THE PERCEPTION OF VOICE ONSET TIME: A CROSS-LANGUAGE STUDY OF AMERICAN ENGLISH AND DANISH

JENS B. CHRISTENSEN

This paper examines the relationship between the production and perception of the voicing distinction for American English and Danish stop consonants in initial position. In a comparison of tne production of /p/ and /k/ for the two languages the Danish stops were found to have longer aspiration. It was therefore hypothesized that perceptually, Danish listeners would have a later cross-over point than American listeners. This was tested in a labelling experiment, using computer-edited, naturally produced stimuli. The aspiration was shortened from right to left to produce a series of stimuli ranging in VOT from +10 to +70 msec of the syllables <u>*pi*</u>*,* <u>*pu*</u>*,* <u>*ki*</u>*, and* <u>*ku*</u>*. The listening tests showed a statistically significant difference in the expected direction for the labial stimuli but not for the velars. This may be due to the stimulus range which proved to be less appropriate for Danish listeners than for American listeners.*

I. INTRODUCTION

Voice onset time, or VOT, the temporal relation between the release of a stop consonant and the onset of vocal fold vibration has been shown to characterize the phonological contrasts of homorganic stops in initial position for most languages. On the basis of spectrographic analyses of naturally produced syllables Lisker and Abramson (1964) proposed three categories into which these languages seem to group their stops: 1) voicing lead, assigned negative values, 2) coincident and short lag, assigned zero or low positive values, and 3) long lag, assigned high positive values. The stops in this last category have traditionally been called aspirated.

Their main objective for setting up these categories was to find "some single best measure" by which to separate the phoneme categories, and do away with the traditional concept of voicing, aspiration, and force of articulation as three mutually independent dimensions. They note that they have not been able to find any language where force of articulation would stand as one single feature, separating phoneme categories of stop consonants. They consider force of articulation, or fortis/lenis, to be closely connected with aspiration. One disconcerting factor which supports Lisker and Abramson's attempt to eliminate fortis/lenis is that phoneticians are still searching to find the physical, or acoustic correlate which will adequately describe this feature. See e.g. Fujimura and Miller (1979) and Kohler (1983).

It is well documented that VOT is a sufficient cue for listeners to differentiate between the phoneme categories found in their native languages according to the categories proposed by Lisker and Abramson: for American English and Thai by Lisker and Abramson (1970), for Spanish, Abramson and Lisker (1973) and Williams (1977), for Polish by Mikós, Keating and Moslin (1976) and Keating, Mikós and Ganong III. (1981).

Other cues have been shown to be operative in the perception of the contrast between stop categories; a cutback of the first formant, Liberman, Delattre and Cooper (1958), and the transition of the first formant, Stevens and Klatt (1974), (see also Lisker, 1975), as well as the fundamental frequency at the onset of voicing, (Haggard, Ambler and Callow, 1970), (Abramson and Lisker 1983).

Common to almost all cross-language studies is that they have compared languages which in their voicing contrasts differ across VOT categories. English has been contrasted with Thai, which shows a three-way contrast, with Spanish and Polish, both contrasting voicing lead with voicing lag. The theory of VOT has influenced work in other areas of linguistics, e.g. child language studies, where VOT was used to investigate possible innate feature detectors in infants. In a study Eimas, Siqueland,Jusczyk, and Vigorito (1971) found that one and four month old babies were able to discriminate between pairs of stimuli of +20 and +40 msec, but were not able to discriminate between pairs of stimuli that both had VOT values of either <20 or >40 msec. Eimas et al. therefore concluded that the 20-40 msec of voicing lag constituted a natural boundary, which incidentally is found in American adult studies. Support for this claim was found in a study on chinchillas (Kuhl and Miller, 1978) where these in a labelling test showed nearly identical category boundaries as those found in humans.

The cross-language studies clearly show that it is possible for humans to learn various sets of categories. Spanish and Polish listeners learn to distinguish between lead and lag, rather than short and long lag, and Thai listeners learn to distinguish among three categories, lead, coincident, and lag for the labial and alveolar places of articulation, as a function of the linguistic input they receive in the acquisition of their native language. Spanish infants of 6-9 months seem to be able to discriminate between the voicing contrast found in Spanish.

A comparison of American English and Danish shows that in terms of VOT categories the same contrast is found in both languages (if the voicing lead is ignored in American English), namely short lag contrasting with long lag. As will be shown below, Danish aspirated stops differ from the American stops in that they show considerably longer VOT values. In the acquisition of the stop system for the two languages it is not necessary for the listeners to learn a completely new contrast. They will have to learn to distinguish between the same categories in both languages, both being near those found in the infant studies, and in the studies on chinchillas.

The question to be addressed in this paper is, whether Danish listeners, because of the longer aspiration, will show a later phoneme boundary than American listeners do, or whether a new contrast is learned in terms of these VOT categories. The question of the effect of formal phonetic training will also be tested to see whether phonetically trained subjects will access a special phonetic mode as participants in perception experiments. Furthermore, the universal difference found in both the production and perception of stop consonants will be considered; whether these differences are still present as reflected in different cross-over boundaries for different place of articulation, even though the stimuli have been constructed in such a way that most of the cues which might account for the perceptual differences have been eliminated.

II. PRODUCTION

In both languages there is a contrast between /*b*,*p*/, /*d*,*t*/, and /*g*,*k*/ in prestressed position. Danish /*d*/ and /*t*/ are often pronounced with a considerable degree of affrication following the release (Fischer-Jørgensen, 1980): [d̥^{s}] and [d̥^{sh}]. This feature is not found in American English, and it is therefore not relevant in a cross-language study of this kind, to compare the production and perception aspects of the alveolar stops.

The Danish /*p*/ and /*k*/ only contrast with /*b*/ and /*g*/ in syllable initial position, when followed by a (sonorant +) full vowel. Danish thus differs from English in that there is no phonological contrast of the kind *rapid* / *rabid*, *bagging* / *backing*. The Danish /*p*/ and /*k*/ are pronounced as voiceless aspirated: [b̥^{h}] and [g̊^{h}]. /*b*/ and /*g*/ are pronounced as voiceless unaspirated: [b̥], [g̊]. The American /*p*/ and /*k*/ are pronounced like the Danish, but are not quite as aspirated as their Danish counterparts. /*b*/ and /*g*/ are sometimes pronounced like in Danish, voiceless unaspirated, and sometimes as fully voiced, having voicing lead. This variant is most often found when preceded by a voiced sound. Since labelling tests have shown that American listeners group stimuli which have voicing lead with those having short lag, it is safe to say that in terms of VOT, Danish and American English both contrast short lag with long lag. In order to see in what way the two languages differ in their production of the stops within the short lag and long lag categories VOT measurements for the two languages will be presented below.

A. DANISH PRODUCTION DATA AND POINTS OF DELIMITATION

The Danish data were obtained from the recordings of two male speakers: JR and NR. The test material consisted of the labials and velars followed by seven Danish vowels: /*i*, *e*, *æ*, *a*, *å*, *o*, *u*/ which were all pronounced long. The test word was inserted in a carrier sentence: *De ska(l) sige* .... (They will say ....). Speaker JR read the list three times and NR four times, giving a total of 49 tokens for each of the four stops.

