced by factors which determine this expansion. The functons illustrated in Figure 2 can be considered representative of a labial stop. For more posterior places of articulation, the rate of increase in supraglottal pressure would be greater and the extent of 'natural' closure voicing would be somewhat less, since the total surface area surrounding the supraglottal cavity – and thus, the total compliance of the supraglottal walls – would also be less (Ohala & Riordan, 1979). Calculations with the model indicate that under otherwise similar conditions, the duration of closure voicing might be some 30% less for a velar stop than for a labial (Keating, 1983). Moreover, if the walls of the supraglottal cavity are assumed to be more lax than they are above (as well as in Figures 5 and 6) – where they are assumed to have the compliance of tense cheek tissue (Ishizaka *et al*, 1975) – the extent of 'natural' closure voicing may be greater. Other calculations indicate that with more compliant walls, voicing could continue for at least 100 ms – i.e. beyond the usual duration of singleton stop closures. On the other hand, if the walls are





assumed to be rigid, effective pressure neutralization (and voice offset) will probably occur within 10 ms of occlusion.

Lengthening an intervocalic closure, as might be appropriate in the simplest case for a homorganic stop cluster (or geminate) bounded by identical vowels, would have no effect on the expected time change in supraglottal pressure which occurs over the initial 80 ms of the closure interval depicted in Figure 2. Rather, lengthening the closure to something on the order of 150 ms would only allow that pressure more time to approximate pressure below the glottis. Thus, we might expect the closure of a relatively long, articulatorily simple intervocalic stop – in effect, a geminate or a homorganic cluster – to be initially voiced and then voiceless. This simple conclusion is strongly reminiscent of an observation by Harms (1978) that the second /d/ in the phrase *mad dog* often seems to be devoiced, though probably for 'natural' reasons, as Harms pointed out, rather than because of any intentional change in the glottal state.

Whether a geminate with a partially-voiced closure, or alternatively a homorganic cluster – which differs from the former in that its closure may span a syllable boundary – might be perceived as voiced, voiceless, or somehow mixed is an interesting though unresolved question. To our knowledge, there are no psychoacoustic data which suggest what effects partial voicing might have on the percept of geminates and stop clusters. Certainly, we might infer from an earlier conclusion by Ohala (1983) that geminates not voiced through release – such as the one we suggest to be articulatorily simplest, and thereby most natural – are generally categorized as voiceless. However, that speculation has yet to be substantiated by parallel, fine-grained analysis of the associated articulatory and acoustic properties of stop geminates, on the one hand, and categorical labelling functions derived from speakers of languages which have them, on the other.

Of course, a medial stop (or homorganic cluster) may be made fully voiceless if articulatory adjustments occur at the level of the larynx, which make the vocal folds less susceptible to oscillation, and/or which hasten neutralization of the pressure drop across them. Tight adduction of the vocal folds, which often occurs in syllable-final voiceless stops in English (Fujimura & Sawashima, 1971; Westbury & Niimi, 1979), obviously has the former effect. Abduction of the vocal folds, which frequently accompanies the closures of voiceless stops in a variety of languages (Hirose, 1977), may have both effects.

Alternatively, a lengthy medial closure may be made more fully voiced than the 60 ms or so suggested by Figure 2 by several methods, including: contracting the expiratory muscles; decreasing average area of the glottis and/or tension of the vocal folds; decreasing the level of activity in muscles which underlie the walls of the supraglottal cavity; actively enlarging the volume of that cavity; or creating a narrow opening between the posterior pharyngeal wall and soft palate (Rothenberg, 1968; Westbury, 1983). These

manœuvres, occurring singly or in combination, will have their greatest effect on duration of closure voicing when they occur during the closure interval itself, in concert with the rise in pressure above the glottis which naturally accompanies vocal tract occlusion. Implementing them in the model involves specifying how each of the relevant control parameters will vary in time.

However, within the context of the breath-stream dynamics model, additional glottal gestures leading to greater voicelessness, or other articulatory manœuvres leading to longer intervals of closure voicing, entail added cost in the sense that both involve changes in articulatory states above and beyond those (presumably) minimally necessary to produce the vowel + stop + vowel sequence. An assumption central to this paper is that (rates of) changes in 'states' of articulators are the fundamental basis which determines relative articulatory cost. If, indeed, speakers seek out the easiest (ways to produce) sounds and sound sequences, then they should do so by minimizing changes in the various articulatory parameters subject to voluntary control. 'Speaking' in that fashion, using a simple model of the breath-stream control mechanism, suggests that a stop sandwiched between two identical vowels should naturally be largely voiced if its closure is relatively short, or voiced–voiceless if its closure is long.

