Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech

Scherer, Klaus R.; Koivumaki, Judy; Rosenthal, Robert

How to cite

SCHERER, Klaus R., KOIVUMAKI, Judy, ROSENTHAL, Robert. Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. In: Journal of Psycholinguistic Research, 1972, vol. 1, n° 3, p. 269–285. doi: 10.1007/BF01074443

This publication URL: https://archive-ouverte.unige.ch/unige:101792
Publication DOI: 10.1007/BF01074443

Journal of Psycholinguistic Research, Vol. 1, No. 3, 1972

Minimal Cues in the Vocal Communication of Affect: Judging Emotions from Content-Masked Speech

Klaus R. Scherer,[1,2] Judy Koivumaki,[1] and Robert Rosenthal[1]

Received August 3, 1971

Vocal expressions of emotions taken from a recorded version of a play were content-masked by using electronic filtering, randomized splicing and a combination of both techniques in addition to a no-treatment condition in a 2 × 2 design. Untrained listener-judges rated the voice samples in the four conditions on 20 semantic differential scales. Irrespective of the severe reduction in the number and types of vocal cues in the masking conditions, the mean ratings of the judges in all four groups agreed on a level significantly beyond chance expectations on the differential position of the emotional expressions in a multidimensional space of emotional meaning. The results suggest that a minimal set of vocal cues consisting of pitch level and variation, amplitude level and variation, and rate of articulation may be sufficient to communicate the evaluation, potency, and activity dimensions of emotional meaning. Each of these dimensions may be associated with a specific pattern of vocal cues or cue combinations. No differential effects of the type of content-masking for specific emotions were found. Systematic effects of the masking techniques consisted in a lowering of the perceived activity level of the emotions in the case of electronic filtering, and more positive ratings on the evaluative dimension in the case of randomized splicing. Electronic filtering tended to decrease, randomized splicing tended to increase inter-rater reliability.

INTRODUCTION

The important role of the paralinguistic channel in the communication of a speaker's emotional state has been illustrated in a large number of studies.

This research was supported by a research grant (GS-2654) from the Division of Social Sciences of the National Science Foundation to Robert Rosenthal.

[1]Harvard University, Cambridge, Massachusetts.

[2]Requests for reprints should be directed to K. R. Scherer, whose present address is: Department of Psychology, University of Pennsylvania, 3815 Walnut St., Philadelphia, Pa. 19104.

This research has demonstrated the ability of listener-judges to identify accurately role-played emotional expression from content-free speech such as letters or numerals, standard text, and content-filtered speech (Kramer, 1963*a*, 1964; Davitz, 1964; Vetter, 1969). Earlier research had identified some of the vocal cues that seem to carry emotional meaning such as pitch level (Ruckmick, 1936), amplitude, rate of speech (Fairbanks and Hoaglin, 1941), and sequential pattern of the speech flow (Dusenberry and Knower, 1939; Knower, 1941). Research by Davitz and his associates (1964) and Constanzo *et al.* (1969) has shown that listeners are able to perceive and isolate these cues. There is little doubt that many of these provide redundant information and that emotional state can be correctly inferred from a very small set of vocal cues. Thus, it has been shown that emotions can be judged with better-than-chance accuracy from whispered speech (Knower, 1941; Pollack, Rubenstein and Horowitz, 1960), from speech samples that are played backwards (Knower, 1941) and from speech samples that have been content filtered by cutting off the high-frequency band of the voice spectrum necessary for speech intelligibility (Soskin and Kauffman, 1961; Starkweather, 1956; Milmoe *et al.*, 1967). The latter technique is of special interest since Mahl (1964) and Starkweather (1967) have argued that the lower frequencies of the voice are particularly important for the communication of emotional meaning, a statement that has received support from findings of Friedhoff *et al.* (1964). However, although electronic low-pass filtering does reduce the voice spectrum to the lower frequencies, there remain a large number of vocal cues that are not at all or only slightly affected by this filtering technique, for example, intonation contour, rate of speech, pause and rhythm (Rogers *et al.*, 1971; Scherer, 1971). It cannot be ruled out, then, that the recognition of emotional expression is based on these sequential, suprasegmental speech variables rather than on the characteristics of the lower frequencies in the voice spectrum. In addition, research by Kramer (1963*b*) has shown that electronic filtering leads to systematic differences in voice ratings by naive listener-judges compared to ratings of normal speech.

The purpose of the present study was twofold. First, an attempt was made to assess the effect of eliminating most of the sequential speech patterns in vocal expressions of emotion on the recognizability of the emotions portrayed, compared to electronic content filtering. This purpose was achieved by content-masking natural speech samples with the "randomized-splicing-technique," developed by Scherer (1971), which results in breaking up most of the sequential aspects of the speech flow (such as stress and intonation contours), shortens lengthy pauses and, to a large extent, masks rate of speech. In a 2 × 2 design, presence and absence of randomized splicing and

presence and absence of electronic filtering in samples of vocal expression of emotion were varied to determine the minimum set of vocal cues which would still be sufficient to communicate the nature of the emotion to the listener. The second purpose of the study was to assess the systematic effects of the two types of content-masking techniques on listener judgments.