When measuring VOT, the point of consonantal release usually does not constitute any problems. Voice onset is a different matter, and various points have been used. Fischer-Jørgensen and Hutters (1981) consider three different possibilities: A) the start of vocal fold vibration, B) the point at which F1 sets in, and C) the point at which the upper formants begin. In VOT studies there seems to be some disagreement among investigators as to which point to use. Lisker and Abramson (1964) do not include "edge vibrations", the vibrations of the vocal cords before they are fully adducted. They are thus considering point B to be the starting point. Point C has been used by e.g. Klatt (1975). Others use the word voice onset in its strictest sense, e.g. Keating et al. (1981) (Keating, personal communication), and thus consider A to be the starting point. This point is used in this study as well, for several reasons, one being that the measurements were made from digitized oscillograms.

In Table I are shown the results for the two Danish speakers. Since the material is limited, the individual values were checked against results found by Fischer-Jørgensen (1980). They were found to differ only slightly from her results, which may be due to e.g. different carrier sentences, and the different position of the test word in the utterance (Lisker and Abramson, 1967), and that Fischer-Jørgensen used point B for voice onset, or vowel onset.

Table I

Danish stops followed by 7 vowels

	Msec VOT			Msec VOT	
b	+	14.9	p	+	102.0
g	+	29.4	k	+	110.8

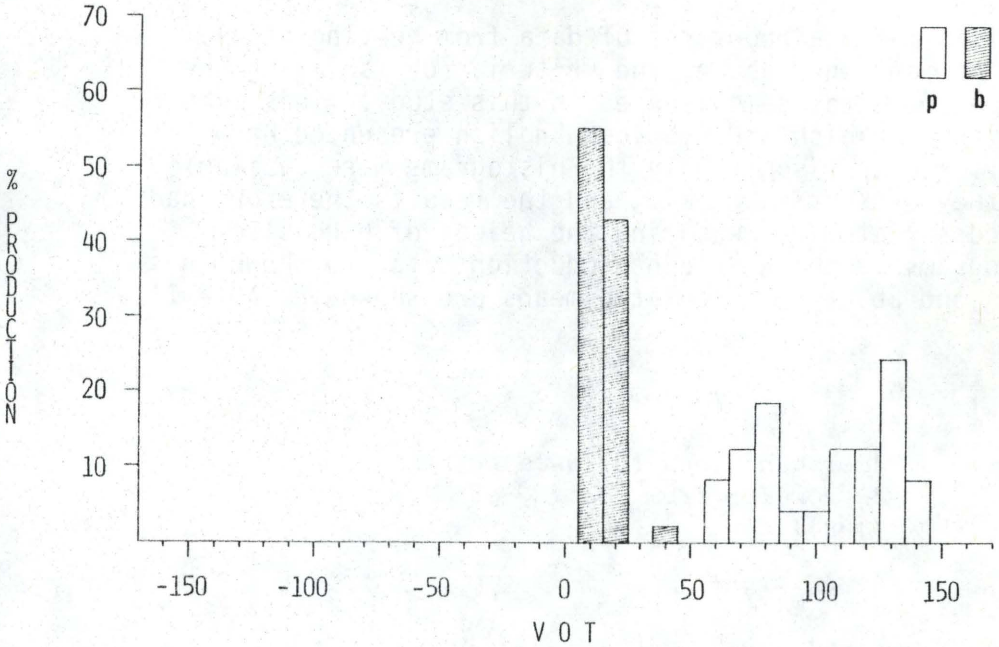

Figure 1a

Danish /b/ and /p/ followed by 7 vowels.

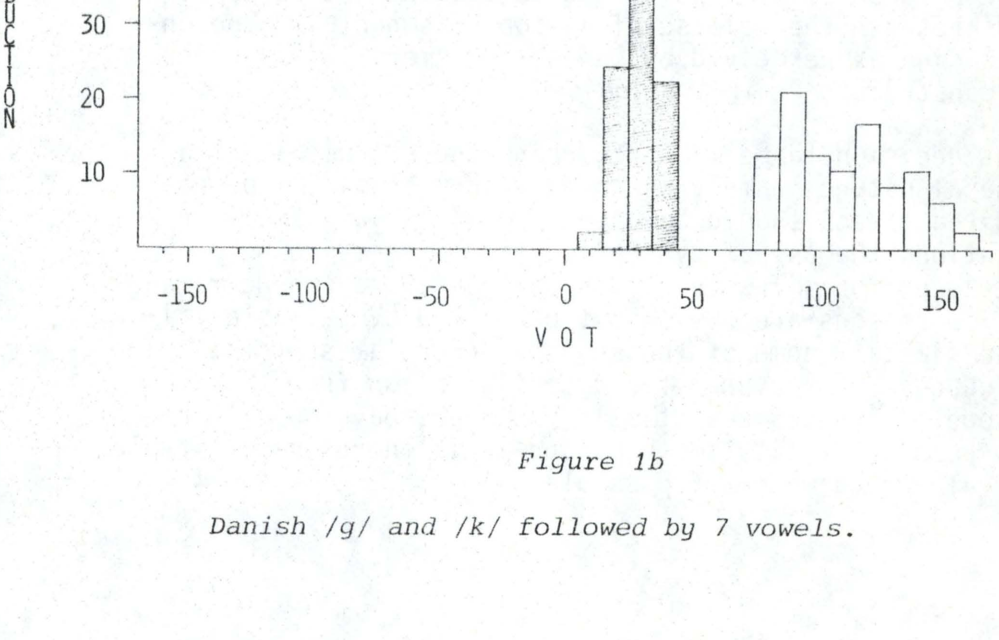

Figure 1b

Danish /g/ and /k/ followed by 7 vowels.

B. AMERICAN PRODUCTION DATA

The American material consists of data from Keating et al. (1981). As mentioned above, the criteria for isolating the voice onset were the same as used in this study. This makes the results of Danish and American English presented here suitable for comparison. Only the histograms were available in the study by Keating et al., and the results therefore had to be reconstructed by measuring the height of each block. The histograms of the American production data are shown in figure 2a and 2b. The calculated means are shown in Table II.

Table II

American stops followed by 12 vowels (from Keating et al. (1981).

Msec VOT					
b	+	5.6	p	+	57.6
g	+	15.5	k	+	71.7

When comparing the data for the two languages it is clear that Danish aspirated stops are considerably longer than the corresponding American ones. In the following it will be shown how much this affects the perception of VOT by Danish and American listeners, reflected in different cross-over boundaries, despite the fact that both languages, in terms of VOT, use the same categories, short lag and long lag.

III. THE EXPERIMENT

A. CHOICE OF MATERIAL

Since the purpose of this study was to examine the temporal relations between the release of a stop consonant and the onset of voicing as perceived by Danish and American listeners, certain constraints limit the choice of test material.

