ON THE PERCEPTIBILITY OF SUB-PHONEMIC DIFFERENCES: THE TOUGH-DUCK EXPERIMENT

Bruce L. Derwing and Terrance M. Nearey

The University of Alberta

1. Introduction. It has long been assumed by proponents of "classical" phonemic theory in linguistics that only contrastive or distinctive segments or features are "heard" by naive (i.e., phonetically untrained) speakers of a language. The following citations are illustrative of this popular view:

> Only two kinds of linguistic records are scientifically relevant. One is a mechanical record of the gross acoustic features, such as is produced in a phonetics laboratory. The other is a record in terms of phonemes, ignoring all features that are not distinctive in the language (Bloomfield 1933: 85).

> In the course of many years of experience in the recording and analysis of unwritten languages, American Indian and African, I have come to the practical realization that what the naive speaker hears is not phonetic elements but phonemes. . . It is exceedingly difficult, if not impossible, to teach a native to take account of purely mechanical phonetic variations which have no phonemic reality for him (Sapir 1949: 47-48).

Though originally buttressed by what can only be described as anecdotal evidence, at best, a few experiments in recent years have lent some measure of empirical support for this traditional view. One of the earliest of these studies was an experiment run by Brown (1958: 213-216), who showed that native English speakers did not spontaneously take cognizance of a (sub-phonemic) distinction of vowel length in categorizing a set of colored chips, whereas native Navaho speakers (for whom the length feature was contrastive) did. The phenomemon of "categorical perception" in experimental phonetics might also be interpreted in support of this view. Thus, for example, in a now classic study of the role of voice onset time (VOT) in the perception of voiced vs. voiceless stops, Abramson & Lisker (1970) showed that only at certain relatively narrow portions of the VOT continuum could reliable distinctions be made and, furthermore, that the number of discrimination peaks involved was directly related to the number of phonemic distinctions made of the language of the speakers tested (i.e., one peak for English, corresponding to /d/ vs. /t/, for example, but two peaks for Thai, corresponding to /d/ vs. /t/ and /t/ vs. /t^h/.) Finally, in a rather different vein, Vitz & Winkler (1973), who ran a series of studies on "judged similarity in sound" between word-pairs, found that their results could be

largely accounted for by a simple phoneme model, in which phonemes were compared on a categorical or all-or-nothing basis. We can conclude from this that their subjects were generally unaware of, or at least relatively insensitive to, predictable or environmentally determined phonetic variations of phonemes.

Other considerations, however, suggest that speakers may well be sensitive to at least some sub-phonemic (or non-phonemic) distinctions. It is a common observation, for example, that speakers can readily detect a wide range of differences in pronunciation (e.g., dialect differences, mispronunciations by foreigners, etc.), few of which conform to the rather highly constrained contrastive patterns of any one language. (Thus it seems rather unlikely that the strong aspiration of the second segment in a word like steam, for example, would pass completely unnoticed by a hearer in any but the most heated of verbal exchanges.) There is experimental evidence, too, that the manipulation of so-called "redundant" phonetic parameters can induce sharp perceptual distinctions (sometimes at some distance from the locus of the manipulated feature). In one well-known study with synthetic speech, for example, Denes (1955) showed that non-contrastive differences in vowel length (recall the Brown study cited above) can cue perceptual distinctions in "voicing" among post-vocalic consonants; this study has now been replicated and extended using digital gating of natural speech by Hogan & Rozsypal (1980). Another relevant and important study is that of Shammass (1980), who produced a complex pattern of perceptual effects by the simple manipulation of segment duration in fricative + stop clusters (see also McCasland, 1977).

Faced with this somewhat conflicting body of indirect evidence, we attempted to construct a more direct test of the Bloomfield-Sapir hypothesis. Specifically, the question we wanted to answer was simply this: "Can naive speakers of English detect differences between allophones of the same phoneme?"