METHOD

Voice Samples

Ten short excerpts from a recorded performance of Arthur Miller's *Death of a Salesman* were chosen in such a way that for each of the two main characters, played by Lee J. Cobb and Mildred Dunnock, there was one sample each of vocal expressions of anger, fear, sadness, happiness, and matter-of-factness. The five male and five female voice segments, each consisting of one or two sentences, lasting between 20 and 30 sec, were recorded from a tape recording of the play.

Four different versions of these ten segments were prepared: (1) the original segments without any change, (2) a randomized-spliced version using the procedure described in Scherer (1971), (3) a filtered version, with all frequencies above 650 Hz removed, using the apparatus described in Rogers *et al.* (1971), and (4) a randomized-spliced/filtered version with both content-masking techniques applied.

Subjects

Sixty-one women between the ages of 17 and 28 (mean = 20.3) were recruited by advertisements in the college newspaper. Most of them were college students or graduates, and most were attending the college summer school. *S*s were paid for their services.

Procedure

The experiment took place in the college language laboratory facility, where each *S* was assigned to a listening booth which was channeled to one of four tapes containing the instructions for the rating experiment and the various versions of the voice segments. The room was divided into four imaginary squares and a 4 × 4 Latin square design was used to control for seating patterns in the room and for possible effects of the four conditions of

a previous experiment with the same *Ss*.[3] This design ensured that equal numbers of *Ss* in each treatment group were assigned to each quadrant of the room, and that no two *Ss* in the same treatment group sat next to each other. At the time of the experiment it was necessary to relocate some *Ss* because of broken headsets and other technical reasons. This resulted in an unequal distribution of the different treatment groups: the left-hand side of the room had more *Ss* in all conditions than did the right-hand side of the room, the differences varying from two more to five more, depending on the condition.

The *E*, standing at the left front of the room, explained how to use the headsets, what to do in case of defective earphones or broken pencils, and which rating sheets to use. *Ss* then heard the following tape-recorded task instructions:

> This is a test of emotional perception. In it we will ask you to judge the emotions you hear in a series of voices. The voices are divided into four parts. These parts are subdivided into 10 numbered segments, each of which is announced by a voice reading the number. When the 10 segments of each part have been read, you will know that the next part is about to begin. The voices in some of the parts will sound very different from normal voices, but don't be surprised, and don't make any special effort to understand the words. Remember, don't worry if you cannot always understand the words. Listen to each voice and then judge it on the appropriate rating sheet on your desk. When you rate, you should judge each voice according to the instructions printed at the top of the sheets: the instructions are exactly alike for all the rating sheets. There will be a pause after each segment, during which you should make your ratings. When the rating pause is over, the voice on the tape will announce the next numbered segment. When you rate, rate just as quickly as you can. If you have a question please raise your hand now. (Pause) Now, please read the instructions at the top of your first rating sheet.

It should be explained that *Ss* rated *all 4* treatment conditions; in other words, each *S* rated all 10 segments in one version, then all 10 segments in another version, and so on. The present analyses are limited to the ratings of the first condition heard by each treatment group. Thirteen *Ss* heard condition 1 first, 16 *Ss* heard condition 2 first, 16 *Ss* heard condition 3 first, and 16 *Ss* heard condition 4 first. The discussion that follows pertains only to this first group of ratings.

Ss rated each of the 10 voice segments on 20 semantic differential rating variables, each of which consisted of a pair of antonyms separated by a seven-point scale. *Ss* were told to place a checkmark on the scale near the adjective that best described the emotion portrayed, and to place it in the

[3]The first part of the experiment consisted of the administration of a standard person perception task under four different conditions of instruction reading. The results of this study are reported in Scherer *et al.* (1971).

middle of the scale if they thought the emotion bore no relation to either adjective. The adjective pairs were chosen from Osgood *et al.* (1957) and Davitz (1964) in such a way as to represent the evaluation, activity, and potency dimensions. In addition, some adjective pairs describing acoustic phenomena were included. *S*s were also asked to provide a short verbal label of their own for each segment.

A post-questionnaire was administered in order to determine the extent to which the actors and/or the play from which the segments were taken were recognized by *S*s. However, this questionnaire was administered after *S*s had rated the segments in all five versions and consequently had been exposed to the verbal content of the segments, rendering recognition of the play and the actors much more likely. As only the ratings of the version heard first in each condition are used in the present analysis, these data are only of limited interest. In the condition where 13 *S*s listened to the original version first, 10 chose the correct actor (actress) from a list of 7 and 10 chose the correct play from a list of 8. *S*s in the remaining three conditions were unable to understand any verbal content, so recognition of the play was rather unlikely, although some *S*s may have recognized the identity of the actors.

