# Vocal cues to speaker affect: Testing two models

Klaus R. Scherer, D. Robert Ladd,[a] and Kim E. A. Silverman[b]
*Department of Psychology, University of Giessen, Otto-Behaghel-Str. 10, D-6300 Giessen, West Germany*

(Received 14 September 1983; accepted for publication 11 June 1984)

We identified certain assumptions implicit in two divergent approaches to studying vocal affect signaling. The "covariance" model assumes that nonverbal cues function independently of verbal content, and that relevant acoustic parameters covary with the strength of the affect conveyed. The "configuration" model assumes that both verbal and nonverbal cues exhibit categorical linguistic structure, and that different affective messages are conveyed by different configurations of category variables. We tested these assumptions in a series of two judgment experiments in which subjects rated recorded utterances, written transcripts, and three different acoustically masked versions of the utterances. Comparison of the different conditions showed that voice quality and $F0$ level can convey affective information independently of the verbal context. However, judgments of the unaltered recordings also showed that intonational categories (contour types) conveyed affective information only in interaction with grammatical features of the text. It appears necessary to distinguish between linguistic features of intonation and other (paralinguistic) nonverbal cues and to design research methods appropriate to the type of cues under study.

PACS numbers: 43.70.Ep, 43.70.Ve

## INTRODUCTION

Past research on the effects of speaker affect on the acoustic speech signal has generally followed one of two broad approaches which might be characterized as experimental versus descriptive. The experimental tradition employs a variety of methods to assess the communicative force of the suprasegmental aspects of speech, such as fundamental frequency, loudness, and voice quality. These methods include (1) the measurement of acoustic variables in natural or simulated affective speech (Williams and Stevens, 1972; Scherer *et al.*, 1973; Bezooijen and Boves, 1983), and (2) the assessment of listener attributions of speaker affect on the basis of auditory samples in which particular cues have either been (a) isolated by means of masking procedures (Starkweather, 1956; Lieberman and Michaels, 1962; Scherer *et al.*, 1972) or (b) artificially manipulated using synthesis techniques (Uldall, 1960, 1964; Scherer and Oshinsky, 1977). The specific associations between acoustic cues and affective messages that have been found in such research are often in disagreement from study to study and even from subject to subject within studies. The few well-established patterns are for the most part rather general (e.g., the association of higher fundamental frequency with greater degree of arousal; cf. Scherer, 1979,1981a,b). It is not clear how listeners use such general associations in arriving at specific attributions like "insulted," "amused," "condescending," etc. (for a review of work in this tradition, see Scherer, 1981a; Fonagy, 1983).

The second general approach is not primarily concerned with speaker affect as such, but with linguistic (phonological and grammatical) description of suprasegmental phenomena. It attempts to apply the usual techniques of linguistic analysis in order to discover minimal contrasts of stress and intonation (such as "fall" versus "fall rise") in otherwise identical utterances, and treats this analysis as a prerequisite to understanding the expression of affective meaning. To the extent that such descriptions have considered speaker affect, they have done so in two ways. First, they have provided lists of affective nuances conveyed by specific intonational choices in specific linguistic contexts (e.g., O'Connor and Arnold, 1961); second, they have often invoked a distinction between "linguistic" (grammatical, categorical) and "paralinguistic" (affective, gradient) aspects of the suprasegmental part of the speech signal, the latter being excluded from most descriptions. General principles underlying the specific affective nuances are usually stated (if at all) in vague, unverifiable terms, and the theoretical basis of the linguistic-paralinguistic distinction has never been made clear (for a discussion see Ladd, 1980, Chap. 5).

The most obvious difference between these two approaches is methodological, and given the unsatisfactory results of both, believers in one approach might be inclined to blame the methods of the other. In our view there is a more substantive difference: the two approaches make different theoretical assumptions about the nature of nonverbal vocal signaling. (We have discussed this question more thoroughly in Ladd *et al.*, in press; cf. also Ladd and Cutler, 1983.) Further progress in this area, and particularly the ability to account for discrepant results, depends on identifying these theoretical assumptions and testing their consequences empirically.

Experimental studies that search for acoustic cues to affective meaning by "controlling" verbal content operate on the assumption that those cues form a kind of *parallel channel* to the text of an utterance, and that the meaning of that channel is superimposed on the meaning of the text in an essentially additive way. This implies that listeners ought to be able to judge affect from the suprasegmental structure

[a] Current address: Cornell University, Dept. of Modern Linguistics and Linguistics, Ithaca, NY 14853.
[b] Current address: MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, England.

 0001-4966/84/111346-11$00.80  

that remains in filtered or otherwise degraded speech signals.

By contrast, many linguistic descriptions implicitly reject the parallel channel approach, assuming instead that intonation conveys specific affective meanings only in conjunction with specific linguistic features of the text; segmental and suprasegmental do not constitute two parallel channels, but are *integral parts of a unified message*. This implies that identical suprasegmental cues may be interpreted very differently depending on the text with which they are used (cf. Cutler, 1977).

A related difference has to do with the way in which the two approaches model the association between acoustic cues and affective message. The kinds of measures usually applied to the acoustic cues in experimental studies (mean $F0$, regression lines through $F0$ contours, etc.) are consistent with a statistical model in which listener judgments are based on the *covariance of continuous variables* with the type and extent of the speaker's affective state. For example, it assumes that if acoustic cues to anger can be identified, then those cues will be present to a greater degree in utterances that convey more anger than in those that convey less. By contrast, almost all linguistic descriptions assume that intonation involves a number of categorical distinctions, analogous to contrasts between segmental phonemes or between grammatical categories. This means that in statistical terms affective judgments are based on *configurations of category variables*, not scalar covariance. Acoustic parameters are not seen primarily as cues to concrete affective meanings like happiness or fear, but to phonological categories like "fall"; the specific affective message conveyed by an utterance is an interpretation, an active inference by the listener based on the total configuration of the linguistic choices in the context (e.g., Pike, 1945; Bolinger, 1984).[1] Of course, to the extent that linguistic descriptions assume the existence of scalar paralinguistic cues, the distinction between the two approaches is blurred. The two models themselves, however, are clearly distinguishable and appear to play a considerable part in influencing the way investigators think about issues related to nonverbal vocal communication. We will refer to the two in what follows as the "covariance model" and the "configuration model," respectively.

The study reported here experimentally tested some of the theoretical assumptions of the two general positions just sketched. For the most part we have concentrated on intonation and voice quality. By the former we mean gross (and, as we argue, linguistically systematized) patterns of $F0$. The latter, voice quality, we operationally defined as those correlates of laryngeal and supralaryngeal settings which appear as characteristic patterns of energy distribution in the spectrum (for a good review, see Laver, 1980). We are aware that other aspects of speech, such as tempo, loudness, frequency and pattern of pausing, as well as regional and social characteristics, all contribute to the overall impression that a spoken utterance conveys. Differences of approach comparable to those we discuss here may be found in the study of these other characteristics as well; we believe that the pattern of our results can be usefully generalized beyond the specific questions investigated here.

