Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures

Scherer, Klaus R.; Banse, Rainer; Wallbott, Harald G.

How to cite

SCHERER, Klaus R., BANSE, Rainer, WALLBOTT, Harald G. Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. In: Journal of Cross-Cultural Psychology, 2001, vol. 32, n° 1, p. 76–92. doi: 10.1177/0022022101032001009

This publication URL: https://archive-ouverte.unige.ch/unige:102078
Publication DOI: 10.1177/0022022101032001009

Whereas the perception of emotion from facial expression has been extensively studied cross-culturally, little is known about judges' ability to infer emotion from vocal cues. This article reports the results from a study conducted in nine countries in Europe, the United States, and Asia on vocal emotion portrayals of anger, sadness, fear, joy, and neutral voice as produced by professional German actors. Data show an overall accuracy of 66% across all emotions and countries. Although accuracy was substantially better than chance, there were sizable differences ranging from 74% in Germany to 52% in Indonesia. However, patterns of confusion were very similar across all countries. These data suggest the existence of similar inference rules from vocal expression across cultures. Generally, accuracy decreased with increasing language dissimilarity from German in spite of the use of language-free speech samples. It is concluded that culture- and language-specific paralinguistic patterns may influence the decoding process.

EMOTION INFERENCES FROM VOCAL EXPRESSION CORRELATE ACROSS LANGUAGES AND CULTURES

KLAUS R. SCHERER
University of Geneva, Switzerland
RAINER BANSE
Humboldt University Berlin, Germany
HARALD G. WALLBOTT
University of Salzburg, Austria

One of the key issues of current debate in the psychology of emotion concerns the universality versus cultural relativity of emotional expression. This has important implications for the central question of the nature and function of emotion. Although there is a general consensus that both biological and cultural factors contribute to the emotion process (see Mesquita, Frijda, & Scherer, 1997), the relative contribution of each of the factors, or the respective amount of variance explained, remains to be explored. An ideal way to study this issue empirically is to compare outward manifestations of emotional reactions with similar appraisals of eliciting situations in different cultures (see Scherer, 1997). Such studies could reveal the extent to which similar expression configurations indicate comparable evaluation and reaction tendencies. Unfortunately, given the difficulty of identifying and systematically studying comparable eliciting situations in different cultures, such studies have yet to be conducted. Instead, researchers in this area have adopted an indirect approach in addressing the issue.

This approach is based on the assumption that, given the indisputable role of emotional expression in social communication, the ability of members of one culture to correctly

AUTHORS' NOTE: This research has been conducted as a collaborative research program in the context of the Coordination Européenne de la Recherche sur les Emotions (CERE), which is formed by the laboratories directed by Matty Chiva, Paris, France; Heiner Ellgring, Würzburg, Germany; Nico Frijda, Antony Manstead, Amsterdam, Netherlands; Pio Ricci-Bitti, Bologna, Italy; Bernard Rimé, Louvain-la-Neuve, Belgium; and Klaus R. Scherer, Geneva, Switzerland. This CERE activity has been supported by the Maison des Sciences de l'Homme, Paris. We thank G. Brighetti, R. Caterina, M. Chiva, R. Dapper, J. F. Dols, N. Frijda, M. Guidetti, A. Kappas, A. Manstead, P. Ricci-Bitti, and Y. Poortinga for collaborating in conducting the judgment studies. We also thank the Westdeutscher Rundfunk, Cologne, Germany, for collaboration in the professional production of the stimuli. Finally, we acknowledge important contributions to the article by Tom Johnstone. Correspondence concerning this article should be addressed to Klaus R. Scherer, University of Geneva, Department of Psychology, 40. Bd. du Pont d'Arve, 1205 Geneva, Switzerland. E-mail: Klaus.Scherer@pse.unige.ch.

JOURNAL OF CROSS-CULTURAL PSYCHOLOGY, Vol. 32 No. 1, January 2001 76-92

identify the meaning of the emotional expressions in another culture provides at least some support for positions claiming a high degree of universality of the emotion process. Much of the evidence to date is based on studies—pioneered by Charles Darwin (1872/1965)—that have demonstrated the ability of members of many different cultures to decode facial expressions (as captured in photographs) of some major emotions as encoded by actors from another (mostly North American) culture (Ekman, 1982; Ekman & Friesen, 1971; Izard, 1994). Just as in the case of facial expression, one can assume that the ability of members of different cultures to decode *vocal* expressions of emotion correctly is an important precondition for the assumption of a psychobiologically determined, universal emotion mechanism (a position that does not deny the existence of powerful modulating factors due to cultural specificity; see Scherer, 1997; Scherer & Wallbott, 1994).

Empirical research on intercultural emotion recognition from the voice is extremely rare. Kramer (1964) showed that American judges could identify content-filtered vocal expressions of emotion from both American and Japanese speakers with better than chance accuracy, although there were differences between emotions with respect to the rate of accuracy. Beier and Zautra (1972) asked American, Polish, and Japanese students to judge American speakers' vocal expressions in standard speech samples of different length. The results showed that recognition rates for intercultural recognition increased with the length of the speech samples.

McCluskey and his collaborators (McCluskey & Albas, 1981; McCluskey, Albas, Niemi, Cuevas, & Ferrer, 1975) had Canadian and Mexican children and adults judge vocal expressions produced by Canadian and Mexican actresses. They found cross-cultural recognition in some cases to be superior to intracultural recognition. In another study, Albas, McCluskey, and Albas (1976) asked male Caucasian and Native American Cree speakers to express happiness, sadness, anger, and love, using any words that came to their mind. These speech samples were electronically low-pass filtered and presented to Caucasian and Cree judges. Here the results showed that each group of judges showed superior performance when inferring emotions expressed by a member of their own group. The authors suggest that language and culture are crucial factors in the transmission of emotion, even on the nonverbal level, but admit that the data are difficult to interpret because the content of the speech material used for encoding was not controlled. The speakers may have used culture-specific expressions, possibly with characteristic suprasegmental cues (intonation contours, rhythm), that might be difficult for outsiders to decode. Few would doubt that there are many language-specific ways of expressing emotional content.

