HOW A FULL ACCOUNT OF SEGMENTAL PERCEPTION DEPENDS ON PROSODY AND VICE VERSA

Quentin Summerfield

ABSTRACT

The synthesis time-base of a synthetic precursor phrase is one factor determining the position of the phoneme boundary on a continuum of synthetic CV target syllables, which vary in the durational cue Voice Onset Times (VOT) and are introduced by the precursor. Reducing the time-base, thereby increasing the rate of speech in the precursor, increases the probability of the consonant in the target being perceived as voiceless. The results of such perceptual experiments are compared to those of a production study in which six adult male speakers of British English produced examples of CV syllables composed of the consonants /b,p,g,k/ and the vowels /i,a/ in a sentence frame. VOTs in productions of /p/ and /k/ were 20 milliseconds shorter at fast than at slow rates of speech, warranting a normalisation of VOT duration in perception like that actually obtained.

VOTs were longer in /ki/ than in /ka/. It has been suggested that voicing onset is retarded before high vowels because the oral constriction is reduced more slowly than before lower vowels, thus delaying the attainment of the transglottal pressure drop required for voicing. However, VOTs in /pi/ were shorter than those in /pa/. In the case of voiceless bilabials the aerodynamic factor appears to be outweighed by a mechanical influence of anticipatory co-articulation of tongue position which leads to larynx elevation and less vocal cord abduction before high vowels. The production result for velars, but not that for bilabials, is paralleled in perception where longer values of VOT are required for stops to be perceived as voiceless before /i/ compared to /a/ at

A. Cohen et al. (eds.), *Structure and Process in Speech Perception*

all places of production. This failure of the perceptual process to follow exactly the constraints in production provides an illustration of a heuristic rather than an algorithmic perceptual strategy presumably designed to allow fast decisions while tolerating some loss of accuracy in exceptional cases.

The 'precursor target' paradigm used in these experiments could be extended to examine prosodic influences on other segmental distinctions, but also, to determine the perceptual substrates of prosodic variables such as "rate of speech" and to measure the precision of perceptual expectations for vowel and consonant durations in different sentential and syntactic environments.

INTRODUCTION

In this paper I shall compare some production data with some perceptual data which, taken together, expose a limit on the precision of the articulatory model which the perceptual process utilises when processing running speech and from which inferences can be made about the 'heuristic' processing of a particular acoustic feature. In the perceptual experiments a synthetic target syllable drawn from an acoustic continuum is introduced by a synthetic precursor phrase on which constrained acoustic variations are imposed. In general, if variation of a particular acoustic aspect of a precursor influences the decision as to the value of a phonetic feature in a target, then we may impute to that aspect a contextual influence on the processing of the feature. I believe that this technique forms a useful complement in our armoury of experimental paradigms to the now popular battering ram of perceptual adaptation. Moreover, by stepping smartly to one side, we may turn the weapon upon itself. If a phonetic feature decision in a target syllable can be shown to be influenced by only one type of information deriving from a precursor phrase, then we may use the magnitude of such effects to determine how that information is extracted from the precursor.

DATA

Summerfield (1974) determined the positions of phoneme boundaries on the dimensions /ga-ka/ and /gi-ki/ on which syllables were distinguished by variation in Voice Onset Time (VOT). The two most interesting conditions were differentiated by the synthesis time-base of the three syllable precursor phrase which introduced the target monosyllables. The phrase, "Why are you...?", was synthesised with durations of 528 and 704 msec. in the two conditions, a ratio difference in overall duration of 1:1.33. Phoneme boundaries were determined by the adaptive algorithm PEST (Taylor and Creelman, 1967) and two major results emerged:

(i) The mean phoneme boundary on the dimension /gi-ki/ occurred at a VOT 6.6 msec. greater than that on the dimension /ga-ka/.

(ii) Averaging over both dimensions, the mean phoneme boundary position was 5.4 msec. greater when syllables were introduced by the longer, slower precursor compared to the faster, shorter precursor.

