Abstract of

A Phonetic Study of a Voicing Contrast in Polish

by Patricia Ann Keating, Ph.D. Brown University, June 1980

A series of experiments on the production and perception of the voicing contrast in Polish was carried out to determine the importance of the phonetic dimension of Voice-Onset-Time (VOT) in that contrast. Polish contrasts prevoiced and short-lag VOT categories, while English contrasts short-lag and long-lag VOT categories. One goal of the study is to explore differences between the two types of contrasts that would account for their relative frequency of occurrence.

Data was collected for 24 Polish speakers in Wrocław, Poland. It was found that the Polish contrast is relatively straightforward in production. Initial and medial stops in isolated minimal pairs and in running speech styles show no overlap in VOT values between voiced and voiceless stops. In running speech, voiced closure duration is used as a measure of VOT for voiced stops. For voiced stops, there is always voicing during closure, while for voiceless stops there is not. Perceptual experiments were carried out using both synthetic and natural stimuli. Closure voicing and voicing lag were found to be strong cues for voicedness and voicelessness, respectively; burst voicing is a weaker cue. However, when synthetic VOT continua are used, listeners' boundaries between the voicing categories are strongly affected by the range of VOT values in each continuum. Only when the range corresponds to that found in natural Polish productions do the isteners' boundaries align with their production categories. For

a range of VOT values corresponding to natural English productions, the Polish listeners' category boundaries are too high. On the other hand, American listeners show consistent boundaries on all three synthetic continua, regardless of the VOT range used. Thus VOT seems to be a less stable perceptual dimension in Polish than in English. It is proposed that Polish listeners respond to non-Polish ranges of VOT on the basis of psychoacoustic categories, giving the higher boundaries. The English voicing categories are already aligned with psychoacoustic categories, so American listeners would not show range effects. Thus the Polish voicing contrast is quite stable in production but less stable in perception. The reverse is true of the English contrast.

In medial stops additional cues are shown to enter into a trading relation that determines the percept. The proportion of closure voicing to closure silence within the total closure duration of a voiced stop has a perceptual effect that corresponds to the patterns seen in production. On the other hand, the duration of the preceding vowel, which has been thought to be universally lengthened before voiced consonants, does not correlate with stop voicing in Polish.

It is proposed that two-category voicing contrasts like those of Polish and English be represented at the phonological level by the feature [±voice] in all languages, so that phonological rules will be expressed equivalently. At the phonetic level, however, the feature representation would be [n VOT], allowing phonetic detail in each language to be described. The results of the experiments indicate that VOT is an excellent dimension for such descriptions.

A Phonetic Study of a Voicing Contrast in Polish

by

Patricia Ann Keating

A.B., Wellesley College, 1974

A.M., Brown University, 1976

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Linguistics at Brown University

June 1980

This dissertation by Patricia Ann Keating

is accepted in its present form by the Department of

Linguistics as satisfying the

dissertation requirement for the degree of Doctor of Philosophy.

Date. July 2, 1979 ..... Sandra Thompson .....

Recommended to the Graduate Council

Date. July 3, 1979 ..... Peter Ladefoged .....

Date. July 3, 1979 ..... Wallace Chafe .....

Approved by the Graduate Council

Date. ....

PATRICIA ANN KEATING

Born July 20, 1952 in Philadelphia, Pennsylvania.

EDUCATION:

June, 1974:	B.A. in French, honors in Russian, Wellesley College, Wellesley, Mass. Phi Beta Kappa.
June, 1976:	M.A., in Linguistics, Brown University
June, 1980:	Ph.D. in Linguistics, Brown University

RESEARCH EXPERIENCE:

1976-1979:	Research assistant to Professor Philip Lieberman, working on child speech acquisition. Research projects have included vowel development, fundamental frequency and vocal register use, cross-language comparison of mothers' and children's use of voice-onset-time, and variation in word duration. Duties include tape recording infants and mothers, making and interpreting sound spectrograms, and designing and implementing research projects, including publication of results.

TEACHING EXPERIENCE:

Spring, 1977:	Teaching substitute for Philip Lieberman's undergraduate course, "The Evolution of Language". Responsibilities included presentation of most class lectures and joint preparation of syllabus and exams.
Summers, 1977 and 1978	Instructor of English, Brown University Summer Program in English as a Foreign Language. Program Director: Frank L. Ryan. Duties included complete responsibility for syllabus and materials for an advanced class: reading, paragraph development, listening, and conversation.

DISSERTATION: "A Phonetic Study of a Voicing Contrast in Polish"

Committee: Sheila E. Blumstein, Linguistics Dept. (supervisor)
Philip Lieberman, Linguistics Dept.
W. Francis Ganong, Psychology Dept.

GRANTS RECEIVED: National Institutes of Health Individual National Research Service Award (Post-doctoral Fellowship)

PAPERS AND PUBLICATIONS:

Refereed and Invited Articles

"Fundamental frequency in the speech of infants and children", J. Acoust. Soc. Am. 63 (2) Feb. 1978 (with R. Buhr).

"The effects of transition length on the perception of stop consonants", J. Acoust. Soc. Am. 64 (1), July 1978 (with S. Blumstein).

"Fundamental frequency and vocal registers in the speech of prelinguistic infants", to appear in Infant Voice, Speech and Language: A Book of Readings, Thomas Murry, ed.

Brown University Student Working Papers in Linguistics

"A preliminary study of vowel length in Czech", VOL. I, May 1976 (with B. Rubin)
"Some observations about Tagalog ng-phrases", Vol. II, June 1977
"Variation in the duration of words", Vol. III, 1978 (with C.A. Kubaska) (text of A.S.A. talk)
"Spectrographic effects of register shifts in speech production", Vol. III, 1978 (with R. Buhr) (text of A.S.A. talk)

Papers Presented

"Mothers' simplification of phonetic input to their children in English and Polish", J. Acoust. Soc. Am. 61, S7(A), 1977 (with B. Moslin)
"Voicing distinction in Polish word-initial stop consonants", J. Acoust. Soc. Am. 62, S27(A), 1977 (with B. Moslin)
"Spectrographic effects of register shifts in speech production", J. Acoust. Soc. Am. 62, S25(A), 1977 (with R. Buhr)
"Variation in the duration of words", J. Acoust. Soc. Am. 63, S56(A), 1978 (with C.A. Kubaska)
"The perception of voice onset time in Polish", J. Acoust. Soc. Am. 63, S19(A), 1978 (with M.J. Mikos and B. Moslin)
"Perception of vowel length in Czech and English", LSA, December 1978
"Fundamental frequency and vocal registers in the speech of prelinguistic infants", New England Child Language Association workshop, April 1979.
P. Lieberman et al., "Speech development in infants--vowel production", J. Acoust. Soc. Am. 59, S43(A), 1976.
P. Lieberman et al., "Development of Vowel Production in Infants", LSA, Dec. 1976.

TABLE OF CONTENTS

Acknowledgements

I would like to thank my advisors, friends, and family for all their help and encouragement throughout the past year. The members of my committee--Sheila Blumstein, Phil Lieberman, and Francis Ganong--are selfless and cooperative above and beyond the call of duty, as all who know them can attest. They deserve special thanks for helping me finish (more or less) on time. Other helpers in the eleventh hour have been Elan Dresher, Barus Lab-ites Cathy Kubaska, Carol Chapin, Jack Ryalls, and John Mertus, and typist Liz Stevens. Not forgotten are past colleagues Bob Buhr, Michael Mikoś, and Barbara Moslin. They have gone on to greater things, as I too now hope to do.

Finally, many thanks must go to all the native speakers of Polish, here and in Poland, for providing the data on which this thesis is based.

CHAPTER ONE -- Introduction

1.1. Phonological Features and Their Phonetic Implementations

1.1.1. Relation of phonological features to speech signal

A primary goal of linguistic phonetics is to relate acoustic signals to units established on linguistic grounds. One such linguistic unit is the distinctive feature, which is motivated within the phonological component of the grammar. Distinctive features describe oppositions among phonemes and allow phonological rules to be represented in terms of natural classes. Since distinctive features were first posited, there has been an attempt to locate them in the speech signal. An important question is whether these linguistic units are in fact present acoustically, and extracted by the perceptual system. It is known that many aspects of the speech signal can serve as acoustic cues to distinctive features, that there are many such cues to each feature, and that these cues vary according to their context (e.g. Lisker, 1977). The issue, then, is whether some cue or cue-complex can uniquely specify a distinctive feature in all contexts (Stevens and Blumstein, 1978). Only if this is the case would distinctive features be "found" in the speech signal. If this is <u>not</u> the case, then some more complex description of how the cues relate to the features is required, allowing for contextual dependencies and cue multiplicity. In the sense that the features would be derived from the cues, they are an abstraction, one step removed from the physical utterance.[1]

1.1.2. Relation of features across languages

The same, limited set of distinctive features appear to be available across languages; each language uses some subset in its phonology. However, across languages, the sounds which these distinctive features describe differ phonetically. In describing possible differences between sounds (even if those differences are never contrastive), a larger, but still limited, set of phonetic features can be used. This set is larger simply because some audible differences in sounds across languages never seem to be used phonologically by any one language. The phonetic features describe all possible human speech sounds, as constrained by human production and perception abilities. In any one language, phonetic features can serve as redundant features providing phonetic detail beyond that conveyed by the distinctive features.

