LINGUISTIC AND LITERARY STUDIES

IN HONOR OF

ARCHIBALD A. HILL

edited by

MOHAMMAD ALI JAZAYERY
EDGAR C. POLOMÉ
WERNER WINTER

Volume I

GENERAL AND
THEORETICAL LINGUISTICS

LISSE
THE PETER DE RIDDER PRESS
1976

ON LEARNING A NEW CONTRAST

LEIGH LISKER

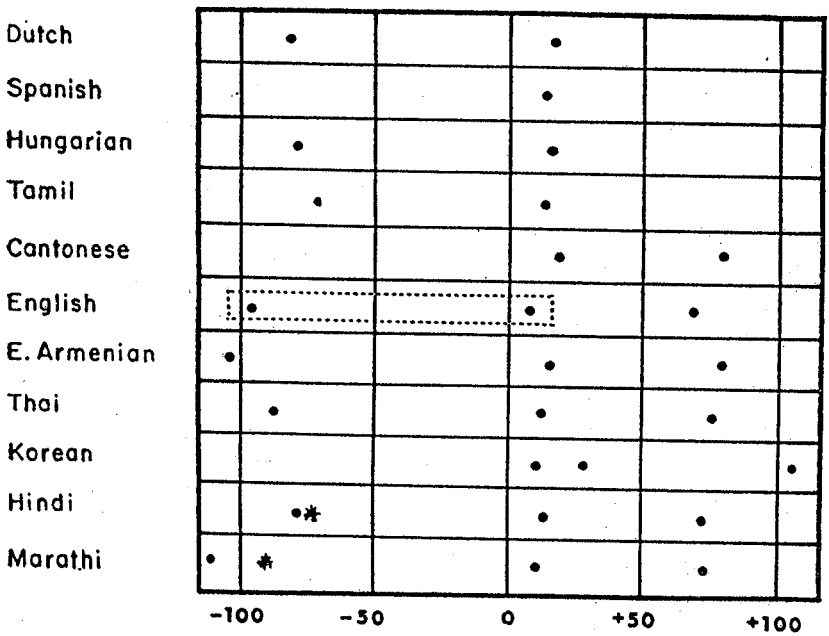
Descriptions of the speech behavior of human beings, both their management of the vocal tract and their perceptual processing of its audible output, require data on a wide variety of speakers if we are to separate the biological and cultural factors which govern speech activity, or, perhaps more realistically, if we are simply to distinguish between the features which characterize speech generally and those specific to particular kinds of speech. One obvious way of checking on the degree of universality of generalizations concerning speech behavior is to compare speakers of diverse languages. We would suppose, if they manage phoneme inventories that differ in size and distinctive phonetic properties, that this is no more to be connected with physical characteristics of the speech-producing and speech-perceiving mechanisms than are grammatical differences in the languages which their speech 'implements'. Phonetic differences between two languages presumably reflect either different choices from some general inventory of phonetic dimensions which are, in principle, equally available to all language users, or different ways of exploiting the same phonetic dimensions. Thus, for example, the feature of glottalization may serve to differentiate consonants in one language and not in another, whose speakers nonetheless might readily distinguish, if they had to,[1] between glottalized and unglottalized consonants. In another case, two languages might make use of very much the same range of vowel sounds, but differ as to just how these are grouped into categories. Here the problem for the speaker of one language learning the other would be to adopt new criteria for deciding which sounds were the same and which different.

In comparing the language behavior of speakers of diverse languages we may follow certain psycholinguistic testing procedures which involve speech or speech-like auditory stimuli. Differences in test performance may in general be taken to reflect differences in the subjects' linguistic backgrounds, i.e., they represent an effect of learning. What is not quite

clear is exactly what it is that was learned, or the extent to which this learning may be said to have affected permanently the ability of speakers of one language to learn to match the performance of speakers of another. Over the past dozen years researchers at the Haskins Laboratories have been testing subjects, for the most part speakers of American English, to determine a relation between their linguistic identification of synthetic speech stimuli and their ability to detect the acoustic differences between individual test stimuli. In tests involving the presentation of steady-state synthetic vowel sounds (Fry, Abramson, Eimas and Liberman 1962, Stevens, Liberman, Studdert-Kennedy and Öhman 1969) no very close connection was found between subjects' identification of stimuli with English vowels and their ability to distinguish them in a conventional ABX test; items labelled alike were almost as easy to discriminate in this test as were items labelled differently. On the other hand, for a stimulus set whose members were distributed among the categories *ba*, *da*, and *ga*, it appeared that subjects were able to distinguish only those items which they assigned to different categories (Liberman, Harris, Hoffman and Griffith 1957). Thus the results of testing for discrimination of isolated vowel and initial stop appeared to be radically different.[2]

