Journal of Phonetics (1989) **17**, 257–282

Incomplete neutralization and pragmatics in German

Robert Port and Penny Crawford

Department of Linguistics, Indiana University, Bloomington, IN, 47405, U.S.A.

Received 26th January 1988, and in revised form 6th April 1989

Earlier studies suggest that the neutralization of the voicing contrast in word pairs like *Bund* and *Bunt* is incomplete, but further research is required to evaluate the competing interpretations. Subjects here were given a range of different speaking tasks: reading continuous text, repeating spoken sentences (in circumstances where the concern of the experiment was carefully hidden), dictating the potential homophones to a German-speaking writer and, finally, reading from a word list. Discriminant analysis was used to combine the five spectro-temporal variables measured from sound spectrograms of these productions to categorize the tokens as voiced or voiceless in each condition. Correct categorization for discriminant analysis varied between 55% and 78% depending on the communicative task but was significant in all conditions. The data show that speakers can control the degree of neutralization depending on pragmatics and that information about the underlying contrast is distributed over much of the word. In Experiment 2, recorded productions from some conditions of Experiment 1 were played for a group of listeners. Through use of signal detection theory (and the statistic *d*′), it is shown that listeners discriminated the intended word with accuracy very similar to that of discriminant analysis. They even tended to make errors on the same tokens. Apparently, the variables we measured capture information that is roughly equivalent to that employed by native listeners. The absence of complete neutralization implies that the German syllable-final devoicing rule cannot be stated in terms of the same [− voice] feature that is employed in the lexical specification of words. Instead, syllable-final devoicing is an effect that resembles implementation rules (since it is graded) and operates directly upon a syllable-like representation.

1. Introduction

Evidence has been mounting that many textbook examples of phonological rules do not work as described in the linguistics literature. In many situations where audition-based phonetics has supported a phonological rule of neutralization, careful acoustic measurements reveal that the neutralization is incomplete. These examples pose a threat to phonological theory by forcing it to deal with a very large number of low-level articulatory or acoustic features, rather than with a small number of abstract phonological features. They imply that there is a difference

0095–4470/89/040257 + 26 $03.00/0

between the distinctive features employed in the lexicon and the features employed in stating rules of neutralization despite the obvious similarity between their effects.

This study pr·.,ents further evidence concerning the incomplete neutralization effect in German. The data support a scalar-valued neutralization effect in the German devoicing rule, and clearly refute a rule using a binary voicing feature. Speakers' communicative intentions and goals, that is, pragmatics, affect details of the execution of a classical phonological rule and do so differently for this "rule" than for lexical [− voice]. The familiar German devoicing rule makes voiced stops and fricatives occurring at the end of syllables become voiceless. Thus, many morphemes alternate their stem-final consonant from, say, the [d] in *Bunde* to the [t] in *Bund* when there is no vowel following the stem. Since *Bunde* contrasts with *bunte* while *Bund* and *bunt* are virtually indistinguishable, an underlying contrast between /d/ and /t/ is supported along with a rule of neutralization. The issue raised by the incomplete neutralization effect is the nature of the neutralization process.

2. The incomplete neutralization effect

Several studies have found that the German devoicing rule does not yield complete neutralization (Dinnsen & Garcia-Zamor, 1971; Port, Mitleb & O'Dell, 1981; Port & O'Dell, 1985; Charles-Luce, 1985). Pairs like *Bund* "association, group" and *bunt* "colorful" have long been said to be homophonous due to a rule changing [+ voice] stops and fricatives to [− voice] when they are final in a syllable, yet the words tend to retain a small difference in certain phonetic parameters. For example, in various experiments, differences havẽ been observed in the duration of the preceding vowel, the duration of the consonant closure, the intensity of the final release burst, the amount of glottal pulsing evident during the final consonant constriction, etc. The differences are not large enough to justify the term "contrast", and are often non-significant statistically when examined one parameter at a time. In addition, there are sometimes large interspeaker differences. Still, in certain situations, the differences are large enough that native listeners can guess better than chance which word a speaker is saying (Port & O'Dell, 1985).

These observations have been criticized by Fourakis & Iverson (1984) as artifactual. They describe two experiments which are claimed to discredit the incomplete neutralization effect for German. In the first experiment, German subjects were coached to produce spontaneous principle parts of verbs, like *meiden, mied, gemieden,* "avoid", and *raten, riet, geraten,* "advise", which include the near-minimal pair *mied* and *riet.* Measurements of vowel and stop closure failed to demonstrate any differences (by *t*-test) between the words due to the underlying voicing. Unfortunately, as pointed out by Dinnsen & Charles-Luce (1984), none of their word sets was minimal pair, thus the effects of the initial consonant or consonant cluster were not controlled for. The absence of an effect on the vowel duration is very likely due to this problem. In their second experiment, they partially replicated the experiment of Port & O'Dell (1985) by having subjects produce word lists containing some minimal pairs. The *t*-tests on individual subjects showed weak effects in the expected direction for both vowel and consonant closure duration, but with $n = 4$, it is not surprising that the significance level is weak. On the other hand, even a simple non-parametric test across mean values for speakers and words (using their Table IV) shows that both vowel duration and consonant duration are significantly different for the voicing categories in the direction expected.[1] Thus, where their data are usable, Fourakis & Iverson replicate the results of Port & O'Dell and show about the same amount of difference between the two classes. Nevertheless, of course, important questions remain about the effect. For example, the possibility of hypercorrect "reading pronunciations" in all these experiments needs to be clearly ruled out.

Studies have also obtained incomplete neutralization in cases of other phonologically supported underlying contrasts in several other languages (Catalan: Dinnsen, 1984; Dinnsen & Charles-Luce, 1984; Charles-Luce, 1987; Polish: Slowiaczek & Dinnsen, 1985; Russian: Pye, 1986; English *ns/nts* clusters: Fourakis & Port, 1986). In fact, another well-known neutralization rule in English—the rule that flaps intervocalic /d/ and /t/ in *bedding–betting* and *rider–writer*— has long been known to be less than perfect. If productions are measured carefully enough (Fox & Terbeek, 1977) or looked at closely in a range of contexts (Hubbel, 1950; Huff, 1980), this rule turns out to produce (a) neutralization in some contrasts (e.g., in New York City *betting–bedding*, Port, 1981), (b) a near-contrast in others (e.g., Northern U.S. *butting–budding*, Fox & Terbeek, 1977) and (c) a very audible allophonic "contrast" in still others (e.g., New York City *writer–rider*).

There are some generalizations that can be drawn across these instances of neutralization. First, the majority of the underlying contrasts are supported by phonemic alternations in the pronunciation of particular morphemes. One exception, however, is the German adverb *weg* "away" which does not alternate, yet is observed to behave like a word with underlying /g/ in contrast to *Weck* (Port & O'Dell, 1985). Although Fourakis & Iverson (1984) suggest this exposes the artifactual character of the incomplete neutralization effect by showing its dependence on orthography, it may simply show that German speakers relate this adverb to the same lexical entry as the noun *Weg* "way, road", despite the fact that the noun contains a long vowel rather than the short one in the adverb. The problem is an interesting one, however. It is raised any time there is a neutralization rule that can apply morpheme-internally—as for example, in words such as English *water*. Is this word learned with underlying /d/, /t/ or some third alternative like an underlying flap? Despite the skepticism of Fourakis & Iverson, it does not seem impossible that orthography could play a role determining speakers' underlying forms in such situations. It is not necessarily an artifact of reading. In this case, of course, the differences should continue to be observed even when subjects are *not* reading the words. In the experiments below we investigate this problem further.

A second generalization about cases of incomplete neutralization is that the context of application of the incompletely neutralizing rules can usually be stated very naturally in terms of position in the syllable. That is, a syllabic representation

[1] As pointed out by Fourakis & Iverson (1984), the data for several words should be discarded for one subject who used nonstandard long vowels instead of short ones. Thus, these items were not minimal pairs and were left out of the sign test. Both the vowel duration ($n = 18$, $x = 2$ where x is the number of pairs going in the opposite direction) and consonant durations ($n = 19$, $x = 4$) are significant at $p < 0.01$. As a matter of fact, Port & O'Dell (1985) did not find a significant effect of voicing on the closure duration whereas Fourakis & Iverson did. In this sense their data are stronger evidence for incomplete neutralization than our own. There are undoubtedly some contexts, especially in fast speech, where complete neutralization will be observed. In fact, even full contrasts are sometimes neutralized in fast speech (e.g. English *prayed–parade*).

of speech seems to provide an appropriate framework for stating all the neutralization rules mentioned above. This suggests the possibility that speakers find it easier to store words in a segmental underlying form and then to implement the neutralizing process using dynamic syllable-based implementation rules, rather than, as supposed by the traditional linguistic interpretation, actually to change segmental features from one value to another. It seems the neutralization process can most naturally be described in a way similar to phonetic implementation rules, like those that govern the temporal effects of voicing contrasts (e.g., Port, 1981; Keating, 1985).