A possible constraint on linguistic theory would be to require that the distinctive features of any language form a subset of the phonetic features. Or linguistic theory could allow the two sets to differ, in which case some mapping between them would have to be provided. All things being equal, the first approach would be preferred for reasons of parsimony, but it will be suggested below that the second alternative has advantages over the first.

Below the phonetic level, there must be some actual physical level, where acoustic cues and the articulations that produce them are described. The phonetic level may be thought of as integrating the production and perception information from the physical level, although the phonetic level itself is not a physical level.

1.1.3. Levels of representation

The question of how linguistic units relate to the speech signal has now been narrowed to the question of how phonological features relate to phonetic features. One form this question has taken is whether phonological features should have "phonetic content", that is, consist of phonetic parameters. One extreme view, held by Fudge (1967) and others, is that there is not necessarily any connection between the phonological and the phonetic features. Fudge in particular was concerned that a single set of features could not serve the two purposes of phonological and phonetic description equally well. Since he assumed that these features had to be either identical or unrelated, he proposed having two unrelated sets of features with an arbitrary mapping between them. The opposite view also assumes that, to be related, phonological and phonetic features must be identical, and so posits this identity. This view is exemplified in Chomsky and Halle (1968) (SPE) and in Ladefoged (1971). This identity is attained by replacing the older notion, Jakobson, Fant and Halle (1952)'s "distinctive" features, with "phonetic" features that can be used as "phonological" features. The phonetic features are designed to provide inherently non-contrastive detail about speech sounds, which distinctive features were not.

The view espoused in SPE warrants special attention, since it is the most explicitly stated, and the most widely-accepted, view of the relation between phonological and phonetic features to date. In this system, the phonological component describes the systematic sound patterns in a language, such as distinctive oppositions between individual sounds, the natural classes of those sounds, and

alternations of sounds in the various realizations of a single morpheme. The input to the phonological component is lexical representations and grammatical information. The phonological rules convert underlying forms to phonetic forms which in turn serve as the input to rules of speech production (or are the output of speech perception devices). The same features are used throughout the system. At the phonological level these features are binary, but at the phonetic level they assume scalar values which provide phonetic detail. Phonological and phonetic forms are represented as matrices, with segments as columns and features as rows. Phonological rules can add or delete segments, or change the value of a feature, represented in a cell of the matrix, but they cannot change the name assigned to any row (feature) in the matrix, at any stage.

This constraint on features is an attempt to ensure that phonological features have phonetic content. The Naturalness Condition formulated by Postal (1968) stipulates that the same features be used at both levels so that phonological rules and representations can be judged for phonetic naturalness. Thus the use of a single set of phonetic features is motivated by a desire to have an evaluation measure that uses phonetic naturalness as one criterion in rating alternative grammars.

One problem facing the SPE system is that linguistic representations are segmental, while articulations and acoustic signals are not. The SPE system is constrained to have segmental representations at all levels, since the phonological level requires them. This forces some ad-hoc descriptions of the phonetic and physical levels (see below) to avoid representing the encoded nature of

speech. The reverse problem imposed by the SPE system is that phonetic detail derived from the physical level is codified as features and as such carried into the phonological level and lexical representations.

A third answer to the question of phonetic content of phonological features is proposed by Lieberman (1970, 1977). Lieberman's "unified" approach requires phonological features to have phonetic content, but not to be identical to phonetic features. Instead, phonological features are implemented as phonetic features, perhaps differently by different languages. However, Lieberman does not specify how the phonetic theory will constrain the relation of the two sets of features. The features are to be optimal for the level at which they are used, but they must be related by the theory.

One purpose of the present study is to consider how different features could be used at different levels of the grammar. The goal is to provide sufficiently abstract, but phonetically-motivated, features at the phonological level, and sufficiently accurate phonetic and physical features. A separate issue is how many types and levels of features are required in the grammar.

In the next section, various treatments of a particular contrast, voicing, are presented. Intuitively, the feature involved is similar across languages. The different accounts are presented, however, to illustrate how linguists are influenced by phonetic variation when describing this contrast. Nonetheless, it will be suggested that, when correctly represented, this phonetic variation can be used to motivate a single phonological feature.

1.2. An Example: Voicing Contrasts

1.2.1. Universal feature of voicing

One feature that seems to occur in contrasts in many unrelated languages is that of "voicing". Almost all languages contrast "voiced" and "voiceless" obstruents; a few languages have sonorant contrasts as well. Many languages have phonological rules which refer to the voicing feature, in particular, rules which assimilate the voicing of one consonant to that of another, and rules which change a voicing feature depending on position within a word. One example is cluster assimilation in Slavic languages and French, where the first of two consonants assimilates in voicing to the second, e.g. Polish /farba/ ~ /farpka/, /kot/ ~ /kod#v#butax/. Another example is plural formation in English, where the phonetic shape of the plural morpheme depends in part on the voicing of a preceding consonant, the final segment of the root, e.g., /kaets/ vs. /kædz/. The second type of example is rules such as word-final devoicing, which occurs in, e.g., Slavic languages, German, and some dialects of English.

The interesting point is that these processes are <u>all</u> known as "voicing assimilation" and "devoicing", even though the so-called "voicing" contrasts involved may differ phonetically. In fact, they differ enough phonetically that most feature systems developed do not treat them as equivalent, as will be seen below. Nonetheless, they intuitively seem to be similar, as evidenced by the tendency to equate the rules, if not the contrasts themselves.

A unified cross-linguistic representation of the "voicing" feature is a goal of the present study.

1.2.2. Feature systems

Traditional descriptions of voicing differ in their treatment of the aspiration in English initial stops. Some authors consider the aspiration grounds for separating the English contrast from other contrasts which do not involve allophonic aspiration.

Heffner (1969)'s account provides the feature-equivalence discussed above. He considers any contrast of the type /b d g/ ~ /p t k/ to be one of voicing, even in English, where, unlike in French, the "voiced" stops are not voiced throughout "their total length". Other languages contrast /p t k/ ~ /p^h t^h k^h/ by a feature of aspiration or "fortis release" (presumably languages with a three-way contrast). The allophonic aspiration of English voiceless stops does not enter into the description of the basic contrast.

Trubetzkoy (1969) used three phonetic features, [±voice], [±tense], and [±aspirated] to describe various voicing categories, but while the same phonetic features could be used for various languages, the distinctive feature for each language must be determined independently by phonological evidence. Slavic languages have distinctive [±voice], redundant [±tense], and no aspiration. French, English, and German have co-varying [±voice] and [±tense], but it is impossible to say which is distinctive and which is redundant. In English, [±aspiration] is an allophonic feature associated with voiceless, tense stops.

Jakobson, Fant, and Halle (1952) decided that in fact tenseness is distinctive, at least for English and French, so that those two languages have redundant voicing, while Slavic languages have distinctive voicing. In Danish, [±tense] is also distinctive, but all stops are voiceless. Tenseness and aspiration are related,

and tenseness is used in describing three-way contrasts, along with voicing.

Another approach is to recognize three or more basic phonetic voicing categories, and describe voicing contrasts as a choice of two or three of these categories. Abercrombie (1967) describes the three categories "voiced", "voiceless unaspirated", and "voiceless aspirated". These categories were given an articulatory and acoustic basis by Lisker and Abramson (1964), who describe five phonetic categories along a single voicing dimension. The phonetic feature is known as voice-onset time (VOT); it will be described in detail in the next section. The five phonetic categories can be combined into the three potentially contrastive categories of Abercrombie. In this view, aspiration is not an independent feature, but a natural concomitant of one of the voicing categories. Ladefoged (1971) also adopted this approach, with five phonetic categories, "fully voiced", "partly voiced", "voiceless unaspirated", "voiceless slightly aspirated", "voiceless aspirated". According to Ladefoged, French voiced stops are "fully voiced", while English ones are "partly voiced". According to Lisker and Abramson, the English voiced stops are usually voiceless unaspirated.

Chomsky and Halle (1968) were aware of Lisker and Abramson's findings and VOT dimension, but they wished to maintain independent binary features, rather than multi-valued scales, at the phonological level. Also, as mentioned above, temporal features are less appropriate when a strict segmental representation is desired. Therefore Chomsky and Halle chose to describe four of the five VOT categories with four binary features, all of them based on articulatory configurations. Since these features are no longer in use

(Sommerstein, 1977) and rather complex, they will not be described here. One point should be made, however. The articulatory detail given by these four features was so fine that they described different articulations with the same acoustic and phonetic output (see Lisker and Abramson, 1971, Figure 2, derived from SPE p. 328). This inherently non-contrastive information was then carried along through to the morphophonemic level. For example, subglottal pressure was used as a phonological feature in languages with underlying aspiration contrasts. The same articulatory detail marks Halle and Stevens (1972)'s feature of stiff/slack vocal cords, which ascribes an invariance to articulation that is probably unwarranted (e.g. Westbury and Niimi, 1979). More recent work in generative phonology uses some version of [± voice] plus [± aspiration] to represent voicing contrasts and phonetic forms.

