Vocal Expression of Emotion: Comparisons of Judge Accuracy, Linguistic Properties, Subjective Reactions, and Acoustical Analyses

Bruce L. Brown
Stanley Feldstein
Daniel Sanderson
Gail Caldwell

Follow this and additional works at: https://scholarsarchive.byu.edu/dlls

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/dlls/vol13/iss1/4

This Article is brought to you for free and open access by the All Journals at BYU ScholarsArchive. It has been accepted for inclusion in Deseret Language and Linguistic Society Symposium by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Vocal Expression of Emotion: Comparisons of Judge Accuracy, Linguistic Properties, Subjective Reactions, and Acoustical Analyses

Bruce L. Brown
Brigham Young University

Stanley Feldstein
University of Maryland Baltimore County

Daniel Sanderson and Gail Caldwell
Brigham Young University

There has been a considerable amount of work in the past twenty-five years on the topic of vocal expression of emotion, as evidenced by a comparison of Kramer’s (1963) classic review and Scherer’s (1986) review twenty-three years later of studies in that area. However, there is much left to be done, and in some ways we are only now beginning to enter an era of extensive careful emotion research. Most of the work has been correlative, and much of it has been crude, loose and somewhat artificial. There are still many more questions than firm answers. Few studies have made use of the precise acoustical analysis methods that have been developed in the past thirty years. Even those that have are lacking either in terms of having a limited sample of vocal emotion (Williams and Stevens, 1972), or only the most rudimentary use of acoustic methods (Van Bezooijen, 1984).

In this paper we will give a very brief preliminary report of a series of studies that employ a very extensive and expensive set of recorded emotion portrayals that was produced by Feldstein in the early 1960s. This set consists of six professional actors’ and six professional actresses’ portrayals of eight emotions each (anger, fear, joy, sadness, depression, hate, nervousness, and neutral) in a naturally-sounding and lengthy monologue. These 96 emotion portrayals (twelve actor/actresses times eight emotions) are superior to those used in previously reported studies in terms of skill of the actors/actresses, quality of the recordings themselves, length of the monologue, and "believability" (how realistic the emotions sound).

With respect to believability, it is important to note that when subjects in our studies listen to one of the emotion portrayals, they invariably think that it is a recording of natural speech. We have used a single emotion testing paradigm that takes advantage of the naturalness and believability of these recordings. That is, since exactly the same monologue is used in each emotion by each actor/actress, a listener subject would immediately know that the recording is not natural if he or she were to listen to more than one emotion and make comparisons. In our single emotion studies, we typically have five subjects assigned to listen to one of the emotion portrayals and then respond to that portrayal on a computer terminal in six ways: (1) the subject describes the speaker’s emotion in his or her own words, (2) the subject chooses which of eight emotions this one is closest to, (3) the subject listens to the recording again and rates the utterance on each of eight emotion scales according to how much of
each emotion is judged to be present, (4) the subject listens to the recording a third time and gives a description in his or her own words of what the speaker is like as a person, (5) the subject rates the speaker on fourteen adjective scales descriptive of personality trait characteristics, (6) the subject listens to the recording a fourth time and gives a description of the vocal properties of the utterance, (7) the subject rates the utterance on twelve adjective scales descriptive of vocal properties.

This produces a great deal of information of a realistic kind (since the recording seems to subjects to be natural rather than portrayed emotion), but it also is expensive in terms of requiring many subjects. Rather than using all 96 emotion portrayals, we used a subset of 64, but with an average of five subjects to each rating group, this still required 320 subjects. Also, in order to test the effect of attributional context we did this under two conditions, one in which subjects are told that they are hearing an excerpt of a psychotherapy interview, and one in which they are told that they are hearing an excerpt of a conversation in a New York City bar. This requires 640 listener subjects, 320 for each of the attributional contexts.