The study proceeded in three stages. In the first stage, we compared affective judgments based on the original recordings of a set of utterances with judgments based on just the transcripts. In the second stage of the study, we considered the assumption of the covariance approach that the nonverbal aspects of an utterance convey information about the speaker's affective state largely independently of the information in the text (the parallel channel assumption). We did this by selecting a subset of the utterances and testing the extent to which affective judgments obtained in the preliminary stage were preserved when the acoustic signals of the utterances were degraded or masked in various ways which also destroy the intelligibility of the text. Finally, in the third stage of the study, we tested two assumptions of the configuration approach: that the interpretation of intonation depends on its categorical linguistic structure, and that the same intonational category may convey different affective meanings in conjunction with different texts. We did this through statistical analyses of the affective judgments obtained in the first two stages of the study.

## I. FIRST STAGE—ESTABLISHING THE AFFECTIVE FORCE OF VOCAL CUES

In this stage of the study we obtained judgments of the affective force of 66 recorded utterances taken from a corpus of spontaneous speech. We also obtained judgments based on transcripts of the same utterances. Comparison of the two sets yielded a number of cases in which the judgments differed considerably; in these we assumed that the nonverbal cues were essential to the message conveyed by the spoken utterance, and from this smaller set we selected the utterances to be used to in the second stage of the experiment. In addition, the preliminary stage was used to assess the reliability of our method for measuring the affective force of utterances.

### A. Method

#### *1. Speech material*

Stimulus utterances were taken from a corpus of tape-recorded interviews between male German social agency workers and two male amateur actors playing the roles of clients (Scherer and Scherer, 1979). The interviews took place in a recording studio set up as an office, and the original recordings were made on high-quality audio tape with professional equipment. For the purposes of this study, excerpts of these recordings were digitized (at 16 kHz with a 7.5-kHz antialiasing filter), which made it possible for utterances to be cleanly cut from their context using a waveform editing program (Standke, 1981). The digitized versions were also used for stimulus preparation in stage two of the study.

A total of 66 utterances spoken by agency workers were selected as stimuli. Utterances from 11 different speakers were used, the number of utterances per speaker ranging from 4 to 11. All utterances were complete sentences between 5 and 25 syllables long (mean length = 11 syllables), spoken rapidly and colloquially. None contained overlapping speech by the client. All the utterances were questions, roughly half wh-questions and half yes/no questions. Most

of the utterances were routine questions dealing with the bureaucratic background of an unemployment compensation case.[2] In several instances identical or nearly identical working was used in two or more utterances.

Our intention in using such relatively unemotional stimuli was to avoid studying the acoustic correlates of extreme rage, grief, or joy. Everyday experience suggests that people can make finely differentiated attributions of speaker affect on the basis of extremely subtle cues; this was the focus of our interest.

### 2. Rating form

So far there are no established measures for the affective force of vocal utterances. Linguists have attempted to *describe* the affective force in terms of speech act analysis employing mostly introspective approaches emphasizing speaker intention (Austin, 1962; Searle, 1969). In psychology emphasis is on the *measurement* of the affective force in terms of the impressions of listeners or observers. Since these impressions are subjective internal events, they can only be measured in the form of self-reports of the observers. The method usually employed to obtain standardized self-reports is the use of rating scales.

We constructed a rating form for use in all stages of the study. From a list of approximately 250 adjectives describing affect, five judges who were familiar with the corpus (two female, three male) selected those that could be applied to any of the corpus interviews. From these selections we chose nine that overlapped as little as possible and provided an adequate range of choice to the subjects, consistent with past work on the measurement of affect. These were *höflich (POLITE); ungeduldig (IMPATIENT); vorwurfsvoll (REPROACHFUL); zweifelnd (DOUBTFUL); freundlich (FRIENDLY); unsicher (INSECURE); gelassen (RELAXED); verständnisvoll (UNDERSTANDING);* and *aggressiv (AGGRESSIVE).*[3]

For each stimulus, subjects were requested to mark with one or two Xs all those adjectives which seemed to them appropriate descriptions of the inferred speaker affect (two Xs signifying that the adjective was seen as an extremely appropriate description). In addition to or instead of selecting from the nine adjectives, they could write a description of the speaker's attitude using their own words. As it turned out, subjects usually selected only one or two adjectives per utterance, used two Xs in only about 9% of these selections, and generally did not resort to free descriptions. Since the number of the latter was so small, we made no further use of them in analyzing the results. For each utterance, we assigned a score on each of the nine adjectives by averaging the responses (one or two Xs) across subjects.[4]

Since rating scales assess internal subjective impressions, it is not possible to assess reliability as for other test instruments. The use of test/retest reliability is not possible since the impressions may change very quickly and are dependent on situational factors. The use of agreement between two judges cannot be used since it is expected that the subjective impressions of different judges will vary as a function of a number of variables including personality, task set, and other conditions. This is not to say, however, that ratings of subjective impression are random. Generally, one uses the average of the ratings of a group of judges in order to obtain a stable estimate for the mean or average subjective impression in that group. The stability of such group means can be assessed by correlating the mean ratings for two groups of different raters for the same set of stimuli.

### 3. Procedure in the full audio condition

There were 32 subjects, predominantly psychology students at the University of Giessen and all native speakers of German. They were divided into two groups of 16, each consisting of eight males and eight females. In this and all subsequent stages of the study, subjects were either paid a small sum or were given credit toward a course requirement that they participate in experiments.

The full audio stimuli, as the term implies, were unmodified recordings of the 66 utterances, prepared digitally as described above. To avoid making the experimental session too long, we split the utterances into two sets and presented each set to only one of the two groups of subjects. Eleven of the utterances were used in both sets, so that the agreement between the two groups of subjects could be subsequently assessed. All stimuli were presented through headphones; each utterance was heard twice and then rated. Half of the subjects in each group heard their respective set of utterances in reverse order.

Subjects were told that the stimuli were questions from social agency interviews that had been recorded with the consent of both the agency worker and the client. They were also told that all the utterances were spoken by the agency workers, not the clients. Subjects were asked to judge each utterance on the basis of its "tone" or "the way it sounds to you." The adjectives on the rating sheet were characterized as "adjectives that are frequently used to describe the tone of spoken utterances."

### 4. Procedure in the transcript condition

For this experimental condition, written transcripts of the 66 utterances were produced. No attempt was made to indicate pauses or colloquial pronunciation. The written sentences neither started with a capital letter nor ended with either a period or question mark, since these might have suggested an intonation.