Thus, in sum, there have been four general approaches to cross-language "voicing" contrasts. One is that all languages have a voicing contrast, and that allophonically English has aspiration. Another is that some languages have a voicing contrast, but that other languages have a tenseness contrast. English then has allophonic aspiration precisely because it has contrastive tenseness rather than voicing. The third approach is to have a more elaborate set of features, like those of SPE, which incorporate explicit phonetic descriptions.

The fourth approach is quite different. It allows the basic phonetic categories to be described by a single feature, VOT, and leaves open how that phonetic feature is to be represented at the phonological level. It also provides a fairly explicit account of the articulatory and acoustic parameters associated with it.

Lieberman (1970) proposes separating the phonologically contrastive feature, [± voice], from the phonetic implementation of that contrast, particular VOT categories for each language. As he says, "It is not necessary to invoke an additional feature of tenseness to explain the differences between English and Russian stops. The difference resides in the way that the feature of voicing is implemented in these languages." (Fn. 3, p. 308). Phonological features are implemented by universal, language-specific, and individual-speaker "implementation rules" which determine the articulation to be used. This approach seems preferable to one which allows low-level phonetic differences to affect phonemic representations. The unity of the abstract feature "voicing" seems to outweigh any effect these differences may have at the phonological level.

The present study is an attempt to develop this two-tiered feature system in a way that will allow cross-language contrasts to be equivalenced at one level but accurately differentiated at another. The feature considered is that of voicing, implemented differentially across languages as VOT categories. In the next section, the phonetic feature VOT is described.

1.3. VOT Voicing Categories

1.3.1. Definitions

Lisker and Abramson, in their pioneering 1964 study, defined VOT as a single production dimension, the time interval between the release of a stop occlusion and the onset of vocal fold vibration. As such it is meant to be a cover-term for various laryngeal events that can affect this timing relation. The acoustic manifestations of this articulatory dimension are diverse. The current

use of "VOT" to describe one of these acoustic manifestations, the time between the burst and the first periodicity in the acoustic signal, is a much narrower use of the term "VOT" than intended by Lisker and Abramson. The entire set of acoustic correlates of the articulatory dimension include not only the timing of periodicity, but also burst intensity, aspiration, F1 cutback, onset frequency of voiced transitions, and fundamental frequency variation (Abramson, 1977). In this view, aspiration is simply the result of turbulent air passing through the glottis and resonating at vocal tract frequencies during a voicing lag, rather than being an orthogonal feature, such as tenseness. The Lisker and Abramson scheme in fact does away entirely with the feature tenseness. In systems with this feature, muscle "tension" for voiceless stops is supposed to inhibit voicing by preventing the pharynx from expanding. Experimental evidence has indicated, however, that the situation can actually be completely reversed. The pharynx can be actively expanded to permit voicing during stop closure (Bell-Berti, 1975), and thus this feature has no physical basis.

The Lisker and Abramson approach defines three overall contrastive VOT categories corresponding to the phonetic categories "fully voiced", "voiceless unaspirated", and "voiceless aspirated". In general, most languages include two or three of these among their stop contrasts. (A few languages, such as Korean and Hindi, have additional stop categories that cannot be accounted for by laryngeal timing.) Fully voiced stops are those with voicing during closure, that is, where voice onset leads before the release ("prevoiced"). Voiceless unaspirated stops are those where voice onset lags after the release by up to 20-25 msec ("short lag"). Voiceless aspirated

stops are those for which voice onset lags after the release by more than 25 msec ("long lag"). VOT is usually represented graphically as a continuum, with voicing at the stop release set equal to 0 msec VOT. Stimuli with voicing lag have positive VOT values; stimuli in which voicing leads the release have negative VOT values.

1.3.2. Cross-language studies

Lisker and Abramson (1964) showed that VOT is an effective measure for separating word-initial stops into phonemic voicing categories in eleven languages. They measured VOT values for productions of absolute-initial stops in each phonological voicing category in Dutch, Spanish, Hungarian, Tamil, Cantonese, English, E. Armenian, Thai, Korean, Hindi, and Marathi. Although the mean VOT values were not identical across these languages, they clustered in ways that allowed the three potentially-contrastive categories to be defined.

Further data on production of VOT has been collected for Spanish by Williams (1977), and for English by several investigators, among them Zlatin (1974) and Moslin (1978). Additional languages studied include Kikuyu (Streeter, 1976), Arabic (Yeni-Komshian, Caramazza, and Preston, 1977), and French (Caramazza and Yeni-Komshian, 1974). The VOT categories of a number of languages are indicated in Table 1. For all these languages, VOT is sufficient to distinguish the word-initial phonological stop categories, except as noted above. The only cases of overlap occur when 1) with a prevoiced-short lag contrast, some tokens that should be prevoiced are not, resulting in short lag VOT values; 2) in rapid conversational style, the English contrast of short lag-long lag is somewhat reduced. However, except for English, running speech

TABLE 1

Use of contrastive VOT categories by several languages, based on Lisker and Abramson (1964). English use of prevoicing is allophonic free variation. Note that the last three languages have stop categories which are not differentiated on the basis of VOT.

	VOT categories		
Language	prevoiced	short lag	long lag
English	(X)	X	X
Spanish	X	X	
Thai	X	X	X
Dutch	X	X	
Hungarian	X	X	
Tamil	X	X	
Cantonese		X	X
E. Armenian	X	X	X
Korean	X	X (X)	X
Hindi	X (X)	X	X
Marathi	X (X)	X	X

environments have not been studied extensively.

Since Lisker and Abramson's demonstration of the usefulness of VOT in describing production, most attention has centered on its usefulness in describing perception. Data on both labeling and discrimination of VOT have been obtained and related to the production data. Typically, a series of stimuli are produced synthetically such that their VOT values vary in small (e.g. 5 or 10 msec) steps along a pseudo-continuum. Listeners are presented these stimuli individually for labeling and in various combinations for discrimination. Such perception experiments have been carried out for a subset of the languages for which production has been studied (see below). It is generally expected that the labeling categories will match the production categories, and that there will be a discrimination peak corresponding to each category boundary. This expectation has been met consistently for English (e.g. Lisker and Abramson, 1970; Abramson and Lisker, 1970; Zlatin, 1974). In fact, this relation is so reliable that it is the basis of much further experimental work in speech perception. All of the other languages studied have shown a somewhat less consistent relation between production, labeling, and discrimination. In particular, production categories are often not aligned with perceptual categories. Table 2 and the following discussion summarize the more important of these studies.

Some data is available for Lebanese Arabic (Yeni-Komshian, Caramazza, and Preston, 1977). This language has a voicing contrast only for apical and emphatic apical stops, with prevoiced and short lag categories. For both pairs, there is a fair amount of production overlap in VOT, since voiced stops are sometimes

TABLE 2

A summary of results of studies on the perception of VOT in languages, with production data given for comparison. All values are in msec VOT.

Place of Articulation	Production Ranges	Labeling Boundaries	Discrimination Peaks
THAI: Lisker and Abramson, 1970 Abramson and Lisker, 1970			
labial	-150/-40;0/+10 +30/+80	-20; +40	-20; +30
apical	-150/-90; 0/+20 +30/+120	-10; +45	-
velar	- ; 0/+40 +50/+150	---; +40	-
THAI: Donald 1976, 1978			
labial	-	-22; +25	-20; +20
velar	-	---; +33	-20; +20
ENGLISH: Lisker and Abramson, 1970 Abramson and Lisker, 1970			
labial	-130/0; +20/+90	+22	+20
apical	-120/+20; +30/+100	+37	---
velar	-150/+30; +50/+130	+40	---
ENGLISH: Kuhl and Miller, 1978			
labial	---	+27	---
apical	---	+35	---
velar	---	+42	---
ENGLISH: Zlatin, 1974			
labial	-210/+50; +10/+220	+32	---
apical	-200/+50; +30/+180	+27	---
velar	-210/+50; +40/+170	+66	---
SPANISH: Lisker and Abramson, 1970 Abramson and Lisker, 1972			
labial	-150/-60; 0/+20	+14	no pooled data given; various numbers of peaks per subject
apical	-150/-90; 0/+10	+22	
velar	-130/-40; +10/+50	+24	

SPANISH: Williams 1974,1977			
labial	-180/-30; +5/+45	-10	---
apical	-160/-40; +10/+40	---	---
velar	-180/-35; +15/+135	---	---
KIKUYU: Streeter, 1976			
labial	x = -65, s = 23.2	---	-15,+20
apical	x = -62, s = 21.2 x = +10, s = 8	---	+10
ARABIC: Yeni-Komshian et al., 1977			
labial	-160/+20; --	---	---
apical	-130/+30; 0/+115	+5	---
(emphatic)	-130/+30; 0/+80	---	---
velar	--- ; 0/+120	---	---
uvular	--- ; 0/+110	---	---
FRENCH: Caramazza et al., 1974			
labial	-150/+10; 0/+25	---	---
apical	-100/0; +15/+35	---	---
velar	-125/0; +20/+40	---	---
FRENCH: Simon and Fourcin, 1978			
apical	---	+15	---

produced without prevoicing. This was found for most of the speakers, in tokens that were judged by other speakers to be perfectly acceptable voiced stops. The speakers were essentially monolingual. Thus it seems likely that prevoicing is not the only correlate of voicedness in Arabic, and further work is needed to clarify the role of VOT there. Labeling was also assessed, through free imitation rather than a forced-choice paradigm, and the category-boundary values given were estimated. The best estimate of the group boundary was about +5 msec VOT, although a wide range of individual boundaries was observed.