More recently, van Bezooijen, Otto, and Heenan (1983) studied vocal portrayals of emotions produced by Dutch adults using standard sentences. Groups of about 40 judges (young adults) each from the Netherlands, Taiwan, and Japan were able to recognize the emotions portrayed with better than chance accuracy. Because the patterns of confusion were similar across judge groups, the authors concluded that there are universally recognizable characteristics of vocal patterns of emotions and that these characteristics are primarily related to the activation dimension of emotional meaning.

The apparent predominance of the activation dimension in vocal emotion has often led critics to suggest that judges' ability to infer emotion from the voice might be limited to basing inference systematically on perceived activation (see Pittam & Scherer, 1993; Scherer, 1979, 1986, for reviews of the issue). In consequence, one might expect that if this single, overriding dimension is controlled for (limiting the possibility of guessing the right answer

category from a small number of available alternatives), there should be little evidence for judges' ability to recognize emotion from purely vocal, nonverbal cues—particularly cross-culturally. However, evidence from a recent study that used a larger than customary set of acoustic variables, as well as more controlled elicitation techniques (Banse & Scherer, 1996), points to the existence of both activation and valence cues to emotion in the voice. Furthermore, these data suggest the existence of *emotion-specific* vocal profiles.

Many of the acoustic parameters involved in these profiles have been theoretically predicted to be based on emotion-specific physiological changes (Scherer, 1986). Consequently, it seems reasonable to assume that the recognition of vocal emotion expressions might work across language and culture boundaries. It should be mentioned at this point that in this article the term *culture* is used as a theoretical concept (in a very general sense), whereas *country* is used as a rough operationalization of culture (as it is the case in many cross-cultural studies). Obviously, it is not suggested that these terms be regarded as synonyms (see below).

This article reports a study involving the comparison of the accuracy of vocal emotion expression recognition across nine different countries (in Europe, North America, and Asia). The stimulus material used has been produced by using professional radio actors and thoroughly piloted elicitation procedures. Furthermore, the stimuli have been extensively pretested and analyzed with the help of digital signal analysis procedures (see Scherer, Banse, Wallbott, & Goldbeck, 1991, for further details).

The aim of the study is to investigate the hypothesis that the vocal expressions of emotions (encoded by actors from a particular culture) can be reliably identified by members of different cultures (here operationalized by *different countries*). The expectation, stemming from theoretical considerations (in particular the assumption of phylogenetic continuity of affect sounds) and from earlier intercultural research on facial and vocal expression, is that judges, independent of their cultural and linguistic origin, will be able to identify the vocally portrayed emotions (using speechlike but semantically meaningless utterances) with better than chance accuracy. Apart from the aim of testing this formal hypothesis using a rather larger set of countries than has been the case in the past, the present study also attempts to determine, in an exploratory fashion, to what extent differences between countries (with respect to geographical location, language, racial origin, history, lifestyle, etc.) will affect the degree of accuracy with which the encoded emotions are recognized. If such differences are found, future work may need to examine the relative contribution of different cultural dimensions to emotion inferences from vocal expression.

METHOD

PROCEDURES

Actors

Four actors (two male, two female) were used. All were professional actors who regularly participated in radio and television productions. Voice recordings were collected in a professional broadcasting studio of West German Radio (Westdeutscher Rundfunk, WDR) in Cologne, Germany. The recording sessions were directed by a professional radio producer and the actors were paid for their services.

Emotions Encoded

Voice samples were elicited for the emotions of joy/happiness, sadness, fear, anger, and disgust. These emotions were selected because many theories of emotions agree that they can be considered as basic and universal (see, e.g., Ekman, 1982; Izard, 1977).

Elicitation of Voice Samples

In research on emotional expression, encoders are often requested to portray the emotional states on the basis of verbal labels only (e.g., "Read this text as if you were very angry"). This procedure presents two problems; first, different encoders may attribute different meanings to such labels, and second, they might envisage different situations as elicitors of these emotions. Such problems can be avoided by using a scenario approach (see, e.g., Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979; Wallbott & Scherer, 1986; Williams & Stevens, 1972). In this approach, actors are provided with situation vignettes or short scenarios describing an emotion-eliciting situation. They are instructed to imagine these situations and act them out as if they were experiencing them.

For the present study, scenarios were not constructed a priori, but instead were selected from actual situation descriptions collected in a large-scale, cross-cultural study (Scherer, Wallbott, & Summerfield, 1986). For each of the five emotions studied, two representative scenarios were selected from this material, based on the criterion that the respective type of situation (e.g., "death of a loved one" for sadness) was reported by a large number of respondents in all countries in the cross-cultural study in response to the request to recall a specific emotion category. Two scenarios for each emotion were used to obtain situational variation (all scenarios are reported in Scherer et al., 1991).

Utterances

To eliminate the potential effects of semantics, the "standard sentence" approach was used. Because the present stimulus set was devised to be used in cross-cultural research, "meaningless multilanguage sentences" were constructed in the following way: A trained phonetician selected two meaningless syllables from each of six European languages, namely, German, English, French, Italian, Spanish, and Danish. These syllables were randomly arranged into several seven-syllable sequences, each containing at least one syllable from each of the six languages. A number of such artificially constructed "sentences" were judged by a group of experts in vocal communication with respect to ease of articulation, language neutrality, and status as "sentences." The two most appropriate (according to agreement of the expert group) sentences were used for the construction of the stimulus set:

Sentence 1: "Hat sundig pron you venzy."

and

Sentence 2: "Fee gott laish jonkill gosterr."

Thus, the complete design for the elicitation of speech samples included four actors (two male, two female) × two scenarios × five emotions × two sentences, resulting in 80 utterances. In addition, each of the four actors uttered the two sentences in a neutral, nonemotional fashion.