The first decision to be made regarding the stimuli is of a more general nature, namely whether synthetic stimuli or natural edited speech should be used. In almost all American investigations the use of synthetic stimuli seems to be prevalent, except for a few (e.g. Winitz, LaRiviere and Herriman 1975). The reasons are obvious: synthesized CV syllables allow us to simulate some of the acoustic correlates of the gradual change of the vocal tract configuration from the stop to the vowel. Synthesized stimuli, however, have up till now had some buzzing quality to them, and it is therefore desirable to use real speech whenever possible.

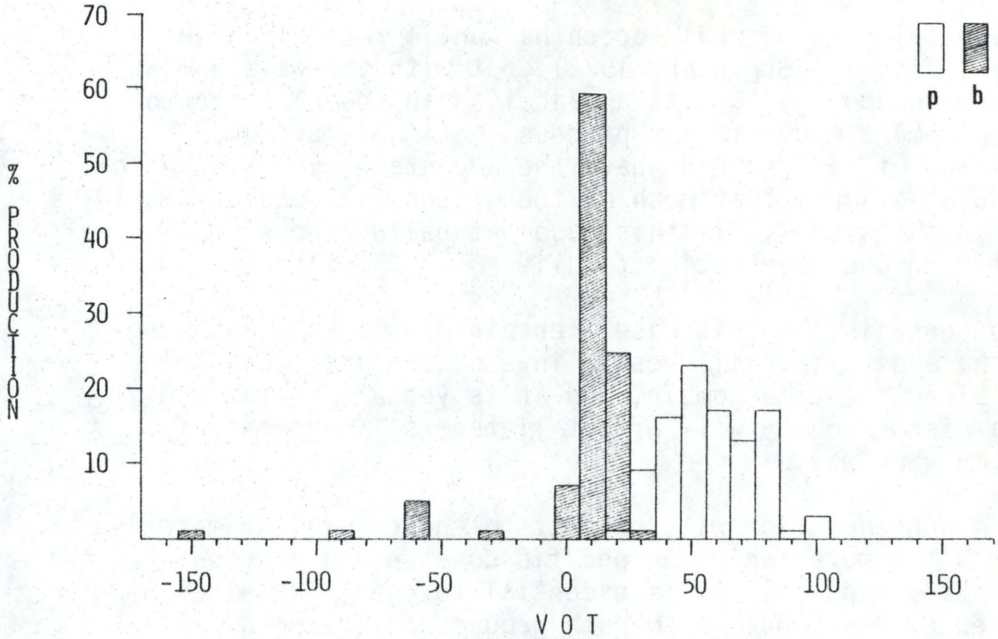

Figure 2a

American /b/ and /p/ followed by 12 vowels. From Keating et al. (1981). Permission by Patricia Keating.

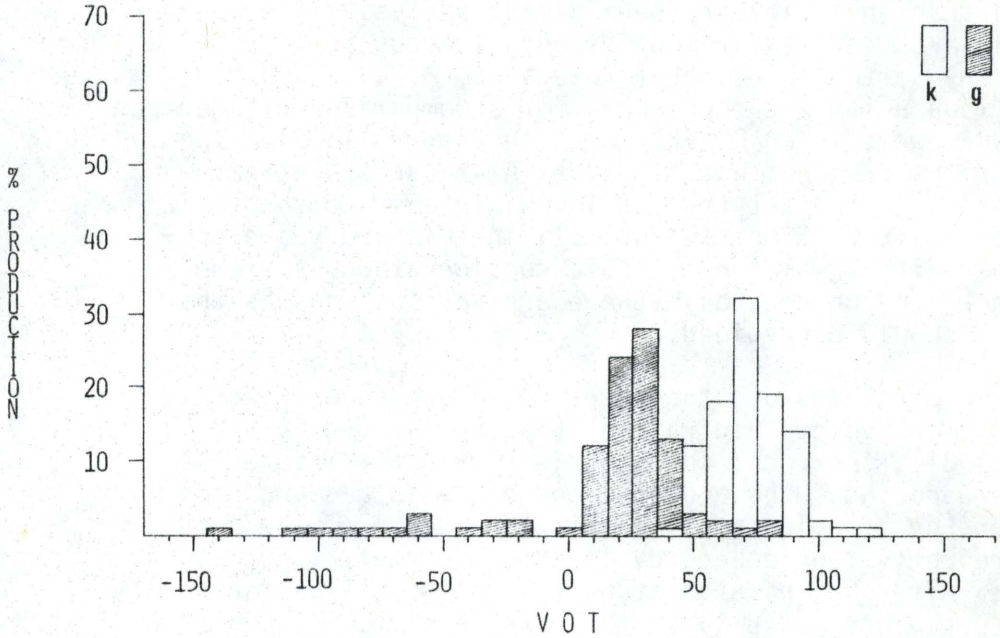

Figure 2b

American /g/ and /k/ followed by 12 vowels. From Keating et al. (1981). Permission by Patricia Keating.

Editing and splicing of real speech has until recently been cumbersome (Fischer-Jørgensen, 1972), but with the development of computer technology and its application in speech research this has become a much simpler process. Stimuli, produced from splicing of real speech have the advantage over synthetically produced stimuli that much of the speech-like quality is retained in the signal. In this case natural edited stimuli were preferred over synthetic stimuli.

The method entails, in this case, certain drawbacks. Removing part of the aspiration will result in a discontinuity in the spectrum along the time domain, and it is yet a question how much this discontinuity will affect listeners' judgement of VOT, and in what way.

The second consideration with respect to choice of test material concerns the more language-specific constraints a cross-language study imposes. It is essential to construct a set of stimuli equally acceptable to both groups of listeners. As mentioned earlier, the Danish alveolar stops differ significantly from the American ones in that the release of this type of stop in Danish is accompanied by a considerable degree of affrication, varying from greatest in most urban dialects to being absent in the rural dialects of especially the west coast of Jutland. The test material therefore includes only labial and velar stops.

The vowels following the consonant require some consideration too. American investigators seem almost exclusively to favour the vowel /*a*/. The requirement of mutual acceptability would hold for any of the three point vowels /*i*, *a*, *u*/ as far as formant frequencies are concerned (for a comparison of American and Danish vowels in an F1-F2 plot, see Disner, 1980). The reason for the frequent use of /*a*/ by American investigators may be found in the fact that it is fairly easy to synthesize with good results. Since the stimuli in this study are produced from edited real speech, this consideration is of no consequence; on the contrary, there are various reasons why this vowel should be avoided.

Following aspirated stops the vowel /*a*/ often shows "edge vibrations" or "voiced aspiration" (Fischer-Jørgensen and Hutters, 1981), the vocal cords begin to vibrate before they are fully adducted. The acoustic correlate is a signal containing energy of a low amplitude in the region of the fundamental frequency, but no energy in the region of F1, or higher formants. After unaspirated stops voicing normally starts simultaneously and abruptly in the whole spectrum. This also seems to be the case for the high vowels /*i*/ and /*u*/ when following aspirated stops. Using /*a*/ might therefore increase the risk that editing aspirated stimuli might produce unaspirated ones with a vowel onset characteristic of a vowel following aspirated stops. This would consequently result in stimuli containing possibly conflicting cues.