Data is also available for Kikuyu (Streeter, 1976). Although data is presented for both labials and apicals, only the latter have contrastive categories. She did not display the actual frequency distribution of VOT measurements, but based on the means and standard deviations she reported, there seems to be little or no category overlap. Discrimination but not labeling was assessed. The observed "peak" in discrimination is extremely wide, covering the entire short lag region of the VOT continuum. Its center is at +10 msec VOT, which is the mean production value for the voiceless category. Thus the discrimination peak is too wide and too high. Streeter concluded that her subjects had performed more on the basis of psychoacoustic than linguistic distinctions.

The most-studied languages other than English have been Spanish and Thai. Spanish is a two-category language with prevoiced and short lag stops. However, the stop voicing contrast is maintained only after a pause or a liquid. In all other positions,a voiced stop is spirantized. Therefore Spanish stop voicing has not been studied in non-initial positions, or in running speech, since the number of tokens would be small. Thai is a three-category language,

except that there is no prevoiced velar stop. All the data obtained for Thai have been for initial position.

Spanish and Thai were the languages selected by Lisker and Abramson, with English, for the first studies of perception of VOT (Lisker and Abramson, 1970; Abramson and Lisker, 1970). The Spanish production data gave a clear category separation in the lead region, but the labeling boundaries fell in the lag region, in the middle of the voiceless category. Subjects were unable to carry out a discrimination task. For Thai, the prevoiced - short lag contrast showed fair agreement between production, labeling, and discrimination data. The short lag - long lag contrast showed labeling boundaries that were too high for the production data, and discrimination peaks that were even higher. Thus, the earliest studies were not entirely in accord with a VOT analysis of these contrasts, although they were largely interpreted as such.

Abramson and Lisker (1972) repeated experiments in discriminating Spanish labial voicing categories. A wide range of peaks is apparent in the individual data. Abramson and Lisker concluded from this that both linguistic and psychoacoustic factors influence discrimination performance.

Williams (1974, 1977) also studied Spanish labial stops, and obtained a good match between production, labeling, and discrimination. The group category boundary fell at about -10 msec VOT. Williams (1977) then considered whether prevoicing is a <u>necessary</u> cue for the voiced category, by editing prevoicing from natural tokens and presenting them to listeners for labeling. Overall, these stimuli were still labeled as voiced, but significantly less often than the unedited stimuli. That is, prevoicing increases

the likelihood of a "voiced" response, but there are other cues to voicedness in the stimulus that remain available when prevoicing is removed. Williams hypothesized that properties of the onset such as absence of a release burst, and continuous voicing through the release, could act as such cues. (Recall that in Spanish most voiced stops actually occur as voiced fricatives.) Williams suggested that the manner and voicing features are conflated in the "property" [± abrupt onset]. Prevoicing, continuous voicing, lack of a release burst, and spirantization would all be [-abrupt onset] and cue voicedness.

Donald (1976, 1978) obtained further labeling and discrimination data for Thai. She found a better fit between perception and production than Lisker and Abramson did, but there was a tendency on the part of some subjects to place the short lag - long lag boundary higher in the labeling task than in discrimination.

Some research has been directed at the question of whether the VOT values within a category are affected when there happens not to be a contrast at a particular place of articulation--that is, when there is a "hole" in the system of voicing contrasts. For example, in Kikuyu, there is a prevoiced but no short lag stop at the labial place of articulation. Streeter (1976) found that the production values for the prevoiced stop were in line with those for the contrastive apical prevoiced stops. (Since VOT varies systematically with place of articulation, we would expect the values for the labials to be slightly lower than those for the apicals, which they are.) In a discrimination task, listeners showed two peaks along a labial VOT continuum--one between the prevoiced and short lag stops, just as if there were a contrast, and also one between

short and long lag values, which are never contrastive in Kikuyu.

Donald (1978) looked at a similar situation in Thai, a three-category language, where there is no prevoiced velar stop, but the velar short lag - long lag contrast is maintained. Again listeners showed a discrimination peak even where there is no contrast, but that peak, between prevoiced and short lag stops, was weaker than the peak found between the contrastive short and long lag velar stops. Dent (1976) measured VOT production values for Spanish voiceless stops in running speech, where, it will be recalled, the contrast is not only one of voicing but also of manner. Again, the integrity of the voicing category was maintained by speakers, who produced voiceless stops with consistent short lag VOT values.

Thus, studies of non-contrastive VOT categories have found that production values remain confined to a single VOT range, as if there were a contrast. Discrimination, however, is improved at the non-linguistic category boundary. That is, when there is no contrast, listeners' performance is influenced by psychoacoustic factors as well.

Overall, the results of studies of VOT have been mixed. In most cases, the problem has been that perceptual boundaries lie about 10 to 20 msec higher than the production boundaries. This problem may be due to non-optimal settings of various parameters in some of the stimuli used, that somehow make the effective VOT value seem perceptually shorter than the measured VOT value.

The precise acoustic manifestations of the VOT dimension, and how an "effective" VOT value is derived by a listener, are still not well understood. Except for Williams' work on Spanish onsets, which was preliminary, work on other voicing cues in initial

position has been focused on English. As stated before, Lisker and Abramson have always insisted that VOT is a complex of co-varying cues. The question arises, however, as to which of these co-varying cues the perceptual system is actually sensitive to. Individual languages and/or speakers may have a hierarchy of cues to which they attend. Some cues which have been shown to be relevant to English perception include aspiration (Winitz et al., 1975; Repp, 1979), voiced transition onset frequency or duration (Stevens and Klatt, 1974; Lisker, 1975; Summerfield and Haggard, 1977), fundamental frequency after release (Fujimura, 1971; Haggard, Ambler, and Callow, 1970),and burst intensity (Repp, 1979). Detailed acoustic descriptions of burst intensity and duration are given by Klatt (1975) and Zue (1976), without perceptual data. The first two of these, aspiration and transition onset, may be of limited use in languages with a prevoiced - short lag contrast. Most short-lag VOT values fall at about +20 msec; since the burst itself can easily comprise the first 10 msec of this interval, the transitions might be just underway at the point of voice-onset. Thus almost all of the transitions' extent will be voiced, so that there is no F1 cutback, and no aspiration (cf. Simon and Fourcin, 1978, for an interpretation of labeling responses of French children).

It is certainly possible that the earliest VOT continua made were simply not satisfactory, and that they were more suited to English than to other languages. However, it is also possible that the discrepancies between production and perception categories may be due to differences in the ranges of VOT values used in various studies. One fact that can be noted is that the mismatched

perceptual boundaries occur in those studies which used a wide range of VOT values in their test continua. The original Abramson and Lisker continua extended from -150 to +150 msec VOT, which covers the range of productions for Thai. Most subsequent studies used most or all of this range. However, the two sets of studies which were successful in matching production and perception, those of Williams for Spanish and Donald for Thai, used much smaller ranges which focused more on the suspected category boundaries: from -40 to +40 msec for Williams, and from -50 to +50 msec for Donald. This factor, then, also warrants consideration in studies of languages other than English.

In sum, it can be said that, although some studies of VOT have certain drawbacks that would make VOT seem suspect as a phonetic universal, other studies have apparently overcome those problems, and give promising results. More work is needed on the influence of other cues besides the timing of the onset of periodicity, and on the effect, if any, of the range of stimuli used on perceptual categories. Nonetheless, VOT seems a reasonable phonetic feature for a study of cross-linguistic voicing contrasts.

1.3.3. Psychophysical basis for VOT contrasts

In an account of cross-language voicing contrasts, another issue which must be addressed is whether one category or one contrast is more highly-valued, and preferred by languages, than another (Stevens, 1972). In fact, all languages with voicing contrasts seem to use the short lag category. In addition, they choose one or both of the other categories. No language contrasts prevoiced and long lag stops alone.

Several lines of research have suggested that there is a psychophysical basis for three VOT categories. One line is the study of discrimination of phonetic contrasts which are not distinctive in a particular language. For example, Streeter and Landauer (1976), in a study of Kikuyu children learning English, found that the first graders had two discrimination peaks for a labial VOT continuum. (Kikuyu has a prevoiced labial stop but no contrast.) One peak corresponded to the Kikuyu contrast of prevoiced - short lag VOT found for apical and velar stops. The other peak corresponded to the English contrast of short lag - long lag VOT. Another example is discrimination of the English contrast by monolingual speakers of Spanish. Both Abramson and Lisker (1972) and Williams (1974) found secondary discrimination peaks at about +35 msec along the VOT dimension which cannot be attributed to linguistic experience, since Spanish does not use this contrast. On the other hand, speakers of English do not appear to discriminate the Spanish prevoiced - short lag contrast; that is, this ability occurs in only one direction.