Unfortunately, in the testing program just referred to, no very extensive cross-language data have as yet been collected. However, what has so far been done in this area does not weaken the notion of a quite different relationship between labelling and discrimination behavior for vowels and stop consonants. In one vowel study (Stevens, Liberman, Studdert-Kennedy and Öhman 1969) American and Swedish subjects were compared in respect to the labelling and discrimination of certain vowel-like stimuli that have very different categorial status in English and Swedish. It was found that linguistic experience, as reflected in differences in the way in which Swedes and American classified the test stimuli, had little apparent effect on their ability to discriminate. A cross-language consonant study (Abramson and Lisker 1970, Lisker and Abramson 1970), of which the present paper is a continuation, involved the comparison of several groups of speakers with respect to the dimension of voicing as this serves to distinguish between categories of stops in synthesized consonant-vowel syllables. Comparison of the labelling and discrimination behavior of groups of English, Spanish and Thai-speaking subjects showed differences in the use of voicing as a feature whereby stop categories in the three languages are phonetically distinguishable. It will be useful here to review briefly the background and findings of this study.

In very many of the world's languages distinctive use is made of two, and sometimes three, categories of initial prevocalic stop consonants which differ, among other things, in the extent to which the larynx participates in their production, as this can be measured by determining the time of voice onset relative to that of release of the occlusion. Spectrographic examination of the word-initial stops in a number of languages (Lisker and Abramson 1964) has shown that there are significant differences in voice onset timing ('VOT') from language to language, but that the placement of category boundaries along this dimension is hardly random. Measurement data (Fig. 1) derived from productions of isolated words in a dozen languages suggest that there are three preferred timing relations between voicing onset and stop release: voicing begins almost one hundred milliseconds before release (VOT $\cong -90$); it begins at or just after the release (VOT $\cong +10$); or it begins well after the release (VOT $\cong +75$). These values, we may assume, correspond respectively to the voiced, voiceless unaspirated and voice-



Mean voicing-onset timings for stops in word-initial position (from Lisker and Abramson 1964). Stop release is at O on abscissa, which represents no. of milliseconds by which voicing onset precedes (negative values) or follows (positive values) release. Starred entries are for voiced aspirates.

Fig. 1

less aspirated stops of classical descriptive phonetics. For all but one of the languages in our sample, Korean, it can be said that they differ essentially in the number and selection of stops from this set of phonetic categories. Thus Dutch, Spanish, Tamil and Hungarian each make use of the two categories at VOT $\cong -90$ and VOT $\cong +10$. Cantonese differs from these languages in that, while it too has two stop categories, they involve VOT values at about +10 and +75. Languages with three categories along the VOT dimension (Eastern Armenian, Thai, Hindi and Marathi, *but not* Korean) simply select all three of the categories described. Two of the languages examined are anomolous, Korean and English. Korean is a three-category language, but in initial position all of its stops are voiceless, i.e., voicing begins only with or following release, with VOT values at roughly +10, +30 and +100. English is peculiar in being a two-category language that utilizes all three VOT values, but does not distinguish initially between stops with voicing lead and those for which VOT = +10. The choice between the two appears to be, for American speakers of English at least, a choice that is partly idiosyncratic (Lisker and Abramson 1964:395) and partly a matter of style (Lisker and Abramson 1967:20-24).

In our comparison of different languages with respect to the timing of voice onset we have been talking as though the only differences among the stop categories being compared are those of voice onset timing. Acoustically, of course, this is very far from being the truth; the effect of a small change in the timing of voice onset is very different, depending on whether onset precedes, coincides with, or follows the stop release. What we have been calling the VOT continuum is, at best, a continuum only in the articulatory domain, provided we make the no-doubt oversimplifying assumption that a series of syllables such as [ba, ḇa, pa, p'a, p^ha] can be produced by executing an articulatory program that is invariant in respect to those components which effect the supraglottal gestures, and which differs only in respect to those which determine when the laryngeal signal begins. This assumption, which derives from Dudley's well-known model of vocal tract operation (Dudley 1940), is most unlikely to be justifiable in detail. However, it is true enough to be convenient, for it provides the rationale for a relatively simple program for the synthesis of stop-vowel syllables. For a particular syllable in which a fixed spectral pattern determines, for example, that it will be identified as consisting of a labial stop and the vowel [a], the voicing state of the stop is controlled by specifying the time at which the signal 'exciting' the pattern shifts from one of aperiodic to one of periodic type. Thus for each place of