A third generalization is that there are often fairly prominent interspeaker differences exhibited in the data of incomplete neutralization (especially Slowiaczek & Dinnsen, 1985, Charles-Luce, 1987). Although it has been suggested that this too is evidence of the artificiality of the effect (Fourakis & Iverson, 1984), speaker differences might also reflect the undeniable fact that there is normally no communicative role for the incompleteness of the neutralization. Speakers should be expected to differ in phonetic detail that is perceptually marginal.

Although there are many important questions to address if the incomplete neutralization effect is valid, the first matter to be addressed is whether or not the whole effect is some sort of artifact. The experiments below attempt to address this question.

3. Rationale for experiments

There are two interpretations of the incomplete neutralization effect. One interpretation is that the tasks employed in these experiments are flawed and that the phonological neutralization as traditionally described is correct. According to this view, speakers in these experiments actually change [+voice] to [−voice] in words like *Bund* resulting in complete neutralization at the phonological level. Then, in response to the abnormal task of reading word lists, the speakers generate abnormal phonetic productions influenced by the orthographic spelling of the lists. Thus, the data reflect a secondary process that might be called "deneutralization". The speakers are either directly influenced by the written spellings or, perhaps, are trying to help out a non-native-speaking experimenter. Thus, they distinguish these true homonyms from each other. From this point of view, the incomplete neutralization effect is "pathological" and of marginal interest to linguistics or phonological theory.

Another intrepretation (Port & O'Dell, 1985) assumes that the effect is not pathological but quite natural. It postulates that there are two distinct kinds of devoicing processes in such languages as German. First, there is the kind of devoicing associated with the [−voice] feature employed in the lexicon (and presumably in lexical phonological rules). This property is observed in words like *bunt*. Secondly, there is a devoicing process associated with the codas of syllables. It might be associated with all obstruent-final syllables or perhaps only with syllables ending in voiced obstruents. The outputs of the two processes resemble each other yet differ. For this reason, *Bund* and *bunt* are phonetically similar yet distinct. If this hypothesis is correct it would have very important implications for the theory of speech production and for linguistic theories of phonology.

What is required to distinguish between these hypotheses? First, it must be determined whether speakers still show incomplete neutralization when they have no way of knowing what the purpose and interest of the experiment is, and where the speech task is more natural than reading lists of words. Obviously, it is not necessary to show that neutralization can *never* be observed, since there will surely be some speaking styles where full neutralization occurs or where any differences are too small to be detected. Secondly, if speakers can actually modify the degree of contrast (as argued by Fourakis & Iverson, 1984), then one should test how well speakers can modify the contrast when they are actually asked to. That is, in addition to a speech task in which speakers will feel no pressure to create an artificial distinction between the apparent homonyms, there should also be tasks in which they are directly asked to make a distinction. In this way, there will be some basis for determining what is artificial and what is not. Finally, it is important to determine whether the effect depends on reading the test items, since the possibility exists that the orthography, just by being looked at, might influence their pronunciations in some way. In the experiments below all these criteria are satisfied.

4. The experiments

Two experiments are reported. In the first, a set of conditions differing pragmatically from one another was employed to study the German voicing contrast. A German assistant asked German speakers to say pairs like *Bund–bunt* in different contexts and in different tasks. The first task disguised the purpose of the experiment by hiding the target words within a long list of sentences. In a later task, subjects attempt to pronounce them distinctly for the German experimenter trying to transcribe the words. Thus, if the subjects are cooperative, we should get some idea of the maximum contrast speakers are capable of producing. In other conditions, attempts were made to eliminate orthographic effects.

Discriminant analysis is used extensively in these experiments to provide a sensitive yet objective means of measuring the degree of contrast present by combining a number of variables. Discriminant analysis is a procedure that obtains the best linear combination of several input variables for distinguishing between groups in the data (Klecka, 1980; Nie, Hull, Jenkins, Steinbrenner & Bent *et al.*, 1975; see Port, Reilly & Maki, 1988 for other applications in phonetics). We use this here as a measure of the degree of contrast. Techniques like analysis of variance and *t*-tests are restricted to evaluation of one variable at a time yet require combining data from multiple trials. Discriminant analysis combines several variables and makes a decision about each trial. This much more closely resembles the task of perception.

A second experiment was done to test the validity of using percent correct categorization by discriminant analysis as an estimate of the contrastiveness of the word pairs for real speakers. Experiment 2 provided a direct comparison of perception performed by native listeners with categorization by discriminant analysis tested on the same set of productions. Such a comparison provides an estimate of the validity of discriminant analysis as a real-valued measure of perceptual discriminability.

5. Experiment 1: production of German syllable-final voicing

5.1. *Methods*

Three pairs of test words which end in final underlying /d/ or /t/ were selected for detailed examination. All are real words. Phonological evidence can be given to support the particular underlying segment for the first two pairs (which alternate with case variants that have a following vowel). The third pair do not show alternation and thus have much weaker phonological evidence for the underlying nature of /d/ *vs.* /t/. Any differences for this pair would have to be due to orthography.

bunt	colorful	*Bund*	group
Rat	advice	*Rad*	bicycle
seit	since	*seid*	be, 2^d *plur*

These words were embedded in natural contexts in three different experimental tasks. The tasks were performed, one subject at a time, in their numerical order. Condition 1A and 1B, however, were randomly ordered across the subjects, that is, half did 1A first and half did 1B first.

5.2. *Condition* 1: *disguised sentences, read and repeated*

The six test words were embedded in 6 test sentences randomly inserted in a list containing 29 other German sentences. For example, two pairs of sentences that incorporate the test words are:

(1) "*Seid sicher*", *sagte der Lehrer*, "*daß Ihr eure Aufgaben lernt*".
"Be sure", said the teacher, "that you learn your lessons".

(2) *Seit sieben Jahren kann ich nicht mehr sehen.*
For seven years I have not been able to see.

(3) *Du sollst dir seinen Rat holen, denn er hat dasselbe Problem.*
You should get his advice since he has the same problem.

(4) *Ich wollte mein Rad haben, aber es war versteckt.*
I wanted to have my bike, but it was hidden.

It can be seen that the test words are embedded in different positions, but there is a close syntactic and phonetic similarity between the contexts of the minimal pair words. Thus, *Rat* and *Rad*, are positioned similarly in their sentences in order to minimize the effects of syntax on phonetic detail. Three examples of the 29 distractor sentences are:

(5) *Zur Abwechslung gehen wir an den Strand, anstatt in die Berge.*
For a change we are going to the beach, instead of the mountains.

(6) *Nachdem er angekommen war, besuchte er seine Freunde.*
After he had arrived, he visited his friends.

(7) *Dieser kleine Vogel ist durch das offene Fenster geflogen.*
This small bird flew through the open window.

It can be seen that the sentences exhibit a wide variety of syntactic patterns and sentence lengths. It is unlikely any of the speakers were able to guess what our particular interests were during Condition 1A or 1B.

This list was presented to the subjects in two forms. In Condition 1A, the subjects were asked just to read the printed list of all sentences twice through. In condition 1B, a native German speaking assistant, who conducted all conditions of the experiment, read each sentence aloud for the subjects to recite from memory. If actually seeing the spelling of the test word has an influence on a subject's production, then there should be a difference between these two conditions. The orders of 1A and 1B were balanced across subjects.

5.3. *Condition* 2: *contrastive sentences*

Here the subject was invited to directly contrast the test words with each other in sentences with "redundant" phrases. That is, we sought to encourage an attempt to distinguish the words by means of phonetic detail. In this condition, the subjects were asked by the German experimenter simply to read the sentences. Sample sentences are:

(8) *Ich habe "Rat", wie Ratschlag, gesagt; nicht "Rad", wie Fahrrad.*
I said *Rat,* as in "bit of advice"; not *Rad,* as in "bicycle").

(9) *Ich habe "Rad", wie Fahrrad, gesagt; nicht "Rat", wie "Ratschlag".*
I said *Rad,* as in "bicycle"; not *Rat,* as in "bit of advice".

Each test word occurred in each sentence position, and the order of presentation within each pair was balanced across positions and subjects with multiple printed lists. Thus, with one repetition of the list, there were 24 sentences produced per subject.