In addition to the single emotion rating studies, we have also used the tape in comparative emotion studies, where a typical subject only has two tasks: (1) to guess the emotion being portrayed, and (2) to rate the portrayal for how much of each of the eight emotions is present. In this paradigm, each subject listens to many portrayals and can thus give comparative ratings with a common baseline. Obviously this common baseline advantage comes with some costs. The subjects are immediately aware that these are portrayals rather than natural emotion, since the monologue is the same for each speaker and for each emotion portrayal.

In addition to the single-emotion paradigm studies and the comparative-emotion paradigm studies using American listener subjects responding to these American actors/actresses, we have also gathered cross-cultural data. We have used the comparative emotion paradigm to examine listener accuracy in identifying emotions and to gather listener impressions both with a Latin American group of listener subjects (Paraguay) and also with an Oriental group of listener subjects (Taiwan). We have also gathered some single emotion paradigm data in Paraguay.

In summary, this series of studies is producing five major kinds of data: single-emotion paradigm subjective ratings data, comparative-emotion paradigm subjective ratings data, cross-cultural subjective ratings data, and acoustic analysis data. In the remainder of this paper we will report some of the results of the first of the five, the single-emotion paradigm subjective ratings data, and describe somewhat the ways in which the acoustical analysis is proceeding. However, first, we will present a secondary analysis of some important recently gathered cross-cultural data on self reports of emotional experience that will provide a context for our results.
About five years ago, Klaus Scherer and Harold Wallbott and their colleagues began a cross-cultural questionnaire of the subjective reminiscences subjects have of emotional experiences. These data are some of the most important ever gathered on emotion and will provide a foundation for much of the emotion work to be done in the coming decades. Essentially, they had each subject identify in his or her life an experience that typifies each of seven emotions (joy, fear, anger, sadness, disgust, shame, and guilt). Each subject then wrote a brief description of each of the seven experiences and made questionnaire ratings of each experience in seven of the eight areas shown in Figure 1: experience/control, physiological symptoms, nonverbal reactions, verbal behavior, situation, responsibility, and coping. (The eighth area in Figure 1, reactions/symptoms is a comparative summary of the information from the physiological symptoms area and that from the nonverbal reactions area.) Figure 1 was produced by our secondary analysis, using a principal components biplot, of the information reported by Wallbott and Scherer (1986, Tables 3, 4 and 5). (See Brown, Williams and Barlow, 1984, Brown and Walters, 1987, and Brown, Williams, Norton and Barrus, this volume, for explanations and examples of the biplot method.)

The first thing to notice from this figure is that the semantic space for emotions changes considerably depending which aspect of emotion one is considering. For example, when one is considering the intensity, duration, control, frequency of occurrence, etc. of emotions (the first biplot in Figure 1), we get quadrilateral extreme points of joy (hi intensity/duration, hi frequency), sadness (hi intensity/duration, lo frequency), shame (lo intensity/duration, lo frequency), and disgust (lo intensity/duration, hi frequency). The vertical dimension is intensity/duration, with those emotions at the top of the biplot being high intensity/duration and those at the bottom being low in intensity duration. The horizontal dimension is frequency (how often each occurs, inferred from average time since occurrence) with the emotions on the right (anger, disgust and joy) being more frequent than those on the left (sadness, fear and shame). Notice that the "hide/control" dimension extends from upper right to lower left, with joy being very low in hide/control, but shame being very high.

For physiological symptoms we get a triad grouping with joy being typified by warm, pleasant bodily symptoms; fear by perspiring, cold, shivering and heart-rate/breathing changes; and sadness and disgust by tense, trembling muscles, lump in throat and stomach troubles. On the other hand, nonverbal reactions have a triad of joy, anger and sadness. Joy is typified by the "moving toward" others kinds of nonverbal gestures and laughing. Anger is typified by "moving against" gestures with screaming/yelling, abrupt bodily movements, gesture changes, voice changes, and facial expression changes. Sadness is typified by withdrawing, "moving away from" kinds of gestures and crying/sobbing. The reactions/symptoms biplot (number 5 in Figure 1) is a comparison of the seven emotions in terms of whether they tend to be typified more by physiological symptoms (more involuntary, or by nonverbal reactions (more voluntary). As might be expected, anger is the emotion most typified by the more voluntary nonverbal reactions (followed closely by joy) and fear is the emotion most typified by the more involuntary
Figure 1. Eight biplots comparing emotion self-report responses across eight domains: a principal components secondary analysis of Wallbott and Scherer's (1986) data from 2235 respondents in 27 countries.
Figure 1 continued.
physiological symptoms.