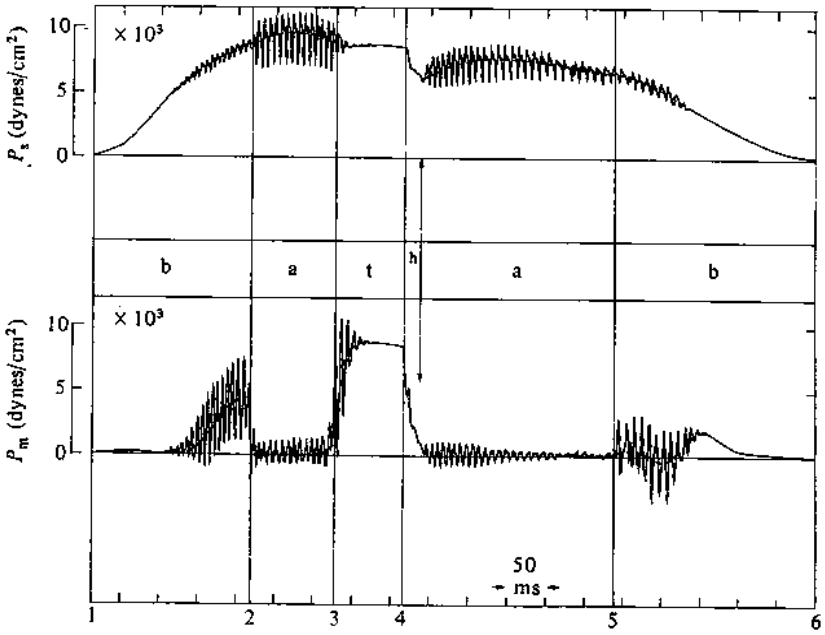


*Figure 3*

Subglottal ($P_s$) and supraglottal ($P_m$) air pressure waveforms recorded during the articulation of the nonsense disyllable /batab/ by one of the authors (JW). Oral closures are judged to occur at moments indicated by 1, 3, and 5, while releases are judged to occur at 2, 4, and perhaps 6.





### 3.2. *The initial case*

The expectations derived from the model and a simple characterization of ease of articulation regarding closure voicing during intervocalic stops and homorganic clusters are different from those regarding closure voicing during stops occurring utterance-initially, or those occurring utterance-finally. One reason for these differences is straightforward. Note from Figures 3 and 4, for example, which show characteristic time functions of air pressures above and below the glottis during isolated disyllables produced by one of the authors (JW), that air pressure below the glottis remains generally high and stable over the middle segments of an utterance, but not over beginning (or ending) segments. Rather, in those respective environments, subglottal pressure rises above and falls toward the air pressure exerted by the atmosphere – zero, in these figures – in a generally linear fashion. Since the incidence of voicing depends in large part upon the difference between subglottal and supraglottal pressures, stop voicing should be more likely utterance-medially than initially or finally, simply because that pressure difference in the former environment tends to be somewhat greater than in the latter environments.

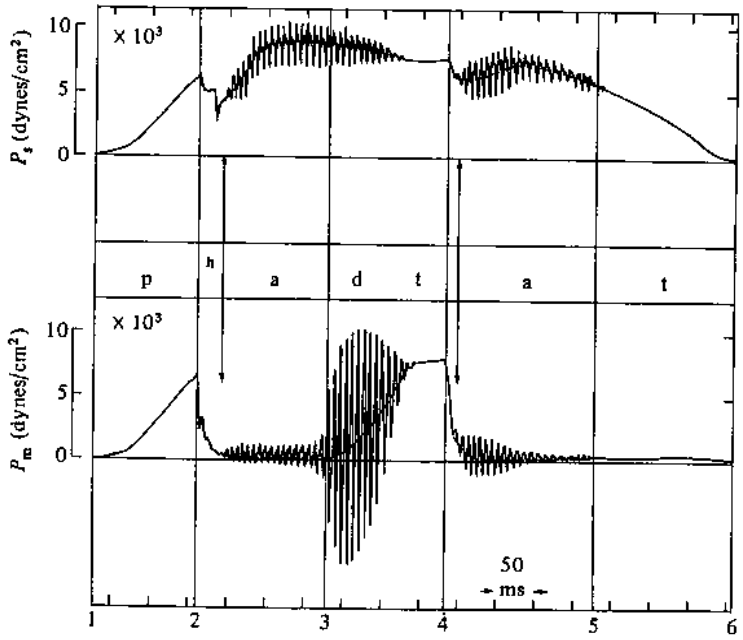


*Figure 4*

Subglottal ($P_s$) and supraglottal ($P_m$) air pressure waveforms recorded during the articulation of the nonsense disyllable /padtat/ by one of the authors (JW). Oral closures are judged to occur at moments indicated by 1, 3, and 5, while releases are judged to occur at 2, 4, and perhaps 6.

Consider then in more detail the same problem for an utterance-initial stop considered previously for an intervocalic one: namely, is it more likely for a stop to be voiced or voiceless in an articulatorily 'simple' string composed of pause + stop + vowel? As before, the model is used to calculate changes in air pressure during the string, determining the likelihood of closure voicing. However, specification of the articulatory conditions which affect the incidence and duration of closure voicing are necessarily different for hypothetical initial and intervocalic stops. As before, the input functions for an initial stop entail no muscularly induced changes in supraglottal volume, or in the mechanical properties of tissues surrounding the lungs and mouth. Moreover, the velopharyngeal port remains fully occluded, and the vocal folds are appropriately and constantly adducted and tensed for voicing. However, cross-sectional area of the mouth opening is varied to release (rather than form) the oral constriction defining the stop. As before, the tissues surrounding the lungs are considered stretched, so that positive subglottal pressure will derive from their elastic recoil, but utterance-initially, the thoracic muscles are not considered quiet. Rather, the net force generated by the inspiratory muscles – whose contractions initially enlarge the thoracic cavity – is assumed to be slowly relaxing, in a roughly linear fashion over the closure interval of the initial stop. That slow relaxation 'checks' the recoil of the stretched thoracic tissues and can thereby provide a slow, smooth rise in subglottal pressure of the sort pointed out in Figures 3 and 4. Moreover, its inclusion among the input conditions for an utterance-initial stop is consistent with experimental observations made some years ago by Draper *et al.* (1959).