5.4. *Condition* 3: *dictation sentences*

In this condition, the subjects were asked to dictate shortened versions of the above sentences (from a page not containing the "redundant" descriptive portions) to the German-speaking assistant. After each production, the assistant wrote down on a sheet of paper which word he thought was spoken. For example, the sentences were as follows;

(10) *Ich habe "Rat" gesagt; nicht "Rad".*
I said *Rat*; not *Rad.*

(11) *Ich habe "Rad" gesagt; nicht "Rat".*
I said *Rad;* not *Rat.*

After each sentence the assistant wrote down his guess as to the order of the words. It was hoped that this procedure would encourage the greatest possible differentiation of the words in each minimal pair.

5.5. *Condition* 4: *word list*

The final condition resembled the isolated word list condition employed in several earlier experiments (Port & O'Dell, 1985; Fourakis & Iverson, 1984). Four occurrences of each of the six test words were placed in a list along with two occurrences of 12 other monosyllabic German words for reading in isolation. Several randomizations of the list were employed to avoid any effect of order in the list.

5.6. *Subjects*

The subjects were five exchange students, age 16–19, from a single town in Hessen, West Germany, located 70 miles north of Frankfurt. These students were near the end of a three week visit to our town. Their spoken English was not good, although all had studied English for 3–5 years in their secondary school. None had ever lived in an English-speaking country for more than a month. Our German assistant was a student from the same town who had been trained in the procedures we employed.

5.7. *Procedure*

All subjects were recorded individually on a single day. Before leaving the laboratory, each subject was told not to discuss the experiment with friends until conclusion of the runs. We believe they cooperated in this. The entire recording session for each subject was taped and the appropriate utterances selectively dubbed onto another tape for making sound spectrograms. All measurements were taken by hand from spectrograms produced by a Voice Identification Series 700 spectrograph. Five measurements of each of the six words were taken in each condition: (1) the duration of the vowel from the apparent release of the initial consonant (/s/, /b/ and /r/) to the closure for the final apical stop, (2) the duration of the final stop closure and (3) the duration of the portion of the burst that was visible on the spectrogram as clearly above the background noise level. Since the longer bursts were always more intense, this can be thought of as a measure of burst intensity. Since *Bund* and *bunt* contain nasals, (4) the nasal closure interval was also measured. Finally in all tokens, (5) glottal pulsing visible on the spectrogram that continued into the stop closure was measured as the number of glottal pulses (not in ms). It is important to note that the measurements were done by hand and that the criteria for most measurements implicitly employed both spectral and temporal aspects. It is difficult to make automatic measurements like those we used. Very likely, however, many other ways of measuring these productions, both automatic and manual, would preserve information equivalent to that obtained here.

The results were analyzed statistically using SPSS. Both analysis of variance and discriminant analysis were employed. As mentioned above, discriminant analysis is a procedure which obtains the best linear combination of the input variables for distinguishing between two or more groups in the data (Klecka, 1980; Nie *et al.*, 1975). A discriminant function is produced by linearly combining the dependent variables and optimizing discrimination power. Each function takes the form:

$$D_k = \sum_{i=1}^{n} w_i c_i$$

where D_k is the value of the kth discriminant function for n dependent variables, c_i, and where w_i are the weights that provide optimal discrimination of the groups in the kth dimension. The coefficients are found by a process of iteration which seeks to predict identity across the entire set of data. This is achieved by maximizing the distance between group centroids along each dimension. In these experiments, $k = 1$ since there cannot be more than $g - 1$ discriminant equations, where g is the number of groups being distinguished (here the two voicing categories). Since the variances were not constant across groups, a criterion of greatest likelihood of group membership was used to classify each token. We use the percent of correct categorizations from this analysis to measure the degree of voiced/voiceless contrast. The scale, therefore, extends from 50% (no significant contrast) to 100% (full contrast). Data were analyzed in several different ways to answer different questions. Since we tested on the same data we trained on, this procedure is, of course, not as difficult as a speech recognition task, where training and testing would generally be on different productions by different speakers.

5.8. *Results*

5.8.1. *Data pooled across speakers.* The basic durational means and standard deviations are shown in Table I.

Analysis of variance. Analysis of variance showed that when the data were pooled across speakers, condition and word pairs, the duration of the burst releasing the final obstruent was most different between the two groups ($F(1, 7) = 49$, $p < 0.001$). No other dependent variable even approached significance in this

TABLE I. Results pooled across speakers with standard deviations for each dependent variable in each condition in Experiment 1

		Vowel dur.		Stop dur.		Burst dur.		Nasal dur.		Clos pulses	
Cond.	Word	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1A	*seid*	97.64	12.42	69.29	14.40	0	0	0	0	2.5	1.18
	seit	81.89	24.89	59.84	11.26	0	0	0	0	1.2	0.42
	Rad	138.19	32.06	54.72	15.35	20.08	9.54	0	0	2.4	1.26
	Rat	145.67	22.27	59.84	10.30	25.59	5.94	0	0	2.3	1.16
	Bund	72.44	12.20	18.50	12.46	7.07	14.35	60.23	14.50	1.9	1.29
	bunt	73.23	13.54	12.20	8.79	2.76	3.24	70.47	18.60	2.6	1.17
1B	*seid*	76.77	10.54	59.06	6.69	0	0	0	0	1.8	1.14
	seit	69.29	11.47	60.24	13.39	0.39	1.25	0	0	1.4	0.70
	Rad	135.03	19.47	50	12.03	16.54	7.61	0	0	2.8	1.14
	Rat	140.94	17.09	54.72	8.60	28.35	12.42	0	0	2.8	1.03
	Bund	71.25	8.60	20.46	10.95	4.73	6.38	59.84	11.41	3	1.15
	bunt	69.29	4.23	32.28	16.47	6.3	7.92	66.54	14.19	2.6	1.08
2	*seid*	186.22	31.73	70.28	32.04	44.69	41.40	0	0	1.8	0.77
	seit	180.51	21.19	68.11	15.33	64.29	31.01	0	0	2.05	0.76
	Rad	178.14	25.70	68.89	17.54	32.87	23.46	0	0	2.8	1.01
	Rat	174.80	41.67	67.52	20.17	39.37	22.20	0	0	2.8	0.77
	Bund	97.63	14.08	32.28	18.14	36.22	17.40	98.43	15.49	1.6	0.75
	bunt	98.03	16	31.30	15.03	48.43	23	89.96	22.86	1.8	0.89
3	*seid*	221.85	28.14	87.40	22.48	33.46	28.92	0	0	1.65	1.04
	seit	203.74	20.35	87.01	13.45	81.89	32.33	0	0	1.6	0.88
	Rad	203.34	26.40	86.22	20.71	25.20	19.89	0	0	2.15	1.04
	Rat	203.57	30.32	83.46	22.18	85.63	36.75	0	0	2.3	0.98
	Bund	112.80	15.66	40.74	21.12	45.28	39.61	139.17	31.61	1.3	0.57
	bunt	112.01	18	49.01	21.77	82.68	40.67	117.52	26.22	1.4	0.60
4	*seid*	237.40	29.46	92.32	20.58	36.61	28.71	0	0	1.7	0.66
	seit	227.56	26.72	85.24	22.24	72.83	37.19	0	0	1.9	0.97
	Rad	235.83	23.55	88.98	33.37	59.06	40.51	0	0	2.2	0.62
	Rat	243.06	20.83	89.93	22.82	66.51	34.07	0	0	2.58	0.84
	Bund	131.50	28.69	54.33	20.34	48.62	34.54	129.96	40.37	1.35	0.59
	bunt	127.36	18.65	57.09	26.41	89.17	39.08	130.91	21.90	1.7	0.80

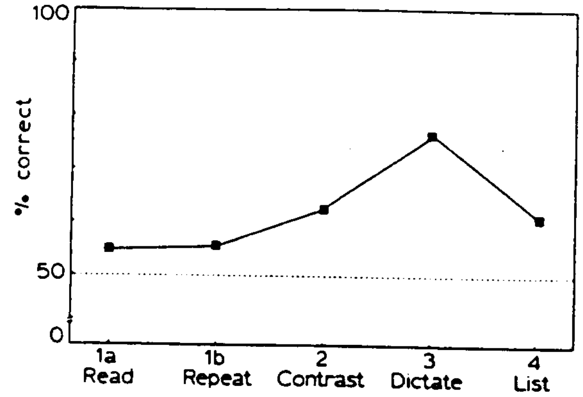

Figure 1. The percent correct classification of the data by discriminant analysis for each condition when trained on data pooled across speakers and conditions in Experiment 1. Condition 1A was Read Sentences, Condition 1B was the same sentences Repeated Orally, Condition 2 were the Contrastive Sentences, Condition 3 the Dictated Sentences and Condition 4 the Word List. The orders of Conditions 1A and 1B were balanced across subjects, otherwise the conditions were conducted in numerical order.

analysis. In particular, no differences were observed in any of the variables between the read *vs.* repeated conditions (1A *vs.* 1B). Additional analysis of variance on the combined data from conditions 1A and 1B showed no significant difference due to condition for any of the variables. This statistic is not ideal, however, since analysis of variance examines the distinctiveness of a single dependent variable at a time. Since listeners are highly trained in speech perception in their native language, they can probably combine many individually weak cues.

