Evidence for the independent function of intonation contour type, voice quality, and $F0$ range in signaling speaker affect

D. Robert Ladd
Department of Linguistics, University of Edinburgh, Adam Ferguson Building, Edinburgh EH8 9LL, Scotland

Kim E. A. Silverman
MRC Applied Psychology Unit, Cambridge CB2 2EF, England

Frank Tolkmitt, Günther Bergmann, and Klaus R. Scherer
Department of Psychology, University of Giessen, Otto-Behaghel-Str. 10, 6300 Giessen, West Germany[a]

(Received 24 September 1984; accepted for publication 22 March 1985)

In three related experiments, listeners judged the affect conveyed by short recorded utterances in which the voice quality, intonation contour type, and fundamental frequency range had been systematically and independently manipulated. (Contour and range were manipulated using digital resynthesis of naturally spoken utterances.) Analyses of variance of the results showed that range and contour, and less clearly range and voice quality, had independent effects on the way the utterances were judged. The results also strongly suggest that these differences are independent of effects due to interspeaker differences and to differences of verbal content. Finally, analysis of the results suggests that differences of $F0$ range, as is commonly assumed, have continuous rather than categorical effects on affective judgments.

PACS numbers: 43.71.Es

INTRODUCTION

In an earlier study (Scherer *et al.*, 1984), we demonstrated that there may be different types of vocal cues to speaker affect, which contribute to the signaling of speaker states and intentions in essentially different ways. By analyzing listener judgments of the affective force of semantically neutral utterances taken from a corpus of natural speech, we provided evidence for a distinction between two types of cues: Those that can be treated as *continuous acoustic variables* whose variation is more or less directly correlated with variations in the affective message [e.g., the higher the average fundamental frequency ($F0$) of an utterance, the more generally aroused the speaker is judged to be], and those that are organized into *linguistic* (and perhaps perceptual) *categories* whose interpretation depends on interaction with other cues in the context, including the verbal content (e.g., a falling intonation contour is judged neutral with a WH question, but aggressive or challenging with a yes/no question). Given the distinction between these two types, we further noted that the continuous variables appear to reflect states of the speaker related to physiological arousal, while the more linguistic variables tend to signal speaker attitudes with a greater cognitive or attitudinal component, such as friendliness or reproach.

The general goal of the present study was to improve on our earlier methodology in order to develop the notion that vocal cues to affect are of different types and have at least partially independent functions. For the present study, we used digital resynthesis of naturally spoken utterances to create sets of stimuli in which certain acoustic variables—specifically, intonation contour type and $F0$ range—were carefully controlled and systematically varied. On the basis of our earlier results, we chose the following hypotheses for testing:

(1) That intonation contour type, overall $F0$ range, and voice quality[1] have independent effects on affective judgments. (The independence of these three variables is assumed by many linguistic descriptions of intonation, explicitly so by Crystal, 1969; cf. also Laver, 1980.)

(2) That overall range and voice quality reflect states of arousal, while differences of contour type signal differences of "cognitive" attitude (cf. the results of a number of studies on vocal cues to emotional stress or arousal, such as Hecker *et al.*, 1968; Scherer, 1979, 1981a,b; Williams and Stevens, 1981; cf. also the descriptions of contour meanings in terms of "attitudes" by linguists such as Pike, 1945, and O'Connor and Arnold, 1961).

(3) That overall range functions as a continuous variable, so that changes in range are directly correlated with changes in the intensity of affective judgments (cf. the distinction made by Bolinger, 1961, between "gradient" and "all-or-none" phenomena in intonation).

We also explored the extent to which verbal content and speaker identity may influence the signaling function of the three main types of acoustic cues under study.[2]

In the first experiment we attempted to assess the relative contribution of intonation contour type, $F0$ range, and voice quality on listeners' affective judgments of three sentences spoken by a single speaker. The second experiment was intended as a partial replication, using three different speakers, in order to determine the generality of the results of the first. The goal of the third experiment was to test whether pitch range variation has continuous or categorical effects on affective judgments.

We did not attempt any direct test of the hypothesis that

[a] Reprint requests should be sent to this address.

 0001-4966/85/080435-10$00.80

contour differences are categorical rather than continuous, since the methodological and theoretical problems in any categorical perception experiment on intonation would be serious. Given the context dependence of intonational meaning, it would be difficult to assign labels to hypothesized contour types (except perhaps in a few cases like "assertion" versus "question"), which would make it difficult to set up a labeling task. More importantly, in the absence of a generally accepted phonological taxonomy of intonation, the choice of a stimulus continuum would force the experimenter to make many decisions on the detail of the contour shapes that could easily invalidate the results. Our view is that the phonological system of intonation needs to be established (and this paper is intended in part as a contribution toward that goal) before tests of categorical perception are likely to be useful (for further discussion see Ladd, 1980, Chap. 5).

I. EXPERIMENT I

In this part of the study we tested the hypotheses that the three acoustic variables CONTOUR, RANGE, and VOICE QUALITY have independent effects in signaling speaker affect, and that RANGE and VOICE QUALITY are more related to speaker arousal than is CONTOUR. We also studied the influence of TEXT on the effects of the three acoustic parameters.

A. METHOD

1. Design

A factorial 2×2×2×3 design was used with two levels of RANGE (narrow, wide), two levels of VOICE QUALITY (normal, harsh), two different CONTOUR types ("uptrend," "downtrend"), and three sentences with different TEXT. The variables RANGE and CONTOUR were manipulated through digital resynthesis; modifications of VOICE QUALITY were produced by the speaker who spoke the three TEXT types.

2. Speech materials

Three sentences were constructed for use in this and subsequent experiments. They are given here with English glosses and with code letters for ease of reference. Major accents are indicated by italics.