Given these input articulatory conditions, air pressures above and below the glottis can roughly be expected to vary with time during an utterance-initial stop as shown in Figure 5. Subglottal pressure begins a steady rise roughly 200 ms before consonantal release, in a fashion similar to that illustrated in Figures 3 and 4. However, supraglottal pressure also rises steadily during closure, virtually in synchrony with subglottal pressure. That fact is due to the assumption that the 'voicing state' exists when the vocal folds are continuously apart, though narrowly so, separated by slightly more than 0.02 cm. As a consequence, the difference between subglottal and supraglottal pressures never exceeds the assumed voice-initiation threshold of 4000 dyne/cm² (cf. Baer, 1975), prior to the consonantal release. Instead, after the closure is released, air flows to the atmosphere, and intraoral pressure drops abruptly while subglottal pressure remains high. Only then is a suitable pressure drop achieved. Thus, the stop in an articulatorily simple pause + stop + vowel string will plausibly be voiceless (and unaspirated). Note that this conclusion does not depend on having the glottis in either a voiceless or a breathing configuration. Either of those configurations would of course make voicing even less likely.

If we suppose, alternatively, that the vocal folds are fully adducted well before the release of an utterance-initial stop, so that air cannot flow from





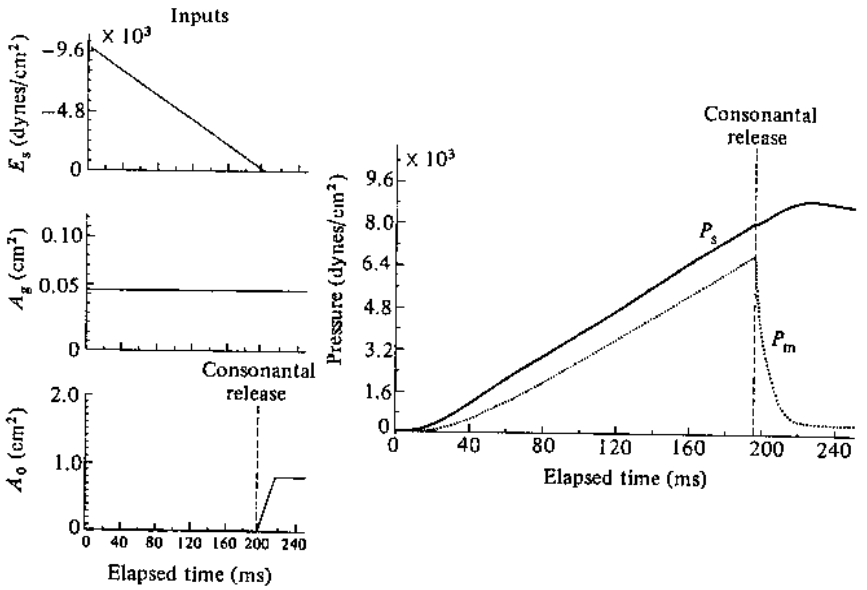


*Figure 5*

Calculated subglottal ($P_s$) and supraglottal ($P_m$) air pressure waveforms during an utterance-initial labial stop consonant. The steadily decreasing inspiratory force ($E_s$), represented here by convention in terms of a negative pressure head returning to zero, checks elastic recoil of the stretched tissues surrounding the subglottal cavity, thereby causing the slow, linear increase in subglottal pressure. The constant glottal area of 0.045 cm² is an approximation of the average area of the glottis during a sustained vowel. This figure suggests that voicing could begin only after release of the oral constriction ($A_o$) for an utterance-initial stop.

the lungs to the mouth, then air pressures above and below the glottis will rise asynchronously, and a different pattern of voicing would probably result. Pressure below the glottis would rise first, and pressure above the glottis would begin rising only after the vocal folds had been separated, perhaps after having been 'blown apart' once some suitable pressure gradient has been reached. Calculations suggest that 30–40 ms of closure voicing might occur, immediately following the moment when the vocal folds are blown apart and the voicing state is established. Most probably, voicing will also then cease some 30–40 ms prior to release. Thus the closure interval of an initial stop articulated under these conditions would be initially voiceless, subsequently voiced, and then voiceless again, all prior to consonant release. Whether a stop such as that would be considered prevoiced or voiceless, by a phonetician and or native listener, is an open question. However, stops with this acoustic pattern are occasionally seen among utterance-initial /b,d,g/ produced by speakers of American English. Such a scenario would be somewhat more consistent with the time functions of subglottal and supraglottal pressures during the utterance-initial /b/ illustrated in Figure 3, which clearly rise

asynchronously. The utterance-initial /p/ in Figure 4, by comparison, shows subglottal and supraglottal pressures rising synchronously, with the vocal folds no doubt relatively widely abducted.

### 3.3. *The final case*

In an articulatorily simple string composed of vowel + stop + pause, air pressures above and below the glottis can be expected to vary somewhat differently than utterance-initially or medially. All but two articulatory inputs necessary for calculation of these functions are the same as those for the intervocalic stop. Cross-sectional area of the mouth opening has been varied to produce a vocal-tract constriction, but (in this case) not to release it. Moreover, elastic recoil of stretched tissues surrounding the lungs is still thought to be responsible for the positive pressure head below the glottis, but that pressure is progressively being opposed by a hypothetically increasing inspiratory force which begins 20 ms or so after the moment of oral occlusion. (The latter input to the model, admittedly *ad hoc*, provides in rather simple fashion a relatively linear decrease in subglottal pressure following implosion for a final stop of qualitatively the same sort that is apparent from either Figure 3 or 4.)