Discriminant analysis. Humans are presumably able to combine many variables in perceiving speech. Discriminant analysis should extract more linguistic information from the signal because it weighs a number of variables and combines them linearly in an optimal fashion to distinguish between two or more groups in the data. The assumption of linearity implies that the effect of a change in any variable will be the same across its entire range of values. By studying the percent correct categorization of underlying voicing produced by discriminant analysis, we should get a measure of how much contrast the speaker groups as a whole maintained in their productions of underlying voiced and voiceless stops.

Discriminant analysis was trained first on the data pooled across speakers, condition and word pair. Then, it was tested separately on each condition. This procedure makes the assumption that whatever might differentiate the underlying voicing contrast should be the same across speakers and word pairs as well as across the various conditions. Thus, it is in agreement with normal linguistic assumptions about the invariance of words (Port, 1986*a*). In this way we should be able to measure the degree of voiced/voiceless contrast subjects made in each pragmatic context. From the percent correct categorizations, it can be seen in Fig. 1 that discriminant analysis succeeded in distinguishing the voiced from the voiceless tokens 64% of the time across all five conditions. In Conditions 1A and 1B, where tokens were embedded in disguised sentences, the underlying voicing was still identified better than chance, 55% and 56%. There appears to be no difference between 1A and 1B, indicating that reading *vs.* verbal repetition of each sentence had no effect on the discriminability of the underlying voicing.

When subjects read contrastive sentences containing paraphrases of the test words (Condition 2), discriminant analysis was able to discriminate underlying voiced from underlying voiceless stops 63% of the time. As expected, when subjects dictated the contrastive sentences for listener identification (Condition 3), the highest degree of contrast, 78%, was found. In the final condition, where lists of words were read, Condition 4, a level of contrast at 62% ·vas obtained.

The contribution of particular variables to the performance of discriminant analysis in classifying the data successfully can be obtained by examining the structure coefficients (normalized to avoid effects of scale). Discriminant analysis employs only variables that make a significant improvement in correct identification of the groups. Values of the structure coefficients closest to zero make the smallest contribution. It can be seen from Table II that all variables except glottal pulsing during closure contributed to the discrimination process. Values with opposite sign contribute inversely to a particular categorization. It can be seen that useful information was not restricted to the stop itself, but was distributed in time across the whole word.

In order to check the possibility that discriminant analysis might be able falsely to report a contrast from a data set that is actually random, discriminant analysis was also applied to attempt to distinguish the first repetition from the second repetition of the test words. If discriminant analysis can find a significant difference here, then our results would be called into serious doubt. In fact, in testing the full data set across all conditions, where the groups were defined as Repetition 1 and Repetition 2, discriminant analysis failed to obtain a significant categorization and obtained a significance level of $p > 0.5$. Thus one need not be concerned that *any* grouping of the data could lead to successful categorization using this technique.

5.8.2. *Individual speakers*. Earlier experiments on incomplete neutralization rules found large variability between individual speakers. If speakers do have idiosyncratic ways of implementing the underlying contrast, analysis of pooled data could not catch these differences. When discriminant analysis is trained on· the pooled data, it seeks to average out any differences due to the speaker (or to the

TABLE II. The normalized structure coefficients for the discriminant analysis of data pooled across speakers and the pragmatic conditions in Experiment 1. The dash means that the variable was not included in the discriminant function since its *F*-ratio was smaller than 0.5

Vowel duration	−0.95
Stop duration	0.31
Burst duration	1.16
Nasal duration	−0.51
Closure pulses	—

TABLE III. Analysis of variance results for effect of underlying voicing on the individual speakers in Experiment 1. For most speakers, only the duration of the burst was significantly affected by underlying voicing

	Speaker				
		2	3	4	5
Vowel duration	ns	ns	ns	ns	ns
Stop duration	ns	ns	ns	ns	ns
Burst duration	***	***	**	*	***
Nasal duration	ns	ns	ns	ns	ns
Closure pulses	**	ns	ns	ns	ns

*** $p < 0.001$.
** $p < 0.050$.
* $p < 0.150$.
ns: $p > 0.300$.

word pair) in order to better categorize the voice feature itself. But if speakers are also analyzed separately and the mean performance obtained, then individual differences in the production of the timing variables should manifest themselves as an improvement in the ability of discriminant analysis to classify correctly in the speaker-by-speaker situation relative to the speaker-pooled results.

First, all the data were pooled across conditions and word pair, yet separated by speaker. Analysis of variance of these data sets (Table III), once again shows little more than that the burst duration tended to be significant. When discriminant analysis is applied, however, performance on the whole data set increases from 14% above chance to 22% above chance. In Fig. 2, the pooled percent-correct categorization and the mean of speaker-dependent categorization results are shown for each condition. In all conditions the average of the speaker-dependent analyses was higher than the speaker-pooled analysis. This improvement in the speaker-dependent analysis over pooled measures the degree of idiosyncrasy between

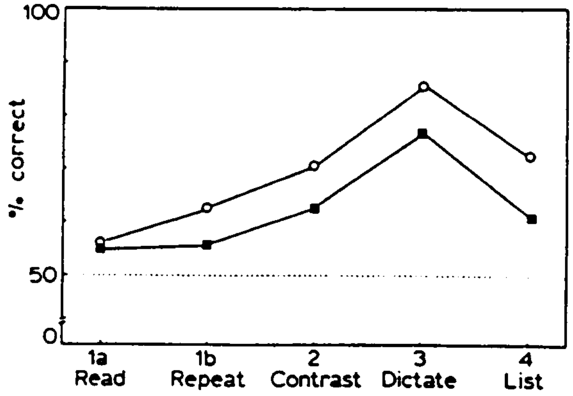

Figure 2. The percent correct classification by discriminant analysis when trained on each speakers' productions across all conditions (O) together with similar classifications when trained across all speakers (■) together for Experiment 1.

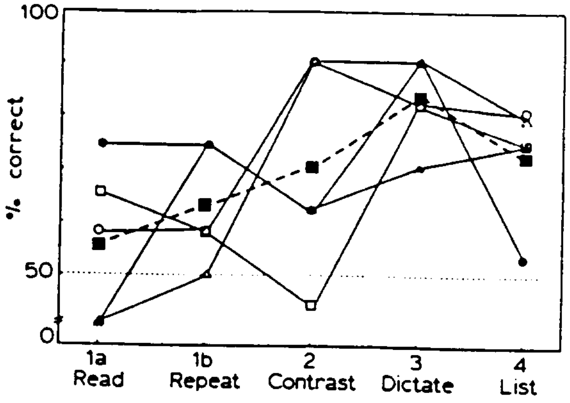

Figure 3. Categorization of voicing by discriminant analysis trained on individual speakers' productions and tested on individual conditions. Speakers: △, 1; ○, 2; □, 3; ▲, 4; ●, 5, ■: mean of five data points for each condition in Experiment 1.

speakers in their implementation of neutralization. Clearly the speakers are different, but this group of speakers, who are particularly homogeneous in dialect (same town, same age), exhibit only modest differences.

We can also show the individual speakers' differences for each condition, as in Fig. 3, where the means for each condition are plotted again. It can be seen that speakers vary somewhat as to which condition shows the greatest contrast (as though they interpreted the pragmatics of the tasks differently). The speakers seemed to treat Condition 2 in two different ways. Two speakers seemed to reduce the contrast here relative to Condition 1 (collapsing A and B which were performed in different orders by the speakers) while three speakers increased the contrast. All speakers do, however, produce a clear contrast in at least one condition. And no speaker does better in any condition than about 90% correct—still much poorer than would be expected for a real phonological contrast like *Bunde–bunte*.