The subjects were 24 students who had not taken part in the full audio condition, divided into two groups of 12 (six males and six females). As in the full audio condition, subjects were told that the sentences were questions asked by social agency workers and were instructed to judge the "tone" of each utterance. The stimuli were divided into two sets exactly corresponding to the two sets used in the full audio condition, and presented to the two groups of subjects, respectively.

## B. Results and discussion

The correlations between the two groups of raters across the 11 utterances they shared, for both the full audio and the transcript conditions, are shown in Table I. As we said above, the agreement between the two groups provides an estimate of stability of group mean. The very high correla-

TABLE I. Correlations between ratings from two groups of subjects across 11 utterances, for each adjective. Probabilities in this and all subsequent tables are two tailed.

| Adjective | Correlation coefficient (df = 9) | |
|---|---|---|
| | Full audio | Transcript |
| Polite | 0.75[b] | 0.82[b] |
| Unsure | 0.89[c] | 0.46 |
| Doubtful | 0.41 | 0.63[a] |
| Relaxed | 0.69[a] | −0.09 |
| Impatient | 0.91[c] | 0.65[a] |
| Friendly | 0.64[a] | 0.66[a] |
| Understanding | 0.77[b] | 0.62[a] |
| Reproachful | 0.83[b] | −0.08 |
| Aggressive | 0.93[c] | 0.82[b] |

[a] $p < 0.05$.
[b] $p < 0.01$.
[c] $p < 0.001$.

tions in the full audio condition show that the mean impression of affective force based on full audio cues is rather stable across groups, at least for the adjectives used here, with the exception of *doubtful*. Because of this low and nonsignificant correlation, the adjective *doubtful* was dropped from all further analyses.

Agreement between the mean ratings was much lower for the transcript condition, as shown by the generally lower level of correlations. This may be considered one first indication of the importance of vocal cues for the judgment of affective force in spoken utterances.

Given the high agreement between the two groups of subjects, we combined the mean ratings for the total set of utterances. For the 11 shared utterances a weighted mean was computed. The data in the transcript condition were treated in the same way and intercorrelations between the full audio and the transcript ratings were obtained. For only one of the adjectives, *aggressive*, was a significant correlation found ($r = 0.43$, $p < 0.001$). This suggests that in this corpus there were verbal, textual cues to aggressiveness as well as vocal cues.

Nevertheless, the lack of a strong intercorrelation pattern between the two sets of ratings for the other adjectives shows that nonverbal vocal cues are essential for the communication of affective force. Also, subjects frequently commented that the transcript task was difficult and unreasonable. This was not true for the full audio condition. All this suggests that nonverbal information played a crucial role in signaling the affect conveyed by the full audio spoken utterances.

While this conclusion is hardly surprising, it should be emphasized that it does not in itself represent evidence for either of the theoretical positions outlined in the Introduction. The fact that the nonverbal part of the utterance is crucial for speaker judgments of affect does not necessarily mean that the signaling is direct and context independent.

## II. SECOND STAGE—INVESTIGATING COVARIANCE ASSUMPTIONS

The second stage of the study attempted to investigate to what extent individual acoustic parameters directly evoke affective judgments and to what extent they have meaning only in conjunction with verbal content. In particular, we wanted to distinguish between $F0$ cues and voice quality cues, since even traditional linguistic descriptions give reason to believe that voice quality is more likely to convey affect according to the parallel-channel assumption of the covariance model. The method chosen was the use of several content masking techniques which degrade or mask specific acoustic cues and isolate or emphasize others.

### A. Method

#### 1. Stimulus material

Since four judgment conditions were to be used in this stage, the number of utterances had to be reduced to keep the judges' task manageable. In order to capture the differences in the judgments of affective force of the utterances obtained in stage one, we based the selection of the reduced set of utterances in part on a multivariate similarity analysis. Cluster analyses were run using three different techniques (Ward's method, single linkage, group average) and using both full audio and transcript ratings in different analysis conditions. The different inclusion criteria made little difference to the overall pattern of the results. For both rating conditions (though with a clearer pattern for the full audio ratings) two main clusters formed. The denser one was broadly positive (i.e., containing utterances characterized by high scores on *friendly*, *relaxed*, and *understanding*), while the more scattered one was broadly negative (i.e., containing utterances characterized by high scores on *impatient* and *insecure*).

On the basis of these cluster analyses, we selected 12 utterances from the broadly positive cluster and 12 from the broadly negative cluster for use in the signal masking experiment. To ensure that our stimuli were utterances in which the attribution of affect was largely dependent on the vocal cues, we chose only utterances that had high scores on either the positive or negative adjectives in the full audio condition and low scores on those adjectives in the transcript condition. We also made our selection so as to achieve a good balance of speakers and a good balance between utterances with rising and falling intonation contours. We prepared four different versions of these 24 utterances for presentation as stimuli: the original full audio, and three different acoustic degradations or distortions chosen in an attempt to mask or eliminate different combinations of acoustic cues.

The three degradations were the following:

*a. Low-pass filtered.* The aim of this condition was to filter out the verbal content and the voice quality, but leave the fundamental frequency ($F0$) contour. Commonly used electronic content-filtering techniques use a single cutoff frequency of about 500 Hz, with a rolloff of between 30 and 40 dB/oct. While this destroys intelligibility, it is probable that it still leaves some voice quality information in the signal. Consequently, we set the cutoff frequency for each utterance at its own highest $F0$ value (usually around 130 Hz, and always with at least 60 dB/oct rolloff). The result of this procedure was still recognizably speech, but the sound was much more "muffled" than ordinary (500 Hz cutoff) content-filtered speech, and it was not possible to distinguish

individual voices unless they were markedly different in overall level and range. As intended, the $F0$ contour was still clearly recognizable in this condition.

*b. Random spliced.* This technique was originally developed to render speech unintelligible while retaining more characteristics of individual speakers' voice quality than is the case with low-pass filtering (Scherer, 1971, 1982). Specifically with regard to nonverbal cues, random splicing destroys the temporal organization and continuity of the $F0$ contour and the overall energy envelope, while retaining information about overall $F0$ level and range, and especially retaining most of the spectral cues to voice quality.

Each digitized utterance was cut into segments of 310 ms (about three phonetic segments), with adjacent segments overlapping by 3 ms. The overlapped portions were linearly attenuated to zero amplitude to avoid the subsequent introduction of transients. These speech segments were then recombined in a random order. As many as eight different random orders were produced, and the one that was judged by laboratory staff to sound the most like continuous natural speech was used as the stimulus. In contrast to the filtered stimuli, individual voices were easily identifiable in this condition.

*c. Reversed.* Random splicing creates its own artifacts, including a "choppy" sound and possibly a new intonation contour; in addition, it occasionally leaves fragments of old words and creates illusions of new ones. A pilot study (Silverman *et al.*, 1983) in which subjects judged more than one random-spliced version of the same utterance suggested that any effects of different intonation contours or word fragments were negligible. However, to control for the effects of the temporal disruption, we included another masking condition which left the voice quality intact, namely reversed speech. This retains both the voice quality and temporal continuity; as with random splicing, the reversed stimuli permitted easy speaker recognition. The reversed stimuli were prepared digitally.