Selection of Utterances

The 88 utterances thus recorded (80 emotional plus 8 neutral) were edited onto audiotape in random order. This tape was played to a group of students at the University of Giessen, Germany ($N = 29$), whose task it was to judge the emotional content expressed in each stimulus. Participants had to rate all stimuli on each of five emotion scales (joy/happiness, sadness, anger, fear, disgust), from 0 (*not at all*) to 6 (*intense*). Furthermore, for each utterance, the participants had to judge the "naturalness" of the stimulus presented on a scale from 0 to 6. Participants took part in groups. After the presentation of each stimulus, the tape was stopped for 15 seconds to allow participants to record their judgments. Multiple judgments for each stimulus were allowed to account for possible emotion blends.

To select the most appropriate stimuli, means were computed across judges for each emotion scale and each stimulus separately. Then, the most unequivocal items were selected by comparing the criterion (i.e., the emotion the actors attempted to encode) with the mean intensity judgments of the respective emotion. Ideally, we aimed at developing a final stimulus set with an equal number of stimuli for each of the levels of the factors actor, actor gender, emotion, and sentence. An additional criterion for selecting items for the final version was that they receive average intensity ratings of at least 2 on the 7-point scale (ranging from 0 to 6), with the other emotions receiving average intensity ratings of less than 1.[1] If a choice was possible between several stimuli with satisfying emotion ratings, utterances with higher naturalness ratings were chosen. To meet these criteria, the emotion disgust had to be dropped because mean ratings for disgust were low and there were many confusions with anger and sadness. Furthermore, it was not possible to balance the number of stimuli over actors (who contributed 9, 7, 8, and 6 utterances, respectively); therefore, the factor actor was dropped in the final version.

Using this selection procedure, a final stimulus set with 30 items was chosen. The application of the criteria did not allow us to obtain an equal number of stimuli per emotion: Sadness was represented with eight items, anger with seven items, joy/happiness with six items, and fear with five items. Furthermore, four items judged as being neutral (criterion: average judgments for all emotion scales less than 1) were included. In terms of actor gender, 16 items were encoded by female actors, and 14 items by male actors. The two nonsense sentences were represented 18 and 12 times, respectively. (A complete list of stimuli and their encoding characteristics are provided in Table 3.)

Preparation of the Final Version of the Stimulus Set

The 30 items were edited in random order onto the final stimulus tape. They were preceded by four practice items (one for each of the four actors and emotions, respectively), allowing judges to get acquainted with the procedure and the nature of the stimuli. The stimulus tape was organized as follows. The stimulus was announced (in English) by its running number ("number *X*"). After a pause of 1 second, the stimulus was presented and after another pause of 1 second, repeated. Then, a pause with a duration of 6 seconds followed, allowing judges to record their judgments. After these 6 seconds, a "beep" sound announced the next stimulus, and the procedure was repeated.

Instructions to Participants

In a series of pretests, instructions were developed that were given to participants in written form. These instructions provided detailed information on the aims of the study, the type

TABLE 1
Sample Characteristics

	City	*Language*	n	*% Female*	*% Male*
Germany	Marburg	German	70	62.9	37.1
Switzerland	Geneva	French	45	66.7	33.3
Great Britain	Manchester	English	40	50.0	50.0
Netherlands	Amsterdam	Dutch	60	63.3	36.7
United States	Berkeley	English	32	43.8	56.3
Italy	Bologna	Italian	43	55.8	44.2
France	Paris	French	51	51.0	49.0
Spain	Madrid	Spanish	49	42.9	57.1
Indonesia	Jakarta	Bahasa Indonesian	38	71.1	28.9
Total			428	57.0	43.0

of actor portrayals used, and the construction of the "sentences." Participants were asked to focus on the emotion expressed rather than trying to understand the utterances. Participants were told of the possibility that more than one emotion might be expressed in a portrayal (as in real-life expressions) and were given the choice of circling two labels for a given stimulus if they thought that a blended emotion was portrayed. They were warned not to use "neutral" to label emotions with weak intensity.

Participants were instructed to answer in a fast and spontaneous manner, and they were told that they should mark a judgment for each individual item, even if they did not feel certain of their judgment.

Participants were requested not to talk or show any nonverbal reaction (such as laughter) during the task, so as not to disturb fellow raters. Then, the four familiarization utterances were presented and it was explicitly mentioned to participants that the four emotions do not necessarily occur with equal frequency in the 30 utterances.

SAMPLE

Recruitment of International Collaborators and Translation of Research Materials

Possible collaborators in different countries were approached (see Authors' Note) and asked whether they would be willing to participate in the study. This constitutes a convenience sample of countries as defined by Lonner and Berry (1986). This approach was adopted because this series of studies was conducted entirely without external support, which required finding volunteers willing to run the experiments in their respective countries. It was felt that the approach was all the more defensible because criteria for the systematic sampling of cultures remain to be established (see Mesquita et al., 1997). In this fashion, data were collected in nine countries (see Table 1; Authors' Note, p. 76).

All research materials were sent to the research collaborators in the different countries in an English-language version for translation. Particular care was requested for the translation of the emotion terms, to keep the connotations comparable. The collaborators in the different countries were instructed to use the emotion labels in their specific language that came closest to the English originals. Similarly, the instructions were to be translated as literally as possible without violating the stylistic requirements in a particular language. A back transla-

tion into English and a comparison with the original was mandatory. Finally, the translated instruction and rating sheets had to be pretested with about six to eight students in each respective country.

Recruitment of Participants

Each collaborator administered the stimulus set to groups of students in his or her university, attempting to balance the number of male and female respondents. The respondents' age was similar between countries, ranging from 18 to 30 years. Foreign students who did not grow up in the particular country were excluded. All participants took part on a voluntary basis or in the form of course requirements. The city in which the study was conducted, the language spoken by the participants, the number of participants per country, and the gender ratios are shown in Table 1.