2. The Experiment. The phonemes chosen for investigation in the present study were English /t/ and /d/. This choice was made both because of the relatively wide range of allophonic variation exhibited by this pair and because of the apparent neutralization (or near-neutralization) of the /t-d/ contrast in at least two environments (viz., #s___V and V́___V). The phones (and presumed phonemes) represented in this study are indicated in Table 1 below, together with the paired sets of real and nonsense words selected for presentation to subjects. (Numbers in parentheses refer to the "forward" presentation order on the test, which was established by a randomization procedure.) The phonetic symbols provided are standard IPA, supplemented by the following set of diacritics: ^{h} for aspiration, ˞ for retroflection (with affrication), ^{w} for labialization, ‿ for "fronting" (i.e., true dental rather than alveolar point of articulation), ^{n} for nasal release, $^{?}$ for (simultaneous) glottal closure, ‾ for an

unreleased stop (sometimes equivalent to [ʔ], though not traditionally treated as such), and a superscript [s] or [z] to indicate a pre- or post-segmental "s or z environment" and any attendant phonetic effects (in word-final consonant clusters only).[1]

As noted above, both a real word and a nonsense word were chosen to illustrate each of the allophones under investigation. The purpose of the nonsense words was to attempt to control for orthographic interference, under the assumption that the "spellings" of nonsense words would be less accessible to subjects than for the real words and hence less likely to influence their judgments. The task was for subjects to evaluate each word for the presence or absence of a specific allophonic "target."

Method

The test items were recorded in a soundproof booth on a TEAC A-7030 tape recorder, using a Sennheiser MD 421N microphone, by a middle-aged male, native Canadian English speaker.[2] A cassette tape was then made from the master tape. The cassette tape was monitored for naturalness and to insure that all the relevant phonetic distinctions were clearly and consistently represented; this tape was then played back on a Sony Tapecorder TC-110A through a high fidelity amplifier and speaker in a variety of reasonably quiet, but otherwise quite ordinary, university classrooms. Subjects were 72 University of Alberta students registered in various sections of an introductory linguistics course. All testing was conducted during the first few days of the course, before any phonetic training had been provided, and all subjects retained were screened to insure that they were native monolingual speakers with no obvious hearing defects and with no prior exposure to formal linguistic or phonetic training.[3]

The subjects were provided with a particular "target" speech sound contained in a "probe" word which was read aloud by the experimenter. Half of the subjects (n=36) were told to focus their attention on the first sound of the probe word tough (i.e., the allophone [t^h] of /t/), while the remaining half were told to do the same for the first sound of the probe word duck (i.e., the allophone [d] of /d/). Subjects were then told that they would hear a list of real and nonsense words which might or might not contain the target sound. For each test item they were told to ask themselves the following question: "Does (the test item) contain the first sound of the word (probe)?" Responses were made on a written answer sheet according to the following scale:

> 0 - NO! = DEFINITELY NOT (i.e., I am quite certain that the target sound does not occur in the word)

Table 1. Test Items With Phones and Phonemes Represented

Phoneme	Phone	Real Word	Nonsense Word
/t/	[t^h]	tub (1)	tupp (8)
		tune (31)	toose (3)
		team (20)	teef (37)
		retain (38)	reteal (17)
		beat$_1$ (54)	lutt$_1$ (55)
	[$\underset{\cdot}{t}^h$]	tree (45)	triz (58)
	[$\underset{\cdot}{t}^{wh}$]	tweak (10)	twif (43)
	[$\underset{\cdot}{t}$]	streak (26)	struff (6)
	[$\underset{\cdot}{t}$]	eighth (16)	naitth (36)
	[$\hat{t}^-$]	beat$_2$ (21)	lutt$_2$ (29)
	[$t^{\cdot n}$]	beaten (57)	hatten (23)
	[t^s]	beats (48)	lutts (39)
	[$^s t^-$]	beast (53)	vist (28)
	[t]	steam (15)	stam (59)
/t/ or /d/	[ɾ]	butter (46)	geater (51)
		buddy (42)	zadey (9)
/d/	[d]	dumb (18)	dupp (47)
		dune (11)	doove (25)
		dean (2)	dobe (27)
		redeem (5)	rediff (30)
	[d^w]	dwell (12)	dweck (41)
	[$\underset{\cdot}{d}$]	dream (32)	drabe (4)
	[$\underset{\cdot}{d}$]	width (19)	medth (7)
	[$\hat{d}^-$]	bead (34)	pudd (44)
	[d^n]	sudden (52)	lidden (56)
	[d^z]	seeds (60)	rudds (24)
	[z_d-]	seized (22)	guzzed (50)
/t ʃ/	[$tʃ^h$]	chief (13)	chuff (35)
/d ʒ/	[dʒ]	jig (33)	jabe (40)
/θ/	[θ]	three (14)	threff (49)