To look more closely at the differences between the speakers, we can again examine the structure coefficients for each speaker as in Table IV. Comparison between columns shows that while some speakers (like Speakers 1 and 5) exhibit an inverse combination of vowel duration and burst duration in maintaining the underlying voicing difference, others varied mainly the duration of the stop closure

TABLE IV. The normalized structure coefficients for individual speakers when discriminant analysis seeks underlying voicing in Experiment 1. Variables with *F*-ratio greater than 0.5 were not included and are marked with a dash

	Speaker				
	1	2	3	4	5
Vowel duration	−1.08	−0.57	—	−1.32	−1.03
Stop duration	0.46	—	−1.26	1.00	—
Burst duration	1.13	1.10	−1.29	1.01	1.26
Nasal duration	−0.46	−0.24	−1.13	—	−0.87
Closure pulses	0.26	0.24	—	0.046	−0.28

Table V. Results of analyses of variance on individual word pairs in Experiment 1

	seit/seid	*Rat/Rad*	*bunt/Bund*
Vowel duration	ns	ns	ns
Stop duration	ns	ns	ns
Burst duration	***	***	***
Nasal duration	—	—	ns
Closure pulses	ns	ns	0.191

ns: $p > 0.300$.
*** $p < 0.001$.

and the burst (Speaker 3), or a balanced combination of all three variables (Speaker 4). In any case, discriminant analysis relied fairly strongly on the final burst measure for all speakers' productions in distinguishing the underlying contrast.

5.8.3. *Analysis by word pair.* Since several previous studies of German voice neutralization obtained prominent differences between individual lexical pairs in the correlates of incomplete neutralization, the data were separated by word pair. In this experiment we also included a "minimal pair" that is supported only by orthography rather than by a phonological alternation (*seit, seid*). Again analysis of variance, as shown in Table V, was not particularly informative. Burst duration was significant for each pair, but little else can be seen. No prominent differences between the word pairs càn be seen.

Another way to test this is to train discriminant analysis on the data pooled across speaker and condition, but separated by the word pairs, *Bund–bunt, Rad–Rat* and *seid–seit.* Word-pair dependent discriminant analysis (for each of which one third of the data are employed) was able to label underlying voice with 67% correct, only 3% better than the analysis on the data pooled across words. The word-pair dependent performance is about the same as the pooled analysis indicating once again only minor differences between the words.

The lack of variability between word classes is further supported by similarities in the structure coefficients shown in Table VI. A combination depending mainly on the burst and vowel duration was important in maintaining the contrast for all the word pairs. Other measures were important as well. The role of closure pulsing and stop-closure duration varied quite a bit across the word pairs while nasal duration was, of course, available only for *bunt–Bund.* Word-pair effects appear to be

Table VI. The structure coefficients for discriminant analyses on individual word pairs in Experiment 1

	seit/seid	*Rat/Rad*	*bunt/Bund*
Vowel duration	0.83	0.63	0.44
Stop duration	—	−0.29	−0.49
Burst duration	−1.23	−1.17	−1.01
Nasal duration	—	—	0.39
Closure pulses	0.49	—	−0.41

minimal in this experiment. This is surprising given the differences between the pairs in segmentation criteria, syntactic position of the test words, as well as differences in neighboring segments, and so forth. It is also surprising since the pair *seit–seid* is not supported by alternations, but only by orthography.

5.9. *Discussion*

The results of this experiment demonstrate several important points about the incomplete neutralization effect.

5.9.1. *The incomplete neutralization effect does not appear to be an artifact due to unnatural speech production.* All the conditions of this experiment showed a significant difference between the two underlying voicing categories—even when the words were spoken in sentences that were not read from orthography and for which the context posed no semantic ambiguity. In Condition 1, which was always performed first, the subjects had no idea what the focus of the experiment was, yet they produced enough difference between the voicing categories that discriminant analysis could identify them with accuracy significantly better than chance. We believe that our precautions to prevent subjects from guessing the primary interest of this experiment were completely effective. Thus any differences between conditions are not explainable as artifacts.

5.9.2. *The effect is not dependent on phonological alternation.* Sometimes neutralization can result in a situation where speakers have a choice of underlying lexical forms. An example in American English is *water* which is normally flapped at the phonetic level, hence could be derived from either underlying /t/ or /d/. In German, such a case occurs with non-alternating words like *seid, und, was,* etc. In this case, the orthography may encourage the choice of a particular underlying spelling. There is no reason to assume that this must be an experimental artifact. Since our subjects treated *seid vs. seit* as analogous to *Rad vs. rat* in all conditions, the data strongly suggest that these words have distinct underlying forms—even though that difference is supported only by the orthography, and not by phonological alternations.

5.9.3. *Speakers can control the degree of neutralization.* When called upon by the pragmatics of the task to contrast the minimal pairs and make a difference between them, speakers increased the differences between the voiced and voiceless categories such that discriminant analysis could do a better job of telling the words apart. This demonstrates clearly that speakers have relatively continuous control over the acoustic parameters that were measured in this experiment. The way in which they implemented this difference when asked to contrast the words seems to be simply an exaggeration of the differences they made when no difference was called for (as in Condition 1). When speakers had a pragmatic reason to keep the words distinct, they apparently modified their implementation of the syllable-final devoicing process so as to differentiate these from the underlying [– voice] segments. This seems possible only if the devoicing is different from [– voice].

In this experiment, we were apparently able to change significantly the communicative task for the speakers such that they made consistent variations in their productions. Condition 1 (both A and B) showed the least difference in underlying voicing, while Conditions 2 and 3 showed much more. The dictation task (Condition 3), in which a listener was trying to guess which word was being pronounced, yielded the greatest difference in underlying voicing. Of course, these pragmatic conditions are partly confounded with the order of their presentation in this experiment (since after Condition 1, the speakers were fully aware of the focus of our attention). This supports the contention of Fourakis & Iverson (1984) that speakers *could* impose a contrast on these pairs if they had a pragmatic reason to do so. Even so, mere awareness of our research interest did not completely determine their phonetic detail, since the word list task (presented as Condition 4) showed much less contrast than the dictation task (in fact, Condition 4 shows about the same contrast as was obtained in Port & O'Dell, 1985). Still, it is methodologically important that this naturally occurring variable can be manipulated experimentally.

5.9.4. *Information is widely distributed across the words.* Discriminant analysis found useful information for this discrimination distributed across most of the spectro-temporal variables in the words. This was clearly demonstrated by the structure coefficients of Tables II, IV and VI which show that the vowel duration, the stop duration and the measure of burst intensity and duration all contribute strongly to the differentiation of the voicing categories. Since analysis of variance tends not to find significant effects, the results show that the speakers are exhibiting a number of weak cues for the contrast, rather than any single feature. The variable of burst duration depends partly on the dynamic range of darkness marking on the sound spectrograph but correlates highly with the peak intensity and integrated energy in the release burst of the stops. It is an acoustic correlate of glottal aperture during the stop (Rothenburg, 1968). Thus, there are two general classes of cues for the underlying voicing: one set of cues indicating differences in the overall timing of the oral syllabic gestures (vowel duration and consonant constriction durations) and another set that reflect glottal aperture (duration of the burst and the number of glottal pulses during closure). Both sets play an important role in differentiating the underlying contrast.

5.9.5. *The perceptual utility of the difference can vary.* Is there any reason to suppose that this 'vestigial' difference is of practical use in everyday speech perception? Almost certainly it is not useful for cases like Condition 1, the condition most similar to everyday use of language. Performance is only slightly above chance. But in Condition 3, where the speaker is trying to maintain a difference, listeners can probably use the information in natural communicative tasks. This is presumably the reason speakers pronounced them in that way. The next experiment will provide some information relevant to this issue.

5.9.6. *Dealing with many weak cues is essential.* For this perceptual problem, to determine the underlying voicing of such pairs, as with many other problems in speech perception, listeners apparently must use many cues, no one of which is sufficient. This orientation contrasts strongly with the traditional linguistic approach that emphasizes necessary and sufficient features for the definition of categories.

Discriminant analysis is a particularly simple and highly constrained learning system that makes categorization from many weakly predictive cues. More powerful techniques exist for combining a large number of variables for categorization, including connectionist networks (Rumelhart & McClelland, 1986; Anderson, Merrill & Port, 1988). Most of these learning-based techniques are not restricted to linear combination of the variables. Discriminant analyis is useful here since it allows a fairly large number of variables to be combined in some optimal way to make a categorical decision. If the system is trained across a group of speakers' productions, it assigns weights to each variable to find a speaker-independent way of making the decision.