Judgment Procedure

To standardize the experimental conditions across countries, identical headphone sets and connection devices for connecting a cassette recorder to five headphones simultaneously were constructed. This equipment was sent to all collaborators, and it was required that all participants listen to the tape by the means of the cassette recorder and the headphones provided. Collaborators received detailed instructions on how to use these devices and were required to use the same pretested loudness (volume) for playback during all stimulus presentations. All headphones were checked before each session to make sure the sound was transmitted binaurally and without distortion.

The presentation of the stimulus set and the judgment procedure took place in a group setting, with a maximum of five participants per group. Participants were seated around a table as far apart from each other as the headphone cables would allow so as to minimize interparticipant influence and interaction.

The first collection of a German sample was interrupted after only 14 participants had been studied.[2] Because we felt that this sample size would be insufficient, the third author ran a judgment study with 70 German students in a lecture-hall setting, with the stimulus set being presented via loudspeakers to the class, the remaining procedure being completely comparable to the description above. The results for the two German samples are very highly correlated, $r = .89$ for the accuracy coefficients (the diagonal of the confusion matrix), $r = .91$ for the correlation between the rows, and $r = .78$ for the column correlations over the off-diagonal entries. Rather than mixing the two samples, it was decided to use only the larger classroom sample for Germany in the analyses.

RESULTS

DATA ANALYSIS

Data analyses were based on confusion matrices—plotting the intended/encoded categories against the inferred/decoded categories. This allows researchers to study the percentages of correct inference (recognition accuracy) through the percentages in the diagonal of the matrix as well as the pattern of errors or confusions in the off-diagonal entries.

TABLE 2
Emotion Recognition and Confusion Percentages for the German and Indonesian Samples

	Encoded Emotions					
Decoded Emotions	*Neutral*	*Anger*	*Fear*	*Joy*	*Sadness*	*Sum (category use)*
a. Germany						
Neutral	**88**	9	2	34	14	147
Anger	5	**79**	3	6	1	94
Fear	1	4	**74**	5	2	86
Joy	0	2	3	**48**	1	54
Sadness	4	4	15	7	**80**	110
Missing	2	1	2	1	1	7
Sum	100	100	100	100	100	
b. Indonesia						
Neutral	**70**	13	7	29	20	139
Anger	13	**64**	1	13	1	92
Fear	3	7	**38**	11	15	75
Joy	4	9	21	**28**	5	67
Sadness	11	7	33	19	**58**	128
Missing	0	0	0	0	0	0
Sum	100	100	100	100	100	

NOTE: Slight imprecision in marginal sums is due to rounding. The values in the diagonal cells of the confusion matrix, representing the accuracy coefficients, are in bold.

In the decoding sessions, judges were allowed to indicate emotional blends by checking two emotion categories if they thought this appropriate. Double ratings were used at least once by 62.6% of the participants; the overall mean frequency of double rating per judge was 2.0. Such double ratings were weighted by .5 and added to the normal confusion matrix. For descriptive purposes, overall decoding accuracy in each country was measured with accuracy percentages and Cohen's kappa was used as a recognition index corrected for guessing and unequal use of emotion categories.[3] To analyze the influence of all factors of the design on emotion recognition, the proportion of hits was calculated for male and female actor portrayals of each emotion for each judge. This procedure resulted in 10 dependent variables of emotion recognition per judge, which were arcsine-transformed and entered in a mixed-model analysis of variance.

Decoding of Vocally Expressed Emotions

To evaluate intercultural differences in the decoding of vocal portrayals of emotional state, we will first present the results for the German sample of decoders. Because German judges share both language and possibly cultural stereotypes of emotion expression with the encoders of the stimulus material, they can be regarded as a reference for optimal emotion recognition. The results of the German sample are presented in Table 2, Panel a).[4]

All four emotions as well as the neutral utterances were recognized with a much higher accuracy rate than would be expected on the basis of guessing (Cohen's kappa = .67, $p < .001$).[5] Whereas the recognition rates for anger, sadness, fear, and the neutral utterances were fairly high (ranging from 74% to 88%), there was a considerable drop in recognition for joy (48%). This result is mainly due to the frequent confusions of the joy stimuli with the neutral

TABLE 3
Mean Emotion Profiles of 30 Stimuli Averaged Across Judges From Nine Countries (*N* = 428)

Encoding					*Decoding*					
Item	*Sentence*	*Actor*	*Actor Gender*	*Emotion*	*Missing*	*Neutral*	*Anger*	*Fear*	*Joy*	*Sadness*
				Neutral						
9	2	1	F		.002	**.875**	.015	.002	.033	.072
23	1	2	M		.005	**.828**	.132	.007	.007	.021
11	2	2	F		.002	**.614**	.011	.047	.019	.307
28	2	1	M		.007	**.655**	.306	.006	.008	.018
				Anger						
19	1	1	M		—	.016	**.943**	.028	.007	.006
8	1	1	M		—	.019	**.936**	.035	.008	.002
13	1	2	M		—	.007	**.924**	.009	.060	—
5	2	2	F		.005	.096	**.728**	.026	.071	.075
26	2	1	F		.005	.251	**.718**	.006	.014	.006
30	1	1	F		—	.248	**.683**	.011	.056	.002
10	1	2	F		.007	.008	**.423**	.250	.022	.290
				Fear						
15	2	1	M		—	—	.011	**.945**	.040	.005
4	1	1	M		.002	.117	.027	**.769**	.075	.011
3	2	2	F		.005	.002	.016	**.702**	.005	.270
18	1	1	F		.012	.171	.079	**.537**	.157	.044
				Joy						
24	2	2	F		.002	.005	.152	.051	**.713**	.077
25	1	2	F		.002	.002	.015	.123	**.561**	.297
14	1	1	M		.005	.081	.026	.209	**.499**	.181
12	1	1	F		.002	.583	.006	.011	**.366**	.033
7	1	2	M		—	.562	.146	.007	**.280**	.005
16	1	2	M		.002	.590	.155	.011	**.159**	.083
				Sadness						
6	2	2	F		—	.054	.006	.046	.009	**.886**
21	1	1	F		—	.084	.008	.036	.007	**.864**
29	1	2	F		.002	.004	.016	.114	.002	**.861**
27	1	1	M		—	.097	.013	.015	.006	**.869**
20	1	2	M		—	.092	.034	.064	—	**.810**
2	2	1	F		.002	.387	.002	.051	.005	**.553**
17	2	2	M		.007	.408	.002	—	.074	**.509**
1	1	2	M		.002	.596	.007	.020	.002	**.373**