Investigations of infant perception also show discrimination abilities which cannot be entirely due to linguistic experience. Both Spanish and Kikuyu infants show discrimination peaks between three VOT categories. (Lasky at al., 1975; Streeter, 1976). Again, however, this ability to make discriminations not in the language does not apply to English. American infants show only one clear peak, corresponding to the English VOT contrast, and merely a tendency to discriminate prevoiced from lag VOT values, given a large enough contrast (Eimas et al., 1971). Also, the Spanish and Kikuyu infants' peaks do not quite correspond to adults' peaks.

For example, Williams (1974) found that Spanish adults have their category boundary at about -10 msec VOT, but Lasky et al. (1975) found that Spanish infants did not discriminate -20 from +20 msec VOT. Rather, their peaks lay beyond -20 and +20 msec.

To explain this non-linguistic discrimination ability, it has been hypothesized that some minimum temporal separation between major acoustic events within a stimulus is required for those events to be distinguishable, for example, stimulus onset and voice onset. This explanation was tested in two experiments with adult speakers of English. Miller at al. (1976) used stimuli that presented a buzz at various time intervals after a noise burst, thus similar to lag VOT stimuli. Listeners showed a category boundary at about 17-20 msec temporal separation. Pisoni (1977) used stimuli with two tones, at 500 and 1500 Hz, where the lower tone led and lagged the higher one in intervals up to 50 msec. These stimuli thus contain a "cue" somewhat like the F1 cutback cue used in synthetic VOT continua. Most listeners showed category boundaries at about +20 and -20 msec temporal separation of the tone onsets.

A fourth line of research has been investigations of animal perception of human speech categories. Perception of VOT has been studied in rhesus monkeys (Waters and Wilson, 1976) and chinchillas (Kuhl and Miller, 1975, 1978). The rhesus monkey study used large differences in VOT (70-msec steps) from -140 to +140 msec VOT. The 0/+70 msec pair was discriminated best, but performance varied depending on the range of stimuli used in a given test. Since Waters and Wilson were looking only for an English-like boundary, they used a forced-choice format (their "discrimination" task is

really more like labeling) that did not allow for two possible boundaries along the "continuum". Thus it is difficult to interpret these results. The Kuhl and Miller studies with chinchillas were more detailed, but they also were not looking for boundaries in the lead region. They used labial, alveolar, and velar continua with VOT values from 0 to +80 msec in 10-msec steps. The chinchillas showed category boundaries like those of English-speaking humans, but their identification functions were less steep. The chinchillas' boundaries varied according to place of articulation, just as the humans' did, from about +25 msec for labials to about +42 msec for velars.

These several studies, taken together, suggest the following:

1) There is a relevant non-speech psychoacoustic boundary at both -20 and +20 msec temporal separation. However, the basic auditory boundary for lag VOT in speech is higher than this, for example, +35 msec VOT for apical stops, corresponding to the English voicing contrast.

2) There is no psychoacoustic support for boundaries from about -10 to +10 msec VOT, although this is the region for boundaries in languages contrasting only prevoiced and short lag voicing categories, such as Spanish.

3) Some infant and animal studies could be tapping either the speech-psychoacoustic or noise-psychoacoustic category boundaries, considering the relatively large step sizes generally used in such studies.

4) The English short lag - long lag category boundaries seem stronger, more salient, than the prevoiced - short lag boundaries. There seem to be three psychoacoustic timing categories for non-speech stimuli, and non-speakers of English evidence three VOT

categories for speech, as do infants from these other speech communities. However, English speakers and their infants do not show a separate perceptual category for lead VOT values. The differential behavior of the English vs. non-English infants is puzzling, but seems to indicate that even a few weeks or months of exposure suffices to activate the prevoiced - short lag contrast for infants.

With the strong correspondence of the English type of voicing categories to basic auditory processing, and the lack of such correspondence for some other languages' contrasts, it is interesting that the prevoiced - short lag contrast is quite common among languages. In fact, it may be more common than the perceptually-easier short lag - long lag contrast. Therefore one object of the present study is to consider why the prevoiced VOT category is so "popular," in particular, by looking at its production and perception in various contexts.

1.4. Study of Polish Voicing

The present study is an investigation of the production and perception of a voicing contrast in Polish. Polish is a West Slavic language spoken by at least 32 million people in Poland, and, as a native language, by a few million others outside Poland. Polish, like other Slavic languages, is described as having voiced and voiceless unaspirated stops, and preliminary observations indeed confirmed that these have lead and short lag VOT values. Thus, as regards voicing contrasts, Polish is a two-category language, which, like Spanish, French, Arabic, etc., uses the phonetic contrast which English does not use. Unlike Spanish, however, Polish

has post-vocalic voicing contrasts, and is therefore more suited to a study using running speech. An additional advantage of Polish is that its spoken dialects are fairly uniform for educated speakers. Also, its phonology has been studied (e.g. Mikoś, 1977), and the processes involving voicing contrasts are understood.

The study is limited to the contrast of the apical stops [t] and [d] in non-cluster environments. Table 3 gives the surface phonemes of Modern Standard Polish, following Mikoś (1977). In this treatment, [t] and [d] are the major allophones of phonemes /t/ and /d/, respectively. Palatalized allophones also occur, before the high front vowel /i/; [t] and [d] occur before the high back unrounded vowel /ɨ/ and before all the other vowels. As the result of historical changes, some /t,d/ alternate morphophonemically with /ć,ǵ/ before front oral vowels. Overall, the voiceless consonant is more common that the voiced, as it enters into many more consonant clusters, and occurs exclusively before pause. There is a Morpheme Structure Condition that obstruent clusters agree in voicing, and most underlying two-obstruent clusters are voiceless. There is a rule of progressive obstruent voicing assimilation within and across words, and a rule of prepausal obstruent devoicing. Thus the Polish [t] - [d] contrast is almost prototypical. There is no question as to the phonological feature involved, as there is for other languages.

The experimental portion of this study is contained in Ch. 2-4. Production and perception experiments aim to arrive at a description of the acoustic nature of the Polish contrast. Ch. 5 uses these results to consider cross-linguistic low-level variation in voicing, and factors favoring each type of contrast. Finally, an

TABLE 3

Surface phonemes of Modern Standard Polish, from Mikoś (1977).

CONSONANTS:

	place of articulation					
manner	bilab	lab-dent	dental	alveolar	palatal	velar
stop	p b		t d			k g
fricative		f v	s z	š ž	ś ź	x
affricate			c ʒ	č ǯ	ć ź	
nasal	m		n		ń	
liquid			l r			(ɫ)
glide					y	w

VOWELS: i ɨ e a o u ẽ õ

attempt is made to justify a multi-level feature system, as discussed above.

FOOTNOTES

[1]This representation of distinctive features as linguistic but not physical units makes no claims at all as to their "psychological reality". It simply excludes them from phonetic processing, but still allows them a role in cognitive activities such as acquiring the phonology as a child, lexical storage, etc.

CHAPTER TWO --

Production and Perception of VOT in Word-Initial [t] and [d]

In this chapter I will examine the effectiveness of VOT in differentiating voiced and voiceless initial apical stops in Polish. In the first section I will consider acoustic measurements of VOT; in the second section I will consider perceptual data.

2.1. Production Measurements

2.1.1. Methodology

2.1.1.1. Recorded speech samples

a. Wrocław minimal pairs

A list of minimal pairs with [t] ~ [d] contrasts was constructed for Polish by MJM.[1] Five pairs were recorded by 24 speakers in Wrocław, Poland. Four of the pairs have word-initial contrasts, and one has a medial contrast.[2] Two of the pairs consist of disyllables with a stressed vowel [a] after the apical stop, and two consist of monosyllables, with the vowels [a] and [u].

The 24 speakers were students at the Institute of Telecommunications and Acoustics; all spoke Modern Standard Polish. Most read the five pairs once; some read them twice. Recordings were made on a cassette through the kindness of Dr. Wojciech Majewski.[3]

b. Polish sentences

A set of sentences was constructed by MJM for a study of Polish intonation. The word-initial tokens of [t] and [d] that happen to occur in these sentences were analyzed for the present study. A recording of five speakers in Poland, who also spoke Modern Standard Polish, reading these sentences was made available by MJM. The recording was made in the radio station in Łódź where

the speakers were employed.

All the word-initial [t] 's happen to occur in the word to, sometimes as a stressed pronoun, sometimes as an unstressed adjective, with a total of four tokens in the sentences. All the word-initial [d]'s occur before the vowel [o], in unstressed prepositions, and in stressed and unstressed syllables of content words, for a total of 14 tokens of word-initial [d] .

c. Providence pairs and sentences

Recordings of minimal pairs and sentences were also made by three native Poles in Providence. Strict standards were set for the inclusion of local residents in this study: that the subjects be near-monolingual (i.e. have difficulty answering simple questions or following instructions in English), and that s/he be visiting other Poles in the US for a limited time. The three Poles who met these criteria were JP, a university researcher about 45 years old, MG, another researcher about 30 years old, and DB, a high school student 16 years old. JP and MG spoke educated standard Polish; DB spoke somewhat less educated Polish.