RESULTS

Rater Reliabilities

The mean inter-rater correlation coefficients averaged over the 20 rating variables are shown in Table I. A 2 X 2 analysis of variance with repeated measures on both factors was computed with the mean inter-rater r per rating variable as observations ($N = 20$). A significant main effect for the randomized-splicing factor ($F = 107.03$, $P < 0.001$, df = 1/19) indicated that the voice segments were rated more reliably when content masked by the randomized-splicing technique. A significant but somewhat weaker main effect for electronic filtering ($F = 7.39$, $P < 0.05$, df = 1/19) suggested that use of

Table I. Average[a] Values of Mean Inter-Rater Reliability Correlations

		Electronic filtering	
		Present	Absent
Random splicing	Present	0.50	0.54
	Absent	0.34	0.42

[a]Averaged over 20 rating variables.

the latter content-masking technique tended to reduce agreement between raters.

Systematic Effects of Masking Techniques

The present study is concerned with comparisons between various groups of raters who have been exposed to different types of content-masked samples of emotional expression. Thus, no attempts are made to assess the accuracy of the raters in identifying the emotion portrayed, especially as there are no valid external criteria for the respective emotions underlying the excerpts from the play except the impressions of two judges who chose the excerpts as examples of specific emotions (based on their knowledge of the plot of the play and the specific rendering by the actors).

The core of the data analysis consists of a 2 X 2 X 10 analysis of

Table II. ANOVA Summary Table[a]

Variables \ Effects	A[b]	B[b]	AB[b]	C[c]	AC[c]	BC[c]	ABC[c]
1 Beautiful/ugly	2.49	1.76	<1	29.16[d]	<1	1.09	<1
2 Strong/weak	<1	<1	2.56	13.41[e]	<1	2.31[d]	1.04
3 Labored/easy	<1	5.38[d]	1.96	26.00[e]	1.05	1.62	<1
4 High/low	<1	<1	3.11	59.44[e]	1.71	1.51	<1
5 Calm/agitated	3.27	4.22[d]	<1	70.98[e]	1.65	<1	1.83
6 Young/old	<1	<1	1.11	14.79[e]	1.09	<1	1.87
7 Kind/cruel	3.50	2.43	<1	37.16[e]	<1	<1	<1
8 Soft/loud	1.77	<1	<1	80.19[e]	1.11	<1	1.99[d]
9 Pleasant/unpleasant	6.01[d]	<1	<1	38.78[e]	<1	1.17	<1
10 Happy/sad	<1	<1	8.28[e]	74.02[e]	1.42	<1	<1
11 Ferocious/peaceful	4.97[d]	8.69[e]	<1	59.65[e]	<1	<1	1.06
12 Relaxed/tense	1.13	2.57	<1	92.82[e]	1.54	<1	1.08
13 Nice/awful	7.68[e]	1.00	<1	33.29[e]	<1	<1	1.05
14 Bass/treble	<1	<1	<1	74.57[e]	<1	<1	<1
15 Active/passive	<1	<1	<1	32.96[e]	1.26	<1	1.91[d]
16 Fast/slow	2.89	2.81	1.58	40.23[e]	<1	<1	1.67
17 Rugged/delicate	<1	<1	<1	38.07[e]	<1	<1	1.15
18 Steady/fluttering	<1	5.54	1.44	20.72[e]	1.38	<1	<1
19 Masculine/feminine	<1	3.74	<1	236.13[e]	<1	<1	<1
20 Mild/intense	<1	<1	<1	24.02[e]	2.01[d]	1.20	1.19

[a]Abbreviations: A, random-splicing factor; B, electronic filter factor; C, voice segment factor, repeated measures.
[b]df ranging between 1/29 and 1/55 due to missing observations.
[c]df ranging between 9/234 and 9/495 due to missing observations.
[d]$p < 0.05$.
[e]$p < 0.01$.

variance with the two content-masking techniques as fixed factors and the 10 voice segments as fixed factor with repeated measures. Table II shows the F ratios and the respective significance levels for all factors and interactions. A number of significant main effects for content-masking techniques and one interaction show that the specific type of masking used has systematically influenced the judges' ratings on specific variables. Table III contains the means for those rating variables showing a main effect for randomized-splicing. The data indicate that the randomized-spliced versions of the 10 voice segments were rated as more pleasant, peaceful, and nice than the unspliced versions, which may be interpreted as a general tendency of the judges to perceive the underlying emotions as more positive on the evaluative dimension.

Table IV contains the means for those rating variables showing a main effect for electronic filtering. The filtered versions of the voice segments tend to be seen as more easy, calm, peaceful, and steady than the unfiltered versions, suggesting a tendency of the judges listening to filtered segments to perceive the respective emotions as closer to the passive pole of the activity dimension.

