Published in final edited form as:

Behav Res Methods. 2008 November ; 40(4): 935–939. doi:10.3758/BRM.40.4.935.

Validation of affective and neutral sentence content for prosodic testing

Jeff B. Russ, **Ruben C. Gur**, **Warren B. Bilker**

University of Pennsylvania Medical Center, Philadelphia, Pennsylvania

Abstract

Conducting a study of emotional prosody often requires that one have a valid set of stimuli for assessing perceived emotion in vocal intonation. In this study, we created a list of sentences with both affective and neutral content, and then validated them against rater opinion. Participants read sentences with content that implied happiness, sadness, anger, fear, or neutrality and rated how well they could imagine each sentence being expressed in each emotion. Coefficients of variation and intraclass correlations were calculated to narrow the list to affective sentences that had high agreement and neutral sentences that had low agreement. We found that raters could easily identify most emotional content and did not ascribe any unique emotion to most neutral content. We also found differences between the intensity of male and female ratings. The final list of sentences is available on the Internet (www.med.upenn.edu/bbl/) and can be recorded for use as stimuli for prosodic studies.

Currently, the field of emotional prosody is relatively underexplored. In a review of the literature on emotion in the face and voice, Edwards, Jackson, and Pattison (2002) divided prosodic research into two main streams: acoustic analysis and cerebral lateralization for prosodic comprehension. Acoustic analysis typically involves measuring various parameters of an audio recording to determine the aspects that convey an emotion. Studies of cerebral lateralization typically include audio stimuli for the activation of brain regions that may be associated with prosodic comprehension. Studies of prosody have also been divided on the basis of whether they measure receptive prosody (the ability to interpret the emotional intonations of others) or expressive prosody (the ability to produce an intended emotion through one's own voice) (Leitman et al., 2005). Emotional prosody, defined as "that faculty of speech which conveys different shades of meaning by means of variations in stress and pitch—irrespective of the words and grammatical construction" (Monrad-Krohn, 1947), needs to be divorced from sentence content to ensure that the emotion being communicated results directly from the intonation rather than from the connotation of the words involved. Separating sentence content from prosody can be done in one of two ways. One can record sentences whose content has been confirmed as neutral, and then vary the expressed emotional prosody. Alternatively, one can record validated sentences with affective content, using a congruent emotional prosody, and then vary the expressed prosody with an incongruent emotion for comparison. The first method is useful for emotion-perception and -

Correspondence should be addressed to R. C. Gur, Brain Behavior Center, University of Pennsylvania Medical School, 10th Floor Gates Building, 3400 Spruce Street, Philadelphia, PA 19104 (gur@upenn.edu).

identification tasks. The second is useful for functional neuroimaging studies, such as in the prosodic fMRI study by Mitchell, Elliott, Barry, Cruttenden, and Woodruff (2003). Both methods require the development and validation of a standard set of audio cues that can be recorded with varying intonations and used to test vocal parameters or prosodic response.

In order to establish a method for measuring the emotional content of a sentence, we asked 12 participants to rate a list of sentences for emotional content, and then calculated the coefficients of variation (CVs) and intraclass correlations (ICCs). Because the definition of affective content is inherently subjective, it was necessary to rely on rater agreement and to quantify its likely variation—thus our use of these measurements. CVs, calculated as the variance divided by the mean, are measures of intersubject reliability. A low CV, therefore, indicates a low degree of variation between subjects relative to the mean (Gomez & Gomez, 1984). We hypothesized that a sentence with content that conveyed a specific emotion would show a significantly smaller CV when it was rated for the content of that emotion than it would when rated for the content of other emotions. By contrast, we did not expect neutral sentences to show any significant difference between CVs for any of the emotional headings, thus indicating that the sentence does not convey any one emotion in particular. It was possible that a sentence that was rated for content in an unintended emotion would also show low variation, and a low CV, if raters agreed that the sentence was clearly not an example of that emotion. For instance, a sentence with happy semantic content might be given unanimously low ratings when it is imagined in an angry context, though we did not encounter this effect in the present study. This problem could be remedied by assessing differences between mean ratings for each emotional heading, using the data in Table 1. For this reason, CVs alone cannot validate emotional content. The concurrent use of ICCs, which also take into account the strength of the rating, can validate the CV results. ICCs, which are similar to Cronbach's coefficient alpha, are also measures of participant agreement that take into account the strength of interrater agreement. For these calculations, we used ICC(2,1), a special calculation of the ICC to be used when raters are selected *randomly* from all possible raters in a population of raters of interest. A high ICC indicates high, strong participant agreement. We hypothesized that we would find a high average ICC for each category of affective sentences, indicating strong participant agreement. We expected neutral sentences to show low ICCs, however, indicating little agreement that any particular emotion was clear. Finally, we also addressed whether differences existed between male and female ratings.