(MD) Mit den *an*deren hat es nur *ein*mal geklappt ('With the others it only worked once')

(AS) Aber *schrift* lich habe ich das *nicht* bekommen ('But I didn't get that in writing')

(DB) Diese *Büch*er muss man aber zu*rück*schicken ('But these books have to be sent back')

The most important formal criteria in the choice of these sentences were that they should have very similar patterns of accent and rhythm (in order to make it easier to compare intonation patterns from one to the other), and that the two major accents should occur on syllables with identical vowel height (in order to minimize "intrinsic pitch" effects; cf. Lehiste, 1970, pp. 68–71). The most important semantic/pragmatic criteria were that the sentences should sound natural and colloquial, and that they should be consistent with a variety of speaker attitudes (surprise, irritation, etc.) depending on context, intonation, and the like.

The two intonation contour types studied are illustrated in Fig. 1. The figure shows the actual contours used on text AS in low range, but for purposes of the experiment it was assumed that these contours represent types in a phonological system. Following recent work on the phonology of intonation ('t Hart and Collier, 1975; Bruce and Gårding, 1978; Pierrehumbert, 1981; Ladd, 1983), we took the *F*0 lev-

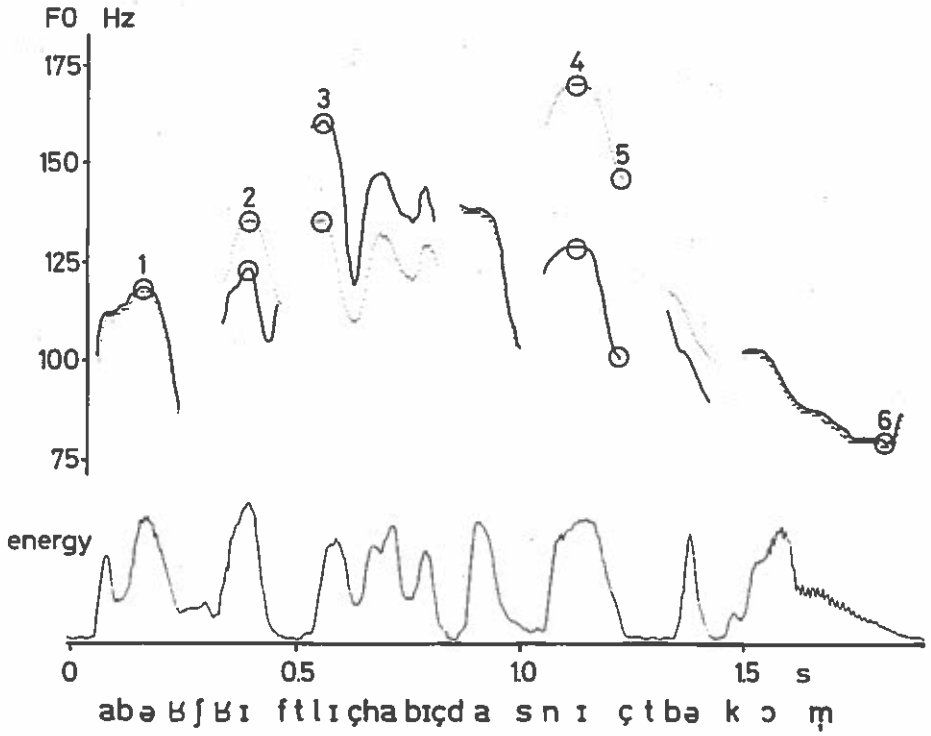

FIG. 1. Examples of the two contour types used in experiments I and II. The top half shows *F*0 tracings for the downtrend (solid line) and uptrend (dashed line) contours, with the anchor points circled. The lower half shows the signal energy, together with a phonetic transcription of the utterance "Aber schriftlich habe ich das nicht bekommen." For further discussion see text, Secs. I A 2 and I A 3.

els at the beginning and the end of the contour and $F0$ movements at the two accents to be the systematically important features of these contour types. Specifically, both types begin in the middle of the speaker's range and end at the bottom of it. Both show a sharp rise on the first accented word and a fall on the second. One contour type gradually rises from the first accent to the second, while the other gradually falls; the one that gradually rises reaches the first accent peak in the first syllable of the accented word, while the one that gradually falls does not reach the peak of the first accent until the accented word's second syllable.[8]

In the terms proposed by Ladd (1983), both types are high ... high/low (i.e., a high accent followed by a high–low or falling accent), but one has a raised second accent peak, while the other has a delayed first peak and a downstepped second peak. For ease of reference we will designate these two contour types "uptrend" (gradual rise; raised second peak) and "downtrend" (gradual fall; downstepped second peak).

Phonetically, the contours were modeled as a sequence of $F0$ values at six "anchor points" related to the phonologically distinctive elements of the contour (one anchor point each at the beginning and the end, and two for each accent movement). The procedure for assigning $F0$ values to these anchor points is described in the Appendix. $F0$ transitions between the anchor points, and contour perturbations due to segmental effects, were assumed to be in the realm of low-level phonetic detail, and were generated by prescribed procedures outlined in the next sections. These procedures, and the phonological approach to data reduction that underlies them, made it possible to create extremely natural-sounding stimuli while at the same time tightly controlling relevant variables in a theoretically well-motivated way.

The sentences were recorded by a male native speaker of German in his thirties. Coached by one of the authors (DRL), the speaker produced the utterances with several different intonation contours (including the two types finally used in the study), and with two different "manners of speaking": A "normal, relaxed, friendly" voice and an "annoyed, irritated, angry" voice. (The main differences between the two speaking styles were that the latter had a higher $F0$ range and a harsh, pressed voice quality.) Before further work on stimulus preparation was attempted, a large number of these utterances were analyzed acoustically and studied for information about the speaker's $F0$ characteristics (e.g., speaker's $F0$ "floor"; see Appendix).

3. Stimulus preparation

Stimuli were prepared by means of digital resynthesis. Contour specifications of the sort described in the preceding section were used to create sets of new $F0$ contours, which then replaced the contours of the originally spoken utterances. This procedure made it possible to produce all the variations of RANGE and CONTOUR by resynthesis from a single "source utterance." That is, we were able to choose a single token of each TEXT type spoken in each VOICE QUALITY type–a total of six "source utterances"—and from them generate all 24 stimuli. This eliminated a good deal of potential variability in rhythm, duration, precision of articulation, etc., which would have been present in the stimuli if we had used many different tokens of the same TEXT type as stimuli.