NOTE: Stimuli are ordered in decreasing typicality. The values in the diagonal cells of the confusion matrix, representing the accuracy coefficients, are in bold.

stimuli (34%). Only two other confusions exceed 10%: Fear was frequently confused with sadness (15%) and sadness with neutral (14%). These recognition rates, including the very low recognition accuracy for joy, are similar to results of a previous judgment study with the same stimulus material (Scherer et al., 1991, Study 2) and to recognition rates obtained with a larger set of different actors and emotion portrayals (Banse & Scherer, 1996). Additionally, recognition rates were calculated itemwise. This analysis showed that within each emotion category there was considerable variation of recognition accuracy for specific utterances, suggesting that some utterances were less extreme or typical. The overall mean recognition rates for each item are reported in Table 3. The inclusion of less typical and therefore more

TABLE 4
Recognition Percentages of Vocal Emotion Stimuli Across Emotions and Countries

	Neutral	*Anger*	*Fear*	*Joy*	*Sadness*	*Total*	*Cohen's Kappa*
Germany	88	79	74	48	80	74	.67
Switzerland	71	79	70	55	71	69	.62
Great Britain	67	83	70	40	82	68	.62
Netherlands	77	86	65	45	69	68	.60
United States	66	80	72	46	73	68	.60
Italy	81	72	77	39	68	67	.57
France	70	69	71	51	67	66	.57
Spain	69	73	65	30	71	62	.52
Indonesia	70	64	38	28	58	52	.39
Mean across Western countries except Germany	72	77	70	44	72	67	
Mean across all countries[a]	73	76	66	42	71	66	

a. The individual recognition rates were first averaged across emotions for male and female judges separately, then averaged within each country, and then across countries. They are therefore independent of different sample sizes.

difficult items is expected to increase the sensitivity for the detection of intercultural differences in emotion recognition. In sum, the stimulus material can be considered adequate for the study of intercultural differences in emotion recognition.

Emotion Recognition Across Countries

The main diagonals of the confusion matrices for each country are reported in Table 4. The countries are ranked in the order of decreasing mean accuracy of emotion recognition. In spite of the multilanguage components of the meaningless stimulus sentences, the vocal emotion portrayals of German actors were best recognized by German judges, thus marking an estimate for the upper limit of emotion recognition. The second highest rank in recognition accuracy is obtained for the French-speaking Swiss sample, followed by Great Britain, the Netherlands, the United States, Italy, France, Spain, and Indonesia, in this order. Although the overall recognition rate of the Indonesian sample is lower than that of the other countries, it is still significantly higher than chance (kappa = .39, $p < .001$). The full confusion matrix for the Indonesian judges, presented in Table 2, Panel b, also shows that the pattern of confusion is similar to the German judges.

Pairwise t tests between the recognition accuracy coefficients of the countries studied show that only the German and Indonesian accuracy rates differ significantly ($p = .03$; one-tailed). Comparing the Indonesian data against the mean of the Western countries, excepting Germany (see Table 4), one finds a marginally significant effect ($p = .08$; one-tailed). There is no significant difference between these mean data and the German data. In consequence, the Indonesian rate of accuracy is lower than the German rate, but only marginally lower than the rate of the other Western countries.

Table 5 shows the correlations over the accuracy profiles (hit rates, i.e., the diagonal entries in the confusion matrices) for the different emotions among the countries. The uniformly high correlations (mean $r = .87$) indicate that the differential ease or difficulty of the decoding of the different emotions is highly comparable across cultures.

In addition to the similarity of the hit rate profiles over emotions, the error patterns observed in the confusion matrices are highly similar across countries. Table 6 shows the

TABLE 5
Correlations of the Accuracy Profiles Over Emotions Between Countries

	D	*CH*	*GB*	*NL*	*US*	*I*	*F*	*E*
CH	.86							
GB	.82	.95						
NL	.87	.96	.85					
US	.81	.98	.98	.88				
I	.94	.84	.75	.80	.80			
F	.92	.90	.85	.83	.89	.98		
E	.94	.96	.96	.90	.95	.91	.95	
IND	.89	.79	.72	.90	.68	.72	.68	.80

NOTE: D = Germany, CH = Switzerland, GB = Great Britain, NL = Netherlands, US = United States, I = Italy, F = France, E = Spain, IND = Indonesia.

TABLE 6
Correlations Between Countries Over Error Patterns in the Confusion Matrices

	D	*CH*	*GB*	*NL*	*US*	*I*	*F*	*E*
CH	.80							
GB	.85	.83						
NL	.90	.93	.87					
US	.85	.94	.93	.91				
I	.89	.84	.77	.88	.81			
F	.86	.92	.84	.91	.88	.88		
E	.90	.88	.92	.91	.93	.90	.88	
IND	.75	.74	.78	.76	.80	.62	.69	.78

NOTE: D = Germany, CH = Switzerland, GB = Great Britain, NL = Netherlands, US = United States, I = Italy, F = France, E = Spain, IND = Indonesia.

profile correlations for the off-diagonal entries in the confusion matrices. Again, the uniformly high correlations (mean $r = .85$) show that the respondents in the different countries tended to make the same kinds of errors. Although the error profile correlations are slightly lower for Indonesia, they are generally situated within the range of variation of the intercorrelations between the other countries.