METHOD

First, 130 different sentences were created and compiled for rating; 30 of the sentences were intended to be neutral (i.e., when drafted, these sentences were supposed to contain no emotional content, such as emotionally relevant words); 100 sentences were intended to be affective—25 sentences in each of four categories (happy, sad, angry, and fearful). These four emotions were selected as four of the six universally recognized emotions that were described by Ekman and colleagues (Ekman 1994; Ekman et al., 1987). (Surprise and disgust were deemed difficult to convey or to recognize in the voice.) "Emotion" words were included in the affective sentences to convey a particular emotion through the sentence content. Emotionally charged sentences, for example, included a relative of one of the four

basic emotions ("I *love* spending time with you"). Also, the sentence might have described a scenario that is often associated with one of the emotions ("They say he was a murderer"). For more examples, see the final set of selected sentences in the Appendix.

The affective sentences were mixed and were presented to participants separate from the neutral sentences. Twelve participants (6 male and 6 female) were chosen to review the sentences. All were undergraduate students, all were fluent in English, and all were committed to being conscientious and diligent in their ratings. Students were given as much time as they needed to complete their ratings, and their commitment to the task was indicated by the absence of missing values and the high degree of rater agreement in the results. All sentences used basic English, and it was assumed that all words were understandable to undergraduates. Participants were asked to read each sentence and then rate, on a scale of 1 to 10, how well they could imagine the sentence being spoken in happy, sad, angry, and fearful contexts. It is important to note that participants were not provided any situational context for the emotional content of each sentence. Instead, participants were asked to imagine the sentence being expressed in their own hypothetical emotional context, in order to determine whether the content exemplified the given emotion. If the participant could very easily imagine the sentence being expressed in the context of one of the emotional headings, they were to rate the sentence as a 10 for that emotion. If the sentence seemed to fit very poorly under an emotional heading, they were to rate the sentence as a 1 for that emotion. If they felt that the emotion fit somewhere in between, they were to assign an appropriate number between 1 and 10.

The ratings for each sentence in each of the four possible emotional expressions were averaged. The mean and standard deviation of the ratings for each group of sentences under each emotional heading are shown in Table 1. Because some sentences were given their highest rating for an emotion that was not intended when they were written, sentences were grouped according to the emotion with the highest average rating. This arrangement also prevented the error that might have arisen from grouping sentences according to participant agreement: the incorrect placement of a sentence because participants had strongly agreed that the sentence was not in the category. Thus, each category of emotion contained only sentences that were rated highest for that particular emotion, indicating positive participant agreement for that emotion. This method of organization yielded 25 happy, 24 sad, 26 angry, 25 fearful, and 30 neutral sentences. The CVs and ICC(2,1)s were calculated for each sentence that was under each emotional heading. These parameters were then compared for each emotional category of sentence (including neutral) that was rated under each emotional heading.

Sentence selection was performed by comparing the ICCs in a similar fashion. Affective sentences with an ICC of .65 or above were deemed valid. For a neutral sentence to be considered truly neutral, there had to be little agreement among participants that any one emotion stood out within the sentence; therefore, sentences with an ICC of less than .35 were considered truly neutral, and those with an ICC above .35 were discarded. Thresholds of .65 and .35 were chosen because the difference between them was large enough to ensure that the remaining sentences belonged to one extreme or the other, yet they still left a reasonable number of each type of sentence.