A significant interaction effect was observed only for ratings on happy/sad. The means, shown in Table V, suggest that the original voice

Table III. Means[a] for Rating Variables with Random-Splicing Main Effect

	Random splicing	
	Present	Absent
Pleasant/unpleasant	4.06	4.46
Ferocious/peaceful	4.12	3.85
Nice/awful	3.66	4.23

[a]Means on 7-point rating scales where 1 is a high rating for the adjective listed first in the pair and 7 a high rating for the second adjective.

Table IV. Means for Rating Variables with Electronic-Filtering Main Effect

	Electronic filtering	
	Present	Absent
Labored/easy	3.81	3.36
Calm/agitated	4.76	5.01
Ferocious/peaceful	4.16	3.81
Steady/fluttering	3.53	4.19

Table V. Means for Ratings on "Happy/sad" [a]

		Electronic filtering	
		Present	Absent
Random splicing	Present	4.59	4.27
	Absent	4.36	4.71

[a]Only significant interaction between masking techniques. $F = 8.28$, $P < 0.01$, df = 1/52.

segments and the versions with both masking techniques applied were seen as expressing less happy emotional states than the voice segments treated with only one of the masking techniques. This finding is rather difficult to interpret and since one might expect 1 out of 20 effects to be significant by chance, the interaction effect might be due to chance variation.

Differential Effects of Masking Techniques

Each of the three groups of listener-judges exposed to content-masked versions of the voice segments had only a restricted set of cues available for the ratings of the position of the respective expression in a three-dimensional emotional space. In addition, the types of cues available were different in each of these groups, as either high-frequency cues or sequence cues or both were removed from the voice sample. If either or both of these cues are important determinants of the recognizability of an emotion via vocal expression, one would expect differences in the ratings between these groups of judges, especially in comparison with the ratings made by the group of judges listening to the original, unmasked voice sample. In the present analysis of variance (Table II) such rating differences should be reflected by interaction effects between the voice segment factor and the content-masking factors. As the F ratios in Table II show, only 2 out of 40 two-way interactions and 2 out of 20 three-way interactions reach the 0.05- level of significance, a number that can be expected to occur by chance given the number of possible effects. Analysis of the differences between the means for those variables showing significant interaction effects did not yield any stable, interpretable patterns. On the whole, then, the type of content-masking used for the stimuli does not seem to have differentially affected the ratings of the emotions expressed. That the emotions were strongly differentiated from each other on each of the rating variables by all groups of raters is demonstrated by consistently large and highly significant main effects for voice segments on each of the rating variables (column C of Table II).

Another way to look at the similarities between the ratings of the

Table VI. Mean Between-Group Correlations Averaged over 20 Rating Variables[a]

	OR	RS	EF
RS	0.90		
EF	0.87	0.91	
RF	0.91	0.94	0.91

[a]Abbreviation: OR, original voice segments; RS; random-spliced segments; EF, electronically filtered segments; RF, random-spliced *and* electronically filtered segments.

various groups of raters would be to analyze the Pearson correlations between the ratings on each variable. Table VI shows the mean correlation coefficients (averaged over 20 rating variables) for the six between-group correlations, indicating a very high degree of agreement between the ratings.

Correlations Between Acoustic Cues and Dimensions of Emotions

A first step toward the identification of the acoustic cues that mediate the recognition of emotion from vocal expression is the analysis of the covariations between the judges' ratings of the acoustic properties of the voice segments and the ratings of the position of the inferred emotions on the three dimensions of emotional meaning. Table VII shows, for each of the four groups of raters, the correlations between the four acoustic rating variables (soft/loud, bass/treble, fast/slow, steady/fluttering) and the three variables most representative of the evaluation, activity, and potency dimensions (pleasant/unpleasant, active/passive, strong/weak). In order to assess the extent to which the correlation matrices for the four groups of judges contain similar perceived relationships, i.e., similar patterns of covariation between acoustic variables and emotional dimensions, the 21 correlation coefficients in each column were ranked and Spearman rank-order correlation coefficients were computed. The results, shown in Table VIII, indicate that there is a very high degree of similarity between the patterns of covariation in the four matrices.

In addition, the three highest positive and the three highest negative correlations, i.e., the three highest and the lowest ranks, are highly similar in magnitude for all four groups, as shown by a very small range of the ranks for these correlations (a range of maximally 2 rank points).