Factors Influencing Emotion Recognition

The influence of the factors emotion and gender of actor (within participants), as well as country and gender of judge (between participants), on emotion recognition was analyzed by repeated measures ANOVA. To normalize the recognition data, hit rates were arcsine-transformed. To test a possible influence of unequal sample sizes for different countries, the same analysis was calculated with a random sample with equal cell frequency for male and female judges for each country. This analysis did not yield any substantial differences compared with using the total sample. Because the results of the total sample are more reliable due to much larger sample sizes for some countries, only these are reported.

Gender effects. The gender of judges factor yielded a small but significant main effect, $F(1, 410) = 4.31$, $p < .04$, $\eta^2 = .01$, because female judges obtained a slightly better recogni-

tion rate than male judges (67% vs. 65%, respectively). However, none of the interactions between gender of judges and other factors of the design reached statistical significance.

Among the four largest effects, we find a gender of actors main effect, $F(1, 410) = 116.07$, $p < .001$, $\eta^2 = .22$, and a Gender of Actors × Emotion interaction, $F(4, 1640) = 245.76$, $p < .001$, $\eta^2 = .37$. Given that only two male and two female actors were used, these results may reflect idiosyncratic skills or deficits of individual actors rather than gender differences in the general population of encoders. In consequence, these results will not be discussed.

Effects of emotion and country on recognition. As expected on the basis of the large differences in recognition rates between emotions, ranging from 76% for anger to 42% for joy, we find a strong main effect for the factor emotion, $F(4, 1640) = 222.72$, $p < .001$, $\eta^2 = .35$. This confirms the well-established finding that recognition accuracy depends largely on the emotion category concerned.

The most important goal of the present study was an empirical test of whether the recognition of vocal emotion portrayals is universal or culture dependent. The considerable main effect of the country factor, $F(8, 410) = 17.68$, $p < .001$, $\eta^2 = .26$, provides clear evidence that there is a substantial country effect. In addition, three interactions of the country factor reached significance: Country × Emotion, $F(32, 1640) = 5.13$, $p < .001$, $\eta^2 = .09$; Country × Gender of Actors, $F(8, 410) = 3.24$, $p < .01$, $\eta^2 = .06$; and Country × Emotion × Gender of Actors, $F(32, 1640) = 2.85$, $p < .001$, $\eta^2 = .05$. Although these interactions indicate a small but reliable cultural influence on the effect of the emotion and gender of actor factors, they are not readily interpretable and will thus not be discussed in detail. All remaining effects did not approach statistical significance.

DISCUSSION

The primary finding of this study, supporting the formal hypothesis, is that judges in nine countries, speaking different languages, can infer four different emotions and neutral state from vocal portrayals using content-free speech with a degree of accuracy that is much better than chance. In addition, differences in hit rates across emotions and the error patterns in the confusion matrices are highly similar across all countries.

As in the case of the parallel work on the intercultural recognizability of facial expressions of emotions, critics might claim that the results could be due to the effects of Hollywood films on the familiarity of particular types of expression (Carroll & Russell, 1997; Mead, 1975). However, because German radio actors were used, it is difficult to imagine that their typical expression patterns have been widely disseminated to other countries. Even the idea that the German actors might have been influenced by their Hollywood counterparts is not very plausible in the case of vocal expression: Because foreign films are almost always dubbed in Germany, the actors are likely to have only heard other German actors speak in Hollywood films.

Critics might also point out potential artifacts due to the experimental design of the study or the use of verbal labels in the recognition task (Russell, 1994; but see Ekman, 1994). Because judges had to decide between a fairly small number of categories, one could argue that recognition is due to a strategy of excluding nonplausible alternatives, for example, positive versus negative emotions, and therefore is not representative of real-life emotion decoding. However, if this were the case the one positive emotion, joy, should have a high recognition rate because it should be possible to identify it as being different from negative

emotions. The fact that joy has the lowest recognition rate militates against this explanation. Furthermore, our data clearly show that the use of fixed-answer alternatives does not preclude finding sizable country effects.

Another potential artifact might be due to the speechlike segments in the content-free voice samples being interpreted differently by judges in countries with different languages, based on perceived semantic similarity (e.g., *jonkil* reminding English speakers of *kill*, *got* reminding German speakers of *Gott* [god], etc.). However, if such effects existed they should reduce rather than increase the similarity of the results between countries. It is unlikely that the differences in the Indonesian data are due to such semantic interpretations, given the dissimilarity of the language of the Indonesian judges from the Indo-European language families (upon which the construction of the speech material was based).

In consequence, until there is more convincing evidence for the operation of artifacts, the finding that judges in different countries tend to interpret actors' vocal portrayals of emotion in a highly similar fashion seems quite robust.

The aims of the study included an exploratory analysis of country effects on the recognition accuracy. A number of such effects, of small to medium magnitude, were found but have proven to be difficult to interpret (as in many other cross-cultural studies, especially when convenience samples rather than systematically planned culture comparisons are used; see Mesquita et al., 1997; Scherer, 1997, for a discussion of the absence of an appropriate "culture comparison grid"). The factor that seemed to receive the most support from the data was language differences—with the exception of the Swiss sample,[6] the rank order of countries with respect to overall recognition accuracy mirrors exactly the decreasing similarity of languages. The languages of the countries with the highest accuracy rates are of Germanic origin (Dutch and English), followed by Romance languages (Italian, French, and Spanish). However, given the small and nonsignificant differences between these countries, one would want to replicate these results before interpreting a potential distance effect with respect to the language of origin of the voice samples. The lowest recognition rate was obtained for the only country studied that does not belong to the Indo-European language family: Indonesia. This difference is significant with respect to the German data and marginally different with respect to the mean of the other Western countries. However, interestingly enough, the profile correlations across emotion-specific hit rates and error patterns show a high degree of similarity with the patterns in the other countries.