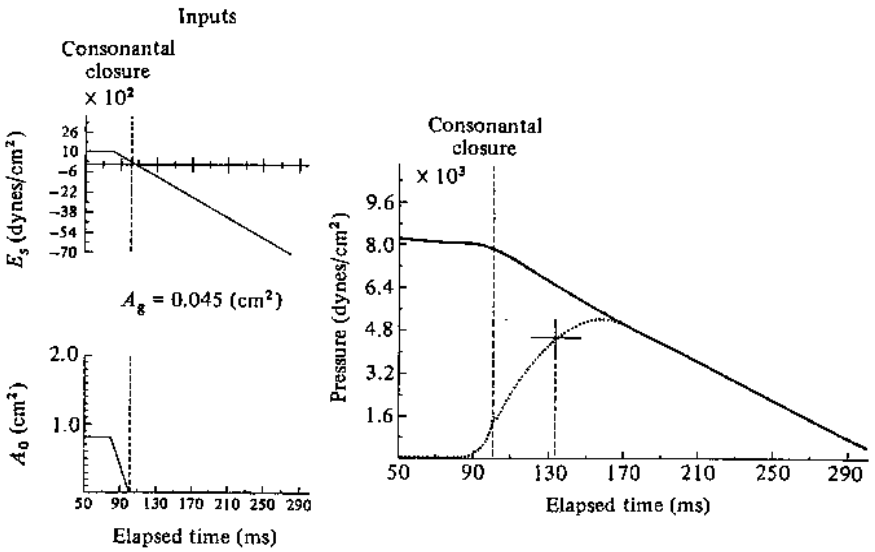


*Figure 6*

Calculated subglottal ($P_s$) and supraglottal ($P_m$) air pressure waveforms during an utterance-final labial stop consonant. The manipulation of respiratory forces ($E_s$) shown here has the articulatory interpretation of being initially expiratory, as might be expected during the latter phase of an utterance (cf. Draper *et al.*, 1959), but quickly becoming inspiratory not long after closure for the final stop. Voice offset would likely occur near the cross intersecting the supraglottal air pressure waveform, *c.* 35 ms following closure of the oral port ($A_0$).





Given these conditions, subglottal pressure can be expected to decrease after implosion for a final stop as shown in Figure 6, in a roughly linear fashion analogous to the final decays in that pressure apparent in Figures 3 and 4. At the same time, intraoral pressure can be expected to increase. The two pressures converge relatively rapidly and voicing offset occurs slightly more than 30 ms after the moment of occlusion. If the stop in an articulatorily simple vowel + stop + pause string is held for something on the order of 100 ms, its closure will then be largely (though not entirely) voiceless. This result does not depend on a laryngeal devoicing gesture, perhaps in preparation for respiration. Even with the vocal folds adducted, voicing will cease relatively early during the stop closure.

Thus, the likelihood of 'natural' consonant voicing should depend in part on position in utterance, to the extent that those positions differ in subglottal pressure functions associated with them. The relatively smaller subglottal pressures in utterance-initial and final positions – whose general character is illustrated in Figures 3 and 4, and which do not depend upon the specification or realization of voicing or voicelessness (Westbury & Niimi, 1979) – presumably make voicing articulatorily more difficult to sustain, and thereby less natural than voicelessness. Relatively high and steady subglottal pressure in medial position, on the other hand, makes voicing more natural than voicelessness. No differences in glottal configurations need be assumed to generate these differences; in all environments, the vocal cords are assumed to be in a 'voicing state' conducive to vibration.

## 4. Comparison with language data

By hypothesis, largely voiceless initial and final stops, voiced intervocalic stops, and voiced–voiceless intervocalic stop clusters are relatively easier to produce than others. These easiest of stops are all articulated with the same 'voiced' configuration of the vocal folds. The extent to which voicing is actually realized during their closures depends then upon characteristic subglottal pressure functions associated with different positions in an utterance.

If languages seek out and prefer segments and utterances which are easiest to produce, and if the characterization of ease of articulation followed herein is fair, then we might expect to find that languages without phonetic voicing contrasts maintain the articulatorily optimal stop system wherein context-dependent phonetic manifestations of the only 'type' of stop – i.e. the natural, easy one – would be variable but also easily predictable. Moreover, we might expect that languages with some phonemic voicing contrasts which evidence defective distributions or contrast neutralization in some environments should show a preference for the easier stops in reduced-contrast slots or at neutralization sites. Neutralization to the optimal singletons and clusters would have the effect of making the utterances containing them easier

to produce than they would otherwise have been. Conversely, maintaining any neutralized phonetic variant other than the optimal one would entail added articulatory cost for the language.

Determining whether these expectations hold across various languages is difficult given information readily available in the literature. Certainly it is well-documented at the PHONEMIC level that voiceless stops are preferred over those that are voiced. Ohala (1983:194) notes, for example, that of the '706 languages whose segment inventories are surveyed by Ruhlen (1975) 166 have only voiceless stops and 4 have only voiced stops'. Maddieson (1984) also finds support for this generalization from the UCLA Phonological Segment Inventory Database. The conventional account for this 'decisive "tilt" toward voicelessness' depends foremost upon the claim that 'there is a well-recognized difficulty in maintaining voicing during a stop' (Ohala, 1983:194). However, the claim underlying this account is too strong. Under certain conditions, there is a well-defined sense in which it is more difficult to terminate than to allow (or maintain) voicing during a stop. Those conditions are not exotic. Rather, as suggested above, they can plausibly be expected quite frequently in at least one, very common phonetic environment, namely, utterance medial position. Thus, expectations regarding stop voicing can be evaluated only by examining, in typological terms, patterns of allophonic variation within and among various languages.