Result (i) was interpreted as reflecting perceptual recalibrations based on perceived vowel identity designed to take account of either (a) the slower rate at which the oral constriction is released in /ki/ compared to /ka/ which results in a slower oral pressure drop and consequently later voicing onset before the high front vowel; or, (b) the greater burst amplitude and duration which generally accompany productions of /ki/ compared to /ka/, in whose absence in these synthetic target syllables, a longer VOT was required to cue voicelessness. Lisker (1974) has since shown that a transitionless first formant increasingly cues voicedness as its frequency is lowered. Whether a steady F1 onset frequency is an additional cue to a voiced F1 transition remains to be determined. Lisker suggests that a low F1 at voicing onset cues voicedness directly through its implication of a closed vocal tract, providing a more parsimonious rationale for certain differences (Summerfield and Haggard, 1974) between various vocalic contexts than our own, being based on an acoustic cue rather than on recalibrations resulting from feedback from at least a parametric representation of the vowel.

Result (ii) was interpreted as a potentially appropriate normalisation to variations in the mean VOTs of (in particular) voiceless stops that would accompany changes in rate of speech in production.

Regardless of how the perceptual adjustments implied by such phoneme boundary shifts are achieved, it is appropriate to determine the extent to which they reflect changes in the mean VOTs of voiced and voiceless stops which accompany variations in contextual factors such as vocalic context and rate of speech in production. Summerfield (1975b) persuaded six adult male speakers of English to produce repetitions of the sentence:-

"Why are you a C_1V_1 when you're a C_2V_2?"

C_1 was always the same as C_2 and was one of /b/, /p/, /g/ and /k/. V_1 always differed from V_2; in each sentence one vowel was /i/ and the other was /a/. Each speaker produced fifteen repetitions of each utterance at each of three rates of speech, nominally identified as "Fast", "Normal" and "Slow". Four measurements were made from oscillograms of each production:-

M1: the time from the start of the utterance to the beginning of the release of C_1.

M2: the VOT of C_1.

M3: the time from the beginning of the release of C_1 to the beginning of the release of C_2.

M4: the VOT of C_2.

We shall concentrate on the data from the productions of voiceless stops. The average values of M1, M2, M3 and M4 for each of /pi/, /pa/, /ki/ and /ka/ are tabulated in Table 1. In Figure 1a values of M2 are plotted against M1 and in Figure 1b values of M4 are plotted against M3. Thus in each figure, average VOT is plotted against the average duration of the four immediately preceding syllables. Table 1 shows that stops in sentence-final syllables were produced with slightly longer VOTs than were stops in sentence-medial syllables. Apart from this, Figures 1a and 1b show essentially the same pattern and illustrate three results:-

(i) As we should expect from Lisker and Abramson's definitive (1967, 1970) spectrographic measurements, /k/s are produced with longer VOTs at each rate of speech and in each vocalic context than /p/s.

(ii) VOTs produced in voiceless stops increase as the rate of speech decreases. This is not surprising, although I do not think that it has been shown before. The pooled data disguise the fact that one of the six speakers did not display this pattern in his bilabial productions which showed no variation in VOT, but the general pattern was displayed in his velar productions and by all the other speakers in both their bilabial and their velar productions.

(iii) The single surprising result is the interchange of absolute VOTs between vowel and place contexts at all rates of speech. For these six speakers, at least, /ki/s were generally produced with longer VOTs than /ka/s while, paradoxically, /pa/s were produced with longer VOTs than /pi/s. An explanation for this result may be found by considering the conditions which must apply for normal voicing to occur. These are that the vocal cords should be adducted and a transglottal

