Recognition of Emotion From Vocal Cues

Johnson, William F.; Emde, Robert N.; Scherer, Klaus R.; Klinnert, Mary D.

How to cite

JOHNSON, William F. et al. Recognition of Emotion From Vocal Cues. In: Archives of General Psychiatry, 1986, vol. 43, n° 3, p. 280–283. doi: 10.1001/archpsyc.1986.01800030098011

This publication URL: https://archive-ouverte.unige.ch/unige:101897

Publication DOI: 10.1001/archpsyc.1986.01800030098011

Recognition of Emotion From Vocal Cues

William F. Johnson, MD; Robert N. Emde, MD; Klaus R. Scherer, PhD; Mary D. Klinnert, PhD

• In two studies investigating the recognition of emotion from vocal cues, each of four emotions (joy, sadness, anger, and fear) was posed by an actress speaking the same, semantically neutral sentence. Judgments of emotion expressed in these segments were compared with similar judgments of voice-synthesized (Moog synthesizer) samples (study 1) or with three different alterations of the full-speech mode (study 2). Correct identification of the posed emotion was high in the full-speech samples. Voice-synthesized samples seemed to capture some cues promoting emotion recognition, but correct identification did not approach that of other segments. Recognition of emotion decreased, but not as dramatically as expected, in each of the three alterations of the original samples.

(*Arch Gen Psychiatry* 1986;43:280-283)

The voice and the face carry much of the affective content of human communication and often form a basis for clinical intuition. Mental health professionals, in particular, rely on skills in recognizing and understanding emotions in the context of diagnosis and treatment. It would seem valuable to design programs to enhance clinical skills of emotion recognition, but to do this it is important to understand the features of normal emotional communication. The basis of such a view can be found in Darwin,[1] who long ago postulated the existence of universal discrete expressions of emotion, an idea that in recent years has received validation from extensive cross-cultural research involving facial expressions.[2,3] The cross-cultural findings implied an innate ability to express and recognize different emotional expressions involving the face, and salient features of the face in each of the basic emotions have now been defined.[4,5] As Darwin also observed, however, discrete emotions are expressed through the vocal channel as well.

Heretofore, the voice has been a relatively neglected aspect of communication. Nonverbal vocal cues have been little investigated, in part because of the difficulties posed by their intimate association with speech. This investigation was designed to explore the accuracy with which emotions are recognized from vocal cues. Intuitively we seem to understand that humans recognize vocal expressions of emotions quite well. The question remaining for research is, What, other than the verbal content of a person's speech, contributes to the hearer's recognition that the speaker is sad rather than happy or angry rather than fearful? In other words, What role do nonverbal cues play in the recognition of emotion from verbal communications? Mixed results from previous experimental studies of emotion recognition from the voice make it difficult to draw any conclusions (see summaries of this research[6,7]). In this study we attempted to overcome a number of problems plaguing previous experimental designs. We used clear signals to give unambiguous expressions of basic emotions. We compared signals that had a recognizable verbal content with nonverbal signals and with those that had the verbal content masked. In previous studies, which concentrated on full speech, it was difficult to determine which cues were important in the judges' decisions. Therefore, in this study, we used a variety of signal-masking techniques to better delineate the acoustic features contributing to emotion recognition. Further, because many previous studies forced respondents to choose between a limited number of listed emotions (thereby potentially introducing unknown constraints), we compared forced-choice responses with a free-response technique.

METHOD

Two studies were done. Both studies used a convenient sample of judges composed primarily of professional trainees in clinical mental health (first- and second-year psychiatric residents and clinical psychology interns). Twenty-one judges were used in study 1 and 23 were used in study 2. The studies were done one year apart, and 16 judges in study 2 had also participated in study 1. Judges were in their late 20s or early 30s and were gathered from the clinical setting in which one of us (W.F.J.) worked. Male judges predominated, with 75% in study 1 and 67% in study 2.

Study 1

In study 1, judgments of the emotion expressed in voice samples were compared with similar judgments of voice-synthesized (Moog synthesizer) samples as used in previous work[8] (also see discussion

Accepted for publication Sept 28, 1984.