By choosing */i/* and */u/* we get a further advantage in the bargain. One of the arguments against the salience of VOT as a perceptual cue has been that listeners include F1 transitions in their judgements (Stevens and Klatt, 1974). Since both these vowels have the lowest F1 of all vowels, we will only get a slight change of this formant during the transition from stop to vowel, and thus minimize the spectral discontinuity in this range.

The stimuli were therefore constructed from real tokens of */p/* and */k/* followed by the vowels */i/* and */u/*.

B. STIMULI

From the recordings of the two Danish male speakers, JR and NR one token of each of the test syllables [b̥ʰi, b̥ʰu, g̊ʰɪ, g̊ʰu] was selected, viz. the first occurrence on the tape which had been measured to have a VOT value greater than 90 msec. Some of speaker JR's tokens did not quite show VOT values of this duration, and in those cases the token with the highest measured value was then used. The original VOT values thus ranged between 80 msec and 136 msec (JR's [b̥ʰu] and NR's [g̊ʰu]).

The entire CV syllable was stored in the buffer of a PDP-8 Digital Equipment Corporation computer and the point of voice onset identified on a digitized oscillogram. The part of the syllable containing glottal periodicity was stored in a separate file. In the process of isolating the voiced portion of the syllable (i.e. the vowel) care was taken to ensure a cut at a zero-crossing in order to avoid offset clicks. The cursor was then positioned 70 msec from the point of consonantal release, and the remaining part of the aspiration was deleted from the buffer. A copy of the file containing the voiced portion could then be read down into the buffer and added to the 70 msec of burst + aspiration. This procedure was repeated, each time shortening the aspiration from right to left in steps of 10 msec, resulting in seven stimuli ranging in length of aspiration from 10 - 70 msec.

This was done with the tokens of both speakers, giving a total of 28 labials and 28 velars. The stimuli were recorded on magnetic tape, using a REVOX A77 open reel tape deck for subsequent tape generation for the listening test.

C. PRODUCTION OF TAPE FOR THE LISTENING TEST

The stimuli were transferred from the magnetic tape and stored in separate files on the disk of a PDP-11/34 Digital Equipment Corporation computer, using a sampling rate of 10.000 Hz and a high-pass filtering of 80 Hz. This sampling rate is, in effect, a low-pass filtering at 5.000 Hz, and it was not assumed that acoustic information above 5.000 Hz would in any way be crucial to the listeners in this experiment.

Four series were generated, the first and second series containing the labials and velars of speaker JR, respectively, the third and fourth series containing the same tokens spoken by NR. In each series the stimuli were duplicated ten times each and recorded on magnetic tape in a randomized order in blocks of 10 with an inter-stimulus interval of 2.5 seconds and an inter-block interval of 4 seconds. They were all recorded on the same tape in the order mentioned above, with a pause of approximately two minutes between series.

D. THE LISTENING TEST

The tape was presented to the subjects for labelling B or P and G or K. Immediately before the test the subjects were given a set of written instructions, explaining to them the inter-stimulus and inter-block intervals. They were also instructed not necessarily to expect an equal number of B's and P's or G's and K's. Orally they were asked not to introduce a third category, but to guess in those cases where they felt unable to decide.

The tape was presented binaurally over headphones. All the subjects took the test individually and had available to them a volume control. The total duration of the test was 28 minutes.

E. SUBJECTS

A total of 16 subjects participated in the experiment, 8 Danish and 8 American listeners. A further subdivision can be made of the two groups, into phonetically trained and phonetically naive listeners, 3 and 5 respectively in the two groups. The criteria according to which the subjects were classified as phonetically trained must almost inevitably differ for the Danish and the American group, for obvious reasons.

The Danish listeners were considered to be phonetically trained if they had taken an intensive three-semester course in ear-training and narrow phonetic transcription at the Institute of Phonetics. In this course, among other things, the students are trained to distinguish between degrees of aspiration and voicing during occlusion, and it could therefore be expected that this group would be more consistent in their responses, i.e. have sharper category boundaries. The rest of the Danish subjects were all students at various levels at the University of Copenhagen.

The American subjects were all graduate students in the linguistics department at Brown University, USA, except one, who was an undergraduate majoring in linguistics. Part of the graduate program in linguistics at Brown University is a one-semester course in ear-training and phonetic transcription (Linguistics 0121). But since it also includes an introduction to basic concepts in phonology, it is not comparable to the intensive course the Danish students take.

Three of the American subjects also worked as research assistants in the linguistics department's speech laboratory. Their duties as such included phonetic transcription of natural discourse between young children and their parents, and it could therefore be expected that their skills in phonetic transcription were beyond the ear-training course. However, the transcription they were required to use is not as narrow as the transcription required by the Danish students. Therefore, in the strictest sense, the American and the Danish phonetically trained groups are not comparable (for the role of formal training in vowel transcription, see Laver, 1965).

All Danish and American subjects were unpaid volunteers. None of the Danish naive listeners had ever participated in perception tests before, whereas all the American listeners had participated in - and were familiar with - such tests. No subject had a known history of hearing loss.

IV. RESULTS AND DISCUSSION

Initial calculations did not reveal any consistent difference between stimuli recorded by the two speakers with respect to cross-over point (i.e. that cross-over points occurred later for one speaker than for the other). Identification results for stimuli from the two speakers were therefore pooled.

A. EFFECTS OF PHONETIC TRAINING

In order to determine whether to treat phonetically trained and naive listeners as one group in the final cross-language comparison, the cross-over points were found for each of these groups. This was done by converting the identification scores for each stimulus to z-scores which could then be fitted to a straight line, using the method of "least sum of squares" (Gilford 1954:123f). The results for the American listeners are shown in Table III.

Table III

Means for identification functions of American phonetically trained and naive listeners.

	American					
	phonetically trained			naive		
stimuli	$\overline{X}$	sd.	N	$\overline{X}$	sd.	N
pi	17.98	8.77	60	19.06	6.55	100
pu	28.85	11.46	60	25.21	7.34	100
ki	61.10	13.19	60	45.59	11.81	100
ku	47.19	12.70	60	42.70	9.45	100

It is seen clearly that the American phonetically trained listeners show later cross-over boundaries than the naive listeners for three out of the four series of stimuli. However, a criterion was set up to determine whether the phonetically trained group should be included: The groups were considered to behave differently if, and only if all of the four cross-over points were found to occur later for one group than for the other, regardless of whether the individual results were statistically significant or not. This did not apply to the American group since *pi* showed an earlier cross-over point for the phonetically trained listeners, whereas for the rest it occurred later. They were therefore included in the final cross-language comparison.

In the case of the Danish listeners, there seems to be a clear tendency for the phoneticians to require shorter VOT before their percept begins to change from one phoneme category to the other. As shown in Table IV, their cross-over points occur earlier for all four series of stimuli than for the naive listeners. A non-directional t-test gave $p < .001$ for all, with the exception of *pi* which proved to be non-significant. The group of Danish phonetically trained listeners were therefore omitted for the cross-language comparison.