The most serious potential artifact with reversed speech is the creation of a new intonation contour. Since many of the original contours had their highest $F0$ on the first accented syllable, the corresponding reversed contours had, in effect, very emphatic high final peaks. Acoustically speaking, that is, reversed speech retains information about overall $F0$ level and range, but perceptually this apparent "final emphasis" may give the impression of a greater range than in the original. We will return to this point below.

### 2. Procedure

Four versions of each of the 24 utterances (full audio, low-pass filtered, random spliced, reversed) were used for a total of 96 stimuli. The experiment was run in two sessions several days apart. In each of the two sessions subjects heard masked versions of half of the utterances and full audio versions of the other half. The full audio versions were not presented in the same session as their corresponding masked versions, to reduce the possibility that subjects would recognize utterances and remember previous judgments. The stimuli were grouped in blocks according to the type of masking: the order of the stimuli within each block and the order of blocks were systematically varied across subjects to reduce order effects. However, the full audio block was always presented last to minimize fatigue, since those judgments had been reported by subjects in the pilot experiment to be the easiest.

The subjects were 18 students who had not participated in earlier parts of the study, divided into two groups of nine. The full group consisted of nine males and nine females. The same rating form was used as in the full audio and transcript conditions, and subjects were given the same information about the nature and source of the recorded utterances. The purpose of the acoustic masking was described as "permitting judgments of the tone of the questions as independently as possible of the words." The three masking conditions were briefly described and an example of each was played before the start of the actual judgment session.

## B. Results and discussion

In order to increase the stability of the ratings, we collapsed those that correlated highly with each other. On the basis of the correlational patterns for both the set of full audio ratings of all 66 utterances and the set of 24 utterances in the four signal masking conditions, we combined three pairs of adjectives, shown in Table II, and chose new labels for ease of reference. As the correlation coefficients in Table II show, these pairs are relatively highly intercorrelated. To ensure that this would not lead to overlooking important patterns of results, we computed all of the subsequent analyses with all of the individual adjectives as well as with the scales. The analyses with the individual adjectives did not show any patterns of results that were different from these combined scales. Consequently, only the results for the latter are reported below, plus those for the two unpaired adjectives *polite* and *insecure*.

### 1. Comparison of the four conditions

The judgments were analyzed to see which aspects of affective force were retained by the various degradations. Table III shows those scales for which the ratings in any of the four conditions correlated with each other. The clearest trends are

(i) One or both of the masking conditions that retain voice quality (reversed and random spliced) correlated significantly with full audio judgments on every scale. Thus all aspects of affect represented in the rating form were to a large extent directly communicated by voice quality cues, independently of the text and despite gross distortions of the $F0$ and energy contours. This is strong evidence that much affective information is conveyed by voice quality, i.e., spectral energy distribution, independent of intonation and text.

(ii) At the same time, although the reversed and random spliced conditions were intended to retain voice quality, they exhibited very little correlation with each other. This presumably reflects the fact that the two masking conditions alter the suprasegmental structure in different ways, as discussed above, and consequently each introduces different artifacts. The significant correlation for *aroused* might well be explainable by a particularly redundant encoding of general

TABLE II. Correlations between adjectives combined together to form scales.

| Scale (combined adjectives) | Correlation coefficients for: 66 Full audio stimuli | 96 Degraded signal stimuli |
|---|---|---|
| Challenging (reproachful + aggressive) | 0.53 | 0.75 |
| Agreeable (friendly + understanding) | 0.73 | 0.73 |
| Aroused (impatient + relaxed) | − 0.55 | − 0.63 |

activation in the speech signal (cf. Davitz, 1969; Scherer *et al.*, 1972; Scherer, 1981a).

(iii) Filtered versions showed only one significant correlation with full audio judgments, for *polite*. This contradicts many previous studies which find that electronically filtered stimuli retain a large part of the affective information contained in the original speech (for a review see Scherer, 1979). This contradiction suggests that normally used filtering techniques still retain enough spectral information for subjects to be able to infer affect, and that our extreme low-pass filtering procedure much more clearly isolates $F0$. At the same time, the significant correlation for *polite* indicates that the very low cutoff frequency used to produce the stimuli did not eliminate all information from the signal. Subsequent analyses reported in the next section suggest that the relevant cue to politeness in the filtered stimuli may be $F0$ level.

### 2. Summary

This experiment shows that even when the text of unemotional speech is artificially rendered unintelligible in various ways, it is possible for much of the affective meaning to remain in the acoustic signal. This tends to confirm the assumption of the covariance model that at least some of the affective force of an utterance can be seen as a parallel channel of nonverbal acoustic cues that convey affect in a direct and context-independent way. However, this pattern of results was only found for those masking conditions in which voice quality cues were retained. These were also the conditions in which $F0$ contours were mutilated or reversed. In the low-pass filtering condition, in which $F0$ contours were left intact and were clearly audible, only one of the five scales showed a correlation with the full audio condition. If we were to interpret these results strictly according to the parallel channel assumption of the covariance approach, we would be forced to conclude that the $F0$ contour contributes little to the affective message of the overall utterance. This is certainly counterintuitive, and contradicts a good deal of past research. This apparent contradiction arises because the use of the signal masking technique does not address one of the basic assumptions of the configuration model, namely that intonational cues signal affect *only in conjunction with the text*. In order to investigate this assumption, we carried out the analyses described in the next section of the paper.

## III. THIRD STAGE—EVIDENCE FOR COVARIANCE AND CONFIGURATION

In this stage of the study we carried out a number of further statistical analyses on the judgments obtained in the first two stages. The primary aim of these analyses was to show two things: (1) that given appropriate hypotheses about the categorical organization of intonation, $F0$ cues can indeed be shown to play a significant role in conveying affect; and (2) that even those aspects of $F0$ that apparently fall outside the realm of categorical linguistic organization, such as overall level and range, may nevertheless act as cues to affect only in the presence of other information in the signal. We restricted these analyses to features of $F0$ for several reasons: first, because it was the one aspect of the signal which was fairly unambiguously isolated in the masking study (viz., in the low-pass filtered condition), and because it

TABLE III. Correlations between ratings in the different masking conditions. Note: $r$ is based on $N = 24$ utterances in each condition.

| | Full audio | | Random spliced | Filtered |
|---|---|---|---|---|
| Reversed | Polite | 0.44[a] | ... | ... |
| | Insecure | 0.41[a] | ... | Insecure 0.49[b] |
| | Challenging | 0.66[b] | ... | ... |
| | Agreeable | 0.61[b] | ... | ... |
| | Aroused | 0.42[a] | Aroused 0.42[a] | ... |
| Random spliced | Polite | 0.43[a] | ... | |
| | Agreeable | 0.82[c] | ... | |
| | Aroused | 0.56[b] | ... | |
| Filtered | Polite | 0.48[b] | | |

[a] $p < 0.05$.
[b] $p < 0.01$.
[c] $p < 0.001$.

conveyed so little information when isolated in that way; second, because it is the main acoustic correlate of putative linguistic categories of intonation; and third, because it is easily measurable and describable, at least in comparison with the various cues to voice quality. This focus on $F0$ is consistent with the overall goal of this paper, which is not simply to identify acoustic cues to affect, but to show that those cues may function in ways not captured by the covariance approach.