From the Department of Psychiatry, University of Colorado Health Sciences Center, Denver (Drs Johnson and Emde); the Department of Psychology, University of Giessen, Giessen, West Germany (Dr Scherer); and the Department of Pediatric Psychiatry, the National Jewish Hospital, Denver (Dr Klinnert).

Reprint requests to Department of Psychiatry, C268, University of Colorado Health Sciences Center, 4200 E Ninth Ave, Denver, CO 80262 (Dr Emde).

Table 1.—Results of Study 1

Acoustic Stimulus and Emotion	% (Proportion) of Correct Identification	
	Forced Choice	Free Response
Human voice		
Joy	100 (84/84)	51.2 (43/84)
Sadness	100 (84/84)	91.7 (77/84)
Anger	100 (84/84)	97.6 (82/84)
Fear	98.8 (83/84)	52.4 (44/84)
Voice synthesizer		
Joy	65.5 (55/84)	40.5 (34/84)
Sadness	92.9 (78/84)	50 (42/84)
Anger	32.1 (27/84)	27.4 (23/84)
Fear	32.1 (27/84)	7.1 (6/84)

below). For the human voice samples, an actress who was familiar with the research question spoke a semantically neutral sentence: "The green book is lying on the table." The sentence was presented in such a way so as to give clear vocal expressions of each of four common discrete emotions: joy, sadness, anger, and fear. While this departs somewhat from what would be encountered in a clinical situation, it serves to reduce potential confusion caused by the content of the sentence. For voice-synthesized samples, nonspeech stimuli were used that in earlier work[8,9] had been shown to generate tendencies toward an attribution of these same four emotions. We thus hoped to assess the relative importance of certain known synthetically encoded cues for emotion recognition.

For study 1, each vocal expression was reproduced four times and randomly ordered, resulting in 16 voice items. Voice-synthesized items were similarly reproduced and randomly ordered. In this study we compared two different judgment procedures: one involving forced-choice labeling and the other involving free-response labeling. In the forced-choice procedure, judges were told the four emotions under study and were asked to select one of the four choices for each auditory stimulus. In the free-response procedure, judges were asked to respond to each auditory stimulus with a word or phrase describing the emotion they heard most clearly expressed. Free-response emotion labels were then categorized into discrete emotions according to a lexicon of emotion words previously validated by Izard.[10] The lexicon allows placement of English words into one of the following categories: joy, interest/excitement, sadness/distress, surprise, disgust, anger, fear, shame, passive-bored/sleepy, and unclassifiable. (The original lexicon was modified slightly and is available on request.) The presentation of voice-synthesized and human voice samples was varied to examine for order effects, but to avoid bias the free-response procedure was always administered before the forced-choice procedure.

Study 2

One year later the above-mentioned voice samples were used in study 2. In this study, however, responses to these stimuli were compared with those of the same voice samples subjected to three separate "masking" conditions designed to obscure certain acoustic factors and to highlight others. The "altered" stimuli were thus composed of the following: (1) content-filtered stimuli (low-pass filtered with a cutoff frequency of 500 Hz; a filter with 45-dB roll-off was used [Krohnhite]); (2) randomly spliced stimuli (with voice sample files segmented into units of about 200 ms and randomly rearranged with the use of a digital speech manipulation system); and (3) reverse speech stimuli with the original voice sample tapes played backward (for details on these three procedures, see the report by Scherer[11]). These masking conditions affect different paralinguistic cues. In the content-filtered condition, voice quality is masked and distorted while loudness is partially masked, but fundamental frequency (pitch), intonation, tempo, rhythm, and pauses are not influenced. With random splicing, the pauses in speech are fully masked; rhythm is distorted, intonation and tempo are partially masked, and loudness, fundamental frequency, and voice quality are unchanged. With backward speech, rhythm, intonation, and voice quality are all distorted, and loudness, fundamental frequency, pauses, and tempo are preserved.[12]

As in study 1, our stimulus presentation involved 16 unaltered voice items (four each of the four emotions). In study 2, however, there were also 48 altered stimuli: four samples of each of three alterations of the four original vocal stimuli. All judgments in study 2 were made according to the free-response labeling procedure and were scored as in study 1.