stop closure a series of syllables could be generated whose initial consonants ranged from fully voiced (with onset of pulsing well before the burst marking stop release) to heavily aspirated voiceless stops (with pulsing onset considerably after the burst). Spectrograms of sample syllables, produced by means of the Haskins Laboratories' parallel resonance synthesizer under computer control (Mattingly 1968), are illustrated in Fig. 2. The upper pattern illustrates the case in which the spectral pattern is excited from start to finish by a periodic signal. Four successive segments of the pattern may be distinguished: an initial segment characterized by a single very-low-frequency formant with a duration of 150 msecs, which corresponds to the articulatory closure; a burst or transient of about 10 msecs, which corresponds to the release of the stop; an interval of about 50 msecs with three formants of shifting frequencies, which 'transition' corresponds to the interval of articulatory movements from closed stop to open vowel state of the oral cavity; a final segment with three formants at fixed frequencies for 450 msecs, corresponding to the vowel [a] as produced with 'steady-state' articulation. The pattern immediately below represents the case where the periodic excitation begins just after the burst, while the lower pattern shows the synthesizer output where the same excitation begins one hundred milliseconds after the burst. In both of the latter patterns the interval beginning with the burst and ending with the onset of pulsing is excited by an aperiodic signal. Particularly in the lowermost pattern, which listeners identify as a syllable beginning with a heavily aspirated stop [p^h], we can observe the feature of 'first-formant cutback', i.e. complete suppression of the first formant over the interval of noise excitation of the burst and upper formants. This feature must be considered *not independent* of the choice of excitation type (Liberman, Delattre and Cooper 1958). We may suppose that the association between this spectral difference and the voicing dimension arises not only because the larynx is a signal source, but because as a part of the cavity system of the vocal tract its state helps determine the resonance properties of the tract. One might then associate with the absence of pulsing a large attenuation of the first formant, provided a fairly large glottal aperture during the interval between release and voicing onset is assumed.[3] Equally well, perhaps, one might suppose the transmission characteristics of the tract to be essentially identical for periodic and aperiodic source signals, but that the aperiodic source is deficient in intensity over the frequency range below the second formant. In any case, a close relation, both in production and in perception, between the onset of pulsing and the fairly rapid

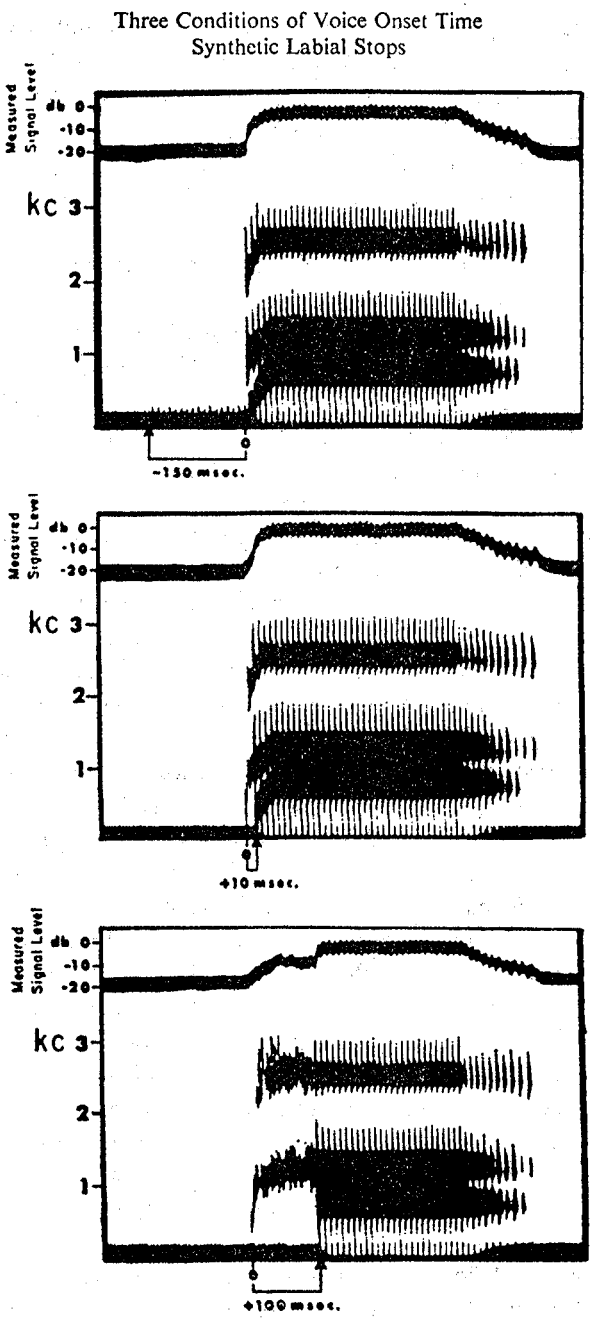

Three Conditions of Voice Onset Time
Synthetic Labial Stops
Measured Signal Level
db 0
-10
-20
kc 3
2
1
-130 msec.
+10 msec.
+100 msec.

Fig. 2

development of first-formant intensity to a level appropriate to the following vowel was established from spectrographic study of real speech and on the basis of preliminary perception testing of synthetic stop-vowel patterns in which the VOT and first-formant cutback features were independently manipulated. Consequently, in our further discussion of what we have been calling the 'VOT dimension', it is to be understood that the onset of pulsing is regularly accompanied by the simultaneous onset of the first formant.