### A. Method

The analyses were performed to investigate and compare the extents to which hypothesized categorical and gradient features of $F0$ could account for the ratings of the affective face of the stimuli. The analysis technique most suitable for this purpose is multiple regression, since it allows the joint assessment of the contributions to the variance in a dependent variable by both categorical and continuous independent variables. Before discussing the procedure chosen, we will describe the definitions used for both categorical and gradient variables in this study.

#### *1. Description of F0 features*

*a. Categorical features.* The $F0$ contours of the 66 utterances analyzed in the full audio stage were grouped into categories based on a linguistic description of German intonation. A distinction was made between "rise" and "fall," based on the final pitch movement of the contour (high versus low boundary tone in the system of, e.g., Pierrehumbert, 1980). This gross categorization obviously captures only the broadest distinctions of a reasonable analysis of German intonation, but it represents two major types that must be included in any description of intonation in questions (cf. Essen, 1956; Pheby, 1975; Cruttenden, 1981; Scuffil, 1982).

*b. Gradient features.* Here, the characteristic of $F0$ under investigation was not contour type, but overall level and range. As we noted in the Introduction, descriptions of intonation frequently draw a distinction between broad categorical contrasts of contour type on the one hand, and, on the other hand, scalar or continuous paralinguistic features which characterize in more detail the manner in which the contour type is realized acoustically. Overall range and level are generally assumed to be of the latter type (cf., e.g., Crystal, 1969).

The most established measures of overall level and range are $F0$ mean and $F0$ standard deviation. In spite of some theoretical misgivings,[5] we use these parameters in the data analyses reported below for the sake of comparability with the earlier literature. The measures of level and range were calculated from smoothed pitch extractions, obtained via autocorrelation of 31-ms segments of the digitized speech signals.

#### *2. General procedure*

The utterances were cross-classified according to question type and contour type. Contours were classified as either rise or fall, as described above; questions were classified as either wh-questions or yes/no questions.[6] This classification yielded a two-factor-design: intonation (fall versus rise) by question type (wh versus yes/no) with 14, 18, 15, and 16 utterances per cell. These factors and their interactions were included in a multiple regression analysis using effects coding (this is mathematically equivalent to an analysis of variance; see Cohen and Cohen, 1975). The only other two variables included were mean $F0$ and $F0$ standard deviation. Stepwise regression with a selection of variables to be entered in the equation based on the highest remaining partial correlation was chosen. Five multiple regression analyses were performed, one for each of the five affect scales.

### B. Results and discussion

#### *1. Multiple regression for utterances in the full audio condition*

Table IV shows those parameters that accounted for a significant amount of the variance in the judgments of the full audio versions of the utterances. The relative contributions of category and gradient variables is not the same in each of the five affect scales; in fact, for only three of the scales do the judgments seem to be based on both types of cues. Taken together, the patterns in the results confirm the assumptions of both the configuration and covariance models, but at the same time reveal them to be differentially appropriate to different types of speaker affect.

Consider first of all the *challenging* scale. As can be seen from Fig. 1, listeners' judgments seemed to be based primarily on the *interaction* of intonational category with question type, irrespective of $F0$ level and range: high ratings on this scale were evoked only by the combination of falling intonation with yes/no questions. The other three combinations had very low ratings. Thus this aspect of the affective force cannot be attributed to intonation alone. (Figure 1 also makes clear that the appearance of "intonation" and "question type" for this affect scale in Table IV is only an artifact because the larger interaction effect is so asymmetrical.) The communication of this type of affect follows the configuration model.

The *aroused* scale shows the opposite type of result. Here $F0$ standard deviation and mean $F0$ account for a sizeable amount of the variance, whereas there is no significant contribution from the categorical variables. Thus, neither intonation, nor question type, nor their interaction has any effect on the judgment of the speaker's arousal. As predicted by the covariance approach, a number of empirical studies have shown that physiological arousal of the speaker, presumably because of increased laryngeal tension, leads to increased mean $F0$ and/or $F0$ variability. The present results seem to imply that judges may utilize these functional relationships in inferring speaker arousal.

A third type of pattern is evident in the *agreeable* and *polite* scales. Here both the interaction between intonation and question type, and also mean $F0$, affect the judgments. The means on which interactions are based are shown in Fig. 1(b) and (c). The most important point about the interactions is that they seem to reflect the traditional descriptions of "normal" or "unmarked" intonation for the two question types. The supposedly "normal" combinations of intonation and question type (i.e., falling wh-questions and rising yes/

TABLE IV. Parameters significantly contributing to the variance in a multiple regression on affect scales—full audio condition with 66 utterances. Note: Only those parameters significantly contributing to the variance not yet accounted for in the equation in a stepwise regression are shown here. The total is the sum of the variance jointly accounted for by the variables shown. Tendencies ($p < 0.10$) are listed in those cases where there are persistent patterns throughout the results.

| Affect scale | Parameter | % of variance accounted for: per parameter | total |
|---|---|---|---|
| Challenging | Intonation × question type | 10.6%[b] | 19.2% |
| | Question type | 7.9%[a] | |
| | Intonation | 6.8%[a] | |
| Agreeable | Mean $F_0$ | 19.6%[c] | 34.6% |
| | Intonation × question type | 15.0%[c] | |
| Polite | Mean $F_0$ | 18.3%[c] | 25.4% |
| | Intonation × question type | 7.1%[a] | |
| Insecure | Intonation | 7.9%[a] | 13.1% |
| | Mean $F_0$ | 5.2%, $p = 0.062$ | |
| Aroused | s.d. $F_0$ | 24.3%[c] | 30.8% |
| | Mean $F_0$ | 6.5%[a] | |

[a] $p < 0.05$.
[b] $p < 0.01$.
[c] $p < 0.001$.

no questions) were judged as *polite* and *agreeable*, while the opposite "marked" combinations (which occurred just as often in our corpus) were rated much more negatively. This does not, of course, mean that the "normal" and "marked" intonations are always associated with positive and negative interpersonal attitudes, respectively, since the pragmatic function of these intonational choices in the original dialogues may have been quite different. However, the fact that subjects' judgments agree in reflecting the distinction between marked and unmarked combinations provides evidence for the validity of this distinction and the categorization of intonation that underlies it.