An important limitation on the generality of the technique employed here is that the system is dependent on a complete table of spectro-temporal measurements for training. Every token must be fitted into the same table. The particular measurements we made are quite arbitrary—even though they were chosen to describe prominent and psychologically salient properties of the utterances. The problem is that there is an indefinitely large number of ways that such measurements might be made, and the kind of boundaries we employed are difficult to find automatically. Yet, discriminant analysis requires a matrix of measured parameters. We measured very few here (only five) but the maximum set would be, say, a set of FFT coefficients for each 5 ms frame for each utterance. Thus, discriminant analysis is a model of speech perception with serious weaknesses. It depends on reliably-made input measurements that are the same for each item to be categorized and is restricted to linear combinations of the variables. This is why we point out that a fully general system, sufficient to solve such perceptual problems for all speech styles in *any* language, must employ many more subtle properties of the speech signal (see Port *et al.* 1988 for further discussion of these issues). Our very success at describing these subtle but linguistically essential properties of the speech signal for a restricted set of minimally contrastive pairs only points to the massiveness of this task that must be solved by any full phonetic perception system.

In brief, this experiment first replicates earlier results by demonstrating incomplete neutralization for the German devoicing rule. Thus it seriously undermines the traditional interpretation of a discrete phonological rule. Second, it shows that pragmatic factors can influence the degree of contrast exhibited by speakers, thereby demonstrating continuous speaker control over the variables. Third, it shows that information relating to the contrast is distributed widely across the words, and, finally, that the differences exhibited are manifested across much of the duration of the words. This raises the question of the extent to which human listeners can actually make perceptual use of information such as that exploited by discriminant analysis, the question addressed in the next experiment.

6. Experiment 2: perception of syllable-final voicing

6.1. *Introduction*

The appropriateness of discriminant analysis in the above experiment is dependent on the assumption that this technique differentiates and categorizes tokens using similar cues to those available to a native speaker when participating in a similar task. The possibility exists, however, that the cues employed by discriminant analysis in its categorization (that is, the variables we measured from spectrograms) might be too subtle for a human to use for accurate categorization in the same task.

A human listener might perceive no difference between the two groups, meaning that even if a contrast were detected by discriminant analysis from our measurements, it would have no communicative value. On the other hand, another possibility is that human listeners could make use of many other cues not measured in our production data. In this case, they might actually outperform discriminant analysis on this task.

Although a direct test of this question is very difficult, since we cannot be sure that human listeners use the variables we measured, the next experiment investigated the relevance of categorization by discriminant analysis as a predictor of the degree of voicing contrast for native speakers. We directly compare human performance in a set of word identification tasks with the results of discriminant analysis on similar tasks. We cannot prove that the same information is used, of course. In the best case, we might demonstrate that analogous information is used if it appears that factors that affect one "perceptual system" also affect the other.

One issue that arises here is an appropriate measure of performance. On reflection, it turns out that percent correct is not a good measure of performance since human subjects may have a bias to respond one way or another. Human subject performance in such a task has two components, first, the ability of the perceptual system to simply discriminate the two categories, and, second, a decision criterion that reflects the degree to which the listener is willing to risk making one kind of error (calling a /d/ a /t/) *vs.* the other error (calling a /t/ a /d/). Thus we employed d' as a measure of discriminability that is independent of response bias (Swets, 1961, or see Kantowicz & Sorkin, 1983, for an accessible introduction to signal detection theory). It is based upon both correct responses and the nature of incorrect ones. It provides a more reliable basis for comparison between human listeners and a bias-free numerical technique like discriminant analysis.

6.2. *Methods*

In the first part of the experiment, native speakers attempted to identify a subset of the productions from Experiment 1. In order to avoid tiring subjects, the data from only two representative speakers from Experiment 1 were used. The speakers were chosen to differ from each other in the ability of discriminant analysis to identify their underlying voicing, but they were neither the worst nor the best of the speakers. In addition, only the recordings from Condition 3 (the dictation condition), and Condition 4 (the word list) were used. In these conditions, the word tokens could be easily isolated from the sentence by waveform editing and there was a large difference in the ability of discriminant analysis to determine underlying voicing (78% correct for Condition 3 *vs.* 62% for Condition 4).

6.2.1. *Materials*. A total of 48 tokens from the original tape (six words, two speakers, four repetitions) from both Conditions 3 and 4 were digitized, carefully edited, copied three times and fully randomized (total 288 stimuli). The tokens were separated by 3 s pauses with a 10 s pause after every 12 tokens and a longer pause between each block of 96.

6.2.2. *Subjects*. The listeners were five native Germans from several areas of the Federal Republic of Germany and Austria who were currently studying as undergraduates or graduates at Indiana University. Thus, unlike Experiment 1, all were excellent speakers of English. Major facts about their home regions and dialect history were recorded. None had ever participated in any of the previous experiments in this laboratory.

6.2.3. *Procedures*. Subjects checked a box on an answer sheet choosing between orthographically spelled /d/-word or /t/-word. The categorization performance on the two intended stop types was obtained for each listener on both speakers in each condition. The discriminant analysis used for comparison with the human listeners was performed by training on all speakers, all word pairs and all conditions in Experiment 1, and then testing categorization performance for just the productions of Speakers 1 and 4 in Conditions 3 and 4. Thus, the system was trained on all speakers and conditions even though training was done on the same data as were tested. Again this procedure seems most closely to resemble normal assumptions about the invariance of words. In the analysis above for Experiment 1, it was found that Speaker 1 maintained a higher level of contrast than Speaker 4, and the average level of contrast for Condition 3 was higher than that of Condition 4.

6.3. *Results*

6.3.1. *Listeners' identification*. The percent correct identification of each listener is shown in Fig. 4 for the two speakers and from both the productions of Condition 3 (where speakers were trying to maximize the contrast) and from the list of isolated words in Condition 4. The listeners had an overall performance of 69.2% (SD = 5.1), about 20% better than chance. For all but one case, the listeners performed better on Condition 3 productions than on Condition 4 regardless of the speaker. All listeners did better on Speaker 1's productions than on Speaker 4's. There was no correlation that we could detect between listeners' performance and the dialect of German they spoke.

In fact, the subjects exhibited strong response bias in Condition 3 (where they

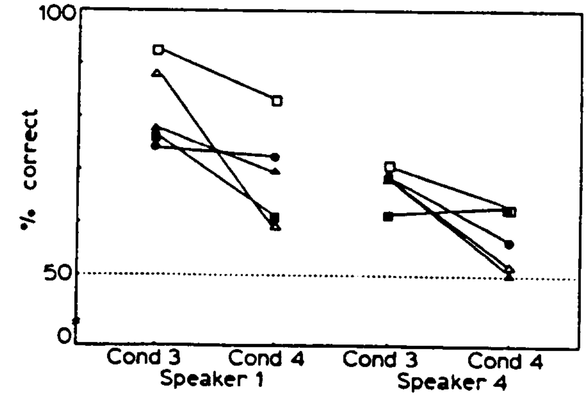

Figure 4. The percent correct categorization of the underlying voicing by the five listeners of Experiment 2 (△: 1; ■: 2; □: 3; ▲: 4; ●: 5) for the two conditions and two speaking voices selected from Experiment 1. Notice that listener 3 does better than 90% correct for Speaker 1's productions in the dictation condition, Condition 3.

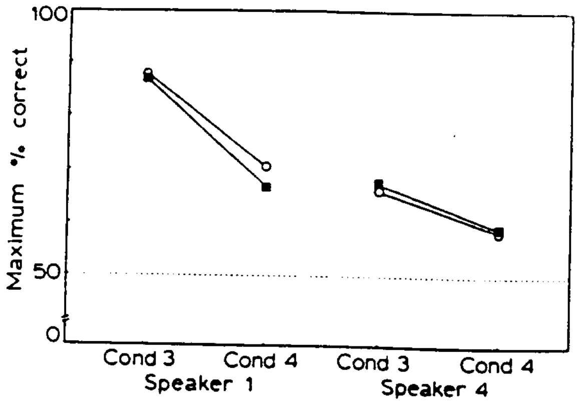

Figure 5. Direct comparison of the maximum percent correct listeners' performance (■)with the performance of discriminant analysis (O) on the same data in Experiment 2. Maximum percent correct is computed using *d'* to correct for listener response bias.

labelled 60% of the productions /t/). This bias can be corrected using *d'* and estimating from it the maximum possible percent correct (Swets, 1961). Accordingly, Fig. 5 shows the maximum possible percent correct identification in solid squares for the four conditions. These scores are somewhat better than the mean of the observed percent correct shown in Fig. 4 since subjects' actual discriminability is sometimes distorted by a response bias for /t/.