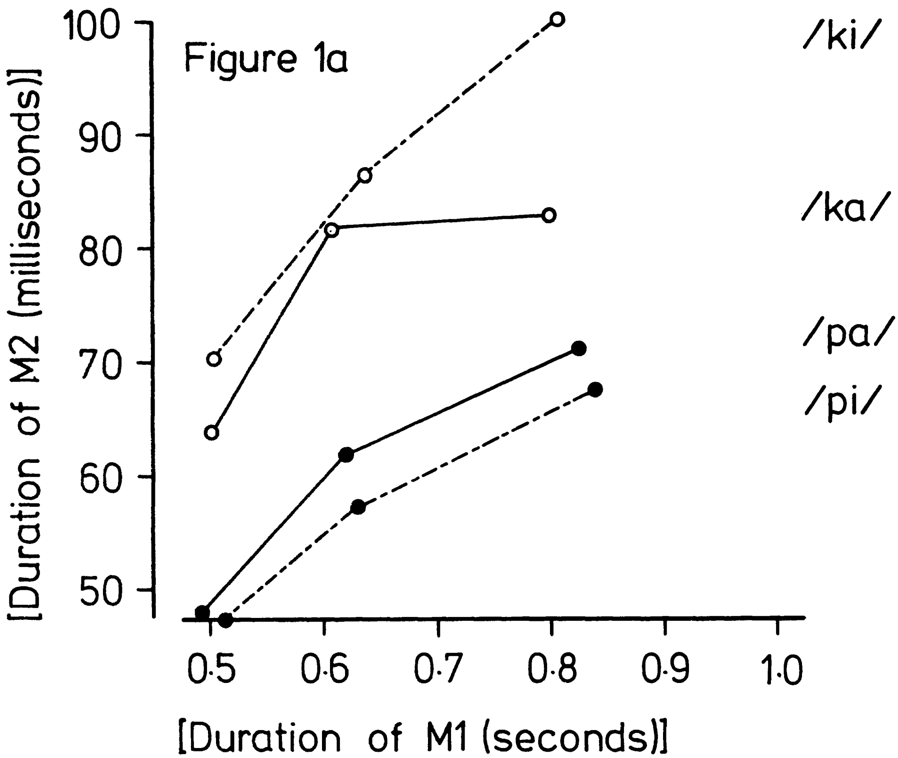

Figure 1a: Plots of data from sentence-medial productions (C_1V_1) of the syllables /pa/, /pi/, /ka/ and /ki/ from the sentence:

"Why are you a C_1V_1 when you're a C_2V_2?"

VOT in C_1V_1 (M2: milliseconds) is plotted against the total duration of the immediately preceding four syllables (M1: seconds) for each syllable at each of three rates of speech. Each point plots the mean of 90 data (15 repetitions by each of 6 speakers). The means are tabulated in Table 1.

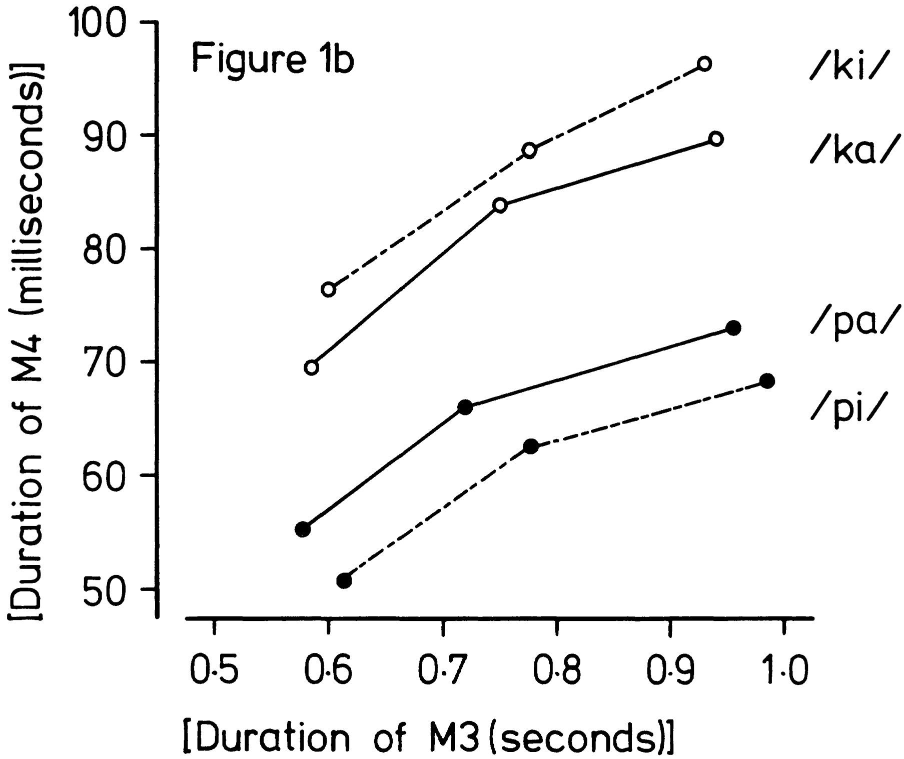

Figure 1b: Plots of data from sentence-final productions (C_2V_2) of the syllables /pa/, /pi/, /ka/ and /ki/ from the sentence:

"Why are you a C_1V_1 when you're a C_2V_2?"