For situation (biplot number 6 in Figure 1) we again get a triad of emotion types, with joy being typified by expected, pleasant situations of positive self-esteem/confidence. Anger and fear are typified by unfair situations in which personal relationships change negatively and plans are hindered. Shame, guilt, disgust and sadness are typified by unexpected, unpleasant situations in which behavioral is seen as immoral and self-esteem/confidence is lost.

Very related to this is the question of responsibility (biplot 7 of Figure 1), and again we have a triad, but a very different one with shame and guilt in one prong of the triad (typified by "self" responsibility), anger and disgust in another prong (typified by "other people" being responsibility, including a potpourri all the way from close friends/relatives to "the authorities), and fear and sadness in the lower prong (typified by nonhuman causes). Notice that joy is not really in any of these three responsibility "prongs".

The question of how one copes with each emotion (biplot 8 of Figure 1) is also somewhat triadic in arrangement. Joy is by itself with "no action necessary." Guilt, shame and anger are off to the left where the person believes he or she could positively influence, or escape, or "pretend nothing important had happened." Sadness and fear are down toward the bottom where the person saw himself or herself as dominated and powerless to do anything.

The only biplot in Figure 1 not yet discussed is biplot 4, verbal behavior, the one potentially most relevant to our emotion/speech project. This one might be easier to interpret if it were rotated 45 degrees counterclockwise so that "lengthy utterances" (which is at about 2:00) would be at the top. Notice that we get two rather major dimensions defining this space. The first is the "lengthy utterances" versus "silence" dimension extending from the top right to the bottom left. Anger and joy are typified by lengthy utterances and all of the others (but especially fear and sadness) by more silence or short utterances. The second dimension, perpendicular to it, and extending from top left to bottom right, is a speech disturbances (typifying shame and anger) as opposed to an implied freedom from speech disturbances (typifying disgust and joy, both of which are very much under control).

These biplots of Figure 1 summarize an incredible amount of information. We computed them on the basis of Wallbott and Scherer’s reported means and proportions from a total of 2235 subjects, and average of about 83 subjects each from each of 26 nations. But the most exciting and important data from the standpoint of our cross-cultural study of mutual intelligibility of emotion portrayals will be the biplots that are constructed to show the contrasts among the 26 nations in their typifying of the seven emotions in each of these eight domains. That will truly create an invaluable foundation for understanding the mutual intelligibility of emotions across cultures. But even at this point, we are now prepared to appreciate that the emotions differ from one another in subtle and multidimensional ways.
Figure 2 presents the data from our single-emotion paradigm subjective ratings study. Shown are the average ratings given to the actors (the "male" end of each connected triad) and the actresses (the "female" end of each connected triad) on each of the eight emotions portrayed. The unfilled circle between these two is their average, that is, the average for each emotion regardless of whether it is portrayed by a male or a female. The three biplots correspond to three kinds of ratings: ratings of vocal properties, personality trait ratings of the speaker, and ratings of each utterance according to how much of each of the seven emotions is judged to be present.