Detailed phonetic data on common allophones – particularly for languages without phonetic voicing contrasts, and to a lesser extent, for languages with contrasts which evidence morphophonemic alteration or surface neutralization of stop voicing – are surprisingly hard to come by. The best source known to us is the recent review by Keating, Linker and Huffman (1983) which provides a general description of allophonic variation among voiced and/or voiceless stops in 51 languages. However, even those data are less than ideal, first because they are largely categorical in nature, and secondly, because they are acoustically based (in the sense that they are derived from transcriptions by phoneticians and/or from spectrographic and oscillographic analysis) rather than physiological.

The problem with 'categorical' data becomes manifest, for example, in a conventional description of modern Polish where utterance-final instances of /b,d,g/ are said to be devoiced. Similarly, final voiced stops in the speech of children acquiring American English are frequently described as devoiced. However, oscillographic analyses have shown that the closures of final /b,d,g/ in both Polish (Giannini & Cinque, 1978) and the developing speech of young children (Smith, 1979) reveal more closure voicing (*c.* 30–40 ms) than do their underlying voiceless counterparts (*c.* 10–20 ms). Whether these differences are perceptible with sensitive test paradigms is not known. If not, these will not be the first cases in which speakers produce systematic differences that they do not seem to perceive. Dinnsen (1985) reviews similar cases and concludes that grammars must be capable of representing such differences. Furthermore,





as Dinnsen stresses, such cases as Polish word-final stops argue against the standard rules of complete neutralization of surface forms, since speakers preserve the underlying distinctions. However, the question of how such processes are to be accommodated within a phonology is beyond the scope of this paper.

At least in acoustic terms, in both cases noted above the underlying voiced consonants are not identical on the surface to the underlying voiceless consonants. Rather – and it is not an uninteresting fact – they are merely labelled similarly. However, that act of categorical labelling – inherent in a segmental inventory, conventional descriptive phonology, or phonetic transcription – may obscure acoustic and physiological details which are crucial for a fair test of our hypotheses. That is because these hypotheses make specific acoustic predictions which derive from specific articulatory manifestations of utterances.

Even noncategorical acoustical data are not ideal for testing claims about articulatory ease. Rather, detailed physiological and acoustic data, in concert, and in terms comparable to the articulatory model, are required for that task. Otherwise, it may be impossible to know whether various acoustic data which correspond to the model's predictions bear out those predictions for the right reasons. Nonetheless, with this reservation in mind, we can at least ask when data from natural languages APPEAR to conform to the expectations derived from the model. After all, if even acoustic correspondences are not found, then there may be little point in entertaining the articulatory hypotheses that have been presented.

Consider first the six languages described in Keating *et al* (1983) – Alyawarra, Hawaiian, Kaititj, Nama, Tiwi, Yidin$^{y}$ – which exhibit no phonemic voicing-related contrast among stops. At least in categorical acoustic terms, we can say that these languages – as a group – prefer voiceless stops in all permissible environments, and show less in the way of allophonic variation regarding stop voicing than languages which have phonemic contrasts. These languages create the impression that variability in the acoustic realization of the same speech sound in different environments is generally undesirable. In simple terms, our most preferable stop 'system' – including voiceless unaspirated stops initially, voiced singletons (and voiced–voiceless clusters) medially, and largely voiceless stops finally – does not seem to occur. Since most of these languages do not allow final stops of any sort, the main difficulty for our proposal comes from medial stops which are generally characterized as voiceless and unaspirated, rather than voiced.

Some consolation may be drawn, however, from the fact that the optimal stop system does seem to be reflected in 'developing languages' without contrasts – in the speech of young children who are acquiring their native tongue. Prior to their mastery of whatever contrasts may exist in the speech of adults, children tend to articulate utterance-initial stops which are most often described as voiceless and unaspirated (Preston, Stark & Yeni-Komshian,

1967; Eilers, Oller & Benito-Garcia, 1984). Similarly, prior to their mastery of final contrasts, children tend to produce 'voiceless' stops before pauses (Smith, 1979). Lastly, children seem to show in their earliest speech a greater preference for voicing during utterance-medial stops than do adults (cf. Smith, 1973), though this last generalization is clouded by considerable individual variation and differences in the temporal control of stop closures (Smith, 1978).

Consider next the more than forty languages described in the Keating *et al* (1983) summary which maintain some system of voicing-related contrasts among stops, in at least some environments. For these languages, we are particularly interested in the phonetic form of stops at sites of neutralization, even when the observed neutralization does not span all places of articulation. Moreover, we are equally interested in the phonetic form of stops where the distributions of phonemes are defective. If, as we have hypothesized, languages prefer stops which are easiest to produce, then we might expect them to do so at least where restriction and/or environmental neutralization of an existing voicing contrast occurs. Maintaining any neutralized phonetic variant other than the optimal one entails added articulatory cost for the language.

Generalizations from the Keating *et al* (1983) review relevant to this expectation are as follows.

First, neutralization of a stop voicing contrast in utterance-initial position is uncommon. There are only four languages in the sample which, for at least some places of articulation, contrast voiced and voiceless stops medially and/or finally but not initially. These include Cuna, Efik, Ewondo, and Tamil. In all but the second of this group, the observed neutralization exhibits the expected 'destination' – i.e. a voiceless unaspirated stop. In Efik, however, a phonemic contrast between labials in medial position is neutralized to a voiced labial utterance-initially.

Secondly, neutralization of voicing contrasts medially appears to be quite rare. In American English, underlying /t/ and /d/ neutralize to the voiced flap [ɒ] in specific medial environments (Zue & Laferriere, 1979), though some contrast between them is maintained initially, finally, and in other medial environments. Similarly, in Zoque, an initial contrast between /tj/ and /dj/ is also neutralized medially, though contrary to our expectation, to a voiceless allophone.