RESULTS

Study 1

For the voice samples, the accuracy of emotion judgment was almost perfect under the forced-choice condition (Table 1). There was only one error among 336 responses. Even with free-response labeling, sadness and anger were accurately recognized. Samples representing fear and joy, however, were identified as such only 50% of the time; judges often "heard" interest/excitement or surprise or made other responses that could not be classified.

The voice-synthesized samples were not as accurately recognized by our judges. In the forced-choice condition, the sadness stimuli were most often correctly identified (93% correct), followed by joy (65% correct). Anger and fear were recognized at levels close to expectable chance probability (32%). With free-response labeling, only 50% recognized the sadness signals, 40% recognized joy, 27% recognized anger, and just 7% recognized fear.

The order of stimulus presentation (synthesized voice vs human voice and vice versa) had no effect on the results.

Study 2

Our results for study 2 (given in Table 2) reinforce the impression that the human voice carries important information about emotional state. The procedure of free-response labeling maximizes the opportunity for judgments involving other emotions; thus, some judges may respond to "blends" (eg, of interest or surprise with joy or fear stimuli). For the unaltered voice, judges correctly identified joy, sadness, and anger more than 80% of the time, and sadness and anger signals were almost perfectly recognized. Fear was identified with 51% accuracy. These results replicate those obtained one year earlier with the free-response labeling procedure of study 1. Test-retest reliability for the 16 judges participating in both studies was 74% (189/256), despite the lengthy interval between study 1 and study 2.

The results for the electronically altered voice samples were a surprise. The three alterations diminished correct identification only slightly, never more than 15%. Sadness and anger were still recognized with almost 95% accuracy. Joy fell from 85% to 75%, and fear slipped from 51% to 35% to 40% recognition. Thus, the pattern of incorrect responses changed little with signal masking. Joy and fear signals were most often misclassified as interest/excitement whether they were original or altered stimuli.

COMMENT

Our judges had little difficulty in recognizing the emotions represented by the unaltered voice samples. When forced to choose between four alternatives in study 1, the proposed emotion was correctly identified more than 99% of the time, as compared with the 25% one might expect by chance. When judges were given no restrictions, they still accurately recognized all emotions more than 50% of the time, and for sadness and anger the recognition surpassed 90%. Similar results were obtained with the unaltered voice samples in study 2. The percent correct identification of sadness, anger, and fear was almost identical to that of study 1, but for joy, identification increased from 50% to 85%. Such improvement led us to wonder about a possible practice effect, even though the studies were conducted approximately one year apart. Had our judges "learned" something about recognizing joy?

The voice-synthesized samples were considerably harder

Table 2.—Results of Study 2

Emotion	Acoustic Stimulus	% (Proportion) of Correct Identification*	Incorrect Responses						
			Joy	Interest/ Excitement	Sadness/ Distress	Passive, Bored, Sleepy	Surprise	Disgust	Other†
Joy	Full voice	84.8 (78/92)	...	4	...	1	4	...	5
	Random-spliced	76.1 (70/92)	...	7	1	3	3	...	8
	Reverse	72.8 (67/92)	...	11	...	...	8	...	6
	Content filtered	77.2 (71/92)	...	8	...	1	1	1	10
Sadness	Full voice	100 (92/92)	...	...	...	...	...	...	...
	Random-spliced	96.7 (89/92)	...	...	...	1	...	...	2
	Reverse	94.6 (87/92)	...	1	...	2	...	...	2
	Content filtered	96.7 (89/92)	...	...	...	2	...	...	1
Anger	Full voice	98.9 (91/92)	...	...	...	...	...	1	...
	Random-spliced	96.7 (89/92)	...	...	...	...	...	1	1
	Reverse	94.6 (87/92)	...	...	...	...	...	1	4
	Content filtered	92.4 (85/92)	...	...	...	...	...	5	2
Fear	Full voice	51.1 (47/92)	1	24	3	...	12	...	5
	Random-spliced	41.3 (38/92)	2	37	2	...	8	...	5
	Reverse	41.3 (38/92)	9	30	2	...	6	...	7
	Content filtered	35.9 (33/92)	7	34	4	...	4	...	10