As stated above, contours were modeled on the basis of six "anchor points" (see Fig. 1). For each of the two CONTOUR types, average $F0$ at each of these points was calculated, averaging across all of the speaker's productions of all utterances with "normal" voice quality. This yielded single standardized representations of the two CONTOUR types. The two "anchor points" in each of the two major accents were then adjusted up or down for each TEXT type to compensate for intrinsic pitch effects, in such a way that /ɪ/ was 0.5 semitone lower than /y/ and 1.0 semitone higher than /a/. Prehead and endpoint $F0$ were not adjusted and were thus the same for all three TEXT types.

$F0$ values for the wide RANGE stimuli were generated from the narrow RANGE values in accordance with the formula given in the Appendix. This formula allows the overall range of an $F0$ contour to be determined by the value of a single range parameter R. This parameter had a value of 1.0 for the low RANGE stimuli, and 1.7 for those with a high RANGE.

The four new $F0$ contours (uptrend and downtrend in wide and narrow RANGE) were inserted into each of the six source utterances. Each source utterance was digitized (at 16 kHz) and analyzed by linear prediction (using 29 filter coefficients, 98% pre-emphasis, and 256-point windows overlapped by 64 points). With the aid of interactive computer graphics (Silverman, 1985), the anchor points in the contour were aligned with the phonetic structure of each source utterance, and a contour interpolated between them using a quadratic spline function (Hirst, 1983). Segmentally related perturbations of $F0$, such as those accompanying obstruents (cf. e.g., Ohde, 1984), were superimposed onto the contours "by hand," following the perturbations in the original utterance as closely as possible.[3] The resulting contours were then substituted into the utterance files and new utterances were resynthesized.

4. Rating form

On the basis of several small pilot studies, we constructed two separate rating forms for arousal-related states and more cognitive attitudes. The arousal form consisted of five bipolar 8-point scales: gelassen/erregt (relaxed/aroused), offen/unaufrichtig (open/deceitful), verärgert/zufrieden (annoyed/content), unsicher/arrogant (insecure/arrogant), gleichgültig/engagiert (indifferent/involved). The cognitive-attitude form consisted of five unipolar 8-point scales: Nachdruck (emphasis), Entgegenkommen (cooperativeness), Widerspruch (contradiction), Überraschung (surprise), Vorwurf (reproach).[4]

The scales were chosen in such a way as to represent speaker attitudes and affect states which are maximally different from each other and cannot be easily subsumed under more general scales or factors such as positive/negative or active/passive. In choosing the scales we used some of the data obtained in earlier studies. Inspection of the intercorrelations between the scales (see Table I, reporting the mean intercorrelations of the scales over three studies) show that

TABLE I. Mean intercorrelations of the scales over the three experiments. Only intercorrelations of the eight scales common to the three experiments are reported. Correlations of scales were also calculated for the three experiments separately, but appeared to be rather similar and are not shown.

	Aroused	Involved	Contradicting	Arrogant	Annoyed	Cooperative	Reproachful
Emphatic	0.465	0.621	0.456	0.288	0.380	−0.018	0.394
Aroused		0.553	0.482	0.251	0.696	−0.219	0.458
Involved			0.464	0.236	0.378	0.056	0.341
Contradicting				0.291	0.450	−0.059	0.467
Arrogant					0.345	−0.099	0.355
Annoyed						−0.312	0.503
Cooperative							−0.253

this aim has been achieved rather successfully. Out of all the intercorrelations only four reach a level of 0.5 or higher, explaining more than 25% of the joint variance. All of the other intercorrelations are lower and many much lower, which shows that there is very little joint variance in the scales. Consequently, we felt that factor analysis or similar data reduction methods would have obscured a large amount of independent variance.

5. Rating sessions

Because of the amount of time required for the task, and because of the two separate rating forms, the subjects rated the stimuli in two separate sessions about one week apart. Half the subjects judged the stimuli on the arousal form first, while the other half received the cognitive-attitude form first. Each session began with a written explanation of the task and three practice stimuli. The instructions were written in such a way that the subjects' attention was not drawn to the distinction between cognitive attitudes and states of arousal until the beginning of the second session (i.e., "Last session you did x, this time what you'll do is y").

Subjects heard the stimuli over loudspeakers in groups of three to six subjects. (Loudspeakers were used instead of headphones in order to minimize the subjects' awareness of the unnatural-sounding aspects of resynthesized speech; in preliminary trials, most subjects were unaware that they were listening to anything but natural recordings when the stimuli were presented in this way.) Each stimulus was heard once. There was no time limit for responding; the experimenter stopped the tape between each stimulus utterance and restarted it when the subjects were ready. The sessions generally lasted about 40 min.

There were 23 paid subjects, mostly students at the University of Giessen and all native speakers of German, ranging in age from 19 to 27. An approximately equal number of male and female subjects took part.

B. Results and discussion

In line with the factorial nature of the design, analysis of variance (ANOVA) procedures were used to analyze the data. In what follows, only those effects reaching F values with $p < 0.01$ are reported, in order to concentrate on strong effects. For greater clarity, Table II shows only F and d values. The latter are an expression of effect size in standard-deviation units, calculated according to the following formula (Cohen, 1977; R. Rosenthal, 1980):

$$d = 2\sqrt{F_{\text{effect}}}/\sqrt{df_{\text{error}}}.$$

These d values are a helpful supplement to the more traditional F statistics, for two reasons. First, d expresses a difference relative to spread of the underlying distributions, thus making it easier to intuitively grasp the extent to which our experimental manipulations have affected subjects' judgments. Second, d values are independent of the number of degrees of freedom. This allows us to directly compare effect sizes across different conditions and experiments.

Cohen (1977) offers guidelines for interpretation of d values by suggesting that $d = 0.2$ is a "small" effect while

TABLE II. Results of experiment I. Only effects that reach a level of significance better than 0.01 are reported.