1 - NO = PROBABLY NOT (i.e., I am fairly confident that the target sound does not occur)

2 - NO? = POSSIBLY NOT (i.e., I am not at all sure, but if forced to decide, I would have to say that the sound does not occur)

3 - YES? = POSSIBLY (i.e., I am not at all sure, but if forced to decide, I would have to say that the sound most likely does occur)

4 - YES = PROBABLY (i.e., I am fairly confident that the target sound does occur in the word)

5 - YES! = DEFINITELY (i.e., I am quite certain that the target sound does occur in the word)

Each word was presented three times, and subjects were given five practice trials on words containing [t^h], [d], or totally extraneous phones in various positions in the word (i.e., no new "test allophones" from Table 1 were used in the practice items). Half of the subjects in each group were given the test items in the order indicated by parentheses in Table 1; the remainder were presented with the test items in the reverse order.

Results

Results exhibited a strong bimodal pattern, viz., they tended to be heavily concentrated at the two extreme ends of the scale. Because of this highly non-normal distribution, it was decided to retabulate the data as a simple proportion of NO (=0 through 2) vs. YES (3 through 5) responses for each test item and probe. (It was in anticipation of this possibility that the response categories were labeled the way they were, i.e., in order to permit a clear bifurcation of positive vs. negative responses.) The resulting data are summarized below for each of the following four categories of stimuli: (1) real words with tough-probe, (2) nonsense words with tough-probe, (3) real words with duck-probe, and (4) nonsense words with duck-probe. Categories (1) and (2) appear in Table 2A below and categories (3) and (4) in Table 2B.[4]

The first observation to be made from these data is the clear distinction between test items containing the target allophone (invariably at the high end for all four categories) and test items containing a distinct phoneme from that represented by the target allophone (heavily concentrated at the opposite end in all cases). This indicates that subjects had little difficulty in distinguishing the target allophone from a phone representing a

Table 2A. Percent YES and Tukey-HSD Groupings for Tough-Probe

Rank Order	Real Words	% YES	HSD Groupings	Rank Order	Nonsense Words	% YES	HSD Groupings
1	tub	100		1.5	$lutt_1$	100	
3	retain	97		1.5	tupp	100	
3	tune	97		3.5	toose	97	
3	team	97		3.5	teef	97	
5	$beat_1$	94		5	reteal	94	
6	tree	92		6.5	twif	89	
8.5	tweak	83		6.5	triz	89	
8.5	streak	83		9	lutts	81	
8.5	steam	83		9	stam	81	
8.5	beast	83		9	struff	81	
11	beats	78		11	hatten	67	
12	beaten	69		12	$lutt_2$	53	
13	butter	56		13.5	naitth	44	
14	$beat_2$	44		13.5	vist	44	
15	eighth	25		15	geater	39	
16	seized	14		16.5	rediff	17	
17.5	chief	11		16.5	pudd	17	
17.5	buddy	11		19.5	guzzed	14	
19	width	9		19.5	chuff	14	
21.5	bead	6		19.5	lidden	14	
21.5	dwell	6		19.5	rudds	14	
21.5	dream	6		22	medth	11	
21.5	three	6		23	dupp	9	
26	sudden	3		25.5	threff	6	
26	redeem	3		25.5	zadey	6	
26	seeds	3		25.5	doove	6	
26	dune	3		25.5	dweck	6	
26	dean	3		28.5	drabe	3	
29.5	dumb	0		28.5	dobe	3	
29.5	jig	0		30	jabe	0	