DISCUSSION

The results of the present study leave little doubt that both the electronic filtering and the randomized-splicing techniques of content masking

Table VII. Within-Group Correlations for Seven Selected Rating Variables

Correlated variables	OR		RS		EF		RF		Range of rankings
	r	Rank	r	Rank	r	Rank	r	Rank	
1 Pleasant-active	−0.51	19	−0.70[a]	18	−0.50	16	−0.33	15	4
2 Pleasant-strong	0.06	9.5	−0.18	12	−0.47	15	−0.21	12	5.5
3 Pleasant-soft	0.80[b]	3	0.83[b]	3	0.83[b]	3	0.60	5	2
4 Pleasant-bass	−0.04	11	0.37	7	0.34	7	0.15	9	4
5 Pleasant-fast	−0.38	17	−0.64[a]	16	−0.52	17.5	−0.45	16.5	1.5
6 Pleasant-steady	0.06	9.5	0.64[a]	4	0.30	8	0.43	6	5.5
7 Active-strong	0.55	5	0.46	6	0.70[a]	4	0.78[b]	3	3
8 Active-soft	−0.88[b]	21	−0.89[b]	13	−0.81[b]	21	−0.91[b]	21	8
9 Active-bass	−0.11	14	−0.50	21	−0.32	14	−0.28	13	8
10 Active-fast	0.92[b]	1	0.97[b]	1	0.91[b]	1	0.93[b]	1	0
11 Active-steady	0.08	8	−0.65[a]	17	−0.02	11.5	−0.30	14	9
12 Strong-soft	−0.40	18	−0.56	14	−0.79[b]	20	−0.81[b]	19	6
13 Strong-bass	0.37	7	0.27	9	0.29	9	0.12	8	2
14 Strong-fast	0.52	6	0.31	8	0.57	5	0.68[a]	4	4
15 Strong-steady	0.70[b]	4	0.24	10	0.47	6	0.02	11	7
16 Soft-bass	−0.07	12	0.20	11	0.14	10	0.03	10	2
17 Soft-fast	−0.76[a]	20	−0.79[b]	20	−0.76[a]	19	−0.83[b]	20	1
18 Soft-steady	−0.14	15	0.48	5	−0.02	11.5	0.17	7	10
19 Bass-fast	−0.36	16	−0.62	15	−0.52	17.5	−0.45	16.5	2.5
20 Bass-steady	0.84[b]	2	0.92[b]	2	0.89[b]	2	0.91[b]	2	0
21 Fast-steady	−0.10	13	−0.73[b]	19	−0.24	13	−0.50	18	6

[a] $P < 0.05$.
[b] $P < 0.01$.

Table VIII. Spearman Rank Correlation Coefficients for Ranked Within-Group Correlations[a]

	OR	RS	EF
RS	0.71[b]		
EF	0.93[b]	0.81[b]	
RF	0.85[b]	0.86[b]	0.93[b]

[a] $N = 21$.
[b] $P < 0.01$.

systematically affect the stimulus quality of speech samples in addition to rendering content unintelligible. These systematic effects on judges' perception of the stimulus material are illustrated both by the reliability differences between the various groups and by differential patterns of ratings on certain variables.

The fact that random-spliced voice segments are rated much more reliably than electronically filtered or original samples may result from the extreme reduction of cues available for inference that is characteristic of this masking technique. As discussed elsewhere (Scherer, 1971), random splicing eliminates virtually all sequential speech variables. In addition to habitual voice quality (cf. Abercrombie, 1969) only pitch and amplitude, both in terms of average level and degree of variation, as well as rate and clarity of articulation on a phonemic level can be distinguished in terms of transitory speech phenomena.

This kind of cue reduction decreases the likelihood of cue or channel discrepancy, a situation in which cues leading to conflicting inferences are simultaneously present in a vocal expression (Mehrabian, 1970; Scherer *et al.*, 1971*a*). For example, the intonation contour discernible in an emotional speech sample may suggest a pleasant emotion to a judge, whereas the pitch level and the degree of pitch variation may seem to indicate unpleasantness. Cue discrepancy of this kind may lead to confusion on the part of the judges, which may result in overdependence on one type of cue or in intuitive guessing without an attempt to isolate analytically specific cues on which inferences can be based. As it is very likely that different judges will react differentially to cue discrepancy, this may substantially reduce the reliability of the ratings of a group of judges.

Cue reduction via randomized splicing may force the judges to focus on one specific set of cues which, in addition, are relatively unambiguous and may not allow too much variation in terms of the inferences that can be drawn. Furthermore, there are fewer possible combinations of cues which again would lead to greater uniformity of judgment, as judges may disagree about the meaning of certain cue combinations or may differentially prefer some of these over others.

This line of argument may, on first view, seem to be contradicted by the finding that electronic filtering, which eliminates content and all those cues carried by frequencies above 500-650 Hz, reduces rater reliability rather than increasing it. However, there are some important differences in terms of cue reduction between the two content-masking techniques. First, electronic filtering preserves many more cues than does randomized splicing, specifically all of the sequential speech variables such as rate of speech, pausing, rhythm, intonation contours, stress patterns, etc. These cues, in addition to the increased likelihood of cue discrepancy and the problem of multiple combinations, may be more ambiguous in terms of inferences that are based on them. It seems rather plausible that paralinguistic variables are strongly affected by social variations of the language code, i.e., subculture or dialect differences, which may be differentially interpreted as to their meaning by judges with

different sociolinguistic backgrounds. By contrast, more basic variables, such as pitch and amplitude levels and gross variations, relatively independent of the linguistic substratum, may reflect emotional arousal unaffected by linguistic conventions, and may be interpreted more uniformly by judges with widely varying backgrounds.