If the present data suggest that the language difference between encoders and decoders may play a major role, the nature of the mechanism is not clear. It seems that as soon as vocal expressions, other than pure nonlinguistic affect bursts, are used,[7] segmental and suprasegmental aspects of language affect encoding and decoding of emotion. Because we used content-free utterances composed of phonological units from many different Indo-European languages, effects must be due either to segmental information (such as phoneme-specific fundamental frequency, articulation differences, formant structure, or the like) or to suprasegmental parameters, such as prosodic cues (intonation, rhythm, timing). To answer these questions, a much more ambitious experimental design has to be used. Rather than using encoders from only one language and culture, encoders and decoders from several different countries would need to be studied, allowing the construction of an encoder-decoder-emotion matrix and to test whether decoders from the countries involved would recognize emotion portrayals by encoders from their own countries most accurately. To understand the nature of these effects, more widely different language families will need to be studied in further work, to allow systematic comparisons of the distinctive features of these languages. Only systematic sampling of languages and cultures, attempting to clearly

characterize both with the help of a system of structural descriptors, will allow one to clarify the joint effects of language and culture. It will be necessary to overcome the logistic and financial limitations, as encountered by the present study, to carry out this type of research.

Obviously, in addition to language dissimilarity, other factors may be involved once one moves away from Western cultures. Among the possibilities are lower familiarity with the type of judgment procedures used in this study and nuances in the translation of the instructions. Alternatively, one might expect a greater role of culture-specific expressions and/or display rules. Clearly, one possibility is also that the nature of the emotions or the underlying mechanisms really differ between cultures, as held by cultural relativists.

One of the most interesting questions arising from the demonstration of robust, universal inference rules concerns the mechanisms underlying such similarities. They may be sought in the existence of universal encoding relationships, based on emotion-specific physiological patterning affecting voice production, as argued by many psychobiologically minded researchers assuming that there is a core of universal meaning in expressive signals of emotion. Unfortunately, because only encoders from one single culture (country and language) were used we cannot draw any conclusions with respect to the universality of vocal emotion *encoding*. Given the general lack of appropriately designed studies, it cannot be excluded at the present time that the cross-culturally stable emotion inferences from vocal expression are due to learning or other cultural transmission techniques.

Yet a strong case can be made for phylogenetic continuity and universality of the vocal expression of emotion. Some of the acoustic features of emotional-motivational signals are quite comparable across a wide variety of mammals, including humans (Morton, 1977; Scherer, 1985, 1991, 1994; Scherer & Kappas, 1988). If this is indeed so, the vocal expression of emotion can be expected to be universal, comparable, and recognizable across human cultures.

Some evidence for this idea is provided by the results of a large number of studies in different countries that have consistently shown correlations between configurations of specific acoustic cues and certain emotional states (Bachorowski & Owren, 1995; Cosmides, 1983; Frick, 1985; Murray & Arnott, 1993; Pittam & Scherer, 1993; Scherer, 1979; Williams & Stevens, 1972). Scherer (1986) has presented a theoretical model linking the physiological changes that underlie certain emotional states to specific patterns of acoustic parameter changes, allowing the explanation of earlier findings and the suggestion of specific predictions for future work in this area (see also Banse & Scherer, 1996).

The results reported here closely parallel the findings reported for the high intercultural recognizability of facial expression, even to the point of showing a comparable size of the overall recognition accuracy. However, there are some important differences with respect to the ease with which different individual emotions can be recognized.

As mentioned above, several studies have shown that joy is a difficult emotion to recognize from the voice, whereas it is by far the best recognized in studies of facial expression (see Ekman, 1994). Conversely, although anger is often badly recognized from the face, it reaches the highest accuracy percentage in the present study. One of the reasons may be that happiness or joy is strongly marked by smiling, the ubiquitous activity of the zygomaticus muscle, in joy-related emotions. No comparably iconic cue may exist for the voice. Rather, changes in parameters such as loudness and fundamental frequency, which are among the most attention-getting vocal cues, may be quite similar in intense joy, fear, and anger (due to high arousal), possibly explaining the patterns of confusion.

Apart from the intrinsic interest of the data reported, this article demonstrates the utility of approaching the universality of expression debate from several angles—taking into account

the multiple specificities of the respective response domains. The phenomenon of vocal expression is of particular interest in this respect, given its likely roots in nonhuman primate vocalizations and the extraordinary variability of language and communication systems that have evolved in different cultures.

NOTES

1. The relatively low criterion of "at least 2" had to be chosen because some of the stimuli were judged to be of low intensity, in spite of being completely unambiguous. However, the large majority of stimuli selected received much higher ratings for the primary emotion.

2. The first author left the German university where the study originated before all data could be collected.

3. For calculating kappa, double codings might have been treated as categories of their own. However, because no emotional blends were encoded by the actors, all double codings would fall outside of the main diagonal and be considered as incorrect. Another possibility would have been to consider the judgment as correct when the encoded emotion was one of the two emotions indicated by the participant. However, this might have inflated the accuracy figure. We chose the conservative option, calculating kappa on the basis of weighted judgments. This means that double codings that included the intended emotion category were treated as partially correct.

4. This table also contains the confusion matrix for Indonesia, which presents the greatest differences from the German one.

5. Given the thorny methodological problems concerning comparisons of accuracy coefficients to expected percentages for chance or guessing (see Banse & Scherer, 1996, for a detailed discussion), we used Cohen's kappa to test for significant departures of the obtained accuracy coefficients from chance.

6. The one apparent exception to this pattern, the French-speaking Swiss sample, requires further consideration. Switzerland has three official languages (German, French, and Italian), German being spoken by a two-thirds majority. Most students of the University of Geneva have French or Italian as their native language. However, familiarity with German is much higher than for the French or Italian samples. All Swiss university students whose mother tongue is not German are obliged to attend German classes by the age of 11. The multilingual context offers many opportunities to come into contact with the German language and most non-German-speaking Swiss have had extensive exposure to spoken German in the media and in interpersonal encounters. Therefore, it is plausible that the Swiss sample is an exception from a general rule of a facilitating effect of mother language similarity for the decoding of vocally expressed emotions.