As part of our cross-language study of stop voicing, three series of synthetic speech patterns of the type illustrated by Fig. 2 were generated, in which VOT was systematically varied over a 300 msec range, from a value of −150 (pulsing onset 150 msec before the burst) to one of +150 (pulsing onset 150 msec after the burst). These stimuli, in various appropriate random orders, were presented to speakers of three of the languages for which we had real-speech VOT measurement data. The languages chosen were English, Spanish and Thai, the first two having two categories of stops each, and the third being representative of 'three-category' languages. Two kinds of data were gathered: labelling responses and something we called 'discrimination' data. The procedure which yielded the labelling data involved asking subjects, native speakers of the languages mentioned, to name the initial stop of a stimulus by identifying it with one or another of the initial stops in their language. The labelling data obtained by presentation of our labial series of stimuli are represented by the curves in Fig. 3. The responses which we took to reflect subjects' ability to discriminate between items of the stimulus set were collected by the following procedure. Stimulus triads were composed of two items that were identical in VOT value and a third differing from these by 20, 30 or 40 msecs along the same dimension. The order of presentation of members was random, with all possible orderings equally represented in the full set of triads submitted to the subjects. The subjects' task was to identify the 'odd ball' as the first, second or third member of the triad. Representative results are given in Fig. 4, which shows discrimination functions for one English-speaking and one Thai-speaking subject.

In the procedure by which labelling responses were obtained as a function of VOT values, the possibility that subjects might recognize more categories than their language possessed was not taken into account. From the discrimination task and the data thereby obtained, it appeared that discriminability is not significantly better than chance for stimulus pairs which are categorially identical, but increases sharply for pairs

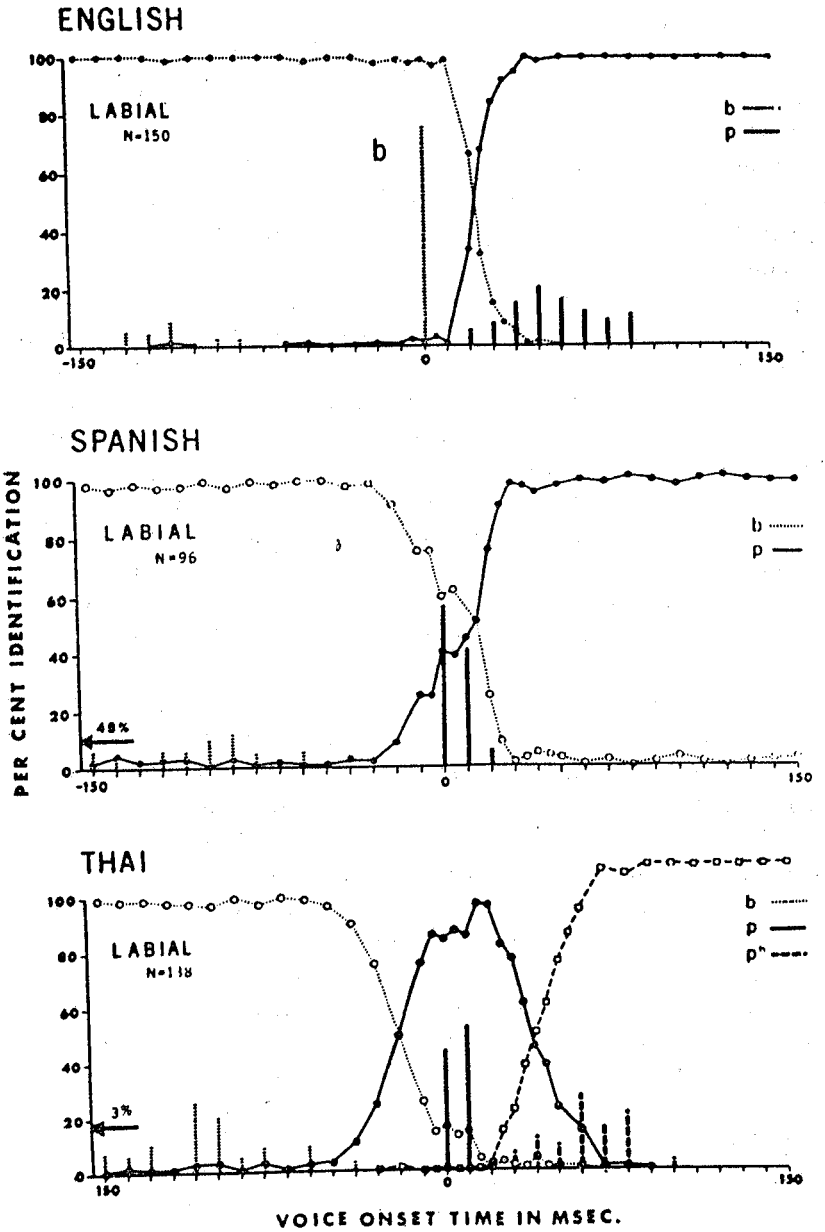

ENGLISH
LABIAL
N=150
b
b
p
SPANISH
LABIAL
N=96
b
p
48%
PER CENT IDENTIFICATION
THAI
LABIAL
N=118
b
p
pʰ
3%
-150
0
150
VOICE ONSET TIME IN MSEC.