Recordings were supervised by MJM; the language spoken during recording sessions was Polish. The minimal pairs were presented, in pairs, on a single sheet of paper in order to highlight the [t] - [d] contrast. Speakers were not told to emphasize those segments, nor what the purpose of the experiment was, but it was hoped that presentation of the words in such a format would encourage distinctiveness in the readings. Speakers were told to pause between words, in order to avoid the effect of running speech. Recordings were made in the home or office of the speaker on a high-quality recorder (Nagra 4.2) at 3 1/2 ips tape speed.

The set of minimal pairs that these speakers read consists of 14 [t] - [d] contrasts. Twelve of the pairs have word-initial contrasts, and two have medial contrasts. The word-initial contrasts occur in both one- and two- syllable words, with the vowels [i u o a] following the initial apical stop.[4] These speakers read the same list of sentences as the speakers recorded in Poland, except that one sentence with initial [t] and one with initial [d] were omitted. Thus there were three tokens of [t] and 13 tokens of [d].

Table 1 lists all the words from the minimal pairs and the sentences; the complete set of sentences is given in the Appendix.

MJM was able to engage one speaker, JP, in a lengthy spontaneous conversation about English borrowings into Polish and vice versa. This conversation was transcribed, and all tokens of word-initial [t] and [d] were considered for inclusion in this study. The [t]-tokens are found in stressed and unstressed syllables of content and function words with all five Polish oral vowels [ɨ e a o u] following the initial apical stop. However, very few tokens of word-initial [d] are to be found in this conversation; all occur before the vowel [o] in unstressed syllables.

It is true of several other conversations recorded by MJM in Polish that the number of word-initial [t]'s is much greater than the number of word-initial [d]'s. The sentences described above, however, contain many more [d]'s than [t]'s. One reason for including these sentences in this study, then, is that they provide tokens of word-initial [d] in running speech to supplement the conversation tokens. A total of 237 tokens of word-initial [t] and 248 tokens of word-initial [d] were recorded.

TABLE 1

Words read in minimal pair list. The * indicates the set read by the 24 subjects in Wrocław, Poland.

to - do
tym - dym
*tama - dama
*tata - data
tom - dom
*ta - da
tomy - domy
tam - dam
*tur - dur
tyle - dyle
tarka - darka
tamy - damy

*rata - rada
roty - rody

Words read in sentences. Capital letters indicate that the word occurred in sentence-initial position, post-pausally. The * indicates words read only by speakers in Łódź, Poland.

Sentence #	
1	to
4	do
5	do
16	To
17	do
	domu
21	do
	domu
25	Dostales
29	do
31	to
*33	To
*35	do
37	dostales
38	do
	domu
40	do
	domu

2.1.1.2. Inclusion of tokens

The data presented in this chapter come only from words with initial [t] and [d] followed immediately by a vowel. No initial consonant clusters or word-internal stops are included here. However, both post-pausal stops and stops embedded in running speech are included.

Not all the 237 tokens of [t] and 248 tokens of [d] that were recorded were actually measured for this study. First, in the read conditions, a speaker sometimes left out a word or did not produce the correct consonant. In the conversation, low speaking volume, background noise, or overlapping speech by the two speakers made it necessary to exclude some tokens. In addition, in very fast speech, vowel deletions sometimes resulted in surface forms with consonant clusters, which were excluded.

Additional tokens could not be included in the analysis because of measurement difficulties. A particular problem was finding bursts in rapid running speech on the tape of sentences from Poland.[5] However, enough tokens could be measured from this tape to justify its use. Measurement problems will be discussed further below. In some cases estimates rather than measurements could be made, providing some additional data.

After exclusions for all of these reasons, 187 of the 237 [t]-tokens and 164 of the 248 [d]-tokens were actually measured for this study. An additional 43 [d]-tokens could be estimated with confidence, for a total of 207 [d]-tokens. These figures can be broken down as follows:

minimal pairs

[t] 120 of 132

[d] 124 of 132

sentences

[t] 23 of 29

[d] 38 of 109 (plus 43 estimates)

JP conversation

[t] 44 of 76

[d] 3 of 7

2.1.1.3. Analysis

a. VOT measurement

All measurements were made on the PDP-11/34 computer in the Phonetics Lab of the Brown Linguistics Department, using the WAVE program written by John Mertus. This system samples at 20 kHz using a 10-bit analog to digital converter, and the input is band-pass filtered from 60 to 10,000 Hz.

VOT was measured from the beginning of the release burst to the beginning of voicing. If the onset of phonation occurs after the release burst, the VOT measurement is positive; if phonation begins before the release burst, the VOT measurement is negative. If both begin at the same point, the VOT measurement is 0. If either the burst or voicing onset cannot be seen in the waveform, no VOT measurement can be made. Conservative criteria biased toward detecting voicing were used in identifying the onset of periodicity, in that any acoustic event representing irregular vocal fold vibration was considered "voicing".[6] For positive VOT, voice onset was measured as the zero-crossing before the first negative peak of

pulsation[7], unless some other point clearly marked the onset of phonation. For negative VOT, voice onset was measured as the low point of the first clear negative peak, rather than to the zero-crossing, because there often was no zero-crossing in the vicinity. Measurements were made to a tenth of a millisecond and rounded to the nearest millisecond.

In the case of voiced tokens, the beginning of the burst was taken from the noise frication, even when the burst appeared to be superposed on a pitch period. The landmark points used in measuring VOT are illustrated in sample waveform displays in Fig. 2-1.

There are systematic differences for [d] across speakers and speaking conditions regarding voicing before and during the release burst that complicate the measurement procedure. For some speakers, there is a tendency for prevoicing to die down completely several milliseconds before the release burst, as shown in Fig. 2-2. In these cases the VOT measurement is still made from the onset of voicing to the burst, ignoring the silent interval.[8]

In the running speech conditions, voicing for a [d] usually continues from a preceding voiced segment through the [d] release and into the following vowel without a break. Therefore there is, strictly speaking, no "onset" of voicing from which to measure VOT. This situation is discussed in the next section.

b. continuous voicing measurement

When voicing is continuous from a preceding segment through stop closure and release, no true VOT measurement is possible. In this case "VOT" was replaced by (or redefined as) a measure of the duration of the closure interval, which for [d] is voiced. This measure is not entirely arbitrary, since a <u>prevoicing</u> measure for

VOT also represents closure voicing. In the case of prevoicing, the closure duration can be thought of as unbounded -- the beginning of closure is seen only in the prevoicing, with no preceding segment to delineate the process of occlusion. In the case of running speech, the onset of closure is clearly defined in the preceding segment, and the closure can be thought of as being bounded by that preceding segment.

Closure was measured from the point where the spectral characteristics of the vowel periods were essentially lost -- where the peaks and troughs disappeared or were smoothed out, etc. -- to the stop burst. In many cases no burst could be found, and so a second set of closure duration values was obtained by estimating closure offset as the point where the amplitude began to increase and the spectral shape began to change for the following vowel. For voiced stops this point is usually quite soon after the burst, and in any event such a measure provided an upper bound on the actual closure duration. These estimates were rounded down to the nearest 5 msec. It will always be stated in the results sections when these estimates are being included.

Examples of closure duration measurements are shown in Fig.2-3.

c. measurement reliability

An assessment was made of the reliability of the VOT measurements reported here. The 12 minimal pairs of word-initial [t] and [d] read once each by MG and JP were measured independently on two occasions. The consistency of these measurement across sessions was tested in two ways. Means and standard deviations were computed for [t]'s and [d] 's for each set of measurements. Then the difference between the members of each pair of measurements (first and

second sets) was computed, and the mean and standard deviations of these differences. The object of this test is to see if the standard deviations for the differences are smaller than those for the measurements themselves--that is, if remeasurement error introduces more variation into the measurements than is already present from sampling. The second way in which consistency was assessed was by calculating the correlation coefficient of each set of paired measurements.

In each case the voicing lead measurements are less reliable than the voicing lag ones, because for each speaker there was one token where the onset of prevoicing was difficult to distinguish from background noise. These measurements varied by tens of milliseconds, and this variation accounts for most of the unreliability. Table 2 shows the mean and standard deviation of each consonant sample for each speaker. The mean and standard deviation of the differences between the samples is also shown, both for 12 values per sample, and for 11 values, with the one unreliable one in each sample left out. The large improvement in the reliability, shown by the decrease in the standard deviations, indicates that these single tokens indeed account for most of the unreliability in the samples.

Correlation coefficients calculated for each sample are also given in Table 2. One way of considering correlation is to ask whether one sample for [t] is like the other sample for [t] for one of the speakers, and whether one [d] sample is like the other [d] sample. This question is relevant when one looks at means and ranges for each consonant. The correlation coefficients for these comparisons are shown in Table 2; the comparisons of the [d]

TABLE 2

Reliability of VOT measurements made for the same tokens (12 minimal pairs spoken by MG and JP) on two different occasions, giving mean and standard deviation (SD) for each set of measurements, and for the differences (Δ) between the paired measurements, for each speaker. The last column gives the SD of the differences when the most different pair in each set is omitted.