7. Even in the case of nonlinguistic affect bursts or interjections, one may find sizable language effects as shown early on by Wundt (1900; see Scherer, 1994).

REFERENCES

Albas, D. C., McCluskey, K. W., & Albas, C. A. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology*, *7*, 481-489.

Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, *6*(4), 219-224.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*, 614-636.

Beier, E. G., & Zautra, A. J. (1972). Identification of vocal communication of emotions across cultures. *Journal of Consulting and Clinical Psychology*, *39*, 166.

Carroll, J. M., & Russell, J. A. (1997). Facial expressions in Hollywood's portrayal of emotion. *Journal of Personality and Social Psychology*, *72*, 164-176.

Cosmides, L. (1983). Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 864-881.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: Murray. (Reprinted 1965, Chicago: University of Chicago Press)

Ekman, P. (Ed.). (1982). *Emotion in the human face* (2nd ed.). Cambridge, UK: Cambridge University Press.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, *115*, 268-287.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*, 124-129.
Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, *97*, 412-429.
Izard, C. E. (1977). *Human emotions*. New York: Plenum.
Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, *115*, 288-299.
Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *Journal of Abnormal and Social Psychology*, *68*, 390-396.
Lonner, W. J., & Berry, J. W. (1986). Sampling and surveying. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 85-110). Beverly Hills, CA: Sage.
McCluskey, K. W., & Albas, D. C. (1981). Perception of the emotional content of speech by Canadian and Mexican children, adolescents, and adults. *International Journal of Psychology*, *16*, 119-132.
McCluskey, K. W., Albas, D. C., Niemi, R. R., Cuevas, C., & Ferrer, C. A. (1975). Cross-cultural differences in the perception of emotional content of speech: A study of the development of sensitivity in Canadian and Mexican children. *Developmental Psychology*, *11*, 15-21.
Mead, M. (1975). Review of "P. Ekman: Darwin and facial expression." *Journal of Communication*, *25*, 209-213.
Mesquita, B., Frijda, N. H., & Scherer, K. R. (1997). Culture and emotion. In J. E. Berry, P. B. Dasen, & T. S. Saraswathi (Eds.), *Handbook of cross-cultural psychology: Vol. 2. Basic processes and developmental psychology* (pp. 255-297). Boston: Allyn & Bacon.
Morton, E. S. (1977). On the occurrence and significance of motivational-structural rules in some bird and mammal sounds. *American Naturalist*, *111*, 855-869.
Murray, I. R., & Arnott, J. L. (1993). Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, *93*, 1097-1108.
Pittam, J., & Scherer, K. R. (1993). Vocal expression and communication of emotion. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 185-198). New York: Guilford.
Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore: Johns Hopkins University Press.
Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*, 102-141.
Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In C. E. Izard (Ed.), *Emotions in personality and psychopathology* (pp. 493-529). New York: Plenum.
Scherer, K. R. (1985). Vocal affect signalling: A comparative approach. In J. Rosenblatt, C. Beer, M. Busnel, & P. J. B. Slater (Eds.), *Advances in the study of behavior* (pp. 189-244). New York: Academic Press.
Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*, 143-165.
Scherer, K. R. (1991). Emotion expression in speech and music. In J. Sundberg, L. Nord, & R. Carlson (Eds.), *Music, language, speech, and brain* (pp. 146-156). Wenner-Gren Center International Symposium series. London: Macmillan.
Scherer, K. R. (1994). Affect bursts. In S. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 161-196). Hillsdale, NJ: Lawrence Erlbaum.
Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology*, *73*, 902-922.
Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, *15*, 123-148.
Scherer, K. R., & Kappas, A. (1988). Primate vocal expression of affective states. In D. Todt, P. Goedeking, & E. Newman (Eds.), *Primate vocal communication* (pp. 171-194). Heidelberg, Germany: Springer.
Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, *66*, 310-328.
Scherer, K. R., Wallbott, H. G., & Summerfield, A. B. (Eds.). (1986). *Experiencing emotion: A cross-cultural study.* Cambridge, UK: Cambridge University Press.
van Bezooijen, R., Otto, S. A., & Heenan, T. A. (1983). Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology*, *14*, 387-406.
Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, *51*, 690-699.
Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, *52*, 1238-1250.
Wundt, W. (1900). *Völkerpsychologie. Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythos, und Sitte. Band I. Die Sprache* [Cultural psychology: A study of the developmental laws of language, myth, and customs: Vol. 1. Language]. Leipzig, Germany: Kröner.

Klaus R. Scherer holds a chair of emotion psychology at the University of Geneva. After studying at the University of Cologne and the London School of Economics, he obtained his Ph.D. in the Social Relations Department at Harvard University. He has taught at the University of Pennsylvania and at the University of Kiel and University of Giessen in Germany. His research is currently focused on experimentally testing his component process model of emotion as well as on the study of vocal and facial expression in relation to psychophysiological reactions.

Rainer Banse obtained his master's degree in psychology from the University of Giessen, Germany, in 1988, and his Ph.D. from the University of Geneva, Switzerland, in 1995. Since 1995, he has been a senior researcher at the Humboldt University Berlin, Germany. His current research interests are emotion psychology, evolutionary psychology, and implicit and explicit cognitive representations of attitudes, persons, and relationships.

Harald G. Wallbott obtained his Ph.D. in 1982 from the University of Giessen, Germany. He is currently a full professor of social and applied psychology at the University of Salzburg in Austria (since 1994). His research interests include nonverbal communication, emotion, impression formation, psychology of the media, new communication media, cross-cultural psychology (especially methodological aspects and the study of the recognition of emotion).