6.3.2. *Discriminant analysis.* Discriminant analysis on the same productions achieved an overall average of 70.8%. Since there the *a priori* probability of each category was 50% and discriminant analysis places category boundaries half-way between group centroids, the technique provides bias-free categorization. The results of the discriminant analysis are compared with the average of the human listeners' results for each condition in Fig. 5. The patterns are very similar and have absolute values that are very close. Of course, discriminant analysis was trained on a set of data that included this test set, and the human subjects were, of course, "trained" on a quite different set. Nor are the human listeners constrained to employ the kinds of measures that we made from spectrograms, nor constrained to linear combination of the variables. Nevertheless, the similarity of performance in these four conditions is highly suggestive that German listeners make use of the kinds of temporal information measured in our first experiment when asked to identify these partly neutralized stops.

One further way of comparing native listeners with discriminant analysis is to examine the particular tokens each made errors on. If they made errors on largely non-overlapping sets of productions, it would suggest that they are not exploiting the same information. If they make mistakes on the same tokens, it would imply reliance on similar information. The results for one randomly chosen listener on Speaker 4, Condition 3 were compared with the performance of discriminant analysis on those same tokens. Table VII shows that for the total of 96 tokens, both classifiers got 60 correct and both got 15 wrong. Thus, 78% were either correctly classified by both or incorrectly classified by both. This shows there is a tendency for native listeners and discriminant analysis to classify the same tokens the same way

TABLE VII. The number of correct and incorrect identifications made by discriminant analysis compared with results by one German listener for the productions in Experiment 2 of Speaker 4, Condition 3

	Listener	
Discrim analysis	Right	Wrong
Right	60	8
Wrong	13	15

and further supports the notion that very similar information is relied upon by the two classification systems.

6.4. *Discussion*

The results from Experiment 2 replicate results from earlier perception experiments demonstrating that listeners were able to perform better than chance in identification tasks of partially neutralized features (Port *et al.*, 1981; Port & O'Dell, 1985). Apparently, discriminant analysis based on the spectro-temporal variables we measured allows categorization at about the same level as human listeners and seems to do better or worse in accord with the same variables. Thus, there is a reasonable empirical basis for using discriminant analysis as a measure of the degree of contrast. Discriminant analysis in this case performs about the same in discriminating the underlying voicing categories as human listeners. It even has a tendency to misclassify the same tokens as are missed by the human listeners. Thus, although there are surely great differences in processing between our procedure of hand-measurements plus discriminant analysis and native-speaking German listeners, apparently the information that is extracted from the speech signal and employed in perceptual decisions is largely equivalent.

7. Concluding discussion

The results of these experiments have implications for a number of issues, including speech perception, speech production and the nature of phonological models.

7.1. *Implications for speech perception*

Although traditional research on speech perception has emphasized a fairly small number of robust "speech cues" (e.g., Liberman, Shankweiler, Cooper & Studdert-Kennedy, 1967; Stevens, 1983), work on speech recognition has made it clear that for most perceptual tasks, a large number of bits of information must be employed (Klatt, 1977, Vaissière, 1985). The question of how temporal information can be used is rather difficult (Port, Anderson & Merrill, 1988). As long as a speech signal can be segmented in some reliable way, then measuring and saving segmental durations in auditory short-term memory is no serious problem. Making such measurements is straightforward when minimal pairs (or near-minimal pairs) are used since easily segmentable words can be selected or the segmentation can be defined *ad hoc* for the particular pair. But how can durations be made in general? For example, what is the "vowel duration" in *Schnee* (no matter what word follows) and how should it be compared with the vowel in *trink* or *Hirsch*? How can the duration of /r/ be measured in these words? These are extremely difficult questions. Unless completely general acoustic segmentation and labelling procedures can be devised that work strictly from the bottom up (something we doubt can ever be achieved), then temporal information could only be saved in auditory memory by retaining an unsegmented, relatively raw form of the speech signal. Such a representation would require a vast number of bits of information in order to support the kind of temporal subtlety implied by the results of these two experiments. The distribution of information over a lengthy window shows that relatively precise information about the temporal location of various events is essential for human-like perception of speech.

This problem is related to the debate going on in cognitive science regarding the appropriateness of classical categories in cognition. As pointed out by many authors (Rosch, 1978, Smith & Medin, 1981; Lakoff, 1987), the classical definition of a category in terms of a set of necessary and sufficient properties is not adequate. Even though linguistic models invariably assume that categories are defined in this way (e.g., /t/ is defined by [− voice, + coronal, etc.]), humans easily recognize many kinds of categories that do not have such defining traits (Wittgenstein's famous example is the notion of a "game"). Lakoff (1987) argues that many, perhaps most, cognitive categories have this property. The point here is that if short-term auditory store is relatively unanalyzed, and the underlying voicing feature is extracted directly from it, then it would appear that this feature cannot be a classical category. It would have to be derived directly from meaningless "subsymbolic features" that have merely numerical values (Smolensky, 1988) rather than simply obtained by pattern matching from previously obtained contentful categories.

In general, then, our data seem to present a problem for classical models of the perception of speech since these results are most compatible with a model that uses relatively meaningless subsymbolic components to directly extract a voicing label. The reason comes down to the fact that general segmentation of speech signals, in a way analogous to our measurements in Experiment 1, is probably impossible. Phoneticians cannot do it from spectrograms unless they restrict themselves to minimal pairs (where *ad hoc* criteria are sufficient), and neither, we suspect, can human listeners.

7.2. *Implementation rules*

Most likely the sensitivity of the temporal phonetic parameters to pragmatic factors does not differentiate them from many other implementation rules. By implementation rules, we mean the "rules" that convert the segmental output of the phonology into graded gestures in time. Presumably, all implementation rules are susceptible in subtle ways to pragmatic and other interpretive factors. What is surprising in these data is only that both linguists and native speakers find it very difficult to detect the phonetic difference between these pairs. Impressionistic transcription allows observation of a very abstract rule of neutralization stated in terms of a phonetic alphabet—a rule that changes a single feature. Yet we find that the articulatory implementation of these lexical objects is based upon their underlying, not superficial, identities.

7.3. *Speech production*

It seems to us that the evidence points toward a control system for speech production that is inherently dynamic (cf. Fowler, 1986; Kelso, Saltzman & Tuller, 1986; Browman & Goldstein, 1986; 1988). In this kind of system, pragmatically or expressively motivated changes in control parameters for the production of speech could have effects that are distributed throughout the production of syllables. In terms of such a control system, it may be easier to see how "phonological rule" effects can be implemented (see Browman & Goldstein, 1988; or Port, 1986*b*). In a dynamic system, control parameters can change slowly, yet have effects that are distributed in time. Since the system directly produces gestures (not specifications for gestures), it is easy to see why it can be highly dependent on speaker idiosyncracies. If such a model can be formally developed, it may handle the effects of many other phonological rules too. Then the linguistic model of a serially ordered rule system that takes ordered symbol strings as input and issues ordered symbol strings as output will need a dynamic level at the lower end that generates real gestures. A devoicing process for German syllable codas should be implemented at that lower level, the level that resembles "implementation rules".

why?

7.4. *Importance of speaking styles*

Our results suggest a new methodology for studying such subtle speech characteristics as speaking style. The small literature of research into speaking styles (Lieberman, 1963; Labov, 1981) shows that subtle properties of the social context influence many kinds of phonetic detail. Similarly, research on speaking rate also demonstrates complex non-linear changes in speech timing (Port, 1981; Miller, 1982). These results only confirm the layman's awareness that the details of how we talk are affected by the social and communicative goals in force at the instant of speaking. It also accords with our experience that actors and other skilled readers can produce speech in ways that both greatly affect our ability to understand a text and affect the nature of the interpretation. Clearly, it is phonetic detail that provides most of this information since recordings and telephones suffice to transmit it. The inability of analysis-of-variance to discover more than a difference in the burst in Experiment 1 shows how much is being lost by looking at a single variable at a time. Listeners can do better than that and discriminant analysis based on our spectro-temporal measurements also does better, but only by looking for complex relations in temporal patterns. When investigating issues like these, analysis–of-variance is nearly useless as a statistical technique.