	1st sample	2nd sample	Δ-12 values	Δ-11 values
	mean (SD)	mean (SD)	mean (SD)	(SD)
MG				
t	18.7 (7.9)	17.8 (9.7)	-.92 (4.1)	(.47)
d	-102.3 (20.3)	-96.6 (19.87)	1.92 (13.95)	(1.46)
JP				
t	35.0 (8.3)	34.9 (8.2)	.08 (.29)	-
d	-79.9 (26.8)	-86.2 (32.0)	-6.25 (24.3)	(2.15)

Correlation Coefficients

MG	2 samples [t]	.91
	2 samples [d]	.80
	2 combined samples	.99
JP	2 samples [t]	1.00
	2 samples [d]	.67
	2 combined samples	.97

samples show much lower correlations than do the [t] samples.

A second way of considering correlation is to ask whether one set of measurements of [t]'s and [d]'s presents a different picture of the distribution of the two consonants than does a second set of measurements. The values for the combination of [t]'s and [d]'s are then compared for each speaker. Table 2 shows these coefficients as well; the combined samples of measurements are quite reliable, even with those values for [d]'s that differ widely.

The values reported in this chapter are those of the first sample, since they show smaller standard deviations.

2.1.2. Results

The data for the post-pausal tokens will be presented separately from the data for the running speech tokens. In the post-pausal environment, prevoicing can easily be measured for [d], while it often cannot in running speech. Also, the speaking rate, etc., may vary in these conditions, and so at the first stage of analysis the data are kept separate. Later, the possibility of collapsing over several conditions is considered.

For each condition, the number of tokens for which there are measurements (N), the entire range the measurements cover in msec VOT, and the mean ($\bar{x}$) and standard deviation (s) for those measurements are presented. Following standard practice, the positive and negative VOT's for a single stop consonant sample are presented separately, since usually there are two distinct distributions involved.

2.1.2.1. Results for post-pausal [t] and [d]

The post pausal tokens are found both in the minimal pair

readings and in initial position in the sentences. Data for each set of readings of the same material (4 or 12 minimal pairs, sentences) are presented separately in Table 3. Sample distributions for individuals and groups, as labeled, are shown in Fig. 2-4.

It can be seen from the frequency distributions that, while the values for an individual speaker (JP, MG, DB) do not cluster around any particular value, the combined values for the 24 speakers from Wrocław give an essentially normal distribution for each phonemic category. The mode for [t] is between +15 and +20 msec, and the mode for [d] is about -120 msec VOT. These distributions are similar to those given by Lisker and Abramson (1964) for other two-category languages contrasting lead and short-lag apicals.

The breakdown by speakers and condition shows that the means and standard deviations for [t] are quite similar across these samples. The exception to this is that speaker JP produces higher VOT values for [t] 's in minimal pairs than do the other speakers. Nonetheless, his higher values, which are all under +50 msec, are within the range reported by Lisker and Abramson, and indeed one Wrocław Pole also produced a +50 msec value. Overall, there is no obstacle to collapsing the data across speakers and conditions for post-pausal [t] .

The values for [d] in minimal pairs are also quite similar, but there is less prevoicing used in the sentence context than in the minimal pairs. Since there are only four prevoiced sentence tokens (the fifth has a lag value), it is difficult to know whether this difference is meaningful, but there is some evidence that it may not be accidental. As will be seen in the running

TABLE 3

Number of tokens (N), range of VOT measurements, mean VOT measurement ($\bar{x}$) and standard deviation (s), in msec VOT, for tokens of [t] and [d] following pause.

1) 24 Wrocław Poles, 4 minimal pairs

[t] range = +3 / +36 msec

N = 90 $\bar{x}$ = 20 s = 7.7

[d] range = -177 / -35, +9 / +13 msec

N = 94 $\bar{x}$ = -120, +11 s = 32.9, -

2) 3 Providence Poles, 12 minimal pairs

[t] range = 0 / +49 msec

N = 36 $\bar{x}$ = 24 s= 11

[d] range = -169 / -34 msec

N = 36 $\bar{x}$ = -102 s = 29

3) 8 readers, sentences, 2 initial [t] and 1 initial [d]

[t] range = +13 / +30 msec

N = 13 $\bar{x}$ = +19 s = 5.5

[d] range = -32 / -27, +22 msec

N = 5 $\bar{x}$ = -30, +22 s = 2.1, -

speech condition, lead VOT values of -30 to -60 msec cover most of the range of voiced closure durations found for [d] in running speech. Thus the shorter values for prevoicing found in sentence-initial position coincide with the values for voiced closure duration found in sentence-internal position. That is, sentence production of [d] may involve a different production strategy from that used in minimal pairs, even in post-pausal position.

Finally, it should be noted that two [d]-tokens were produced without prevoicing by one of the Wrocław Poles, and one [d]-token was produced without prevoicing in a sentence.

2.1.2.2. Results for [t] and [d] in running speech

As in the preceding section, data for the sentences is combined across speakers. The data for JP's conversation is presented separately. Table 4 gives the number of tokens, the range, mean and standard deviation for each sample of the two consonants.

All but three [d]-tokens in the sample from which data for this section is derived contain voicing continuous from a preceding segment through the closure and release portions. For these three tokens, voicing was not sustained from the preceding segment through the apical stop; voicing died down a few milliseconds before the burst. Of these three, one could not be measured, one had a VOT of 0, and the other a VOT of 16.7 msec. The values given in Table 4 are closure duration measurements, and so do not include these last two values. Closure durations will be discussed in more detail in Ch. 4.

It can be seen from Table 4 that the data from the two conditions, both in the means and the standard deviations, are quite similar, and could be collapsed.

Speaker JP contributes a disproportionate share of tokens to this sample of data, and some of his [t] values are slightly higher than those of other speakers. In this sample, however, it is not always because of longer lag or aspiration, but often because he tended to produce double bursts. VOT is always measured from the <u>first</u> burst, if there are more than one, and so the VOT values measured for this type of token will be somewhat inflated. See Fig. 2-5 for an illustration of double bursts.

TABLE 4

Number of tokens (N), range of VOT measurements, mean VOT measurement ($\bar{x}$) and standard deviation (s), in msec VOT, for tokens of [t] and [d] embedded in speech <u>not</u> following pause.

1) sentences, 12/13 embedded [d], 2 embedded [t]

[t]	N = 10	range = 0 / +26 msec
	$\bar{x}$ = +15	s = 8.8
[d]	N = 33	range = -20/ -83 msec
	$\bar{x}$ = -50	s = 15.5
	with <u>estimated</u> values:	
	N = 78	range = -15/-83 msec
	$\bar{x}$ = -45	s = 16.3

2) conversation

[t]	N = 44	range = 0/+53 msec
	$\bar{x}$ = +28	s = 11.1
[d]	N = 2 measured, 1 estimated	
	range = -58/-100 msec	
	$\bar{x}$ = -73	

A few of the [d] values are noticeably longer than the others; these are all taken from the phrase "zaraz do domu", in the unstressed preposition. Although it might be expected that the closure would be quite short in this context, the closure after the fricative [z] contains some frication and lasts longer than other closures. Presumably this is due to the homorganic place of articulation, which allows the closure for the [d] to be coarticulated with the [z].

2.1.3. Observations and Discussion

2.1.3.1. Comparison across conditions

There are two factors that might affect the distribution of VOT values in the present samples. The first is post-pausal vs. non-post-pausal (embedded) position; the second is word-list (minimal pairs) vs. continuous speech context. Of the four possible combinations of these two crossed factors, only three are found in these samples, as there are no explicit minimal pairs embedded in running speech. The ± post-pausal contrast occurs only in the sentences; the ± context contrast occurs only in post-pausal position. Consider each factor in turn. A comparison of Table 3, #3 with Table 4, #1 indicates that there is no substantial difference in VOT for [t] and [d] in post-pausal and non-post-pausal positions. On the other hand, a comparison of Table 3, #1 and 2 with #3 indicates that the speech context does affect VOT, in that post-pausal tokens in sentences have a more restricted range of VOT values than do minimal pairs read in word lists. This is especially true of [d]-tokens, which have much lower VOT values in sentences, even in post-pausal position, than isolated minimal

pairs.

In the presentation of the data, it was noted that values within the post-pausal and running speech conditions could be partially collapsed over speakers. Comparison across conditions indicates that all samples are similar enough in means and standard deviations to allow collapsing, except for the ± minimal pairs factor for [d]. Here, it will be recalled, values are much larger for the minimal pairs condition. In fact, the two sets of [d]-values, for ± minimal pairs, form two distinct distributions, with modes at about -120 and -60 msec VOT.

Fig. 2-6 presents all the production data, collapsed over most conditions. The data for [t] is quite systematic, with a quasi-normal distribution. The data for [d] is separated into ± minimal pairs, and the outline of the <u>combined</u> distribution is given in dots. Since this combined distribution is, at best, bimodal, it seems best not to collapse over these conditions without further data for [d] to clarify the nature of the distribution.

In English, a fair amount of category overlap is found for voiced and voiceless stops in running speech, especially in casual conversation (Lisker and Abramson, 1967; Moslin, 1978). For example, Fig. 2-7 shows a distribution from a casual English conversation, taken from Moslin (1978). It is therefore remarkable how little category overlap there is in Polish. Even in running speech, where [d] is represented by the less distinctive voiced closure duration measure, the voicing categories are quite separate. The Polish distinction "voicing before release"/"voicing after release" seems to be an easy one to produce consistently.