	Main effects											
	Contour				Range				Voice quality			
Variables	F	d	down	up	F	d	wide	narrow	F	d	normal	harsh
Relaxed/aroused	47.7	2.94	4.67	5.34	140.7	5.06	5.84	4.18	182.2	5.76	4.25	5.76
Open/deceitful	...	...	...	...	...	...	...	...	11.8	1.46	3.35	3.72
Annoyed/content	22.0	2.00	3.72	3.29	36.5	2.58	3.12	3.90	85.1	3.93	4.08	2.94
Insecure/arrogant	...	...	...	...	...	...	...	...	19.5	1.88	4.33	5.00
Indifferent/involved	12.9	1.53	5.14	5.47	72.9	3.64	5.84	4.77	70.0	3.57	4.74	5.87
Emphasis	11.7	1.46	4.53	4.90	55.7	3.18	5.21	4.23	110.6	4.48	3.79	6.06
Cooperativeness	15.9	1.70	2.74	2.44	...	...	...	...	40.8	2.72	3.07	2.11
Contradiction	12.0	1.48	3.97	4.31	18.5	1.83	4.42	3.87	21.4	1.97	3.67	4.61
Surprise	...	...	...	...	15.1	1.66	2.96	2.49	...	...	...	...
Reproach	...	...	...	...	13.0	1.54	5.22	4.63	31.5	2.39	4.30	5.47

$d = 0.8$ could be considered "large." These criteria are somewhat liberal—an effect size of 0.8 only explains 13.8% of the variance. These conventions at least underline the size and clarity of the differences we report below. Means and effect size values are not reported for effects involving SPEAKER and TEXT, since, as discussed below, these effects are largely irrelevant to the questions investigated here.

The variable TEXT—the verbal content of the three different sentences—has a significant effect on nine of the ten judgment scales, although the size of these effects is generally not very large (d values around 1; only for the scales emphasis and contradiction are d values of 1.72 and 1.54 achieved). Moreover, there are also quite a number of interactions between TEXT and either CONTOUR or VOICE QUALITY. This indicates that, even in the absence of obviously "loaded" verbal content, the affective interpretation of an utterance is a joint effect of the text and the various acoustic variables. This is consistent with other findings (e.g., Scuffil, 1982, experiments XIII, XIV, XVII; Scherer *et al.*, 1984), and fits with the assumption, now widespread in the literature on linguistic pragmatics, that the interpretation of an utterance is the result of an active process of inference based on all the information—verbal, paralinguistic, and contextual—available to the listener. (For a discussion of this assumption as it applies to the study of intonation and affect, see Ladd, 1980, Chap. 6; Ladd *et al.*, 1985; cf. also footnote 5.) However, since the experimental techniques used here are not very appropriate for investigating pragmatic differences based on differences of verbal content, we will not discuss the effects of TEXT further at this point.

The major result of the experiment is that, as predicted, RANGE and VOICE QUALITY had a strong effect on judges' inference of speaker arousal: Harsh voice quality and wide range are seen as signals of arousal, annoyance, and involvement. The d values show that the largest effects in the experiment were those of these two acoustic variables on the arousal judgments (more than five standard deviations, explaining about 93% of the variance). Yet there are also significant effects of VOICE QUALITY and RANGE on more cognitive attitudes (emphasis, contradiction, and reproach), though these effects are all weaker than for arousal states. This may show that there is an arousal component inherent in those cognitive attitudes; alternatively, it may suggest the difficulty of mapping psychological categories of emotion directly onto acoustic cues. We will return to this question below.

As for the difference between VOICE QUALITY and RANGE, the results suggest that RANGE may be more strictly related to arousal, while VOICE QUALITY has a component of positive–negative valence (the speaker's positive or negative evaluation of the interlocutor or semantic content) as well. Specifically, harsh voice quality leads to the attribution of negative states (less cooperative, more deceitful, more arrogant) that are apparently unrelated to RANGE. Note also that surprise, which is relatively neutral from the point of view of valence, seems to be related only to RANGE and not VOICE QUALITY.

We had predicted that CONTOUR should affect the rating of cognitive attitudes rather than arousal. On three of the five cognitive-attitude scales, CONTOUR does have a significant effect: Uptrend signals greater emphasis, stronger contradiction, and less cooperativeness. However, CONTOUR also has effects on the arousal scales that are even stronger than those on the cognitive attitude scales. Uptrending intonation, like wide range and harsh voice quality, is interpreted as signaling arousal, annoyance, and involvement. Once again, an obvious conclusion is that the distinction between arousal and cognitive attitudes is not directly reflected in the acoustic cues; however, other explanations are possible.[6]

The only clusters of significant interactions were in TEXT by CONTOUR and TEXT by VOICE QUALITY. Only two of the remaining 90 interaction effects were statistically significant. In particular, there were no RANGE by CONTOUR interactions. This suggests that the acoustic variables studied here—CONTOUR, RANGE, and VOICE QUALITY—may indeed operate independently.

II. EXPERIMENT II

The purpose of this experiment was to replicate some of the findings of experiment I, and to assess their generality across speakers. The VOICE QUALITY manipulation, which had been done by speaker simulation in experiment I, was dropped from experiment II, since we could not be confident that the different speakers would produce comparable changes of voice quality. This means that the replication was restricted to the resynthesized variables RANGE and CONTOUR. However, the use of more than one speaker made it possible to test for interaction effects between SPEAKER and the acoustic variables. A further difference between experiment I and II was the improvement of the rating procedure, described in this section.

A. Method

1. Design

A $2 \times 2 \times 3 \times 3$ factorial design was used, with two levels of RANGE (narrow, wide), two types of CONTOUR (uptrend, downtrend), three different TEXTs (the same three used in experiment I), and three different speakers (FT, GB, and CL).

2. Speech materials

The same TEXT and CONTOUR types were studied as in experiment I (but see footnote 8). The speakers (two of whom, FT and GB, are co-authors of this report) were again coached by DRL to produce the intonation contours under study. Also as in experiment I, a considerable number of utterances were analyzed acoustically in order to arrive at a rough picture of each speaker's $F0$ characteristics.