*Free response.
†"Other" includes anger, fear, shame, and unclassifiable categories: 68 of 69 incorrect responses labeled "other" were "unclassifiable"; one was "fear."

for our judges to "decode" emotionally. Some of the essential cues for specific emotion recognition seemed to be contained in the samples because, in the forced-choice situation, joy and sadness were recognized with greater than chance frequency. The voice-synthesized sounds, however, did not contain a sufficient number of unambiguous, emotion-specific cues to allow accurate emotion recognition, a conclusion further demonstrated by the fact that correct identifications fell off dramatically when judges were free to answer with any emotion word. Some samples did seem to capture more of the essential cues than others. Even with the free-response technique, 50% of our judges correctly identified the voice-synthesized sadness samples, whereas only 7% recognized "fear."

The remarkable ease with which emotions were recognized in the human voice in study 1 led us to the signal manipulations of study 2. A high percentage of correct judgments persisted even when acoustic cues present in the original voice samples were degraded by these techniques. On the basis of previous theoretic and experimental work,[8,13] we had expected judgment accuracy of an emotion to fall off as relevant acoustic factors were masked by each signal alteration. For each type of alteration, the percent of correct judgments fell slightly. Because none of these resulted in a major drop of accuracy, however, we could not form any hypotheses concerning the relevance of specific cues that were masked.

Overall, sadness was the emotion most clearly recognized; even voice-synthesized samples were correctly identified 50% of the time (90% in the forced-choice procedure). The nonverbal cues for this emotion seem to be basic: slow tempo and little pitch variation that yield an impression of an energyless or passive speech style. Tempo is partially masked in the randomly spliced stimuli, but these samples were no less accurately identified than the other two. Enough essential cues survived the manipulations that produced the altered stimuli, and such cues were fairly well captured in the voice-synthesized samples.

Anger was consistently recognized with a high degree of accuracy from the human voice-based signals, but fared poorly in the voice-synthesized samples. On the basis of previous experimental work, salient acoustic characteristics of anger were thought to be many harmonics, fast tempo, high pitch level, small pitch variation, and rising pitch contours. Although each of the masking techniques used in this study would substantially degrade selected characteristics, recognition of anger in the altered stimuli remained high. These results suggest that there is no unique combination of paralinguistic features necessary to recognize anger and that there is enough redundancy in the features remaining to allow identification. Perhaps there is a special "angry" voice quality, eg, a growling harshness, that persists even in the electronic filter condition and allows for the clear identification. This voice-quality impression might be based on pitch perturbation.[13]

Fear was the emotion our judges had most difficulty recognizing. Although there was only one error in the forced-choice condition of study 1, correct identification of unaltered voice samples of fear fell to 50% in the two (studies 1 and 2) free-response conditions. All other emotions in study 2 were recognized over 80% of the time. Fear signals were often classified as interest/excitement and, less often, as surprise; this may reflect a common element of "arousal."

Arousal (sometimes referred to as "stress" or "anxiety" in previous studies of nonverbal communication) has been shown to affect emotion recognition by means of its association with an increase in mean fundamental frequency.[14,15]

While this mechanism may have influenced fear judgments and although there is increased arousal in anger as well, our judges did not misidentify anger.

Despite the masking conditions, joy was correctly identified by approximately 75% of our judges. This figure is substantially less than for anger and sadness but is more than for fear. While there are aspects of a joyful voice that were not rendered unrecognizable by the signal degrada-

tions, the judges' confusion with interest/excitement and surprise hints that arousal is again an important cue. Joy is an emotion whose paralinguistic features have been poorly studied (in part because much of the research has focused on detecting stress and/or deception).

CONCLUSION

Our studies have demonstrated that the voice is indeed a powerful source of information about emotion and that it is a source that is difficult to "disguise." Vocal cues seem to be strong and stable, yet subtle. We do not yet understand the degree of redundancy in such signals or how the various characteristics contribute to particular emotion recognition.