This does not explain, however, why the original segments are rated more reliably than the electronically filtered segments, although even more cues can be perceived in the former case. A possible explanation is that the judges in the electronic filtering condition may try very hard to understand the content and read meaning into the unintelligible expressions, a phenomenon that is obviously subject to widely differing tendencies on the part of individual raters. Electronically filtered speech sounds like "a kind of mumble as though heard through a wall" (Starkweather, 1956, p. 396) and it often tends to become intelligible after repeated exposures. It is quite likely, therefore, that in this situation judges will make an attempt to understand the content and to "fill in" the cues missing compared to natural speech, whereas random-spliced speech samples are clearly unsuitable for such recovery attempts. The consequent improvement of inter-rater reliability is of particular importance for research in this area, given the special role of ratings by naive or expert judges in person perception studies.

Systematic effects on the ratings of specific dimensions of emotional meaning may also be due to the differential type of cue reduction resulting from the two masking techniques. Cutting of all high-frequency components of the voice spectrum seems to lead to the perception of a lowered activity level on the part of the judges as indicated by higher ratings for easy, calm, peaceful, and steady—all of which seem to connote a general dampening of perceived variability in the speech signal. It has been shown (Davitz, 1964) that pitch, amplitude, timbre, and rate of speech are related to the subjectively rated activity level of emotions inferred from vocal behavior. It can be argued that the activity level of an emotional state is communicated by pitch and amplitude variations rather than their absolute levels or possibly by combinations of levels and variations. Electronic filtering seems to depress pitch and amplitude variations as the low-pass filters remove energy concentrations in the higher frequencies and consequently only those variations falling in the range of the spectrum below approximately 500 Hz can be perceived. In addition, content-filtering speech may result in a lowering of the perceived pitch level. There is a trend in the present study for judges in the electronic filter conditions to rate the voice segments as more masculine than other rater groups ($F = 3.74$, $P < 0.10$, 1/31). As the judges seem to associate the lower frequencies (or "bass") with lack of variability or "steadiness" in the voice sample (cf. Table VII), the perception of lower pitch may provide a further

explanation for the lower activity level ratings in the case of the electronic filter conditions.

Unfortunately, most of the research on the perception of filtered speech is concerned with intelligibility (Licklider and Miller, 1951) and does not concern itself with the effect of filtering on the perception of intonation contours or pitch level and variation in general (Starkweather, 1967). Thus one can only speculate on the relationship between pitch and amplitude variations under electronic filtering conditions and the perceived activity level of the emotion communicated to the listener. However it seems safe to point out that researchers using electronic content filtering in person perception research should be aware of the technique's systematic effect on ratings of the respective stimuli's position on a general activity or variability dimension.

The systematic effect of random splicing in the direction of more positive ratings on the evaluative dimension are less easily interpreted. This finding does seem to contradict Davitz' (1964) assertion that only the perceived activity level of emotions is communicated by nonverbal aspects of speech whereas the evaluation and potency dimensions of emotional meaning are represented by verbal cues. If this assumption held, one would expect interaction effects in the present results, as the ratings of the evaluational dimension of the inferred emotions should differ between the rater group listening to the original, unmasked speech samples and all other groups of raters. This is not the case, nor is there an effect for content masking by electronic filtering. These results seem to suggest that verbal content is less important as a mediator of the evaluational or pleasantness component of emotion than those types of nonlinguistic speech variables that are eliminated by random splicing, i.e., mainly the sequential aspects of speech. The present data do not allow strong inferences as to which paralinguistic cues specifically carry evaluational information, but the patterns of correlations in Table VII seem to suggest that amplitude and rate of speech are related to ratings of the pleasantness of an emotion. In spite of his numerical results, Davitz' (1964) listing of the vocal cues associated with certain emotions tends to indicate the important role of amplitude, rate of speech, and, in addition, pitch level and intonation contour (both in terms of pitch variations and upward or downward slope) in relation to evaluative aspects of the respective emotions. Of these variables, only amplitude and pitch level and variation remain unaffected by random splicing whereas rate of speech is partially masked and the shape of the pitch contour is broken up completely. One might speculate that the more positive ratings in the random-splicing conditions indicate a trend for the former set of cues, i.e., amplitude and pitch level and variation, to communicate positive evaluation, and for the latter set of cues, i.e., rate of speech and intonation contour, to mediate negative evaluation. However, the relation-

ships between pleasantness of an emotion and the mediating vocal cues are probably much more complex and may involve different patterns of cue combinations as well as interactions with specific emotions.