3. Stimulus preparation

As in experiment I, one utterance of each TEXT by each speaker was selected as a source utterance for resynthesizing all the stimuli. Standardized values for the anchor points in the contours were derived in the same way as in experiment I. Narrow and wide ranges were generated by using R values of 0.75 and 1.25 in the range formula. The contours derived

in this way were resynthesized using the same methods as in experiment I.

4. Rating form

Instead of using two separate rating forms, one for states of arousal and one for cognitive attitudes, as we had done in experiment I, a single rating form was constructed that included both types. A total of eight 8-point unipolar scales were used, four for states of arousal: erregt (aroused), ärgerlich (annoyed), arrogant (arrogant), and engagiert (involved), and four for cognitive attitudes: nachdrücklich (emphatic), entgegenkommend (cooperative), widersprechend (contradicting), and vorwurfsvoll (reproachful). It should be noted that the scale *cooperative* is negatively poled with respect to the other seven scales.

5. Rating sessions

Because of the new rating form, only a single rating session was necessary. In all other respects the rating procedure was unchanged from experiment I. The subjects for this experiment were 17 students (an approximately equal number of males and females), ranging in age from 18 to 25, who had not taken part in experiment I.

B. Results and discussion

As in experiment I, ANOVA was used to analyze the data. Again, only F values with a $p < 0.01$ are reported. F values, effect-size parameters (d values), and the means for the main effects of CONTOUR and RANGE are shown in Table III.

First, we consider SPEAKER and TEXT effects not reported in the table. As might be expected, there are strong main effects for SPEAKER on all judgment scales but one (arrogant). That is, independent of the various experimental manipulations, different speakers are heard as differing on the various scales used in the rating form. This does not in itself necessarily limit the generality of the findings concerning the effects of the acoustic variables; that would be true only if there were strong interactions between SPEAKER and the vocal variables. But there are few such interactions in the present data,[7] which indicates that any main effects for acoustic variables are likely to be generalizable over different speakers.

As in experiment I, we find strong main effects for TEXT. Interactions between TEXT and the acoustic variables are even smaller than in experiment I, reaching significance only in the one four-way interaction (see footnote 7). None of the interactions between TEXT and CONTOUR that were found in experiment I were replicated. There are, however, some interactions between SPEAKER and TEXT, perhaps reflecting individual differences in the actual tokens chosen as the source utterances for resynthesis. The interactions with SPEAKER were all among the smallest effects found in the experiment, being less than half the size of the main effects for SPEAKER and RANGE on the same scales. The absence of sizeable interactions between SPEAKER or TEXT and either of the acoustic variables means that we can treat any replicated effects for the acoustic variables as fairly general effects.

The results obtained for RANGE in experiment I are in general clearly replicated. Wider range is heard as a signal of the speaker's being more aroused, annoyed, involved, emphatic, contradicting, and reproachful. As in the previous experiment, the largest effect of RANGE was on the judgments of arousal. Nevertheless, the difficulty of distinguishing acoustic correlates of arousal from those of various cognitive attitudes is even greater than in experiment I, since in experiment II there is a less decisive difference in effect size between the two. At the same time, however, it is tempting to look for a single dimension, such as arousal, that might be common to all these scales; in any case, it cannot simply be asserted that the subjects responded to wider range with more extreme judgments on all the scales, since wide RANGE is associated with *lower* cooperativeness.

For CONTOUR, there is only one significant main effect, with uptrending intonation being judged as more emphatic. The fact that we find fewer such main effects in this experiment may reflect the changes in the rating procedure (viz., running a single rating session and not separating the two sets of rating scales). To the extent that CONTOUR involves categorical linguistic distinctions rather than continuous variables, the failure to replicate some of the findings of experiment I may also reflect the general inappropriateness of rating scales as a means of expressing the pragmatic effects of different contour choices.[5]

TABLE III. Results of experiment II. Only effects that reach a level of significance better than 0.01 are reported.

	Main effects							
	Contour				Range			
Variables	F	d	down	up	F	d	wide	narrow
Aroused	...	...	...	...	49.2	3.51	2.80	1.74
Annoyed	...	...	...	...	25.3	2.52	2.67	2.03
Arrogant	...	...	...	...	...	...	...	...
Involved	...	...	...	...	20.3	2.25	3.11	1.98
Emphatic	13.9	1.87	3.64	3.96	13.2	1.82	4.25	3.35
Cooperative	...	...	...	...	...	...	...	...
Contradicting	...	...	...	...	21.32	2.31	3.33	2.36
Reproachful	...	...	...	...	31.39	2.80	3.57	2.89

As in experiment I, we do not find significant interaction effects between the two acoustic variables that were experimentally manipulated here. This again underlines the independence of these factors and their additive effect.

III. EXPERIMENT III

Given the clear correlation of RANGE with judgments of speaker arousal in the first two experiments, our goal in the third experiment was to investigate in more detail the effects of changes in RANGE. In particular, we wanted to see whether, as is commonly assumed, gradual increases in RANGE lead to gradual changes in affective judgments, or whether there are categorical effects as well such that one might distinguish discrete levels (e.g., "normal range" versus "raised range").

A. Method

1. Design

To reduce the number of stimuli to be judged, we concentrated only on range, using a $2\times2\times5$ design with two TEXT types, two SPEAKERs, and a continuum of five levels of RANGE. In comparison with experiment II, we eliminated one TEXT type (MD, eliminated because of technical difficulties in obtaining natural-sounding resynthesized versions), one SPEAKER (eliminating CL, the one with the relatively narrow overall range), and the CONTOUR variable (only the "downtrend" contour was used).

2. Speech material and stimulus preparation

The recorded utterances used in experiment II were used again as the basis for the resynthesized stimuli in this experiment. The $F0$ values of the downtrend contours on texts AS and DB as spoken by speakers FT and GB were transformed into five RANGE settings, using R values of 0.6, 0.8, 1.0, 1.2, and 1.4 in the formula introduced in the Appendix. These $F0$ contours were resynthesized onto the source utterances, yielding a sequence of five stimuli with continuously increasing ranges for each combination of SPEAKER and TEXT. As in experiment II, these were presented in different random orders.