VOT in C_2V_2 (M4: milliseconds) is plotted against the total duration of the immediately preceding four syllables (M3: seconds) for each syllable at each of three rates of speech. Each point plots the mean of 90 data (15 repetitions by each of 6 speakers). The means are tabulated in Table 1.

Table 1 Values in milliseconds of the measures M1, M2, M3 and M4 for each syllable averaged over six speakers.

M1 and M2 refer to sentence-medial productions (C_1V_1). M3 and M4 refer to sentence-final productions (C_2V_2). M1 and M3 are the total durations of the four syllables preceding C_1V_1 and C_2V_2. M2 and M4 are the VOTs of C_1V_1 and C_2V_2.

Sentence Medial Productions (C_1V_1)

RATE of Speech		/pa/	/pi/	/ka/	/ki/
FAST	M1	493	513	503	503
	M2	48.1	47.4	63.8	70.3
MEDIUM	M1	620	630	611	636
	M2	61.8	57.2	81.9	86.4
SLOW	M1	824	841	800	808
	M2	71.2	67.8	83.0	100.2

Sentence Final Productions (C_2V_2)

RATE of Speech		/pa/	/pi/	/ka/	/ki/
FAST	M3	578	614	583	600
	M4	55.3	50.8	69.6	76.4
MEDIUM	M3	718	777	750	777
	M4	66.1	62.8	84.1	88.0
SLOW	M3	955	984	939	931
	M4	73.1	68.4	89.7	96.4

pressure drop should exist. Thus, three factors will influence the timing of voicing onset relative to oral closure release: (i) most importantly, the relative timing of the commands for adduction relative to release; (ii) the physical separation of the vocal cords at the moment when adduction is initiated; and, (iii) the rate at which the oral pressure falls to establish the transglottal pressure drop. Klatt (1973) suggests that factor (iii) underlies the occurrence of longer VOTs in /ki/ compared to /ka/, since, before /i/ the tongue movement at release is almost tangential to the point of closure, while before /a/, it is almost perpendicular. It may also be an important determinant of the development of the VOT differences between the three places of production although these are presumably maintained actively in adults through factor (i). Kim (1970) has shown that factor (ii) accompanies the variations in VOT in the three categories of stops in Korean, at least in his own speech. It is this factor which is responsible for shorter VOTs in /pi/s compared to /pa/s despite aerodynamic considerations favouring the reversed pattern. With bilabials, in running speech, the tongue position for the vowel can be anticipated during the stop closure. With velars, much less anticipatory coarticulation can occur because the tongue body is involved in the closure gesture itself. Summerfield (1975b) argued that the high front position of the tongue for /i/ raises the hyoid which results in increased vocal cord tension. The same effect here results in less complete glottal opening and faster adduction in /pi/ compared to /pa/. In consequence, the aerodynamic factors which delay voicing onset in /ki/ are outweighed by mechanical factors which advance voicing onset in /pi/ relative to /pa/.

While this phenomenon and its explanation are of some interest, per se, the effect is small. It is important, however, because such a mechanical influence is only likely to appear in the production of stressed /p/ before a high vowel, while the aerodynamic factors should dominate for the other places of production and for all stops, including /p/, in consonant clusters. It is precisely the sort of 'exceptional' situation which a heuristic perceptual procedure would not allow for. If, as Lisker has suggested, one of the influences of a low frequency F1 at the onset of voicing is to cue voicedness through the implication of a more closed vocal tract, then we may test whether this cue is processed algorithmically by examining its effect in different vocalic contexts at each of the three places of production. If the closure cue is processed algorithmically then we should expect phoneme boundaries to occur at shorter VOTs on the continuum /bi-pi/ compared to /ba-pa/. If, on the other hand, the cue is processed heuristically, as seems likely, then we should expect the reverse pattern so that the boundary will always occur at a longer VOT in /i/ compared to /a/ context, just as both the present production data and the perceptual data of Summerfield (1974) find for velars.