Thirdly, in languages which maintain voicing contrasts of one sort or another among stops occurring initially and/or medially, final position is far and away the most common site for neutralization. Eighteen of the fifty-one languages surveyed by Keating *et al* (1983) exhibit at least some neutralization of voicing-related contrasts among stops in that environment. Those include Basque, Bulgarian, Cantonese, Choctaw, Cuna, Dutch, Efik, Ewondo, Gaelic, German, Korean, Polish, Russian, Spanish, Thai, Tikar, Vietnamese, and Zoque. In fifteen of these, neutralization proceeds to preferred targets,





but other, 'unexpected' results obtain in German (aspirated stops), Spanish (continuants), and Zoque (aspirated stops).

Together, these generalizations show that the majority of known cases of contrast neutralization involving singleton stops seems compatible with expectations derived from the model. It must be stressed, however, that in most cases the domain under consideration in the literature survey is the word, not the utterance. For example, devoicing of phonemically voiced stops may obtain in final position within a syllable, a word, or an utterance in various languages. The expectations from modelling do not account for positional effects other than those associated with position in utterance. Other effects might be considered language-specific generalizations of the utterance patterns. However, it should be noted that some effects commonly reported as 'word' effects are in fact constrained by pause – i.e. they are utterance effects. For example, Mikos (1977) has discussed the tendency for the Polish rule of 'word-final' consonant devoicing to apply only before pause. Before a vowel-initial word, and apparently, generally utterance-medially (cf. Slowiaczek & Dinnsen, 1985), neutralization of the voicing contrast does not occur. In fact, in some dialects voicing is truly utterance conditioned: not only will devoicing not apply, but word-final voiceless stops will voice before a vowel. This is precisely the sort of situation our modelling leads us to expect.

The survey in Keating *et al.* (1983) does not provide any data regarding contrast neutralization in medial clusters. However, we feel that it can safely be said that assimilation of voicing in medial stop clusters – yielding neutral allophones from two or more sources – is thought to be quite common across languages. Certainly, our prediction regarding the most preferable medial cluster is not borne out by the few facts known to us. The optimal cluster – which in acoustic terms, might be described as having fully dissimilated members, but which in physiological terms has fully assimilated members – is clearly not the customary destination of voicing neutralization (or assimilation) rules. Languages which have medial clusters underlyingly typically either maintain all voicing-related contrasts among them at the surface (e.g., American English, Punjabi, Hindi), or partially collapse underlying contrasts to yield two distinct surface forms whose respective members are at least acoustically assimilated (e.g. Russian, Dutch, French, Hungarian). Thus, the number of surface contrasts among medial clusters is either the same as or slightly reduced relative to the number of underlying contrasts. But, as far as we know, no language reduces all underlying contrasts among clusters to a single surface form.

These facts suggest two things. Foremost, there must be considerable pressure not to collapse entirely underlying systems of contrast. Reducing the voicing contrasts among all underlying stop clusters to a single, maximally simple surface form might indeed be preferable from a physiological point of view, but that reduction would at the same time be decidedly less than optimal from what we might call an information-transfer point of view. 'We

speak to be heard in order to be understood' (Jakobson *et al*, 1963: 13). The utility of contrasts for conveying information – i.e. for making ourselves understood – is obvious. Surely, then, the extent to which we allow articulatory principles to govern our linguistic behaviour depends upon the extent to which they impinge on its primary function.

Secondly, the sheer prevalence of languages which exhibit voicing assimilation in medial clusters suggests – following the generally intuitive line that ease of articulation provides a fundamental basis for distributions of sounds and sound sequences – that acoustic voicing assimilation in medial clusters may be relatively easier, in some physiological sense, than certain types of acoustic dissimilation which depend upon parallel differentiation at a physiological (articulatory) level. That is, making a medial cluster fully voiced or fully voiceless may be somehow articulatorily easier than making one initially voiced and finally voiceless, by some means other than the 'natural' method, or initially voiceless and finally voiced. One line of reasoning which maintains the ease-of-articulation principle introduced earlier in this paper, and which 'prefers' clusters of the former rather than the latter sort, is the following:

In general terms, assimilation can be thought of as a reduction in the rates of articulatory changes which are specified to occur between adjacent or temporally-proximal segmental states in the underlying representation of an utterance. At the phonetic surface, that reduction can be manifest as (1) a slowing of articulatory transitions, so that they (or perhaps, the segmental 'steady' states themselves) spread out in time, and/or (2) a reduction in the differences between adjacent states, so that one or all are undershot. If – for reasons unknown – voicing must be acoustically and physically maintained during the latter portion of the lengthy closure of a medial stop cluster, we know from experience with the breath-stream control system model that a speaker must expend extra articulatory effort. That effort might take several articulatory forms, but if its acoustic effect is to be local only to the latter portion of a lengthy closure, it must be expended in a ballistic or 'step' fashion. By hypothesis, expenditures of the latter sort are articulatorily costly. Alternatively, a speaker might make qualitatively similar articulatory adjustments to maintain voicing cluster-finally, but begin them sooner, in effect sacrificing acoustic (and physiological) integrity of the cluster-initial stop in favour of easing articulation of the string which contains it.

Similar reasoning can be used to argue that 'regressive assimilation' of voicelessness in underlying voiced–voiceless clusters might be preferable to maintaining acoustic dissimilarity between clusters' members. If voicelessness local to an underlying voiced–voiceless cluster's final portion is to be insured by vocal fold abduction, and if subsequent glottal abduction for a following sonorant must be largely complete before the cluster is released, then the sequence of stops comprising the cluster may be relatively easier to articulate if the changes in glottal state are slowed by 'spreading backward in time'.