3. Rating form and rating sessions

The rating procedures were identical to those in experiment II. The subjects were 25 students (an approximately equal number of males and females) ranging in age from 18 to 27, who had not taken part in the earlier experiments.

B. Results and discussion

Results are shown in Table IV. As in the first two experiments, we find main effects for SPEAKER and TEXT, although they are much less pervasive than before (only two effects for SPEAKER and three effects for TEXT reach significance). It is possible that the number and strength of such effects was reduced by eliminating the specific text MD and speaker CL—i.e., that the SPEAKER and TEXT main effects were, at least to some extent, explainable artifacts due to CL's narrow range and to the technical problems in resynthesizing text MD. None of the interactions between SPEAKER and TEXT that were found in experiment II were replicated.

TABLE IV. Results of experiment III. Only effects that reach a level of significance better than 0.01 are reported. Means of the five range levels are presented graphically in Fig. 2.

	Main effects		Interactions	
	Range		Text × range	
Variables	F	d	F	d
Aroused	44.3	1.36	...	...
Annoyed	19.8	0.91	4.07	0.82
Arrogant	17.4	0.85	5.08	0.92
Involved	16.8	0.84	...	...
Emphatic	11.9	0.71	...	...
Cooperative	9.0	0.61	6.60	1.05
Contradicting	8.4	0.59	...	...
Reproachful	34.7	1.20	6.10	1.01

Whatever the explanation for the SPEAKER and TEXT effects, we find again a clear replication of the strong effects of RANGE on all the affective judgments. Once again, as in both of the previous experiments, the largest effect of RANGE was on the arousal judgments. To investi-

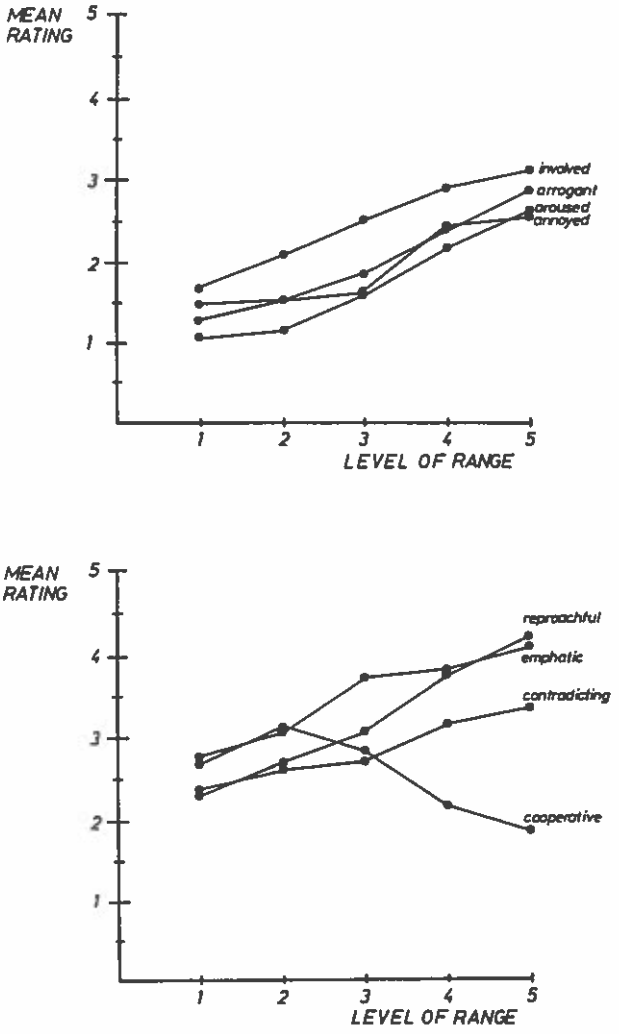

FIG. 2. Ratings as a function of the five range levels used in experiment III.

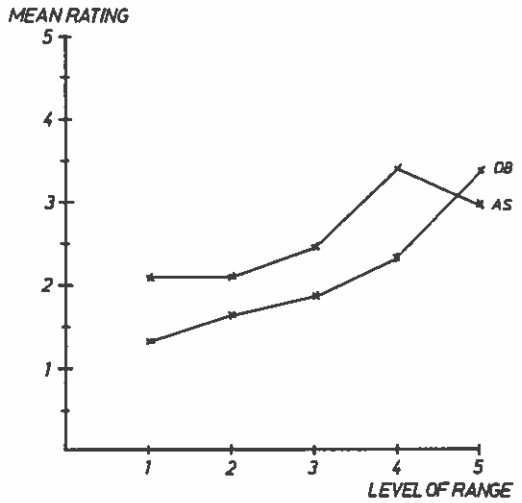

FIG. 3. Means of the interaction between the two TEXT types and the five range levels averaged over the scales annoyed, arrogant, and reproachful (DB = Diese Bücher muss man aber zurückschicken; AS = Aber schriftlich habe ich das nicht bekommen). The crossover at the highest range was present on each of the three scales that are averaged together here.

gate the question of primary interest for this experiment, namely whether the relation between RANGE and affective judgments is categorical or continuous, we plotted the judgment means for the five different levels of range (Fig. 2). Figure 2 suggests strongly that a continuous and not a categorical relation must be assumed. This impression was confirmed by partitioning the mean square of the variance in the ANOVA and testing for trends. The linear trend for all eight scales was highly significant, with the smallest linear trend being obtained for the "contradicting" scale reading $p = 0.0014$. None of the higher-order nonlinear trends was significant for six of the scales. The quadratic component reached $p < 0.01$ for cooperative, and the fourth-order component $p < 0.05$ for annoyed (see Fig. 2). Given the large number of higher-order component tests the latter two effects could well be due to chance.

A new and unexpected result was the finding of significant interactions between TEXT and RANGE for the scales arrogant, annoyed, cooperative, and reproachful. These means, averaged over the scales arrogant, annoyed, and reproachful, are plotted in Fig. 3 (the scale cooperative is excluded, since it has a negative slope). It can be seen that the interaction effect is due to a sudden decrease in the attribution of these attitudes for the text AS as range rises to its highest level. The explanation for this interaction is not clear; it could be largely methodological (e.g., inadequacies in the range formula might create increasingly unnatural-sounding utterances as the value of R increases), or it could be of direct relevance to the problem of continuous versus categorical (e.g., affective judgments may be affected categorically as the $F0$ range reaches a ceiling for a given voice). Further experimentation would be necessary to address this question.