Table 2B. Percent YES and Tukey-HSD Groupings for Duck-Probe

Rank Order	Real Words	% YES	HSD Groupings	Rank Order	Nonsense Words	% YES	HSD Groupings
2.5	buddy	100		2	zadey	100	
2.5	redeem	100		2	dweck	100	
2.5	dwell	100		2	dobe	100	
2.5	dune	100		5	pudd	97	
5.5	dumb	97		5	dupp	97	
5.5	dean	97		5	doove	97	
7	width	94		8	rediff	94	
7	sudden	94		8	rudds	94	
7	dream	94		8	drabe	94	
10	seized	72		10	lidden	86	
11	bead	69		11	medth	78	
12	butter	58		12	geater	69	
13	seeds	31		13	guzzed	67	
14	tune	17		14	stam	56	
16	retain	8		15	jabe	25	
16	eighth	8		16	naitth	22	
16	beats	8		17	vist	14	
18.5	beaten	6		18	triz	11	
18.5	$beat_2$	6		19.5	$lutt_1$	8	
22.5	jig	3		19.5	lutts	8	
22.5	beast	3		24	chuff	3	
22.5	$beat_1$	3		24	hatten	3	
22.5	streak	3		24	$lutt_2$	3	
22.5	tweak	3		24	reteal	3	
22.5	tub	3		24	struff	3	
28	tree	0		24	tupp	3	
28	steam	0		24	toose	3	
28	team	0		29	teef	0	
28	three	0		29	twif	0	
28	chief	0		29	threff	0	

completely different phoneme category. (For several such items the responses were, in fact, unanimous across the 36 subjects in each probe class.) Another straightforward observation is the highly similar response patterns provided for corresponding real vs. nonsense words with a given probe. (Again, for some of these pairs, the responses were almost or precisely identical, e.g., team vs. teef on both tests.) The Spearman rank-order correlations for these two categories of items were, in fact, .89 (real vs. nonsense for the tough-probe) and .81 (real vs. nonsense for the duck-probe), both significant at the .001 level. Both tables also exhibit a rather consistent gradation of responses between the two extremes already noted, but a statistical test is required to indicate whether any of these differences are significant.

However, proportional data of this kind do not satisfy the assumptions of such standard statistical tests as analysis of variance. In order to remedy this situation, therefore, the following additional steps were next taken:

(1) The 36 subjects in each probe group were randomly assigned to six sub-groups of six members each, in order to provide a within-groups variance for the analysis;

(2) The proportions of YES responses within each such sub-group were then transformed by the angular transformation recommended by Bock & Jones (1968: 72, equation 3.23a), in order to stabilize the within-group variances across the test items.[5]

The transformed data were then subjected to the Tukey A or HSD (honestly significant difference) procedure (Winer 1971: 198), which is a relatively conservative statistical test for significant differences among the mean scores for all possible pairs of items. (A significance level of .05 was used.) The results of this test are summarized at the right in Tables 2A and 2B above by the use of vertical lines to link together those items which are not significantly different. (Any pair of items not contained within the range of a single vertical line thus differ significantly, e.g., the pair tub vs. beaten in Table 2A, but not the pair tub vs. beats.)

Discussion

By and large, the results of this study are consistent with the Bloomfield-Sapir hypothesis, since the scores for most of the allophones of /t/ are not significantly different from those for the target allophone [t^h] of the tough probe, whereas scores for all of the allophones of /d/ and the other phonemes (including the affricates /tʃ/ and /dʒ/) have significantly lower scores. Similarly, for the target [d] sound in the duck-probe, scores for most of the allophones of /d/ do not differ from those for the target, as opposed to allophones of /t/ or any of the remaining

phonemes. However, there are a few "fuzzy allophones" in both cases, as well as some inconsistencies in the treatment of the ambiguous or neutralized [t] and [ɾ] phones; there are also some discrepancies in the treatment of the real vs. the nonsense words which warrant discussion.