Table IV

Means for identification functions of Danish phonetically trained and naive listeners.

	Danish					
	phonetically trained			naive		
stimuli	$\overline{X}$	sd.	N	$\overline{X}$	sd.	N
pi	21.30	7.82	60	22.59	5.84	100
pu	32.42	5.86	60	39.72	7.72	100
ki	47.82	10.28	60	53.37	9.58	100
ku	36.06	10.10	60	45.66	7.83	100

It may be argued that the reasons for excluding the Danish but including the American phonetically trained listeners may be somewhat vague or arbitrary. One alternative would be to exclude both groups (rather than including the Danish listeners as well). However, the number of subjects in this study is small anyway, and including the American phonetically trained listeners only moves the cross-over points of the total group towards that expected for the Danish listeners. Consequently, the results influence the cross-language comparison in such a way that the expected difference between Danish and American listeners is reduced. The reason why phonetically trained persons were used as subjects in the first place in addition to naive listeners was that I was interested in looking into the effects of formal phonetic training.

When comparing the results obtained from the groups of Danish listeners it is interesting to see that the expected effects were not found, viz. that trained phoneticians would show sharper category boundaries than naive listeners.

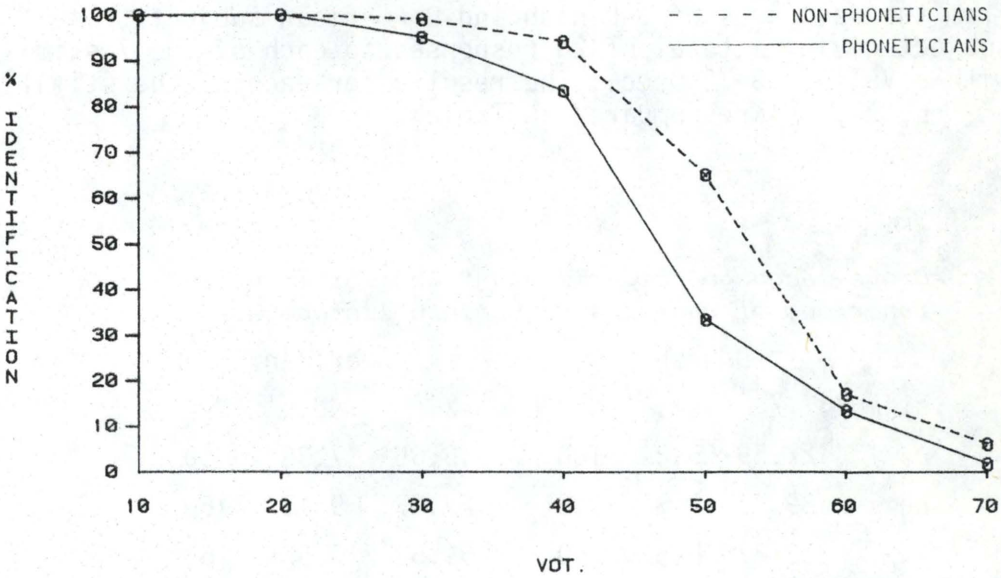

Figure 3

Identification functions for phonetically trained and naive listeners of the *<u>ki</u>* *stimuli.*

In Figure 3 the percentages of responses from the phonetically trained and naive listeners are plotted against the stimuli of *ki*. It is clear from this figure, as was seen in Table IV, that the cross-over point occurs much earlier for the phonetically trained listeners, which must be ascribed to the effect of formal training. What it does not tell us is how the same phonetically trained listeners would perceive VOT in normal discourse. One possible hypothesis may be that their category boundary would be identical to that of the naive listeners, and that the nature of a perception test will trigger a special "phonetic mode" of listening in the trained group of subjects.

If the identification functions for the two groups are compared with the standard deviations given in Table IV, it is interesting to note, since the standard deviation is reflected in the steepness of the curve (the slope), that the phonetically trained listeners do not show sharper category boundaries than those found for the naive listeners. This is in a sense surprising since it could be expected that phonetic training, among other things, would teach the listeners to "latch on" to one specific acoustic variable and in that way be able to identify the individual stimuli more consistently.

B. CROSS-LANGUAGE COMPARISON

Since the Danish phonetically trained listeners were omitted from the study, the final cross-language comparison was made from the responses of 5 Danish and 8 American subjects, each subject giving a total of 20 responses to each of the 7 stimuli with a VOT of 10-70 msec. The results for each of the stimuli *pi*, *pu*, *ki*, *ku* are compared in Table V.

Table V

Cross-language comparison. Means for identification functions of Danish and American listeners.

	Danish			American		
stimuli	$\overline{X}$	sd.	N	$\overline{X}$	sd.	N
pi	22.59	5.84	100	18.65	7.38	160
pu	39.72	7.72	100	27.67	9.15	160
ki	53.37	9.58	100	51.53	13.55	160
ku	45.66	7.83	100	45.49	10.48	160

Looking first at the means, i.e. cross-over points given in Table V, it is clear that the perceptual shift from one phoneme to the other takes place earlier for the American listeners, as was expected according to the original hypothesis. If the boundary between phonemes in the American production data is estimated by visual inspection of the histograms, figure 2a and 2b, it will be - for the labials - approximately 20-30 msec and for the velars approximately 40-45 msec. The corresponding phoneme boundaries would be, for the Danish data, figure 1a and 1b, 40-45 for the labials and 55-65 msec for the velars. For both places of articulation the Danish perceptual cross-over boundary would therefore be expected to occur much later.

For the labial place of articulation this is quite clearly confirmed in Table V. Although the Danish cross-over point occurs much earlier than predicted from the production data, it still differs from the American one as expected: a t-test gives $t = 4.528$ $p < .001$ for *pi*, and for *pu* $t = 10.952$ $p < .001$ when compared to the American cross-over points. In figure 4a and 4b

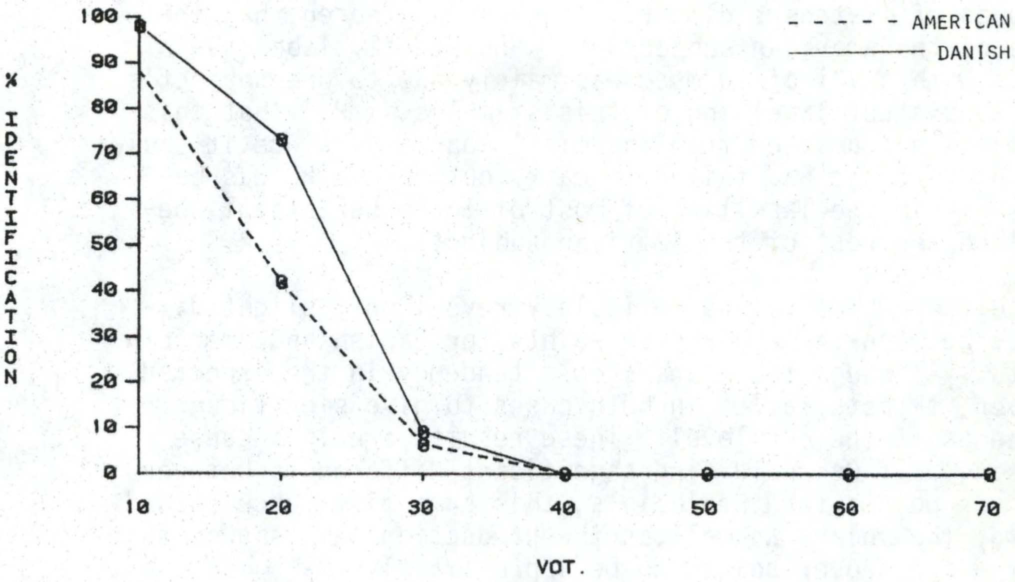

Figure 4a

Identification functions for American and Danish listeners of the pi stimuli.