The present discussion has repeatedly emphasized the need for further research on the nature of the relationship between specific vocal cues and the inference processes leading to the recognition or identification of the emotional state of the speaker. The contribution of the present study consists mainly in the finding that only a very limited number of cues is necessary to communicate the position of an emotion in a multidimensional space of emotional meaning to untrained listeners. The present results reaffirm the earlier assertion (Starkweather, 1956, 1967; Mahl, 1964) that the lower frequencies of the voice spectrum are sufficient to communicate the affective state of a speaker. In addition, it has been demonstrated that virtually all sequential speech variables can be eliminated from emotionally toned speech samples without greatly reducing the recognizability of the underlying emotion in terms of its position in a multidimensional space of emotional meaning. As the present results show, judges listening to electronically filtered and/or random-spliced versions of emotional speech segments agree extremely well with judges listening to the original samples, on the differentiation of the 10 different segments on almost all of 20 different rating scales. In other words, in spite of the large differences in terms of the number and the nature of the vocal cues available for inferring the respective emotion, judges in all four conditions agreed on a level far beyond chance expectations on the position of the ten emotions judged in the multidimensional space of emotional meaning.

This finding can be interpreted to mean that the number and types of cues observable in the condition with the most extreme cue reduction, in this case the electronic filter plus random-splicing condition, are sufficient to communicate the essential information necessary to identify the basic dimensions of emotions expressed vocally. It is quite possible that a more complete set of vocal cues are required to enable listeners to make finer distinctions and to differentiate between emotions that are located in close vicinity to one another on one or more of the basic three dimensions of the emotional space, i.e., evaluation, potency, and activity.

However, only a very small set of vocal cues seems necessary to distinguish a number of basic emotions that differ widely from each other on a number of dimensions. More research is needed to identify more clearly this basic set of vocal cues that seems to carry the bulk of information about a speaker's kind and degree of affective arousal. The present findings suggest that those vocal cues that remain unaffected by both electronic filtering and random splicing, which basically seem to consist of pitch level (fundamental

frequency) and pitch variation, amplitude (loudness) level and variation, and possibly rate of articulation, constitute all, or at least an important part, of the minimal set of vocal cues communicating basic emotional meaning. In order to fully understand the process by which judges infer emotions from vocal cues, a more detailed knowledge of which dimensions of emotional meaning are communicated by which specific vocal cues or cue combinations are needed; in other words, a mapping of a space of acoustic or vocal dimensions into the space of emotional meaning is needed. Scherer (1971) has suggested that the evaluative dimension of emotional meaning might be communicated by pitch level, the potency dimension by amplitude level, and the activity level by the degree of pitch and amplitude variation.

Only the assumed relationship between potency and amplitude is supported to some extent by the correlations between the judges' ratings of the acoustic characteristics and the emotional dimensions shown in Table VII. Pleasantness of the inferred emotion seems unrelated to the bass/treble rating, and activity level shows a significant correlation with steady/fluttering, which may be seen as related to variability only in the case of the random-splicing group.

The present data suggest different kinds of relationships between specific vocal cues and specific dimensions of emotional meaning. For all groups of judges, there are significant correlations between perceived pleasantness of an emotion and ratings of softness and slowness, and between perceived activity level and ratings of fast and loud, indicating more complex, interactive relationships between vocal cues and the dimensions of emotional meaning.

CONCLUSIONS

The present data are clearly inadequate for drawing any conclusions in this respect. In addition to the lack of a more adequate sampling base in terms of number and type of emotions and the problems associated with using role-played rather than real emotions (although the present samples were drawn from a coherent context of a play rather than asking actors for isolated portrayals), the present correlational approach is rather inconclusive. In spite of the technical difficulties, a manipulative approach will have to be adopted in order to gain a more clear-cut understanding of the effects of vocal cues in isolation and combination on the judgment of emotions from vocal expressions. Experienced actors or speech experts can be used to vary specific vocal cues in rendering standard texts, using a factorial design. These renditions can then be played back to judges to determine main and interaction effects of the manipulations on the judges' ratings (cf. Addington, 1968). Even more

controlled are cue manipulations via electronic equipment. Standard recordings of speech samples can be rerecorded with differential recording levels to change amplitude level and variation (cf. Scherer *et al.*, 1971*b*), pauses can be spliced into standard speech samples or continuously variable tape speed can be used to manipulate rate of speech (although this is confounded by simultaneous pitch changes). Changes in pitch level and intonation contours require rather sophisticated speech synthesis equipment (cf. Udall, 1960) but are not impossible to manipulate in standard speech samples. In addition to the manipulation of vocal cues in emotional expression, further extensive acoustic analyses of recordings of natural, interactive expressions of emotions in everyday life, possibly using computerized methods of acoustic analysis (Starkweather, 1967; Ostwald, 1963; Scherer *et al.*, 1971*a*), are necessary to gain more complete understanding of both the indicative and the communicative functions (Ekman and Friesen, 1969) of nonlinguistic vocal cues in the expression of emotions.