In Table 2A, for instance, the real words ordered 1-11 are all contained within the first HSD grouping indicated and clearly seem to belong in the target class; by the same token, the real words 16-29.5 are all together in the last grouping and thus seem to be clearly outside of the target class. Real words 12-15, however, have scores which are significantly different from some members of one or the other extreme groups. The phones involved for these items are [$t^{ʔn}$] (in beaten), [ɾ] (in butter), [t^{-}] (in beat$_2$) and [t͡] (in eighth). The results for the nonsense words with the same tough-probe are virtually identical, except that the additional phone [$^{s}t^{-}$] (in vist) is added to the intermediate or "fuzzy" category. Much the same general response pattern is exhibited by the results in Table 2B for the [d]-target of duck, where [ɾ] (in butter and geater) and [$^{z}d^{-}$] (in seized and guzzed) occupy the "fuzzy" area throughout. Similarly, [d^{z}] (in seeds) and [d^{-}] (in bead) are not clearly classified for the real words, but the situation is different, note, for the nonsense words! These intermediate items thus provide evidence that the Bloomfield-Sapir hypothesis is not strictly correct, as it fails in the cases of these few allophones of /t/ and /d/, which subjects judge to be different from at least some of the other allophones of the phonemes in question. Some sub-phonemic distinctions thus can be heard by phonetically untrained hearers, at least under the conditions imposed by this experiment.[6]

Apart from this important finding, however, the most interesting results involve our subjects' treatment of the phones which appear in those two environments where the /t/ vs. /d/ contrast is effectively neutralized in this language. The first of these concerns the unaspirated and voiceless phone [t], which appears in the general environment #s____V. Judging from the results of the tough-probe, this phone is interpreted as a relatively clear case of /t/, as strongly positive responses are given to both the real (i.e., steam) and nonsense (stam) items which contain this phone. With the duck-probe, however, the results are much more ambiguous: the real word steam provides the expected and clear (in fact, unanimous) negative or "non-/d/" response pattern, but the nonsense word stam produces a positive response rate of over 50%, which has the effect of pushing the [t] phone completely out of the clear "non-/d/" HSD grouping. We might initially have thought that the spelling of the real words might influence our hearers' judgments of this phone, but what accounts for the wide disparity in the interpretation of the nonsense word, such that the [d]-target presents such a radically different picture than the [t^{h}]-target?

The second neutralizing environment is $\acute{V}$____V, where (in this dialect) both /t/ and /d/ are realized as the flap phone [ɾ]. One of the most outstanding and consistent features of the results summarized in both Tables 2A and 2B involves our subjects' treatment of this phone: both buddy and zadey are heard as absolutely clear (100% positive) cases of /d/ under the duck-probe and as quite clear cases of "non-/t/" sounds under the tough-probe (89% and 94% negative, respectively). The flaps in both butter and geater are, however, rated as uncertain "fuzzy" sounds with either probe. Again, we might consider the possible influence of orthography in the case of the real words, but can this explanation apply to the nonsense words, as well? (Perhaps there are "preferred spellings" for the two particular nonsense items we used here which help direct the interpretations in a way parallel to the real items. We did not test for this, but it is a factor that might merit investigation in future studies of this kind.) Another possibility, of course, is that there might be systematic differences in, for example, the duration of the pre-flap vowels - or even in the production of the flap elements themselves - which might serve as marginally distinctive cues for /d/ vs. /t/ (cf. Derwing 1973: 209). These possibilities, too, ought to be investigated further.

3. Summary and Conclusions. In this exploratory study we have taken a quite straightforward approach to the question of the perceptibility of sub-phonemic differences in English. We have selected what we felt were the archetypical allophones of the English /t/ and /d/ phonemes and used these as comparison models or "targets" for a wide range of additional allophones. In most cases the "new" allophones were judged to be repetitions of the targets, at least insofar as our relatively conservative statistical test would indicate. The raw data do exhibit a tantalizing gradation in response level, however, from the target allophones themselves (nearly 100% positive throughout) down to rates as low as 25% for some of the more markedly distinct of the allophones. For some of these, in fact, such as those involving "fronting," a flap articulation, or a nasal release or non-release, these differences were statistically significant even by the conservative (Tukey HSD) test. We can thus conclude that the Bloomfield-Sapir hypothesis has been falsified, at least in its strong form, which states that only contrastive or distinctive phonetic differences can be perceived by phonetically untrained monolingual speakers. It is also important to note that although our test showed only a few of the allophones to yield statistically significant results, we may not conclude from this that the remaining allophones are all interpreted as the same. It is quite possible that a more powerful experimental design aimed at more specific questions might well show that some of the other phonetic distinctions are also readily perceptible to speakers and, as already noted above, even our data exhibit a rank ordering which is highly suggestive in this regard.