The vertical dimension on the vocal properties semantic space has to do with how dynamic the voice sounds, with the top of the biplot being typified by articulate, loud, fast, and high-pitched speech (primarily anger, but to a lesser extent joy and fear), and the bottom of the biplot being typified by slurred, soft, slow, and low-pitched speech. The horizontal dimension is primarily a reflection of breath control problems at the left (fear and nervousness) as opposed to the more controlled, full and fluent voice at the right (neutral and hate and male joy). It is interesting that male joy, male sadness, and male depression (as well as the male average in general) is more free of the breath control problems than the corresponding female emotions, but for anger this is reversed with female being more free of breath control problems. It is also interesting to notice that hate in this biplot is right in the middle of a cluster with joy and neutral. Does that mean that hate sounds like joy and neutral emotion to subjects. Most certainly not. It is the case that they are not readily distinguishable on the particular vocal properties scales used for this biplot. Our challenge is to find the dimensions (considered either vocally or acoustically) that do differentiate them. Notice in the personality trait ratings biplot that hate is at opposite ends of the semantic space from joy and neutral, showing that subjects can certainly differentiate them when given the right rating scales to do so.

The vertical dimension in the personality trait ratings space is primarily speaker competence, with positive competence at the top (strong, self-assured, intelligent, etc.) and negative competence at the bottom (weak, self-doubting, unintelligent, but guileless) at the bottom. The horizontal dimension reflects benevolence/kindness/goodness, with the positive end of this dimension at the right (kind, polite, flexible, likeable, placid) and the negative end at the left (unkind, impolite, rigid, unlikeable, temperamental). Of course, the portrayals of hate and anger are at the extreme left, the negative end of this dimension, and neutral and joy are at the right. With respect to the vertical dimension, a positive competence impression is given by the joy, the neutral and the anger portrayals, while a negative competence impression is given by the nervous, depression, sadness and fear portrayals.

The last semantic space to be considered is a curious one. It deals with the question of how much of the eight emotions is judged to be present in each of the others. That is, how much "crosstalk" is
Figure 2: Three biplots comparing male and female speakers in the received ratings of their eight emotion portrayals in each of three domains: vocal properties ratings, personality trait ratings and ratings on emotion scales.
there among the emotions. We are not surprised to see that joy and neutral are together and are at the end of each of the other dimensions that reflects little of any of the negative emotions. We are also not surprised to see depression and sadness together and opposite to joy. But the third corner of the triad is an interesting one. Not only are anger and hate together, but fear is with them, with all three being typified by the highest ratings on four dimensions: much anger, much fear, much hate, and much nervousness. That means that to these listener subjects, the fear portrayal seems to also have quite a bit of anger and hate in it; and the anger portrayal seems to also have quite a bit of nervousness and fear in it, etc. Note that nervousness is midway between the sadness/depression cluster and the anger/fear/hate cluster.

We can now clearly see the difficulty in linguistically, acoustically, and semantically/subjectively typifying the emotions. It is like the old story of the blind men describing the elephant (the one touching the tail says an elephant is like a rope, the one touching the a leg says that an elephant is like a tree trunk, etc.) What we find linguistically, subjectively and acoustically is very much tied to the dimensions of focus that we choose. But, being basically optimistic types, we will go ahead anyway.

As a first pass on the acoustic analysis of these tapes we are making long term power spectra of each emotion portrayal for each of the actors and actresses. We will also compute straightforward dynamic properties such as mean fundamental frequency, variance of fundamental frequency, mean amplitude and variance of amplitude. We also have a couple of more exotic methods we plan to try. It may be a long shot, but we think that there may be systematic differences in the way steady state vowels are pronounced in portraying each emotion: more closed in the sadness/depression cluster and more open in anger and joy and perhaps even fear. We think that this will show up in Labovian type F1 by F2 plots (see for example, Labov, 1972, p. 76). The multivariate graphical methods we have developed over the past twenty years (Brown, Williams and Barlow, 1984; Brown and Walters, 1987), of which the biplots of Figures 1 and 2 are an example, will be used to create holistic spatial representations of each of the total pattern of measured acoustic dimensions. We also are beginning work with a new multivariate graphical technique developed by Clifford A. Pickover (see a description by Peterson, 1987), that uses a lattice walk procedure for converting waveforms into six-petalled flower-like symmetric dot patterns that are amazingly subtle in differentiating the acoustic properties of sounds and making similarities in pattern visually accessible.

Perhaps a year from now we will know a great deal more than we know now about the acoustic concomitants of emotion and their subjective significance cross-culturally.
References