For both the *agreeable* and *polite* scales, mean $F0$ accounted for a high percentage of the variance, with higher voices generally rated as less *agreeable* and less *polite*. It is likely that this effect is culturally mediated, i.e., that voice pitch may have different meaning in different cultures. For example, Laver attributes very low pitches for American males to cultural stereotypes (Laver, 1975, p. 268); Loveday (1981) suggests that higher pitch is seen as *more* polite for Japanese females.

The results for the *insecure* scale are more difficult to interpret. There is a significant main effect for intonation, and in addition a nearly significant contribution from mean $F0$. However, additional analyses discussed below, in which the first of these two effects could not be replicated, show that this may be an artifact of the correlation between measures of $F0$ level and intonation ($r = -0.32, p < 0.01$) in the corpus of utterances.

### 2. Multiple regression for 24 utterances in the signal masking conditions

Multiple regression analyses of the same type were run for the ratings in the signal masking conditions. Since only a subset of 24 utterances was used in this study, the distribution of question types was slightly skewed, and we decided not to use question type and interaction with intonation in the regression analyses. The independent variables used in the regression analyses were therefore mean $F0$, standard deviation of $F0$, and intonation (rise versus fall). The results of these regressions, again listing only the parameters with a significant contribution to the variance, are shown in Table V.

Looking at the column for the full audio condition in

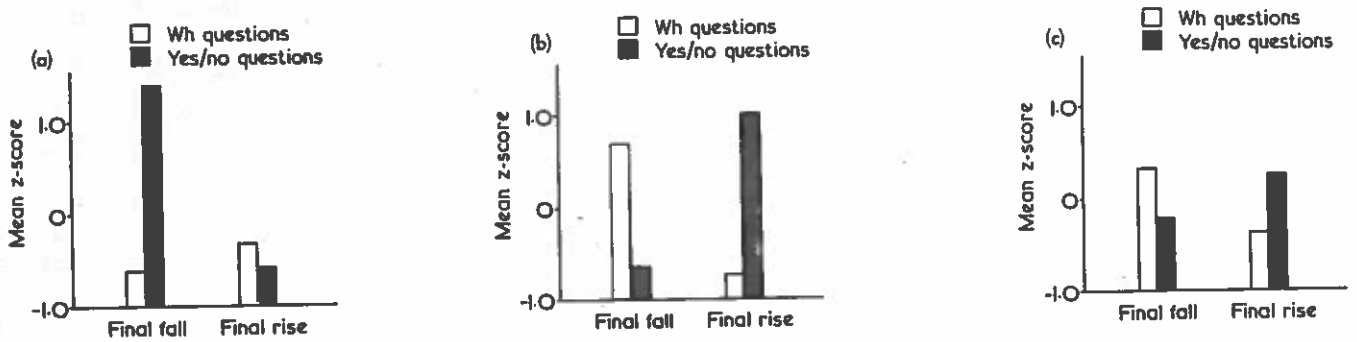


FIG. 1. Ratings of affect in utterances with final falls versus those with final rises, for the three scales where intonation interacted with question type. (a) Challenging, (b) agreeable, (c) polite.

TABLE V. Parameters significantly contributing to the variance in a multiple regression on affect scales—signal masking conditions. Note: The table shows the proportion of the variance accounted for by those parameters significantly contributing to the variance not yet accounted for in the equation of a stepwise regression. Tendencies ($p < 0.10$) are listed in those cases where there are persistent patterns throughout the results.

| Affect scales | Full audio | Reversed | Random spliced | Filtered |
|---|---|---|---|---|
| Challenging | Intonation 22.9%[a] | s.d. $F_0$ 17.1%[a] | ... | ... |
| Agreeable | Mean $F_0$ 32.8%[b] | ... | Mean $F_0$ 17.0%[a]<br>Intonation 16.3%[a] | ... |
| Polite | Mean $F_0$ 28.0%[b] | ... | ... | ...[d] |
| Insecure | Mean $F_0$ 23.3%[a] | Mean $F_0$ 30.2%[b]<br>s.d. $F_0$ 25.5%[b] | Mean $F_0$ 22.0%[a] | ... |
| Arousal | s.d. $F_0$ 23.8%[a]<br>Intonation 11.8%, $p = 0.064$<br>Mean $F_0$ 14.0%[a] | s.d. $F_0$ 18.1%[a] | Mean $F_0$ 12.9%, $p = 0.085$<br>Intonation 14.8%, $p = 0.051$ | ... |

[a] $p < 0.05$.
[b] $p < 0.01$.
[c] $p < 0.001$.
[d] Floor and high/floor (see footnote 5), though not included in the regressions reported in this table, together accounted for 39.8% ($p < 0.01$) of the variance in judgments of filtered stimuli on the *polite* scale.

this table, one can see that the pattern of data found for the judgments of the full audio condition with the complete set of utterances (shown in Table IV) is replicated remarkably well (except for question type and interaction which could not be analyzed in this condition). This replication provides further evidence for the stability of the ratings of affective force used in this study as well as for the effects discussed above. As for the signal masking conditions, mean $F0$ and standard deviation account for a high percentage of the variance in both the reversed and random-spliced conditions, indicating that judges made use of these parameters in judging some of the scales, particularly *insecure* and *aroused*. As in the full audio conditions, higher $F0$ mean and standard deviations are seen as signs of insecurity and arousal. In the random-splicing condition, as in both full audio conditions, higher mean $F0$ is associated with lower agreeableness ratings.

It is striking that even though the $F0$ parameters accounting for a sizeable proportion of the variance in these judgments were retained and are clearly audible in the low-pass filtered stimuli, they did not account for the ratings in this condition. This strongly suggests that judges find it difficult to use $F0$ related parameters in isolation as cues to speaker affect. It seems highly likely that the voice quality correlates of increased laryngeal tension, changes in spectral energy distribution arising from differences in glottal pulse shape, are perceptually more relevant than pure $F0$ for the inference of affective information in speech.

The absence of correlations in the filtered condition may lead to the impression that the low-pass filtering procedure used in this study was too extreme to retain any of the information-carrying $F0$ features. However, some of the other $F0$ related parameters measured but not reported here (see footnote 5) account for a rather high proportion of the variance in the politeness judgments in the filtered condition (floor and high/low, together accounting for 39.8% of the variances $p < 0.01$). This may well explain why *polite* was the only scale for which we found a significant correlation between the ratings in the filtered and full audio conditions (see Sec. II B 1; Table III). It is highly likely, then, that our low-pass filtering procedure did retain at least some information-carrying $F0$ related cues.