ACKNOWLEDGMENT

We thank Mr. Bruce Humphrey, Director, Modern Language Center, Harvard University for providing access to the language laboratory facilities.

REFERENCES

Abercrombie, D. (1969). Voice qualities. In N. N. Markel (ed.), *Psycholinguistics,* Dorsey, Homewood, Ill.

Addington, D. W. (1968). Voice and the perception of personality: An experimental study. Oklahoma State University Monographs, *Social Sciences Series,* No. 15.

Constanzo, F. S., Markel, N. N. and Constanzo, P. R. (1969). Voice quality profile and perceived emotion. *J. Counsel. Psychol.* **16**: 267-270.

Davitz, J. R. (1964). *The Communication of Emotional Meaning.* McGraw-Hill, New York.

Dusenberry, D. and Knower, F. H. (1939). Experimental studies of the symbolism of action and voice. II. A Study of the specificity of meaning in abstract tonal symbols. *Quart. J. Speech* **25**: 67-75.

Ekman, P. and Friesen, W. V. (1969). The repertoire of non-verbal behavior: Origins, usage, coding, and categories. *Semiotica* **1**: 49-98.

Fairbanks, G., and Hoaglin, L. W. (1941). An experimental study of the durational characteristics of the voice during expressions of emotions. *Speech Monographs* 8: 85-90.

Friedhoff, A. J., Alpert, M., and Kurtzberg, R. L. (1964). Infra-content channels of vocal communication. In *Disorders of Communication,* Vol. XLII: Research Publications, A.R.N.M.D.

Knower, F. H. (1941). Analysis of some experimental variations of simulated vocal expressions of the emotions. *J. Soc. Psychol.* **14**: 369-372.

Kramer, E. (1963*a*). The judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychol. Bull.* 60: 408-420.

Kramer, E. (1963*b*). Judgment of portrayed emotion from normal English, filtered English, and Japanese speech. *Diss. Abstr.* 24: 1699-1700.

Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *J. Abnorm. Soc. Psychol.* 68: 390-396.

Licklider, J. C. R., and Miller, G. A. (1951). The perception of speech. In S. S. Stevens (ed.), *Handbook of Experimental Psychology,* Wiley, New York, pp. 1040-1074.

Mahl, G. F. (1964). Some observations about research on vocal behavior. In D. McK. Rioch (ed.), *Disorders of Communication,* Proceedings of ARNMD, Vol. 42, William & Wilkins, Baltimore.

Mehrabian, A. (1970). When are feelings communicated inconsistently? *J. Exptl. Res. Person.* 4: 198-212.

Milmoe, S., Rosenthal, R., Blane, H., Chafetz, M., and Wolf, I. (1967). The doctor's voice: Postdictor of successful referral of alcoholic patients. *J. Abnorm. Psychol.* 72: 78-84.

Osgood, C. E., Suci, G., and Tannenbaum, P. (1957). *The Measurement of Meaning.* Univ. of Illinois Press, Urbana, Ill.

Ostwald, P. F. (1963). Soundmaking: The acoustic communication of emotion. Charles C. Thomas, Springfield, Ill.

Pollack, I., Rubenstein, H., and Horowitz, A. (1960). Communication of verbal modes of expression. *Lang. and Speech* 3: 121-130.

Rogers, P. L., Scherer, K. R., and Rosenthal, R. (1971). Content-filtering human speech. *Behav. Res. Methods Instrumentation* 3: 16-18.

Ruckmick, C. A. (1936). The psychology of feeling and emotion. McGraw-Hill, New York.

Scherer, K. R. (1971). Randomized-splicing: A note on a simple technique for masking speech content. *J. Exptl. Res. Personality* 5: 155-159.

Scherer, K. R., London, H., and Wolf, J. J. (1971*a*). The voice of confidence: Paralinguistic cues and audience evaluation. Unpublished manuscript, Harvard University.

Scherer, K. R., Rosenthal, R., and Koivumaki, J. (1971*b*). Mediation of experimenter expectancy effect via differential speech intensity in instruction reading. Paper read at the 42nd Annual Meeting of the Eastern Psychological Association, New York, N.Y., April 1971.

Soskin, W. F., and Kauffman, P. E. (1961). Judgment of emotion in word-free voice samples. *J. Communication* 11: 73-80.

Starkweather, J. A. (1956). The communication value of content-free speech. *Amer. J. Psychol.* 69: 121-123.

Starkweather, J. A. (1967). Vocal behavior as an information channel of speaker status. In Salzinger, K., and Salzinger, S. (eds.), *Research in Verbal Behavior and Some Neurophysiological Implications,* Academic Press, New York, pp. 253-262.

Udall, E. (1960). Attitudinal meanings conveyed by intonation contours. *Lang. and Speech* 3: 223-234.

Vetter, H. J. (1969). *Language Behavior and Communication.* Peacock, Itasca, Ill.