Fig. 3

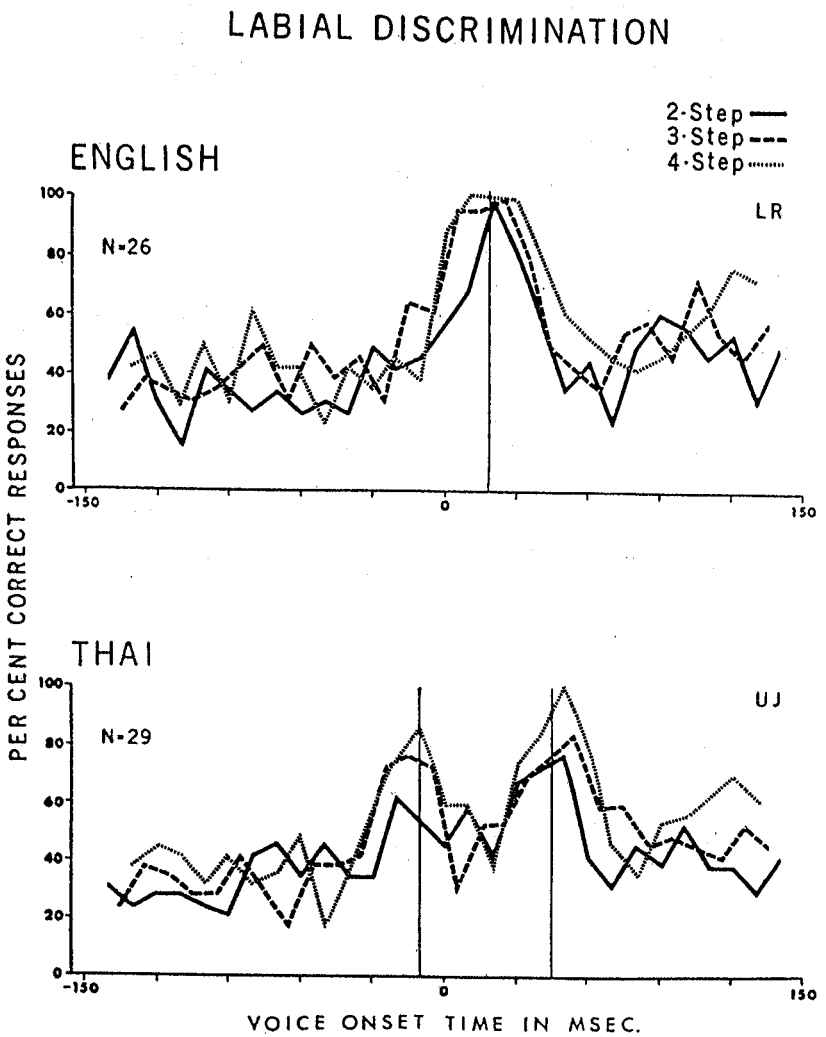


Fig. 4

located near the boundary between categories along the VOT dimension. Insofar as the locations of these discrimination peaks differ for the speakers of different languages, and indeed sometimes for individual subjects, matching thereby the boundaries between linguistic categories established by the labelling tests, it would seem difficult to decide whether the failure to make subcategorial discriminations means that subjects cannot or simply that they did not make such discriminations, given a test in which

some comparisons were across a category boundary and others were not. One major purpose of some additional tests carried out at the speech research laboratory of the Pavlov Institute near Leningrad[4] was to learn whether speakers of Russian, a two-category language with voiced and voiceless unaspirated stops, can readily distinguish between stimuli which, for speakers of English, are categorially different, but which for Russians are of the same category. There was, unfortunately, no opportunity to make VOT measurements of spoken Russian stops comparable in quantity with the large body of cross-language data presented in Lisker and Abramson 1964, but a modest quantity of labelling and discrimination data was collected.