Clinicians may underestimate how much the vocal channel contributes to one's intuitive "feeling" about a patient. Voice quality is less obvious than physical appearance but seems to be an important aspect of impressions about another person. Vocal cues may be especially relevant when our intuition is apparently at odds with the words of the patient. Astute clinicians may be able to "hear through" attempts to disguise affect. Although it has been widely claimed that nonverbal communication of emotion is primarily accomplished by the visual channel,[16] there is some evidence that personality judgments based on the vocal channel correlate more highly with whole-person judgments.[17] As our results indicate, even limited vocal cues allow the correct identification of some emotions.

Clinical sensitivity is likely to remain an important aspect of psychiatric practice, despite the increasing sophistication of special diagnostic procedures. A crucial part of our clinical skill is the ability to recognize both visually and vocally communicated affects and to interpret the words of the patient within these contexts. Our study has demonstrated the considerable communicative power of the voice and has suggested that clinicians may respond to that source of information more than had been appreciated. As the important features of nonverbal vocal communication are further delineated, sensitivity to this channel of information about emotion will be better understood, and methods to impart such sensitivity to clinicians can be developed. We await further research to integrate these findings about the voice with data about facial expression as we strive to better understand the phenomenon of clinical intuition.

References

1. Darwin C: *The Expression of Emotions in Man and Animals.* London, John Murray, 1872 (reprinted University of Chicago Press, Chicago, 1965).
2. Ekman P, Friesen WV, Ellsworth P: *Emotion in the Human Face: Guidelines for Research and an Integration of Findings.* New York, Pergammon Press, 1972.
3. Izard CE: *The Face of Emotion.* New York, Appleton-Century-Crofts, 1971.
4. Scherer KR, Ekman P (eds): *Handbook of Methods in Nonverbal Behavior Research.* New York, Cambridge University Press, 1982.
5. Izard CE (ed): *Measuring Emotions in Infants and Children.* New York, Cambridge University Press, 1982.
6. Scherer KR: Nonlinguistic indicators of emotion and psychopathology, in Izard CE (ed): *Emotions in Personality and Psychopathology.* New York, Plenum Publishing Corp, 1979, pp 495-529.
7. Scherer KR: Speech and emotional states, in Darby J (ed): *Speech Evaluation in Psychiatry.* New York, Grune & Stratton Inc, 1981, pp 115-135.
8. Scherer KR, Oshinsky J: Cue utilization in emotion attribution from auditory stimuli. *Motivation Emotion,* 1977;1:331-346.
9. Scherer KR: Acoustic communicants of emotional dimensions: Judging affect from synthesized tone segments, in Weitz S (ed): *Nonverbal Communication.* New York, Oxford University Press Inc, 1979, pp 105-111.
10. Izard C: *Patterns of Emotion: A New Analysis of Anxiety and Depression.* New York, Academic Press Inc, 1972.
11. Scherer KR: Methods of research on vocal communication: Paradigms and parameters, in Scherer KR, Ekman P (eds): *Handbook of Methods in Nonverbal Behavior Research.* New York, Cambridge University Press, 1982, pp 136-198.
12. Scherer KR, Feldstein S, Bond RN, Rosenthal R: Vocal cues to deception: A comparative channel approach, in press.
13. Lieberman P: Perturbations in vocal pitch. *J Acoust Soc Am* 1961;33:597-603.
14. Williams CE, Stevens KN: Emotions and speech: Some acoustical correlates. *J Acoust Soc Am* 1972;52:1238-1250.
15. Streeter LA, Krauss RM, Geller V, Olson LL, Apple W: Pitch changes during attempted deception. *J Pers Soc Psychol* 1977;35:345-350.
16. DePaulo BM, Rosenthal R, Eisenstat RA, Rogers PL, Finkelstein S: Decoding discrepant nonverbal cues. *J Pers Soc Psychol* 1978;36:313-323.
17. Ekman P, Friesen WV, O'Sullivan M, Scherer K: Relative importance of face, body and speech in judgments of personality and affect. *J Pers Soc Psychol* 1980;38:270-277.