7.5. *Implications for phonology*

The theoretical significance of this effect for phonology is considerable. The results implicate a level of rules that resemble distinctive phonological features but which operate at a very low level where speaker differences are considerable. One can apparently only write accurate rules for German devoicing by making them speaker-dependent and by employing a very large set of articulatory features to capture the detailed dynamic differences between speakers' implementation of the contrast. But, in that case, every speaker will have his own rule, yet *none* of the rules will actually neutralize the contrast! Ironically, if the German devoicing rule is to specify *all* the phonetically controllable parameters of speech production, then it will be quite incapable of capturing the linguistically significant fact about German that there is practical neutralization of voicing in syllable-final position. The most important and far-reaching implication of these experiments is that German clearly does *not* have a syllable-final devoicing rule. This fact cannot be avoided or wished away. Thus, apparently one of the simplest and most familiar examples of a phonological rule must be accounted for in some other way, and in a way that is messy. German speakers do not simply change [+ voice] to [− voice] for obstruents at the end of syllables. The tantalizing question remains then "Where *should* the practical neutralization in German be stated?" We do not have an answer to this question. We know only that individual speakers do not employ a feature-changing neutralization rule. Practical neutralization is a fact, but it is apparently not a rule.

Beyond this, our data show that phonological alternations are not a prerequisite for the creation of an underlying form that is abstract. Apparently, given the presence of a general neutralizing process at implementation, speakers have the option of choosing either unit of the neutralizing pair as the underlying form. It appears that orthography may sometimes be sufficient to encourage speakers to choose the phonologically more abstract form as underlying.

In conclusion, the results of these experiments demonstrate clearly that the incomplete neutralization effect is not an artifact. German apparently does not have an abstract phonological rule of neutralization, despite almost a hundred years of assertions by linguists and German pedagogists that it does. Our results suggest a new way of evaluating the notion of "degree of contrast" by use of an optimization technique such as discriminant analysis. We have shown that German speakers can control the degree of distinctness of syllable-final stops depending on communicative contingencies. The question of what neutralization rules really are and how they work is apparently much more mysterious and less well understood than has been assumed within modern linguistics.

We are grateful especially to Matthias Weimer for helpful assistance with the conduct of this study and to the other exchange students from Dillenburg, Hessen, West Germany. We also thank Amy Rominger, Sven Anderson, John Merrill, Stuart Davis and Dan Maki for advice and consultation. This research was supported in part by grants from the National Institutes of Health HD12511 and the National Science Foundation DCR85-05635 and DCR85-18725, and by the office of Research and Graduate Development, Indiana University.

References

Anderson, S. R., Merrill, J. W. L. & Port, R. F. (1988) Dynamic speech categorization with recurrent networks. In *Proceedings of the 1988 connectionist summer school* (Morgan-Kauffmann), pp. 398–406.

Browman, C. & Goldstein, L. (1986) Towards an articulatory phonology, *Phonology Yearbook*, 3, 219–252.

Browman, C. & Goldstein, L. (1988) Tiers in articulatory phonology, with some implications for casual speech. In *Papers in laboratory phonology I: between the grammar and the physics of speech* (J.

Kingston & M. E. Beckman, editors). Cambridge: Cambridge University Press.

Charles-Luce, J. (1985) Word-final devoicing in German: effects of phonetic and sentential contexts, *Journal of Phonetics*, **13**, 309–324.

Charles-Luce, J. (1987) An acoustic investigation of neutralization in Catalan. Indiana University, Department of Linguistics, doctoral dissertation.

Dinnsen, D. A. (1984) A re-examination of phonological neutralization, *Journal of Linguistics*, 21, 265–279.

Dinnsen, D. A. & Charles-Luce, J. (1984) Phonological neutralization, phonetic implementation and individual differences, *Journal of Phonetics*, **12**, 49–60.

Dinnsen, D. A. & Garcia-Zamor, M. (1971) The three degrees of vowel length in German, *Papers in Linguistics*, **4**, 111-126.

Fourakis, M. & Iverson, G. K. (1984) On the 'incomplete neutralization' of German final obstruents, *Phonetica*, **41**, 140–149.

Fourakis, M. & Port, R. F. (1986) Stop epenthesis in English, *Journal of Phonetics*, **14**, 197–221.

Fowler, C. (1986) An event approach to the study of speech perception from a direct-realist perspective, *Journal of Phonetics*, **14**, 3–28.

Fox, R. & Terbeek, D. (1977) Dental flaps, vowel duration and rule ordering in American English, *Journal of Phonetics*, **5**, 27–34.

Hubbel, A. F. (1950) *The pronunciation of English in New York City*. New York: King's Crown.

Huff, C. (1980) Voicing and flap neutralization in New York City English. In *Research in phonetics, report no. 1*. pp. 233–256. Bloomington: Indiana University Department of Linguistics.

Kantowicz, B. H. & Sorkin, R. (1983) *Human factors: understanding people-system relationships*. New York: Wiley.

Keating P. (1985) Universal phonetics and the organization of grammars. In *Phonetic linguistics, Essays in honor of Peter Ladefoged* (V. Fromkin, editor), pp. 115–132. Orlando: Academic Press.

Kelso, J., Scott, A., Saltzman, E. & Tuller, B. (1986) The dynamical perspective on speech production: data and theory, *Journal of Phonetics*, **14**, 29–59.

Klatt, D. (1977) A review of the DARPA speech understanding project. *Journal of the Acoustical Society of America*, **62**, 1345–1366.

Klecka, W. (1980) *Discriminant analysis, Sage University paper series qualitative applications in the social sciences, no. 07-019.* Beverley Hills and London: Sage Publications.

Labov, W. (1981) Resolving the Neogrammarian controversy, *Language*, **57**, 267–308.

Lakoff, G. (1987) *Women, fire and dangerous things*. Chicago: University of Chicago Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. (1967) Perception of the speech code, *Psychological Review*, **74–76**, 431–461.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech, *Language and Speech*, **6**, 172–187.

Miller, J. (1981) Effects of speaking rate on segmental distinctions. In *Perspectives in the study of speech* (P. Eimas & J. L. Miller, editors) pp. 35–74. Hillsdale, NJ: L. Erlbaum.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K. & Bent, D. (1975) *Statistical package for the social sciences*. New York: McGraw-Hill.

Port, R. F. (1981) Linguistic timing factors in combination, *Journal of the Acoustical Society of America*, **69**, 262–274.

Port, R. F. (1986*a*) Invariance in phonetics. In *Invariance and variability in speech processes* (J. S. Perkell & D. H. Klatt, editor), pp. 540–559. Hillsdale, NJ: L. Erlbaum.

Port, R. F. (1986*b*) Translating linguistic symbols into time. In *Research in phonetics and computational linguistics* 5, 155–174. Indiana University, Bloomington; Department of Linguistics.

Port, R., Anderson, S. & Merrill, J. (1988) Temporal information and memory in connectionist networks. Technical Report No. 265, Department of Computer Science, Indiana University.

Port, R. F. & O'Dell, M. (1985) Neutralization of syllable-final voicing in German, *Journal of Phonetics*, **13**, 455–471.

Port, R. F., Mitleb, F. M. & O'Dell, M. (1981) Neutralization of obstruent voicing is incomplete, *Journal of the Acoustical Society of America*, **70**, S10. Also in *Research in Phonetics*, **4**, pp. 163–175. Department of Linguistics, Indiana University.

Port, R. F., Reilly, W. T. & Maki, D. (1988) Use of syllable-scale timing to discriminate words, *Journal of the Acoustical Society of America*, **83**, 265–273.

Pye, S. (1986) Word-final devoicing of obstruents in Russian. In *Cambridge Papers in Phonetics and Experimental Linguistics* 5, pp. P1–P10. Cambridge University; Department of Linguistics.

Rosch, E. (1978) Principles of categorization. In *Cognition and categorization* (E. Rosch & B. B. Lloyd, editors) pp. 27–48. Hillsdale, NJ: L. L. Erlbaum.

Rothenburg, M. (1968) The breath-stream dynamics of simple-released-plosive production. *Biblioteca phonetica*. Basel: Karger.

Rumelhart, D. E. & McClelland, J. L. (1986) *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*. Cambridge, MA: Bradford Books/MIT Press.

Slowiaczek, L. M. & Dinnsen, D. A. (1985) On the neutralizing status of Polish word-final devoicing. *Journal of Phonetics*, **13**, 325–341.

Smith, E. E. & Medin, D. L. (1981) *Categories and concepts*. Harvard University Press.

Smolensky, P. (1988) On the proper treatment of 'connectionism'. In *The brain and the behavioral sciences.*

Stevens, K. N. (1983) Design features of speech sound systems. In *The production of speech* (P. F. MacNeilage, editor), pp. 247–261. New York: Springer-Verlag.

Swets, J. A. (1961) Is there a sensory threshold? *Science*, **34**, 168–177.

Vaissière, J. (1985) Speech recognition: a tutorial. In *Computer speech processing* (F. Fallside & W. A. Woods, editors) pp. 191–242. Englewood Cliffs, NJ: Prentice Hall.