Accordingly, Summerfield (1975a) determined phoneme boundaries on the nine synthetic VOT continua formed by the combination of three places of production and the vocalic contexts /a/, /ɜ/ and /i/ with steady-state F1 frequencies of 664, 504 and 277 Hz respectively. To demonstrate the generality of the syllabic rate influence on voicing perception, target stimuli were again presented under two conditions differentiated by the duration of the steady-state portions of precursor phrases. Two precursors were used. The longer, slower precursor had an overall duration of 1832 msec. and the shorter, faster precursor an overall duration of 1464 msec. The ratio of durations was, therefore, slightly smaller than in the previous experiment, being 1:1.26. In Table 2, the average phoneme boundaries of the sixteen subjects and the average shifts in the positions of these boundaries

Table 2 Average phoneme boundary positions and average phoneme boundary shifts between two rates of speech on nine VOT continua in milliseconds.

CONTINUUM	AVERAGE PHONEME BOUNDARY POSITION	AVERAGE PHONEME BOUNDARY SHIFT	DIFFERENCE BETWEEN /i/ AND /a/ BOUNDARIES	No. OF SUBJECTS
/ba-pa/	15.5	2.53		16
/bɜ-pɜ/	22.8	2.33	6.4	16
/bi-pi/	21.9	4.31		16
/da-ta/	17.1	1.90		16
/dɜ-tɜ/	23.3	1.94	9.2	16
/di-ti/	28.3	3.00		13
/ga-ka/	25.3	1.45		16
/gɜ-kɜ/	29.8	1.89	8.3	16
/gi-ki/	33.6	4.45		13

are tabulated in milliseconds of VOT.

We may note firstly that for every place of production the phoneme boundary in /i/ context occurred at a longer VOT than that in /a/ context. Table 2 tabulates the differences between /a/ and /i/ contexts which averaged 8.0 msec. and are statistically equivalent to one another. This suggests that the F1 closure cue is assessed heuristically in the sense that its interpretation is not influenced by higher order decisions concerning vowel identity as Summerfield (1974) had presumed. Secondly, Table 2 shows that in each place/vowel context, the effect of increased syllabic rate in the precursor is to cause a greater number of voiceless percepts; i.e., the phoneme boundary shifts, appropriately, toward lower VOT values.

While the 'heuristic/algorithmics' distinction is of some interest, for the purposes of the present symposium the demonstration of a prosodic influence on the perceptual interpretation of a segmental feature which parallels an equivalent variation in production, is more important. It is a moot point whether the size of the boundary shifts found in these perceptual experiments would be large enough to normalise differences in VOT occurring in production. In Figures 1a and 1b the ratio of the durations of the four syllables immediately preceding the CV syllables between the "Fast" and "Slow" speaking rates is 1:1.62. The difference in average voiceless VOT was about 20 msec. If variation in rate of speech does not significantly affect the duration of VOTs produced in voiced stops, which was the case here, in so far as it could be assessed, then we should expect a shift in the boundary position of half the shift in mean voiceless VOT; i.e., a shift of about 10 msec. in phoneme boundary should accompany a change in on-going syllabic rate of 62%. We have already noted that Summerfield (1974) found that an average shift of 5.5 msec. resulted from a syllabic rate variation of 33%, which is of the right size, therefore. However, the sizes of the boundary shifts tabulated in Table 2, which resulted from a syllabic rate variation of 26%, are rather small. I do not think that any hard and fast inference can be drawn from these comparisons. However, it would clearly be interesting to couple synthetic targets to real speech precursors spoken at different rates where I should expect the production/perception correlation to be more precise.

DISCUSSION

The precursor/target paradigm can obviously be extended and, I predict, would demonstrate speech rate influences on most other feature distinctions subsumed by durational cues. It would be of greater interest to determine whether the value of a particular duration cue in a precursor could act as the dominant criterion-setting influence for the same cue in the target. However, given that phoneme boundary shifts on the VOT dimension can indicate changes in the overall perceived rate of speech, the paradigm can be turned upon its head and then exploited in three ways.

Firstly, the paradigm could be used to determine the perceptual substrates of what I have rather vaguely been calling "syllabic rate" or "rate of speech". One component may indeed be the rate of syllables over time. Another could be the rate of fundamental frequency changes over time, though we have yet to test this notion. Before doing so, it is necessary to establish whether other contextual variables carried in precursor phrases, apart from syllabic rate, can influence the perception of voicing in the target. So far, I have only examined one such possibility. Summerfield (1975a) manipulated the vocal tract lengths implied by precursor phrases and target syllables in addition to the syllabic rate of the precursors.