### 3. Summary

The results of this stage of the study provide clear evidence of interactions between contour type and text in communicating several important aspects of speaker affect. The existence of such interactions shows the limitations of content-masking techniques as a method for investigating nonverbal affect cues: to the extent that these cues operate only in combination with verbal content, experiments that attempt to determine the affective force of isolated individual cues limit the usefulness of their results from the outset. This stage of the study also suggests a distinction between categorical and continuous variables in the way $F0$ is used to communicate affect. This provides some justification for the distinction between "linguistic" and "paralinguistic" aspects of intonation that is widely assumed in linguistic descriptions.

## IV. CONCLUSION

The present study has investigated some of the underlying assumptions of past work on how affect is signaled in speech, and has shown that features of both the "covariance" and the "configuration" models must be included in any adequate general account.

The preliminary stage of the study, in which judgments of recorded utterances were compared with judgments of written transcripts, made clear that the nonverbal aspects of the speech signal contribute crucially to the communication of speaker affect. The second stage, in which the acoustic stimuli were artificially rendered unintelligible, showed that affective force may be conveyed by the nonverbal features alone. As assumed in the "covariance model," the nonverbal cues appear to function independently of, and in parallel with, any affective information in the text. At the same time, the experiment also showed that it is important to distinguish between different types of nonverbal features. In parti-

cular, it showed that the parallel-channel model may apply to voice quality better than it does to $F0$ contours.

In the third stage of the study, further analyses showed that the affective signaling functions of $F0$ depend in part on specific combinations of sentence type and contour type. This provides evidence for the existence of categorical linguistic organization of $F0$, and more generally for the assumption that affective signaling may depend on configurations of category variables. However, it is also important to distinguish linguistic and paralinguistic features of $F0$, since overall $F0$ level and range, unlike contour type, do show "covariance" effects on affect judgments.

In order to make progress beyond these rather general conclusions, it will be necessary to go beyond such methods as signal degradation and correlational analyses, and to manipulate acoustic stimuli in a much more focused way. In particular, the existence of extensive interactions between different nonverbal variables and between the nonverbal variables and the text would appear to require experimental techniques that modify nonverbal variables while leaving other aspects of the speech signal intact. For $F0$ this can now be done relatively easily and successfully by means of linear predictive resynthesis (Markel and Gray, 1976). Using this technique, it is possible to precisely modify the $F0$ contour of an utterance without changing the timing, energy, or segmental structure, so as to create systematically differentiated stimuli for judgment experiments. We are currently using resynthesized stimuli to study the interaction of voice quality and $F0$ range, and preliminary results suggest that the technique has considerable promise. Given the important effects of discourse context on may aspects of $F0$ range (Menn and Boyce, 1982), it will be necessary to extend such studies in the direction of more complex types of discourse.

In addition to more appropriate research paradigms, this field of study urgently needs clearly stated hypotheses. A prerequisite for the development of such hypotheses would seem to be a clarification of the conceptual muddle that characterizes the notion of speaker affect. The pattern of results reported in this paper leaves little doubt that there are several clearly distinct types of affective messages and that there are major differences between these types in terms of the acoustic cues involved and the nature of the inference processes. At the very least, we see a need for a distinction between transitory physiologically based states and conscious intentions to communicate through the use of a spoken utterance. Somewhat simplified, one might assume the covariance model to be more adequate in cases where the influence of biological factors on the acoustic realization of an utterance can be expected, whereas the configuration model, on the other hand, may be more appropriate when sociocultural and linguistic conventions seem to dominate.[7]

The derivation of hypotheses, the design of specific research operations, and the method and techniques to be used in a study, all need to be tailored to the underlying model and its assumptions. Since it seems highly likely that one cannot choose to accept one and reject the other model, we need to design studies in which it is possible to jointly assess the respective contribution of both of these models. Given the association between the two models and the methodological preferences of different disciplines (in particular, psychology and linguistics), there is a clear need for interdisciplinary research.

## ACKNOWLEDGMENTS

The work reported here was supported by the Deutsche Forschungsgemeinschaft. We thank Arvid Kappas, Klaus-Peter Ningel, and Isabell Reinhardt for their assistance, and Jurek Karylowski for his advice on statistical analysis.

[1]This general point of view is in line with recent theoretical developments in pragmatics, in particular with the notion of conversational implicature, e.g., Grice, 1975; Levinson, 1983, Chap. 3.

[2]Examples of transcripts of the stimulus utterances as presented in the "transcript" condition, with rough English glosses. The letter in each utterance number identifies the speaker: A8. bis wann haben Sie zuletzt Ihr Arbeitslosengeld bekommen (when was the last time you got your unemployment compensation), B3. und das Arbeitsamt hat Ihnen ja bestimmt etwas angeboten (well the employment bureau must have offered you something), C2. wie entwickelt sich hier die Belastung für ihr Haus (how do you break down these expenses for your house), C4. aber schriftlich haben Sie es nicht vom Arbeitsamt (but you don't have it from the employment bureau in writing), H1. wie alt sind die Kinder (how old are the children), J2. sie sagten Sie sind verheiratet (you said you're married), J4. was ist in diesen 1300 DM Belastung für das Haus eigentlich enthalten, Herr Horn (so what's included in this 1300 Marks housing costs, Mr. Horn).

[3]Throughout the paper we will refer to these categories by the English labels given in parentheses, but it should be kept in mind that these are only approximate equivalents of the German originals. *Vorwurfsvoll* and *gelassen* are particularly difficult to translate with single English terms; the former suggests "critical" and the latter "unruffled" or "calm." Note also that *unsicher* can mean "insecure" or "uncertain."

[4]Although this method involves assigning a value of 1 to one X and 2 to two Xs, it does not necessarily assume that two Xs meant *twice* as much affect as one X, only that two Xs mean *more* affect than one X. For a discussion of this method of converting ordinal data to interval level, see Cohen and Cohen (1975).

[5]These measures are not theoretically neutral, but bring with them a number of implications. First, they give equal statistical weight to every sample point in digitally extracted contours. This runs against recent empirical studies showing that certain *target points* (specifically the relative height of accent peaks) are of particular communicative importance in the interpretation of contours (Menn and Boyce, 1982; Liberman and Pierrehumbert, 1983). Second, $F0$ mean and standard deviation are often correlated with each other (in the present corpus, $r = 0.32$), suggesting that they are not independent. This conclusion is consistent with recent findings that the bottom of the $F0$ range is a speaker constant (Maeda, 1976; Menn and Boyce, 1982; Liberman and Pierrehumbert, 1983); a constant $F0$ floor means that range expansion (greater standard deviation) necessarily entails higher overall level (higher mean). Finally, there are two related points about the ability of the two measures to distinguish different factors affecting $F0$: mean $F0$ cannot distinguish between inter- and intra-speaker differences of level, and $F0$ standard deviation does not distinguish relatively flat contours high in the speaker's range from relatively "bumpy" or "changeable" contours lower in the speaker's range. In this study, we explored four other, theoretically justified measures of $F0$ level and range: (i) *floor*, an estimate of each speaker's normal speaking level, based on the average $F0$ of several low utterance-final contour end points; (ii) *high*, the $F0$ of the highest point in each contour; (iii) *high /floor*, the quotient of the first two measures; and (iv) *high /low*, the quotient of the highest peak and the lowest valley in each contour. Space limitations do not permit the discussion of the interesting interrelationships between these measures and $F0$ mean and standard deviation. While our additional analyses showed that none of the results reported for the latter variables would be contradicted, the four measures described above show somewhat different relationships to affect judgments in some cases, warranting further investigations along these lines.