More data on this problem are thus strongly to be desired, not only to clarify the main issue of the perceptibility of allophones, but also to shed light on the role of factors (such as orthographic interference) which have not been adequately controlled in the present study. Other phonemes besides English /t/ and /d/ need to be investigated, and the present study, too, would certainly benefit from the application of additional experimental designs and new data-collection techniques. What is surprising, in a way, is that even some of the most obvious and simple experimental approaches to a theoretical problem can often yield results which are both interesting and highly suggestive of further study.

NOTES

1. The acoustic cues for stop segments in the environment of fricatives are quite different from those in the environment of vowels. In the absence of any conventional diacritic to indicate these differences, we have resorted to the ad hoc notational device indicated above to represent the phones involved. We should perhaps also point out that although clear acoustic differences also exist among the aspirated stops as a function of their environment (initial, medial, or final), no attempt at all has been made to represent these latter differences in the phonetic transcriptions used here.

2. Thanks to Dr. John T. Hogan for recording the stimuli for us.

3. Thanks are also due to the following graduate students, who all provided us with valuable assistance in the presentation of the stimuli and/or the scoring of the results: Justin Chen, Bruce Connell, Maureen Dow, Patricia Hunter, Nobuya Itagaki, Richard Jehn, Mary Kolic, Wendy Rollins, and Shaunie Shammass.

4. As indicated in Table 1, the items labeled $beat_1$ and $lutt_1$ in Tables 2AB were recorded with strong aspiration of the final segment, whereas $beat_2$ and $lutt_2$ were recorded with simultaneous alveolar and glottal closure (no release).

5. A subsequent series of Cochran's C tests for homogeneity of variance (cf. Winer 1971: 208) indicated that this condition was satisfied, provided that items with unanimous YES or NO responses were not included in the analysis. There was some evidence of mild heterogeneity when the unanimous items were included, but we have examined the results of both the full and partial analyses and found that they do not differ in any substantial way. For convenience of discussion, therefore, we have decided to report the results from the full analysis here.

6. It should be noted, for example, that at least some of the responses could have been affected by perceptual errors, particularly in the case of the final unreleased allophones, due to the less than ideal listening conditions employed.

REFERENCES

Abramson, A.S., & L. Lisker. 1970. Discriminability along the Voicing Continuum: Cross-Language Tests. Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967. (Prague: Academia), pp569-573.

Bloomfield, L. 1933. Language. New York: Holt, Rinehart & Winston.

Bock, R.D., & L.V. Jones. 1968. The Measurement and Prediction of Judgment and Choice. San Francisco: Holden-Day.

Brown, R. 1958. Words and Things. Glencoe, Ill.: The Free Press.

Denes, P. 1955. Effect of Duration on the Perception of Voicing. Journal of the Acoustical Society of America 27: 761-764.

Derwing, B.L. 1973. Transformational Grammar as a Theory of Language Acquisition. London: Cambridge University Press.

Hogan, J.T., & A.J. Rozsypal. 1980. Evaluation of Vowel Duration as a Cue for the Voicing Distinction in the Following Word-Final Consontant. The Journal of the Acoustical Society of America 67: 1764-1771.

McCasland, G. 1977. English Stops after /s/ and at Medial Word-Boundary. Phonetica 34: 218-228.

Sapir, E. 1949. The Psychological Reality of Phonemes. In D.G. Mandelbaum, ed., Selected Writings of Edward Sapir in Language, Culture and Personality. (Berkeley & Los Angeles: University of California Press), pp46-60.

Shammass, S. 1980. An Experimental Investigation of Segment Duration and Intensity in English Juncture. The University of Alberta, [Unpublished M.Sc. Thesis.]

Vitz, P.C., & B.S. Winkler. 1973. Predicting the Judged "Similarity of Sound" of English Words. Journal of Verbal Learning and Verbal Behavior 12: 373-388.

Winer, B.J. 1971. Statistical Principles in Experimental Design. New York: Mc-Graw Hill.