The acoustic consequence of that spread, of course, would be that the cluster-initial stop would appear less voiced than if no change in glottal state were to occur.

In general, then, if the closures of medial clusters must encompass certain articulatory manœuvres to insure voicing and voicelessness, initially or finally, and if their closure durations cannot be extended appreciably, then it may be easier to spread those manœuvres in time, 'causing' assimilation of voicing (or voicelessness) among the clusters' members, than to make the manœuvres rapidly enough to limit the temporal domain of their effects.

## 5. Summary and conclusions

How well does aerodynamic modelling predict the distribution of acoustically voiced and voiceless stop consonants? (1) In the pre-contrast (or developing contrast) stages of children's speech, it seems to do well, at least for singleton consonants. We know of no relevant data concerning consonant clusters. (2) In languages with no stop consonant voicing contrast, stops tend to be voiceless in all positions. Clearly, then, we are confronted with a number of languages which maintain articulatorily more difficult stops utterance-medially than is necessary. We note that, as the price of this greater articulatory effort, these languages maintain phonetic similarity among their positional allophones. It is as though these languages sacrifice, in part, ease of articulation in favour of limiting acoustic variability in the realization of their segments. A similar case is seen in the speech of those speakers of American English who prevoice utterance initial /b,d,g/, which requires 'extra' articulatory effort (relative to voiceless, unaspirated stops) but makes such initial stops more like the phonetically voiced medial /b,d,g/. Thus, results derived from modelling suggest a case in which ease of articulation cannot be the only factor in a sound pattern. The value of modelling, of course, is that it provides one means for identifying phenomena that may be 'explained' in terms of an articulatory ease principle, and those that must derive from other, equally powerful principles. (3) In languages with a voicing contrast of some sort, we find relatively few instances of variation. Overall, contrast languages maintain contrasts of articulatorily simpler stops with somewhat more costly stops. It is perhaps significant that in initial position, virtually all languages utilize the voiceless unaspirated stops. If our modelling is taken to predict a strong preference for one and only stop category initially, then of course it fails; but it does suggest which of the categories available for contrast would most frequently be chosen by languages. Similarly, in medial position, we find no strong preference across languages for any stop consonant category, either the voiced one predicted by modelling, or any other. Some languages favour voiceless unaspirated stops medially, while some favour voiced stops. The model finds more correspondence with natural language in predicting the occurrence of voiceless unaspirated stops in utterance-final position, where

neutralization of voicing contrasts is most common. Limited acoustic data suggest that in some cases at least the details of model-based predictions – that final stops should be somewhat, but not greatly, voiced, and therefore acoustically distinct from both completely voiced and completely voiceless stops – are borne out. Here, however, the lack of parallel articulatory and acoustic studies of final neutralization of voicing limits our ability to interpret the language survey data with regard to the model's predictions.

The articulatory model described in this report provides a forum within which a notion like ease of articulation can be explicitly defined and tested. When the model's predictions and the facts of language disagree, what does that mean? We conclude that, with regard to stop consonant voicing, ease of articulation is probably not the primary determinant of phonetic form, implicitly assuming that the model and the necessary implementation of an ease of articulation principle are adequate for our purposes. The model allows us to investigate the limits of the influence of that principle on phonetic form, and particularly to identify those few cases where it seems to play some clear role. We then know which cases remain to be explained, and can hypothesize further principles relevant in such cases, such as communicative efficiency, acoustic invariability, and perceptual requirements.[1]

## REFERENCES


Baer, T, (1975). Investigation of phonation using excised larynxes. Unpublished doctoral dissertation, M.I.T.

van den Berg, J. (1958). Myoelastic theory of voice production. *JSHR* 1. 227–244.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Crothers, J. (1978). Typology and universals of vowel systems. In Greenberg, J. (ed.), *Universals of human language*, Vol. 2. Stanford, CA: Stanford University Press. 93–152.

Dinnsen, D. (1985). A re-examination of phonological neutralization. *JL* 21. 265–279.

Draper, M., Ladefoged, P. & Whitteridge, D. (1959). Respiratory muscles in speech. *JSHR* 2. 16–27.

Eilers, R., Oller, D. K. & Benito-Garcia, C. (1984). The acquisition of voicing contrasts in Spanish and English learning infants and children: a longitudinal study. *JCL* 11. 313–336.

Flanagan, J., Ishizaka, K. & Shipley, K. (1975). Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *Bell System Technical Journal* 54. 485–506.

Fujimura, W. & Sawashima, M. (1971). Consonant sequences and laryngeal control. *Ann. Bul. Res. Logoped. Phoniat. (Tokyo)* 5. 1–6.

Gianinni, A. & Cinque, U. (1978). Phonetic status and phonemic function of the final devoiced stops in Polish. *Speech Laboratory Report I. Laboratorio di Fonetica Sperimentale (Naples)*.

Harms, R. (1978). Some nonrules of English. In Jazayery, M. A., Polomé, E. C. & Winter, W. (eds.), *Linguistic and literary studies in honor of Archibald A. Hill. II: Descriptive linguistics* The Hague: Mouton. 39–51.



[1] The research described in this paper was carried out in the Speech Communication Group at M.I.T., under the direction of Professor Kenneth N. Stevens, with support from post-doctoral fellowships from NIH. We are indebted to Prof. Stevens for his help. Further work and preparation of the manuscript was supported at UNC by the National Institute of Dental Research, and, at UCLA, by the UCLA Academic Senate research program and by a grant from the National Institute of Neurological and Communicative Disorders and Stroke to Peter Ladefoged. We also thank Howard Colub and Gunnar Fant for advice.