[6]The wh-questions included one wh-question imbedded in a yes/no question. Three with wh-words in noninitial position were excluded from the analysis. The yes/no questions included those with statement syntax.

[7]For a more detailed taxonomy of different types of speaker affect and some hypotheses about the relevant acoustic cues and appropriate inference models, see Scherer (in press).


Austin, J. L. (1962). *How to do Things with Words* (Harvard U.P., Cambridge, MA).

Bezooijen, R. van, and Boves, L. (1983). "The relative importance of vocal speech parameters for discrimination among emotion," paper presented at the Tenth International Congress of Phonetic Sciences, Utrecht.

Bolinger, D. (1984). *Intonation and its Parts* (Stanford U.P., CA).

Cohen, J., and Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (Wiley, New York).

Cruttenden, A. (1981). "Falls and Rises: Meanings and Universals," J. Ling. 17, 77–91.

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

Cutler, A. (1977). "The Context-Dependence of 'Intonational Meanings'," Chicago Linguist. Soc. 13, 104–115.

Davitz, J. R. (1969). *The Language of Emotion* (Academic, New York).

Esen, O. von (1956). *Grundzüge der hochdeutschen Satzintonation* (Henn, Ratingen bei Düsseldorf).

Fonagy, I. (1983). *La Vive Voix* (Payot, Paris).

Grice, H. P. (1975). "Logic and Conversation," in *Speech Acts: Syntax and Semantics, III*, edited by J. Morgan (Academic, New York).

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

Ladd, D. R., and Cutler, A. (1983). "Models and Measurements in the Study of Prosody," in *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd (Springer, Heidelberg).

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

Laver, J. (1975). "Individual Features in Voice Quality," Ph.D. thesis, University of Edinburgh.

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

Levinson, S. C. (1983). *Pragmatics* (*Cambridge Textbooks in Linguistics*) (Cambridge U.P., Cambridge, UK).

Liberman, M., and Pierrehumbert, J. (1983). "Intonational Invariance under Changes in Pitch Range and Length," in *Language Sound Structure*, edited by M. Aronoff and R. Oehrle (MIT, Cambridge, MA).

Lieberman, P., and Michaels, S. B. (1962). "Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to Emotional Context of Speech," J. Acoust. Soc. Am. 34, 922–927.

Loveday (1981). "Pitch, Politeness, and Sexual Role: An Exploratory Investigation into the Pitch Correlates of English and Japanese Politeness Formulae," Lang. Speech 24, 71–89.

Maeda, S. (1976). "A Characterization of American English Intonation," MIT Doctoral dissertation.

Markel, J. D., and Gray, A. H., Jr. (1976). *Linear Prediction of Speech* (Springer, Heidelberg).

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).

Pheby, J. (1975). *Intonation und Grammatik im Deutschen* (Akademie-Verlag, Berlin).

Pierrehumbert, J. (1980). "The Phonology and Phonetics of English Intonation," MIT Doctoral dissertation.

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

Searle, J. R. (1969). *Speech Acts* (Cambridge U.P., Cambridge, UK).

Scherer, K. R. (1971). "Randomized Splicing: A Note on a Simple Technique for Masking Speech Content," J. Exp. Res. Pers. 5, 155–159.

Scherer, K. R. (1979). "Nonlinguistic Vocal Indicators of Emotion and Psychopathology," in *Emotions 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. (1982). "Methods of Research on Vocal Communication: Paradigms and Parameters," in *Handbook of Methods in Nonverbal Behavior Research*, edited by K. R. Scherer and P. Ekman (Cambridge U.P., Cambridge, UK).

Scherer, K. R. (in press). "Vocal Affect Signalling: A Comparative Approach," in *Advances in the Study of Behavior Vol. 14*, edited by J. Rosenblatt, C. Beer, and M.-C. Busnel (Academic, New York).

Scherer, K. R., Koivumaki, J., and Rosenthal, R. (1972). "Minimal Cues in the Vocal Communication of Affect: Judging Emotions from Content-Masked Speech," J. Psycholinguist. Res. 1, 269–285.

Scherer, K. R., London, H., and Wolf, J. (1973). "The Voice of Confidence: Paralinguistic Cues and Audience Evaluation," J. Res. Pers. 7, 31–44.

Scherer, K. R., and Oshinsky, J. S. (1977). "Cue Utilization in Emotion Attribution from Auditory Stimuli," Motiv. Emotion 1, 331–346.

Scherer, U., and Scherer, K. R. (1979). "Psychological Factors in Bureaucratic Encounters: Determinants and Effects of Interactions between Officials and Clients," in *The Analysis of Social Skills*, edited by W. T. Singleton, P. Spurgeon, and R. B. Stammers, NATO Conference Series, Series III (Human Factors), Vol. 11 (Plenum, New York).

Scuffil, M. (1982). *Experiments in Comparative Intonation: A Case-Study of English and German* (Niemeyer, Tübingen).

Silverman, K. E. A., Ladd, D. R., and Scherer, K. R. (1983).-"Intonation and Attitude: Empirical Tests of Theoretical Assumptions," in *Bericht über den 33. Kongress der Deutschen Gesellschaft für Psychologie, Mainz 1982*, edited by G. Lüer (Hogrefe, Göttingen).

Standke, R. (1981). "GISYS: Ein Software-Editor zur fileorientierten digitalen Sprachverarbeitung im Zeitbereich," in *Bericht über den 32. Kongress der Deutschen Gesellschaft für Psychologie in Zürich 1980 (Band 1)*, edited by W. Michaelis (Hogrefe, Göttingen).

Starkweather, J. A. (1956). "Context-Free Speech as a Source of Information about the Speaker," J. Abnorm. Soc. Psychol. 52(3), 394–402.

Uldall, E. (1960). "Attitudinal Meaning Conveyed by Intonation Contours," Lang. Speech 3, 223–234.

Uldall, E. (1964). "Dimensions of Meaning in Intonation," in *In Honour of Daniel Jones*, edited by D. Abercrombie (Longmans, London).

Williams, C. E., and Stevens, K. N. (1972). "Emotion and Speech: Some Acoustical Correlates," J. Acoust. Soc. Am. 52, 1238–1250.