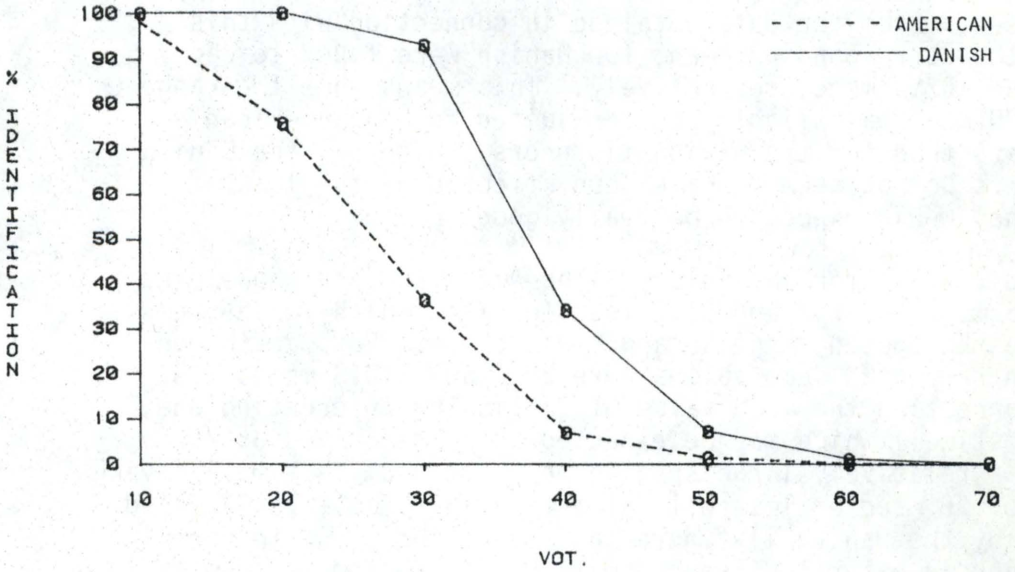

Figure 4b

Identification functions for American and Danish listeners of the pu stimuli.

the percentages of b-responses for the individual stimuli are plotted, and it is perhaps clearer from this figure how the two groups of listeners differ. It should be noted that the failure of the American subjects to consistently label the *pi* stimulus with a VOT of 10 msec was mainly due to one subject's almost consistent labelling of this stimulus as P. That this subject was not omitted from the final analysis, since it could be argued that she had failed to carry out the task, was because she - in the labelling of most of the other stimuli behaved like the rest of the American subjects.

The results for the velars in Table V reveal only slight differences between the cross-over points for Danish and American subjects. Although the means show a tendency in the expected direction, t-tests failed in both cases to give significant differences at the .05 level. These results are in a sense surprising. If we could find significant differences between cross-over points for the labials, this same difference should exist for the velars as well,as the production data showed the estimated cross-over points to be approximately 40-45 msec for the Americans, and 55-65 msec for the Danish data.

The question of range effects and listeners' strategies may help explain the failure to show any significant difference of cross-over points. That category boundaries may be shifted around due to range effects has been shown for English by e.g. Brady and Darwin (1978), but the shifts they found were relatively small, on the order of 5-7 msec. This led to a reexamination of the ranges employed in this experiment to see if they were in any way inappropriate for one or both groups of listeners. The computed means, on the basis of results from Keating et al. (1981), were for */b/* 5.6 msec and for */p/* 57.6 msec. From the data obtained in connection with this study the corresponding means for Danish were found to be 14.9 and 102.4 msec, respectively. This means that the range of 10-70 msec was slightly better suited for the American listeners than for the Danish listeners. However, the Danish listeners do not seem to have been affected by the lack of what they would expect to be really good */p/*'s.

It is different for the velars. The American data show means of 15.5 msec for */g/* and 71.7 msec for */k/*, which again seems to make the chosen range an appropriate one. For Danish, on the contrary, the same values were 29.4 and 110.8 msec. In this connection the mean value of */g/* is the interesting one, as the stimuli which may be expected to be perceived as */g/* occupy a relatively large span in the continuum. Stimuli having a VOT of 30 msec or less will almost without doubt be labelled as */g/* by the Danish listeners, and since they, due to normal variation in natural discourse will hear intended */g/*'s with longer VOT than the 30 msec, they are more likely to label them as G than American listeners. The 10-70 msec range must therefore be said to be less appropriate for Danish listeners.

In figure 5a and 5b are shown the identification functions for *ki* and *ku*. It is interesting that the Danish listeners con-

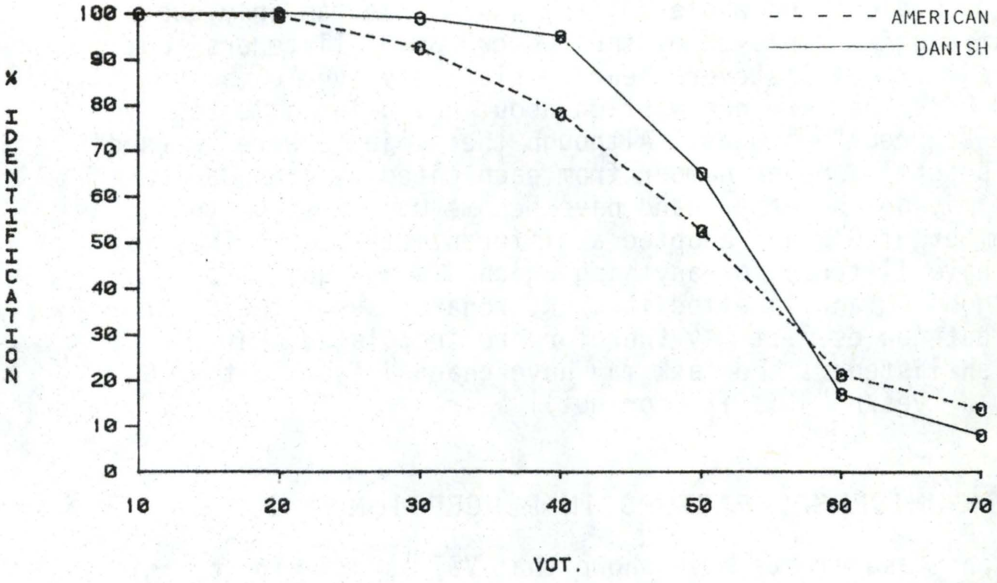

Figure 5a

Identification functions for American and Danish listeners of the <u>*ki*</u> *stimuli.*