I found that neither the position of the phoneme boundary on the dimension /bi-pi/, nor the size of the shift in its position which resulted from variation in precursor syllabic rate, were influenced either by variation in the vocal tract length implied by the precursor or by the vocal tract length relationship between the precursor and the target. The result implies either that the voicing cues in the target syllable are not interpreted in relation to the current perceptual vocal tract length reference, or that the target syllable itself supplies that reference. In either case, it seems that syllable timing information in the precursor does influence voicing perception in the target while other articulatory variables do not. The result encourages the use of the paradigm in investigating the perception of the parameters of 'syllabic rate' since it suggests that only these parameters influence the perception of target voicing.

Secondly, a parametric study which titrated the duration of the portion of the precursor abbutting the target on which duration variations were imposed, could determine the "time-window" over which syllabic rate is assessed. Summerfield (1975a) carried out a pilot study directed toward this end. It appears that the effective time window is less than one second in duration since synthesising the first half of a 2 second precursor at one rate (a) and the second half at another rate (b) produced exactly the same effect on the perception of voicing in the target as did synthesising both halves at the second rate (b).

Thirdly, and finally, the paradigm could be used to determine what perceptual expectations apply to syllabic durations in different sentential and syntactic environments. In the production data reviewed here, it was found that some speakers produced consistently longer VOTs in sentence-final syllables compared to sentence-medial syllabes. If the same precision of perceptual expectation applies here as in the situations reported by Klatt and Cooper (1975) then I should expect phoneme boundaries to occur at longer VOTs in sentence-final compared to sentence medial stop-vowel syllables. Supporting this hypothesis, Summerfield (1971) found that increasing the duration of the vowel increased the probability of the stop which preceded it being perceived as voiced. In natural speech vowel steady-states are the main site of pre-pausal lengthening (Lindblom, 1968; Oller, 1973). In the same way, I should expect even longer vowel durations to be required to cue voicedness post-vocalically in sentence-final syllables than in sentence-initial and sentence-medial syllables.

In all these situations, considerable interest lies in determining the precision of the perceptual expectation for variations in duration, which, given the demonstrably low variance of timing variation in production (Kozhevnikov and Chistovich, 1965; Lackner and Levine, 1975), can be expected to be high. Nevertheless, as with the effects of vocalic context reported here, the nature of perceptual processing may be powerfully exposed by a few exceptional cases, where perceptual effects and production constraints do not coincide.

REFERENCES

KIM, C.W. (1970). A theory of aspiration. Phonetica, 21, 107-116.

KLATT, D.H. (1973). Voice onset time, frication and aspiration in word-initial consonant clusters. M.I.T. Research Laboratory of Electronics, Quarterly Progress Report No. 109, 124-136.

KLATT, D.H. and COOPER, W.E. (1975). Perception of vowel duration in sentence contexts. Paper presented to the 89th meeting of the Acoustical Society of America, Austin, Texas.

KOZHEVNIKOV, V.A. and CHISTOVICH, L.A. Speech Articulation and Perception. JPRS 30, 543. Washington: U.S. Department of Commerce, 1965.

LACKNER, J.R. and LEVINE, B.K. (1975). Speech production: evidence for syntactically and phonologically determined units. Perception and Psychophysics, 17, 107-113.

LINDBLOM, B. (1968). Temporal organisation of syllable production. Speech Transmission Laboratory Quarterly Progress Report, 2/3, 1-5. (Royal Institute of Technology, Stockholm.)

LISKER, L. (1974). Is it VOT or a first formant transition detector? Paper presented to the autumn meeting of the American Association of Phonetic Sciences.

LISKER, L. and ABRAMSON, A.S. (1967). Some effects of context on Voice Onset Time in English stops. Language and Speech, 10, 1-28.

LISKER, L. and ABRAMSON, A.S. (1970). The voicing dimension: some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences. Prague, 1967. Prague: Academia.

OLLER, D.K. (1973). The effect of position in utterance of speech segment duration in English. Journal of the Acoustical Society of America, 54, 1217-1247.