In order to obtain VOT labelling data from Russian speakers the same synthetic speech stimuli tested previously with American, Puerto Rican, and Thai subjects were presented to a group of fifteen members of the Pavlov Institute staff. Their responses to these stimuli are given in Fig. 5. Although Russian is a two-category language, with stops resembling those of Spanish and Hungarian, stimuli with VOT values greater

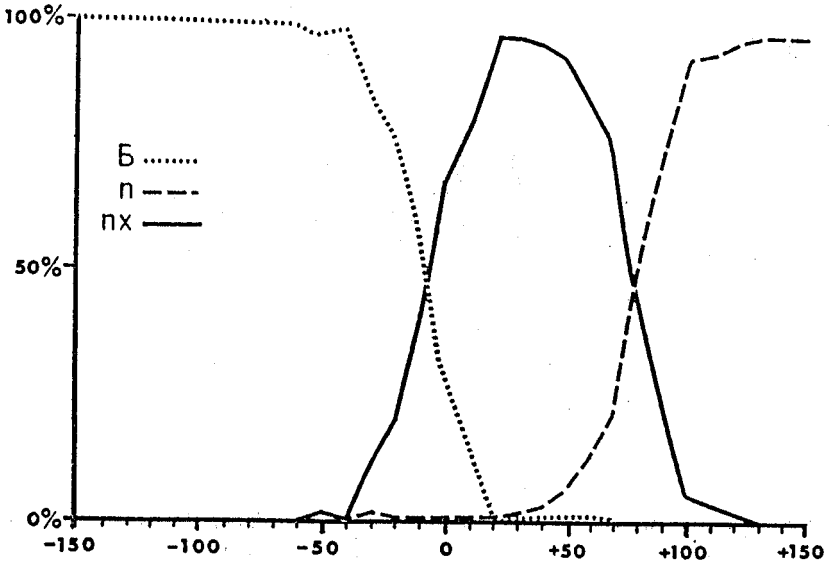


Fig. 5

than +80 were judged to begin with the cluster *px*, i.e. the voiceless bilabial stop followed by a voiceless velar fricative. The labelling behavior of our Russian-speaking subjects may be compared with the data obtained from speakers of the three other languages previously mentioned (Fig. 3). The differences from language to language are considerable, even if we allow for variations, probably minor, due to the fact that listening conditions could not be rigorously controlled. Presumably the fact that English speakers divide the VOT space into *b* and *p* categories at +25, while for Russians the crossover point dividing *b* from *p* is close to −10, is of significance, particularly in view of the fact that this difference is observed in actual speech production as well. At the same time it should be remarked that the match with production data is not always very good, specifically in the Spanish case and in the *p*-*p*^{h} boundary in Thai.

Since English speakers identify stimuli with VOT values in the range −10 ... +25 with items of lower (i.e. more negative) values, while Russians identify them with items having higher values, the question arises as to whether this difference means that acoustic cues available to both groups are assessed according to different strategies, or whether the cues to which one group attends are simply not available to the other. Is it the case, for example, that Russian listeners are quite capable of distinguishing between items at +20 and +50, although both are *p* for them, and that Americans can hear the difference between VOT values of −30 and 0, though both are labelled *b*? If one group is able to discriminate between stimuli which are categorially different only for the other, then the implication would be that there is a psychoacoustic basis for the category boundary.

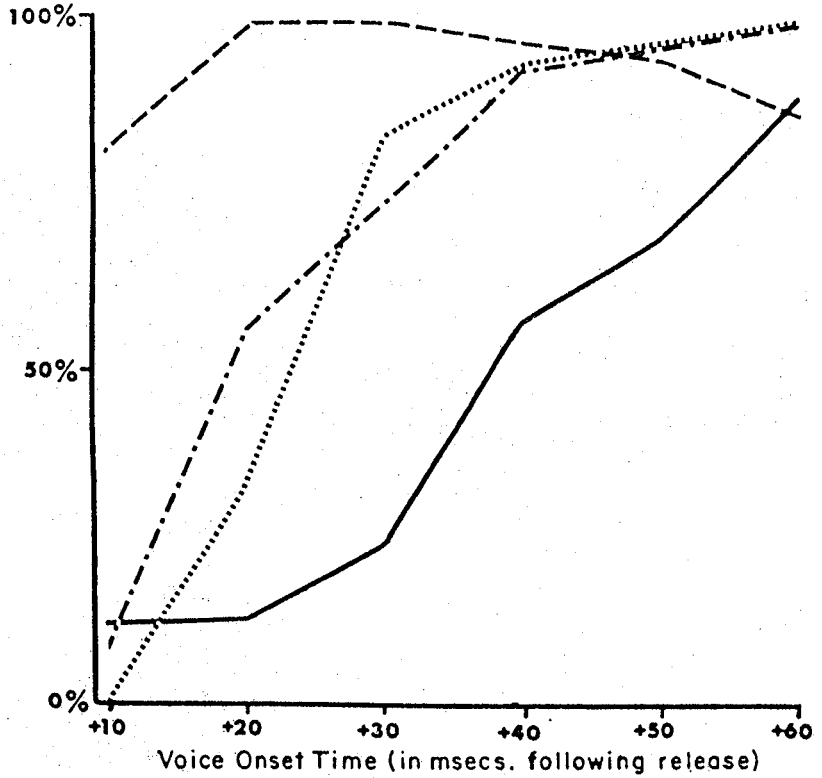
In order to learn whether the boundary at VOT = +25 between English *b* and *p* is susceptible of detection by speakers of Russian, the following experiment was carried out. Items having VOT values of +10 and +60, both of them *p*, were presented to a group of Russian speakers who were trained to assign different labels to them, i.e. to move a toggle switch one way for +10 stimuli and the other way for +60 stimuli. Subjects' success in learning this task was ascertained by presenting the two stimuli many times in random order, simultaneously registering their responses by means of an electromechanical recording system. It appeared that the test group was able to do significantly better than chance, with a majority of the six subjects tested getting above 90% correct in identifying the +10 items. Identification of the +60 stimuli was less good, but still better than 75% correct for all but one subject.