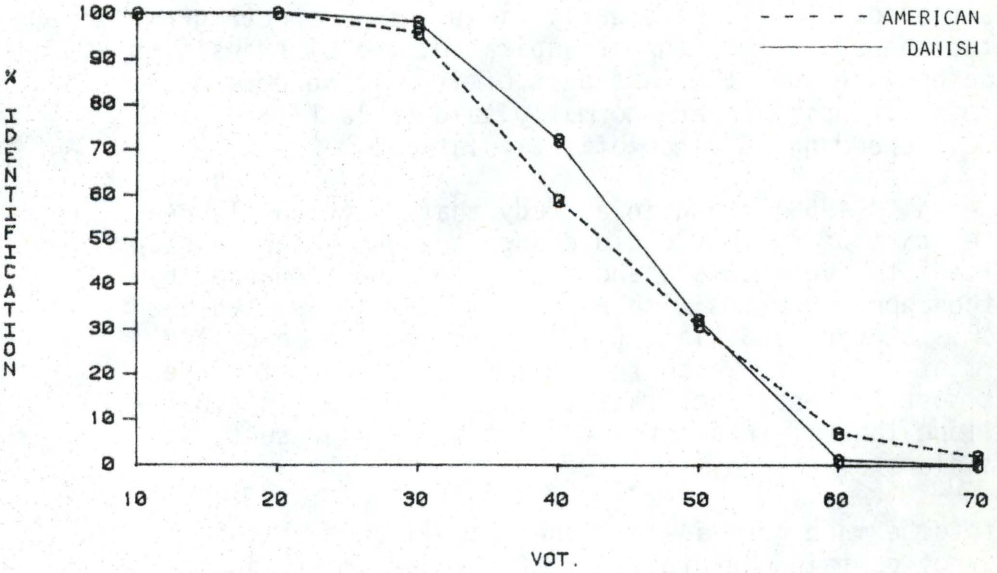

Figure 5b

Identification functions for American and Danish listeners of the <u>*ku*</u> *stimuli.*

sistently label stimuli, especially *ki* (figure 5a) of 10-40 msec, as G's. The Americans, on the other hand, show a more gradual drop along the whole VOT continuum. We can only guess at the strategies employed by the two groups of listeners, but since the American listeners hear a relatively even number of *gi*'s and *ki*'s they are not worried about not being able to give the "correct" answers. Although the subjects were "warned" about a possibly uneven number from each category, the Danish subjects may on the other hand have become uneasy about the large number of G's and adopted a different strategy. They may in fact have listened for anything which did not quite sound like a "good" G and labelled it as K, regardless of their normal phoneme categories. It may therefore be speculated that for the Danish listeners the task may have changed from "either G or K" to a "yes/no" (is it G or not).

C. UNIVERSAL FACTORS IN PERCEPTION

Production measurements have shown that VOT is affected by place of articulation (Lisker and Abramson, 1964, Fischer-Jørgensen, 1980). VOT has been found generally to be longer following the release of a velar than a labial stop. The explanation given to this phenomenon has been that the movement of the tongue body away from the passive articulator is slower, resulting in a delayed drop in oral air pressure relative to the release, which is necessary for vocal fold vibration to begin.

The acoustic correlate is consequently slower formant movements during the transition from the stop to the vowel. If glottal pulsing starts immediately after the release (i.e. short lag), the transition of F1 will be clearly visible on a spectrogram. On the other hand, if the stop is aspirated, the F1 transition will be absent, or only the last part of it will be present. These formant transitions are normally said to last for about 40-80 msec, depending on place of articulation.

Liberman et al. (1958) found in a study that by gradually delaying the start of F1 they could change the percept of a stop from "voiced" to "voiceless", and that effect was enhanced by filling the upper formants with noise simultaneously with the F1 cutback. Stevens and Klatt (1974) proposed the so-called First Formant Detector, which they claim can account for the fact that most listeners hesitate to identify stops as "voiceless" or long lag, if this formant transition is present, almost regardless of the physical VOT.

The point to be made here is that the stimuli used in this study did not contain measurable first formant transitions at all. This was checked on spectrograms of [g^{h}i], recorded by both speakers. Both the Danish and American listeners required longer VOT for the velar stimuli, before their percept changed from /*g*/ to /*k*/. This is shown in figure 6 where the responses for the two groups are compared for place of articulation.

The results therefore suggest that the perception of VOT is influenced by place of articulation, even though the stimuli have been stripped of the cues which are normally used to explain the later cross-over boundary for velars. Listeners must therefore have some expectation about the longer VOT for velars which is quite consistently reflected in the responses of both Danish and American listeners.

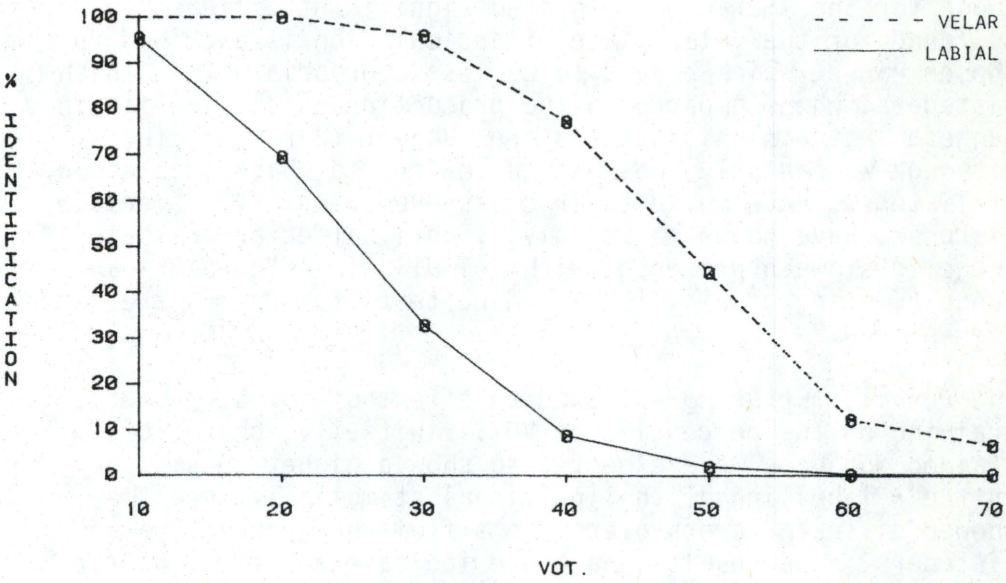

Figure 6

Identification function for Danish <u>*and*</u> *American listeners, describing universal difference in the perception of VOT for place of articulation.*

That this is not only the case for place of articulation of the stop consonant, but also for rounding of the vowel following the stop, may be seen from Table V. In her study of Danish production Fischer-Jørgensen (1980) found that rounded vowels following labial stops had a longer "open interval" (which she defines as the interval from the release of the consonant to the onset of the vowel, point B according to Fischer-Jørgensen and Hutters, 1981). Since the tongue is already in position for the vowel at the release of a labial stop, the place of oral constriction for the vowel cannot affect the pressure drop in the oral cavity, and the longer VOT in /*bu*/ and /*pu*/ must therefore be due to the coarticulation of the vowel. The cross-over points (in Table V) for labials followed by /*i*/ were compared to those followed by /*u*/ for the Danish and the American listeners in a t-test. Both groups were found to have later cross-over points for *pu*, $p < .001$, which again indicates that listeners expect rounded vowels to delay the onset of voicing and therefore use this knowledge in their responses, even though the stimuli did not contain information which would effect a later cross-over boundary.