IV. CONCLUSION

By using digital resynthesis, we have been able to manipulate three different acoustic variables independently of one another in a theoretically motivated way. The factorial design resulting from this procedure allowed us not only to assess the effect of changes in those acoustic variables, but also to investigate their interactions. As might have been expected from earlier correlational studies, our results clearly show the effect of range, voice quality, and contour type on judgments of speaker affect. A more significant aspect of our findings is the absence of interaction effects, suggesting that the acoustic variables in question function largely independently of one another.

The second important finding of the study is that, despite the presence of significant SPEAKER and TEXT effects, there are virtually no interactions between these factors and the acoustic variables manipulated. This gives rise to the hope that the effects found here can be generalized over a wide range of speakers and utterances.

The third important finding of the study is that there are pervasive effects of RANGE on affective judgments, particularly on attributions of arousal. Moreover, these effects appear to be a continuous function of changes in range. The possibility of categorical effects at extreme ranges, suggested by some of the data from experiment III, needs further study.

Results for CONTOUR and VOICE QUALITY are less conclusive. The problems interpreting the results for CONTOUR are probably due in part to inadequate understanding of the linguistic structure of intonation, and to the likelihood that rating-scale methods are inappropriate tools for investigating intonational nuances. In the case of VOICE QUALITY, the biggest problem is our inability to manipulate voice quality variables as we have manipulated $F0$ variables. This inability is due in part to inadequate understanding of the acoustic parameters involved in signaling voice quality distinctions, and in part to technical difficulties, at the present time, in modifying the relevant parameters (e.g., glottal wave shape) by means of pitch-asynchronous digital resynthesis.

Perhaps the most important weakness of this study, and indeed of the whole general area of research, is the absence of a widely accepted taxonomy of emotion and attitude. Not only does this make it difficult to state hypotheses and predictions clearly, but (on a more practical level) it makes it difficult to select appropriate labels in designing rating forms. At the same time, however, we have been able to avoid some of the difficulties caused by these theoretical inadequacies, because our goal was not so much to identify the acoustic cues to such-and-such a (hypothesized) emotional state, but to study the *kinds* of signaling functions of the various acoustic cues. This difference in emphasis from certain earlier studies (e.g., Uldall, 1960; Williams and Stevens, 1972; Streeter *et al.*, 1983) means that our results were more robust, in particular in surviving the substantial modifications in rating form from experiment I to II.

In summary, we feel that our study has demonstrated both the usefulness of resynthesis as an experimental technique, and the possibility of associating distinct functions with different putative paralinguistic features such as range, voice quality, and contour type. The results reported here seem to warrant further investigation along these lines.

ACKNOWLEDGMENTS

The research reported here was supported by the Deutsche Forschungsgemeinschaft. We thank Arvid Kappas and Ludwig Lammer for their assistance.

APPENDIX

$F0$ range was manipulated according to a simple model, which is based on recent published work (notably Bruce, 1982; Menn and Boyce, 1982; Liberman and Pierrehumbert, 1984) and on preliminary studies of our own. The central assumption of this model is that $F0$ targets are scaled relative to an idealized "floor" or bottom-of-speaking-range, which is a speaker constant. This floor serves as the x axis, so to speak, for time plots of contours. The y axis is assumed to be logarithmic.

Contours can thus be represented as a sequence of values of the form

$$\log F0/Fr,$$

where Fr is the speaker floor. The value of Fr for any given speaker can be approximated (as in the present study) by taking the average value of the low endpoints of contours ending with an ordinary declarative fall (cf. Menn and Boyce, 1982; Liberman and Pierrehumbert, 1984).

Given such a representation, the relation between one contour and the same contour spoken with a wider or narrower overall range can be stated by multiplying each of the sequence of values by a single range factor R. That is

$$\log [F0\,(\text{range } 2)/Fr] = R \log [F0\,(\text{range } 1)/Fr],$$

or

$$F0\,(\text{range } 2) = Fr\,[F0\,(\text{range } 1)/Fr]^R.$$

The latter form of the equation was used to modify the $F0$ range in the experimental stimuli. For this purpose the logarithm base is irrelevant because the logarithm terms cancel out.

This formula is an empirical model, based on the data provided by the works just cited, and though it is undoubtedly oversimplified (for example, it somewhat exaggerates the effect of range expansion on target points that are already fairly high), it produced acceptable-sounding output for use in the present study.

[1]The acoustic correlates of differences in "voice quality," and an explicit definition of that term, are not entirely clear, but they certainly include such things as breathiness, $F0$ perturbation, differences in formant positions, and in gross energy distribution in the spectrum. The details are not really at issue in the present study, but it should be noted that most of the acoustic correlates of the two different voice qualities in experiment I were captured in the linear predictive filter coefficients, and survived the manipulations of $F0$, so that the two voice qualities were clearly distinct in the results.

[2]In what follows we will use the terms CONTOUR, RANGE, VOICE QUALITY, TEXT, and SPEAKER to refer to the independent variables in the various experiments. The cover term "acoustic variables" will be used to refer to CONTOUR, RANGE, and VOICE QUALITY, as distinct from the other two.

[3]These perturbations in the $F0$ contours were predominantly V-shaped dips, about 15 Hz deep and 70 ms wide, centered over the voiced consonants. They improved the naturalness of the resynthesized speech (by removing the "mechanical" sound of long stretches of otherwise smoothly varying $F0$ values) and increased the intelligibility of the consonants and some of the unstressed vowels.

[4]Naturally, the English glosses are only approximate; attitudes and emotions are often exceedingly difficult to translate with a single term.

[5]For example, the uptrending contour with text DB might indicate "I'm surprised you hadn't realized that these books have to be sent back." Such an exact nuance is obviously very difficult to capture on quantifiable rating scales.