2.1.3.2. Individual differences

There is considerable variation in the production of voicing contrasts across the 32 speakers represented in the corpus used here. In this section these differences are illustrated in order to show that a single "VOT" value can be realized acoustically in a number of different ways. Also, very little explicit information on acoustic details of the prevoicing-short lag VOT contrast is available, and another purpose of this section is to provide such information.

a. variation for [t]

The three basic types of [t]'s to be found in the corpus are illustrated in Fig. 2-8. With the first type of [t], voicing begins immediately or shortly after the burst, giving a low VOT value. However, in this case the burst is quite prominent, with frication and possibly some aspiration, and a relatively high amplitude. Of course, care must be taken in interpreting relative intensity in waveform displays, since the overall recording level may vary across tokens. A more detailed analysis based on a carefully controlled recorded sample would be required to be certain of these differences. Still, comparing burst portions to the surrounding portions in a single token, it seems that some bursts and lag intervals are more prominent than others, even with similar VOT values. This sort of variation does not accord with Lisker and Abramson's model of voice timing, and indicates that some other production parameter may be involved (e.g., subglottal pressure, muscle tension). It also indicates that VOT may not be the only cue to voicelessness. Other aspects of the acoustic signal may vary in a way that <u>compensates</u> for low VOT values, rather

than co-vary with them.

With the second type of [t], voicelessness is seen in the amount of aspiration present after the burst, with a higher VOT value, from +20 to +50 msec. The third type of [t], however, has a fairly long lag VOT but little or no aspiration during the voicing lag. Especially common is some combination of aspirated and silent lag. The point to be made is that speakers vary in how much aspiration fills the lag interval, judging from inspection of waveform displays. The degree of aspiration does not seem to be directly correlated with the VOT value. The perceptual aspect of this question will be addressed in the next chapter.

b. variation for [d]

In section 2.1.1.3a some patterns of prevoicing were described. Fig. 2-9 illustrates further the variation found in Polish [d]. Of the three types of productions, by far the least common is the one where the prevoicing stops before the burst, and several msec of silence intervenes between them. (See Fig. 2-2) This pattern is used by only a very few speakers, in the minimal pairs condition. In Ch. 4 we will see that this pattern also occurs with <u>medial</u> [d] in minimal pairs. It does not seem to occur in running speech. When prevoicing is measured, the silent interval is included in the VOT value, and these VOT values are typically longer than those for tokens where the voicing does not stop before the release. That is, the amount of actual voicing is similar, but the VOT is greater because of the <u>added</u> silence.

The most prevalent pattern in minimal pairs is what Lee Williams (1977) called "interrupted" voicing: the prevoicing continues up until the release, but with lower energy immediately before it;

the burst is voiceless, and voicing recommences after the burst. This pattern is quite rare in the running speech condition. (See Fig. 2-9a.)

The usual pattern in running speech, and which also occurs consistently for some speakers in the minimal pairs, is continuous voicing up to and through the release. There are two sub-types here, which Williams does not distinguish. The first is illustrated in Fig. 2-9b: the burst appears to be superposed on a pitch pulse. This type is rare for Polish, although it may be the dominant pattern in other prevoicing languages.[9]

Most instances of continuous voicing in Polish involve a burst which is clearly distinguishable from prevoicing, but which contains several short pitch pulses within it. Fig. 2-9c illustrates such a pattern. The components of the burst are evenly spaced and the burst frication appears superposed on them; compare the normal noisy-looking burst shown in Fig. 2-1b or the fricated [t]-burst in Fig. 2-8a.

The effect of these various patterns is that no simple statement can be made about a [d]-burst being characteristically voiced. In minimal pairs it is often voiceless, while in running speech it is typically voiced. In both conditions there are tens of milliseconds of voicing during the closure interval before the burst. However, in minimal pairs this voicing is much longer than it is in running speech. It is in the minimal pairs, where the lengthy prevoicing is an unmistakeable correlate of voicedness, that the release itself is ambiguous, since a <u>voiceless</u> burst can occur in either a [t] or a [d] in Polish. Where there is much less closure voicing, in running speech conditions, the release is unambiguously

voiced.

2.1.4. Summary and Conclusions

Data has been presented on voice-onset measures for Polish word-initial [t] and [d] . In post-pausal position, where both lead and lag VOT can be measured, VOT is extremely effective in distinguishing the two stop categories. Values for [t] fall in the short lag region, and values for [d] fall in the lead (especially long lead) region, of the VOT continuum. In running speech, VOT values for [t] also fall in the short lag region, but no point of voicing onset can be identified for [d], as voicing is continuous from a preceding segment. Therefore, for [d] in running speech, a measure of the (voiced) closure duration was substituted for a true VOT measure. This measure was always negative, and so "VOT" was still effective in distinguishing the stop categories. The overall result of the study is that for [d], there is always <u>some</u> voicing during closure, while for [t], there is never <u>any</u> voicing during closure. Also, for [d] the burst may be voiced or voiceless, but for [t] it is always voiceless.

Many speakers have voiceless bursts for [d] in minimal pairs. Were this not so, the voicing distinction could be characterized by burst voicing alone -- possibly in terms of spectral properties of the release. And because some [d] 's have voiceless bursts, [d] cannot be defined by continuous voicing, vs. interrupted voicing for [t]. However, in those situations where the burst itself is not voiced, there is more voicing (longer VOT) during the closure, possibly providing more information about the voicing category of the consonant.

The voicing contrast in Polish is thus not the simplest one

that could be imagined for a prevoiced - voiceless unaspirated language. The simplest situation would be one in which voicing was never interrupted for [d], and so [d]-bursts were always voiced, with continuous voicing from closure through the release, and in which [t]-bursts were always voiceless with lag VOT (as they are in Polish). In this simple situation, the voicing categories could be distinguished by referring to any <u>one</u> of: the closure interval, the burst, or the segment immediately after the burst.

The Polish situation, as we have seen, is more complicated. The burst in itself is ambiguous. In the portion immediately before the burst, voicing information is also ambiguous, since prevoicing can die down before the release. Immediately after the burst, frication and aspiration may obscure a gradual voice-onset, making the VOT unclear. Portions well before and slightly after the burst should, however, be unambiguous with respect to voicing: [t]'s show some voicing lag after the burst, and [d]'s show some voicing before the burst.

In terms of <u>contrasting</u> Polish [t] and [d], VOT provides more information than is actually needed. A numerical VOT value is not needed to clearly separate the [t]-category from the [d]-category; the "+" or "-" sign attached to the number is sufficient. The effectiveness of these signs, however, crucially depends on the measurement procedures set up earlier. For example, it must be understood that an interruption in voicing, followed by a <u>second</u> voicing onset <u>after</u> the burst, will be ignored. If we were first to look for a voicing onset <u>after</u> the burst, then tokens of [d]with interrupted voicing would be mis-categorized as [t].

Polish [t] and [d] can be distinguished by noting whether or not there is any voicing before the burst. Because voicing sometimes dies down immediately before the burst, one must look for voicing well before the burst. The best place to do so is probably immediately after the point of occlusion, at the onset of closure. If there is voicing at and after the onset of closure, then the stop is voiced; if voicing ends at closure, then the stop is voiceless. In this scheme, the release burst plays no particular role as a reference point relative to the onset of voicing; rather, the occlusion -- the onset of closure -- plays that role.

Why then use VOT at all? First, the burst seems to be important for categorization along several phonetic dimensions, for example stop/fricative, and probably place of articulation (Stevens and Blumstein, 1978). Thus there is no advantage in not involving it in this voicing categorization. It is not clear what independently-motivated role is required for the onset of closure.

Second, numerical VOT values are useful in describing the compensatory relationship that may exist between closure voicing duration and burst voicing. As described earlier, there seem to be more voiceless [d]-bursts in minimal pair readings than in running speech. It is also in the minimal pairs that long intervals of closure voicing -- prevoicing -- occur, and it was speculated that the longer prevoicing may help compensate for the voiceless bursts. If this is so, then only numerical values can represent the fact that the prevoicing is longer than the closure voicing in running speech. However, this account is purely speculative, and so cannot be used for or against any particular measure at this point.

Third, VOT is useful in providing non-contrastive acoustic detail about Polish stop voicing. The VOT distributions shown here are systematic, and so a VOT measure is not arbitrary. More importantly, VOT is useful in non-contrastive cross-language comparisons, for example between Polish and English. We will return to this point in the final chapter.

2.2. Perception Data

2.2.1. Introduction

In the preceding section on production data a fairly straightforward account of the voicing contrast in Polish apical stops was given. For [d], there is vocal fold vibration before and often through the release burst. For [t], vocal fold vibration begins after the release burst. These generalizations hold for minimal pairs, read sentences, and spontaneous conversation. VOT is a useful measure for representing this contrast and especially for describing acoustic details.

However, once an acoustic correlate of a contrast is found, it is necessary to investigate the value of that correlate as a <u>cue to</u> the perception of that contrast. It could be that VOT happens to co-vary with some perceptually-relevant parameter in natural speech, such that the correlation obtained for the production data is misleading. It might be that, as has happened with VOT (Lisker and Abramson, 1970) the production and perception categories are not quite matched indicating that other parameters play a role in perception. Or, on the other hand, it could instead be the case that VOT is the most reliable metric for describing the perception of voicing.