Thus it seemed that the subjects as a group could both distinguish the two test stimuli and also apply two different labels to them in a reasonably consistent way.

A second labelling test was next constructed in which was presented a set of stimuli covering the range from +10 to +60 in steps of 10 msecs. The test subjects' task was to identify each stimulus by judging whether it was more similar to the +10 or the +60 item. Each stimulus was represented five times in the random order presentation, and the entire set of 30 items was administered to the same group of subjects repeatedly over several days.

The responses of our subjects in the labelling test just described are represented by the solid line in Fig. 6, which shows the percentage of +60 identifications as a function of VOT value. This labelling function is to be compared with the two functions representing the behavior of a group of English-speaking subjects tested subsequently in the United States: the dotted line in Fig. 6 gives percentage *p* judgments derived from tests in which stimuli covering the full −150 ... +150 range were presented, while the dot-and-dash line in the same figure gives responses to the restricted set ranging from +10 to +60 along the VOT dimension. (The dashed line, representing the Russians' *p* labellings for the full VOT range, is included in Fig. 6 for ease of comparison.) It is apparent that the Russian and the American English data do not closely match, and one is tempted to believe that, by and large, the Russian listeners were making continuous rather than categorical judgments, i.e., that they were estimating the magnitude of the difference between a given stimulus and each of the standard stimuli rather than deciding whether or not it shared some feature with one of the standards. The Americans' judgments were very much the same in the two labelling tests; whether they were dividing the full VOT range into *b* and *p* categories or the restricted range by matching with the +10 as against the +60 stimuli, their judgments were evenly divided at about +20. The Russian judgments, by contrast, show a crossover somewhere between +30 and +40, i.e. at about the midpoint of the +10 ... +60 range; within this same range, of course, the curve representing the partition of the full VOT range into *b* and *p* categories does not approach the 50% value on the ordinate. When we look at the behavior of individual subjects, moreover, we find a marked difference in the degree of variability for the two groups; the American subjects are noticeably more alike in their division of the restricted range of stimuli than are the Russians (Fig. 7), who place their boundaries anywhere between +20 and +50. It may possibly be true that

Comparison of Russian and American Labelling of Full and Restricted Ranges Along the VOT Dimension



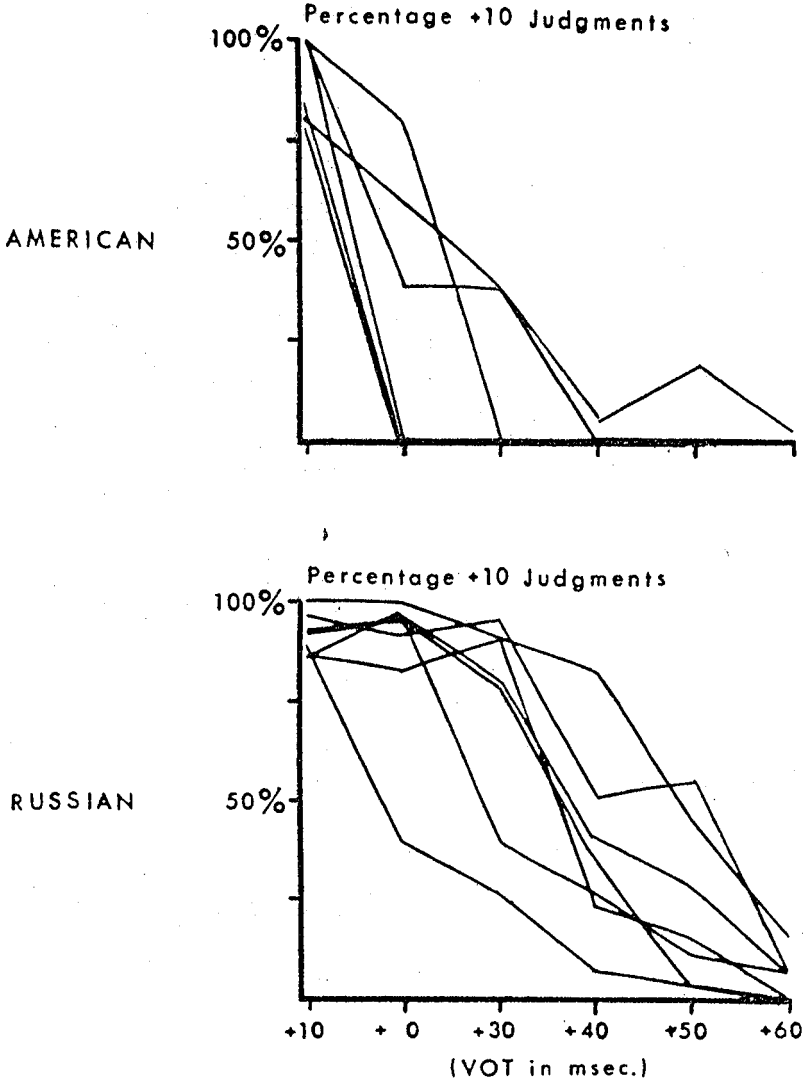
Key ——— : Russian identifications of restricted range (+10..+60) as +60