PETERSON, G.E. and BARNEY, H.L. (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 175-184.

SUMMERFIELD, A.Q. (1971). Some tests of an information-flow model of speech perception. Unpublished Bachelor's Thesis, University of Cambridge.

SUMMERFIELD, A.Q. (1974). Towards a detailed model for the perception of voicing constrasts Speech Perception No. 3, 1-26. (Progress Report, Department of Psychology, The Queen's University of Belfast.)

SUMMERFIELD, A.Q. (1975a). Cues, contexts and complications in the perception of voicing contrasts. Speech Perception No. 4, (in press). (Progress Report, Department of Psychology, The Queen's University of Belfast.)

SUMMERFIELD, A.Q. (1975b). Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables. Speech Perception No. 4, (in press). (Progress Report, Department of Psychology, The Queen's University of Belfast.)

SUMMERFIELD, A.Q. and HAGGARD, M.P. (1974). Perceptual processing of multiple cues and contexts: effects of following vowel on stop consonant voicing. Journal of Phonetics, 2, 279-295.

TAYLOR, M.M. and CREELMAN, D.D. (1967). PEST: Efficient estimates on probability functions. Journal of the Acoustical Society of America, 41, 782-787.

DISCUSSION of "How a full account of segmental perception depends on prosody and vice versa"
SUMMERFIELD

KLATT:
Your physiological explanation for the slightly shorter VOT in /pi/ than in /pa/ in the production data is not consistent with other data. It is probably true that the larynx is more raised in the /pi/ environment and that the vocal folds are thereby stiffened, but this seems to inhibit voicing rather than facilitate an earlier voicing onset. Evidence includes the devoicing of Japanese high vowels in these environments, studies of laryngeal preparations, and theoretical arguments (Stevens, VIIIth Congress of Phon. Scs., Leeds 1975).

SUMMERFIELD:
I agree. The /pi/ - /pa/ inversion is paradoxical. However, the intention of my paper was not really to provide a detailed physiological explanation for such paradoxes in production but, rather, to suggest the existence of a continuously estimated perceptual duration, or rate, parameter. Not only could this be useful in maintaining a rate-dependent phonetic distinction as here, but it could interact with knowledge of the number of syllables in a word to predict the level of that word in the parsing tree or vice versa. While the perceptual paradigm appears to provide a very sensitive measure of this parameter's extraction (as also in SIEB NOOTEBOOM's paper), it would be interesting to cross-calibrate it against a prosodic measure.

LEHISTE:
I would like to suggest an explanation for the fact that you got such a short VOT for voiceless plosives in isolated syllables. It is possible that people expected prevoicing for voiced plosives produced in syllables presented in isolation.

SUMMERFIELD:
This question relates to some data which I presented verbally at the Symposium but which was not included in the text above. I had determined the positions of phoneme boundaries on a synthetic /gi-ki/ VOT continuum both when the target stimuli were introduced by pre-

cursor phrases and when they were presented in isolation. The rate of speech of the precursors was varied over a considerably greater range than the variations referred to in the experiments reported above and exposed the limits on the size of phoneme boundary shift that could be obtained by such speech rate variation. Surprisingly, for 5 of my 6 subjects, boundaries estimated in isolation coincided with boundaries estimated when the same target continuum was introduced by the precursor phrase at the fastest rate of speech. The 6th subject's isolated boundary coincided with her boundary estimated at the slowest rate of speech. While I realise that stop consonants have different VOT characteristics when produced in isolation as opposed to sentence contexts. I think that this result reflects a major difference between the perceptual processing of speech relevant durations in and out of sentence context as well as a dramatic difference between individual perceivers, rather than a specific perceptual expectancy for prevoicing.

FISCHER JØRGENSEN:

I am not sure that the relation you have found between the VOT-values of pi and pa is universal, and I doubt whether your physiological explanation is correct. If it is, pu should also have a shorter VOT-value than pa, but according to my experience pu has normally a longer VOT-value than both pa and pi.

I also think that the difference between k and p have aerodynamic and physiological explanations: the smaller subglottal cavity and the slower movement of the tongue compared to the lips will have the effect that it takes longer time for the supraglottal air-pressure to come down to a level permitting vocal chord vibrations.