[6]One possible explanation has to do with the fact that two separate rating sessions were held for the two types of affective messages. It is possible that subjects felt compelled to use the clearly perceptible manipulation of CONTOUR type in arriving at their judgments on the arousal related scales, even though in a more natural situation CONTOUR might play little role. A second possible explanation has to do with the assumption that RANGE modification affects the entire contour. It is possible that in short utterances such as these perceptions of RANGE are most reliably based only on the height of the nuclear accent (i.e., in the present case the second accent). Since in the stimuli used here the height of the nuclear accents forms a continuum (narrow- and wide-range downtrend, narrow- and wide-range uptrend), there may be some RANGE effects implicit in the distinctions of CONTOUR type.

[7]The one exception is the existence of two moderately sized interactions between SPEAKER and RANGE. The d values indicate that these effects both accounted for differences of less than 0.85 deviations, smaller than any other main effects in either this or the previous experiment. Closer inspection of the data suggested that these interactions were due to the fact that one of the speakers (CL) had a generally narrower range than the other two, which resulted in much smaller differences in the ratings for the speaker. Of the remaining 48 possible interaction effects, only one four-way interaction is significant at the 0.01 level.

[8]However, the distinction in "alignment" of the first accent with the stressed syllable was not made in experiment II.

Bolinger, D. (1961). *Generality, Gradience, and the All-or-None* (Mouton, The Hague, The Netherlands).

Bruçe, G. (1982). "Developing the Swedish Intonation Model," Work. Pap. Ling. Univ. Lund 22, 51–114.

Bruce, G., and Gårding, E. (1978). "A Prosodic Typology for Swedish Dialects," in *Nordic Prosody*, edited by E. Gårding, G. Bruce, and R. Bannert (Travaux de l'Institut de Linguistique de Lund no. 13), pp. 219–228.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (Academic, New York).

Crystal, D. (1969). *Prosodic Systems and Intonation in English* (Cambridge U.P., Cambridge, UK).

Hecker, M. H. L., Stevens, K. N., Bismarck, G. V., and Williams, L. E. (1968). "Manifestation of Task-Induced Stress in the Acoustic Speech Signal," J. Acoust. Soc. Am. 44, 993-1001.

Hirst, D. (1983). "Structures and Categories in Prosodic Representations," in *Prosody: Models and Measurements*, edited by A. Cutler and D.R. Ladd (Springer, Heidelberg, Germany).

Ladd, D. R. (1980). *The Structure of Intonational Meaning: Evidence from English* (Indiana U. P., Bloomington, IN).

Ladd, D. R. (1983). "Phonological Features of Intonational Peaks," Language, 59, 721–759.

Ladd, D. R., Scherer, K.R., and Silverman, K.E.A. (1985). "An Integrated Approach to Studying Intonation and Attitude," to appear in *Intonation and Discourse*, edited by C. Johns-Lewis (Croom Helm, London).

Laver, J. (1980). *The Phonetic Description of Voice Quality* (Cambridge U.P., Cambridge, UK).

Lehiste, I. (1970). *Suprasegmentals* (MIT, Cambridge, MA).

Liberman, M., and Pierrehumbert, J. (1984). "Intonational Invariance Under Changes in Pitch Range and Length," in *Language Sound Structure*, edited by M. Aronoff and R. Oehrle (MIT, Cambridge, MA).

Menn, L., and Boyce, S. (1982). "Fundamental Frequency and Discourse Structure," Lang. Speech 25(4), 341–383.

O'Connor, J. D., and Arnold, G.F. (1961). *Intonation of Colloquial English* (Longmans, London).

Ohde, R. N. (1984). "Fundamental Frequency as an Acoustic Correlate of Stop Consonant Voicing," J. Acoust. Soc. Am. 75, 224–230.

Pierrehumbert, J. (1981). "Synthesizing Intonation," J. Acoust. Soc. Am. 70, 985–995.

Pike, K. L. (1945). *The Intonation of American English* (Univ. of Michigan P., Ann Arbor, MI).

Rosenthal, R. (1980). Personal communication.

Scherer, K. R. (1979). "Nonlinguistic Vocal Indicators of Emotion and Psy-

chopathology," in *Personality and Psychopathology*, edited by C.E. Izard (Plenum, New York).
Scherer, K. R. (1981a). "Speech and Emotional States," in *The Evaluation of Speech in Psychiatry and Medicine*, edited by J. Darby (Grune and Stratton, New York).
Scherer, K. R. (1981b). "Vocal Indicators of Stress," in *The Evaluation of Speech in Psychiatry and Medicine*, edited by J. Darby (Grune and Stratton, New York).
Scherer, K. R., Ladd, D. R., and Silverman, K. E. A. (1984). "Vocal Cues to Speaker Affect: Testing Two Models," J. Acoust. Soc. Am. 76, 1346–1356.
Scuffil, M. (1982). *Experiments in Comparative Intonation: A Case-Study of English and German* (Niemeyer, Tübingen).
Silverman, K. E. A. (1985). "FRED: An Interactive Graphics Program to Modify Fundamental Frequency Contours in Resynthesized Speech," to appear in Behav. Res. Methods Instrum. Comput.
Streeter, L. A., MacDonald, N. H., Apple, W., Krauss, R. M., and Galotti, K. M. (1983). "Acoustic and Perceptual Indicators of Emotional Stress," J. Acoust. Soc. Am. 73, 1354–1360.
't Hart, J., and Collier, R. (1975). "Integrating Different Levels of Intonation Analysis," J. Phonet. 3, 235–255.
Uldall, E. (1960). "Attitudinal Meaning Conveyed by Intonation Contours," Lang. Speech 3, 223–234.
Williams, C. E., and Stevens, K. N. (1972). "Emotion and Speech: Some Acoustical Correlates," J. Acoust. Soc. Am. 52, 1238–1250.
Williams, C. E., and Stevens, K. N. (1981). "Vocal Correlates of Emotional States," in *Speech Evaluation in Psychiatry*, edited by J. Darby (Grune and Stratton, New York).