Hirose, H. (1977). Larnygeal adjustments in consonant production. *Phonetica* 34. 289–294.

Hooper, J. (1976). *An introduction to natural generative phonology*. New York: Academic Press.

Ishizaka, K., French, J. & Flanagan, J. (1975). Direct determination of vocal tract wall impedance. *IEEE Trans. Acoust., Speech Signal Process.*, ASSP-23. 370–373.

Ishizaka, K. & Matsudaira, M. (1972). Fluid mechanical considerations of vocal cord vibration. *Speech Com. Res. Lab. Monograph 8*.

Jakobson, R., Fant, G. & Halle, M. (1963). *Preliminaries to speech analysis*. Cambridge, Mass.: MIT Press.

Keating, P. (1983). Physiological effects on stop consonant voicing. *JAcS Supplement* 1, 73. S47. Also in *UCLA Working Papers in Phonetics* 59, 1984.

Keating, P. (1984). Aerodynamic modeling at UCLA. *UCLA Working Papers in Phonetics* 59. 18–28.

Keating, P., Linker. W. & Huffman, M. (1983). Patterns of allophone distribution for voiced and voiceless stops. *Journal of Phonetics* 11. 277–290.

Ladefoged, P. (1964). Comment on 'Evaluation of methods of estimating subglottal air pressure'. *JSHR* 7. 291–292.

Ladefoged, P. (1982). *A course in phonetics* (2nd edition). New York: Harcourt Brace & Jovanovich.

Liljencrants, J. & Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Lg* 48. 839–862.

Lindblom, B. (1978). Phonetic aspects of linguistic explanation. *Studia Linguistica* 32. 137–153.

Lindblom, B. (1983). Economy of speech gestures. In MacNeilage, P. F. (ed.) *The production of speech*. New York: Springer Verlag. 217–246.

Lindqvist, J. (1972). Laryngeal articulation studied on Swedish subjects. *Speech Trans. Lab. Quarterly Progress Report*. 2–3. 10–27.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Mikos, M. J. (1977). *Problems in Polish phonology*. Unpublished doctoral dissertation, Brown University.

Muller, E. & Brown, W. (1980). Variations in the supraglottal air pressure waveform and their articulatory interpretation. In Lass, N. (ed.), *Speech and language: Advances in basic research and practice*, Vol 4. New York: Academic Press. 317–389.

Ohala, J. (1974). Phonetic explanations in phonology. In Bruck, A., Fox, R. & LaGaly, M. (eds.), *Papers from the parasession on natural phonology*. Chicago: Chicago Linguistic Society. 251–274.

Ohala, J. (1976). A model of speech aerodynamics. *Report of the Phonology Laboratory (Berkeley)* 1. 93–107.

Ohala, J. (1983). The origin of sound patterns in vocal tract constraints. In MacNeilage, P. F. (ed.), *The production of speech*. New York: Springer Verlag. 189–216.

Ohala, J. & Riordan, C. (1979). Passive vocal tract enlargement during voiced stops. In Wolf, J. & Klatt, D. (eds.), *Speech communication papers presented at the 97th meeting of the Acoustical Society of America*. New York: Acoustical Society of America.

Preston, M. S., Yeni-Komshian, G. & Stark, R. (1967). A study of voicing in initial stops found in the pre-linguistic vocalization of infants from different language environments. *Haskins Laboratory Status Report on Speech Research* 10.

Rothenberg, M. (1968). The breath-stream dynamics of simple-released-plosive production. *Bibliotheca Phonetica* 6.

Ruthlen, M. (1975). *A guide to the languages of the world*. Stanford, CA: Stanford University Press.

Schachter, P. (1969). Natural assimilation rules in Akan. *IJAL* 35. 342–355.

Schane, S. (1972). Natural rules in phonology. In Stockwell, & Macauley, R. (eds.), *Linguistic change and generative theory*. Bloomington, Ind.: Indiana University Press. 199–229.

Scully, C. (1969). Problems in the interpretation of pressure and air flow data in speech. *University of Leeds Phonetics Department Report* 2. 53–92.

Slowiaczek, L. & Dinnsen, D. A. (1985). On the neutralizing status of Polish word-final devoicing. *Journal of Phonetics*. (in press).

Smith, B. (1978). Temporal aspects of English speech production. A developmental perspective. *Journal of Phonetics* 6. 37–67.

Smith, B. (1979). A phonetic analysis of consonant devoicing in children's spech. *JCL* 6. 19–28.

Smith, N. (1973). *The acquisition of phonology*. London: Cambridge University Press.
Stampe, D. (1979). *A dissertation on natural phonology*. Bloomington, Ind.: Indiana Linguistics Club.
Vennemann, T. (1972). Sound change and markedness theory: on the history of the Germanic consonant system. In Stockwell, R. & Macauley, R. (eds.), *Linguistic change and generative theory*. Bloomington, Ind.: Indiana University Press, 230–274.
Westbury, J. R. (1983). Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *JAcS* 73. 1322–1336.
Westbury, J. R. & Niimi, S. (1979). An effect of phonetic environment on voicing control mechanisms during stop consonants. In Wolf, J. & Klatt, D. (eds.), *Speech communication papers presented at the 97th meeting of the Acoustical Society of America*. New York: Acoustical Society of America.
Zue, V. & Laferriere, (1979). Acoustic study of medial /t,d/ in American English. *JAcS* 66. 1039–1050.