V. CONCLUSION

The results found in this pilot study confirm the original hypothesis in showing a clear correspondence between the production and perception of VOT. It has been shown that the cross-over points for Danish listeners did, on the whole, occur later than those for the Americans. That no significant difference could be found for the velar place of articulation is ascribed to the chosen range, which proved to be less appropriate for Danish listeners, when compared to the production data. The findings suggest that Danish listeners are subject to range effects, although we can only guess at the extent to which this group of listeners have moved their cross-over boundary. American listeners have shown to be only slightly affected in their judgements, when presented with stimuli of an "un-American" range (Keating et al., 1981). More tests for Danish are therefore needed.

The results furthermore showed an effect of formal phonetic training on the perception of VOT. Initially, phonetically trained subjects were expected to show a higher consistency in their labelling of the individual stimuli; instead they showed different cross-over points from the group of naive listeners. The results therefore indicate an effect of formal training which is different from what was expected. This calls for further experiments of phoneticians' perception of VOT since in other studies they have been shown not to behave differently when compared to non-phoneticians.

Differences in VOT which are physiologically and aero-dynamically determined (i.e. the differences found for labial and velar stops) are quite clearly reflected in the responses both by Danish and American listeners, even if the stimuli contained none of the cues (vowel formant transitions, etc.) that are normally used to explain these perceptual differences. The results therefore support the hypothesis that VOT and place of articulation are not processed as two independent cues (see e.g. Sawusch and Pisoni, 1974). In this experiment place of articulation was given on the answer sheet and the listeners may have assigned different "expected" VOT values accordingly.

ACKNOWLEDGEMENTS

This paper is a revised version of a thesis submitted in partial fulfilment of the requirements for the degree of Exam. art. in the Institute of Phonetics at the University of Copenhagen. I would like to thank Niels Reinholt Petersen who served as my advisor, and whose helpful suggestions have been of invaluable help to me, Jørgen Rischel for helpful comments, and the staff for moral support. Special thanks to Philip Lieberman and

Sheila Blumstein for the use of computer facilities in the linguistics laboratory at Brown University, USA, and to Christian Boysen, Tectronics, and Carsten Henriksen for their help with the graphs.

REFERENCES

Abramson, Arthur S. and Lisker, Leigh 1973: "Voice time perception in Spanish word-initial stops", *J. Phonetics 1*, p. 1-8

Abramson, Arthur S. and Lisker, Leigh 1983: "Relative power of cues: F0 shifts versus voice onset time", *Abstracts 10th Int. Congr. Phon. Sc.*, p. 505

Brady, S.A. and Darwin, C.J. 1978: "A range effect in the perception of voicing", *J. Acoust. Soc. Am. 63*, p. 1556-1558

Disner, Sandra F. 1980: "Evaluation of vowel normalization procedures", *J. Acoust. Soc. Am. 67/1*, p. 253-261

Eimas, P.D., Siqueland, E., Jusczyk, P. and Vigorito, J. 1971: "Speech perception in infants", *Science 171*, p. 303-306

Fischer-Jørgensen, E. 1972: "Tape-cutting experiments with Danish stop consonants in initial position", *Ann. Rep. Inst. Phon. Univ. Cph. 6*, p. 104-168

Fischer-Jørgensen, E. 1980: "Temporal relations in Danish tautosyllabic CV sequences with stop consonants", *Ann. Rep. Inst. Phon. Univ. Cph. 14*, p. 207-261

Fischer-Jørgensen, E. and Hutters, B. 1981: "Unaspirated stop consonants before low vowels, a problem of delimitation - its causes and consequences", *Ann. Rep. Inst. Phon. Univ. Cph. 15*, p. 77-102

Fujimura, O. and Miller, J.E. 1979: "Mandible height and syllable-final tenseness", *Phonetica 36*, p. 263-272

Gilford, J.P. 1954: *Psychometric Methods*, 2nd ed. (McGraw-Hill)

Haggard, M.S., Ambler, S. and Callow, M. 1970: "Pitch as a voicing cue", *J. Acoust. Soc. Am. 47*, p. 613-617

Keating, P.A., Mikós, M.J. and Ganong III., W.F. 1981: "A cross-language study of voice onset time in the perception of initial stop voicing", *J. Acoust. Soc. Am. 70*, p. 1261-1271

Klatt, D.H. 1975: "Voice Onset Time, frication and aspiration in word-initial consonant clusters", *J. Speech and Hearing Research 18/4*, p. 686-706

Kohler, K. 1983: "Phonetic explanation in phonology", *Abstracts 10th Int. Congr. Phon. Sc.*, p. 275-283

Kuhl, P.A. and Miller, J.D. 1978: "Speech perception by the Chinchilla: Identification for synthetic VOT stimuli", *J. Acoust. Soc. Am. 63*, p. 905-917

Laver, J. 1965: "Variability in vowel perception", *Language and Speech 8*, p. 95-121

Liberman, A.M., Delattre, P.C. and Cooper, F.S. 1958: "Some cues for the distinction between voiced and unvoiced stops in initial position", *Language and Speech 1*, p. 153-166

Lisker, L. 1975: "Is it VOT or a first-formant detector?", *J. Acoust. Soc. Am. 57,6, part II.*

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

Lisker, L. and Abramson, A.S. 1968 (1965): "Some effects of context on voice onset time in English stops", *Language and Speech 10,1*, p. 1-28

Mikōs, M.J., Keating, P.A. and Moslin, B.J. 1976: "The perception of voice onset time in Polish", *J. Acoust. Soc. Am. Suppl. I*, p. S19

Sawusch, J.R. and Pisoni, D.B. 1974: "On the identification of place and voicing features in synthetic stop consonants", *J. Phonetics 2/3*, p. 181-194

Stevens, K.N. and Klatt, D. 1974: "Role of formant transitions in the voiced-voiceless distinction for stops", *J. Acoust. Soc. Am. 56/3*, p. 653-659

Williams, L. 1977: "The voicing contrast in Spanish", *J. Phonetics 5*, p. 169-184

Winitz, H., LaRiviera, C., and Herriman, E. 1974: "Variations in VOT for English stops", *J. Phonetics 3*, p. 41-52.