– – – – : Russian identifications of full range (-150..+150) as /ɲ/

—·—·— : American identifications of restricted range (+10..+60) as +60

····· : American identifications of full range (-150..+150) as /p/

Fig. 6



Individual Variation in Identification of Stimuli with +10 as against +60

Fig. 7

the single Russian listener who observed a crossover value near +20 was following the Americans' strategy, but the Russians as a group were certainly not attending to the same cues as the latter. On the other hand, it is possibly only accidental that the Russian crossover near +40 is very nearly at the midpoint of the +10 ... +60 range, so that we cannot be certain that they were estimating difference magnitudes rather than responding categorically to some acoustic cue. There is the possibility, moreover, that this crossover value near +40 is to be related to one of the crossover values determined for our Thai subjects in the full-range labelling test (Fig. 3). What we can be reasonably certain of, on the basis of our present data, is that our Russian listeners did *not* generally observe the boundary which served the American subjects in both labelling tasks. It would be appropriate to determine Russian crossover values for several additional VOT ranges, e.g. +20 ... +70 and +30 ... +80, that fall within the *p* category, in order to learn whether the crossover values remain fixed or tend to move with the range boundaries. In the first event we should be in a position to assert that the Russian listeners were evaluating the stimuli, in categorical fashion, according to acoustic criteria other than those motivating the American listeners; in the second event we should have to suppose that a continuous kind of perception and comparison was being practiced.

University of Pennsylvania

NOTES

[1] As in a phonetic or psycholinguistic exercise, for example. If subjects discriminate between items they call 'the same' so far as differentiating words of their language, then this counters the view that linguistic coding always intervenes between the peripheral processing of the acoustic signal and the execution of the discrimination task.

[2] Whether these differences can be taken as evidence for the motor theory of speech perception may be regarded as doubtful, since the acoustic variables involved in the vowel and consonant studies differ markedly. Nor can one readily assume that conclusions based on tests using steady-state vowel patterns will be valid for the perception of vowels in running speech.

[3] Current work involving motion picture photography of the glottis via fiberoptics indicates that this is regularly the case during production of voiceless aspirated stops, at least for English.

[4] These tests were conducted by the writer as a guest researcher at the speech laboratory of the Pavlov Institute of Physiology, which he visited under the auspices of the cultural exchange program of the National Academy of Sciences of the U. S. and the Academy of Sciences of the U.S.S.R. The work reported here could not have been accomplished without the generous cooperation of Dr. L. A. Chistovich and her colleagues of the Pavlov Institute.

REFERENCES

Abramson, A. S., and Lisker, L. 1970. Discriminability along the voicing continuum: cross-language tests. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967.

Dudley, H. 1940. The carrier nature of speech. Bell System Technical Journal 19:495-515.

Fry, D. B., Abramson, A. S., Eimas, P. and Liberman, A. M. 1962. The identification and discrimination of synthetic vowels. Language and Speech 5:171-189.

Liberman, A. M., Harris, K. S., Hoffman, H. S. and Griffith, B. C. 1957. The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psych. 54, No. 5:358-368.

Liberman, A. M., Delattre, P. C. and Cooper, F. S. 1958. Some cues for the distinction between voiced and voiceless stops in initial position. Language and Speech 1:153-167.

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

—. 1967. Some effects of context on voice onset time in English stops. Language and Speech 10:1-28.

—. 1970. The voicing dimension: some experiments in comparative phonetics. Proceedings of the 6th Congress of Phonetic Sciences, Prague, 1967, 563-567.

Mattingly, I. G. 1968. Experimental methods for speech synthesis by rule. IEEE Transactions. Audio 16:198-202.

Stevens, K. N., Liberman, A. M., Studdert-Kennedy, M. and Öhman, S. E. G. 1969. Crosslanguage study of vowel perception. Language and Speech 12, Pt. 1:1-23.