For affective sentences, one-way, within-group, repeated measures ANOVAs were performed to compare the CVs of sentences that were imagined within their intended context with the average CVs of sentences rated in the hypothetical contexts of the other emotions. For neutral sentences, an ANOVA was performed to look for an effect between the CVs of each emotional heading. To compare ICCs, an ANOVA was performed to look for an effect between the ICCs of the neutral sentences and the average ICCs of the other emotional categories. This was also done to look for an effect between the ICCs of happy sentences and the average ICCs of sad, angry, and fearful sentences. Another ANOVA, sex × emotion, was performed to investigate whether there were differences in the ratings that could be attributed to the sex of the raters. A preset alpha level of significance of $p < .05$ was used. Analyses were carried out within Microsoft Excel and SAS.

RESULTS

For sentence-selection purposes, ICCs were compared and tested for significance. A repeated measures ANOVA comparing the ICCs of the neutral sentences with the average ICCs of the happy, sad, angry, and fearful sentences showed a significant difference ($p < .001$) between the neutral and affective sentences. An ANOVA comparing the ICCs of happy sentences with the average ICCs of sad, angry, and fearful sentences showed significantly higher ICCs for happy sentences. Table 2 shows the differences among the average ICCs for each category of sentence. ICC thresholds were then used for sentence selection. Selecting for affective sentences with an ICC above .65 and neutral sentences with an ICC below .35 resulted in 24 happy sentences, 18 sad sentences, 25 angry sentences, 17 fearful sentences, and 14 neutral sentences.

After sentences that did not meet the ICC thresholds had been removed, CVs for each category of emotion under each emotional heading were tested for significant differences. For happy sentences, a repeated measures ANOVA was performed that compared the CVs of happy sentences that were rated within a hypothetical happy context with the average CVs of happy sentences that were rated in sad, angry, and fearful emotional contexts. Happy sentences that were imagined in a happy context had significantly lower CVs than they did in any other context ($p < .001$). For sad sentences, an ANOVA was performed that compared CVs of sad sentences in a hypothetical sad context with the average CVs of sad sentences in every other emotional context. It was found that sad sentences that were rated in an imagined sad context had significantly lower CVs than they did in any other context ($p < .001$). Similar repeated measures ANOVAs were performed for angry and fearful sentences. It was found that angry sentences had significantly lower CVs in a hypothetical angry context than they did in any other context ($p < .001$) and that fearful sentences had significantly lower CVs in a hypothetical fearful context than they did in any other context ($p < .001$). For neutral sentences, an ANOVA was performed to look for a significant effect between CVs when they were rated for expressiveness under each emotional heading; none was found ($p = .31$). Table 3 shows the average CVs for sentences under each emotional heading and the results of the ANOVAs. It is noteworthy that CV comparisons were made after sentence selection by ICC, thus resulting in the lower number of sentences for each category.

A sex × emotion ANOVA determined that there was a main effect of rater gender [$F(1,10) = 5.54$, $p = .0404$], with females having overall higher ratings than those for males. There was also a main effect of emotion [$F(3,30) = 21.79$, $p < .0001$], indicating variability in overall intensity ratings, with the highest ratings for anger and the lowest for happiness. Finally, no interaction between sex and emotion was found [$F(3,30) = 1.96$, $p = .1414$]. The means and standard deviations of male and female ratings are given in Table 4.

DISCUSSION

We created a list of affective and neutral sentences and selected as valid sentences those that could be labeled reliably as conveying an intended emotion or neutrality. Narrowing down the sentence list to the final set of affective and neutral sentences required a comparison of the ratings of 12 participants.

Sentences in all affective categories showed acceptable CVs when they were considered in their respective hypothetical contexts. This indicates that given the chosen content of each affective sentence, there was considerable agreement among participants that the sentence actually conveyed the emotion that was intended. Neutral sentences that were considered for each affective heading showed no significant differences between average CVs, which implies that there was little agreement among participants that any one emotion stood out. Because no one particular emotion was most evident in this sentence set, it is fair to deem it neutral.

Analysis of the ICCs also showed strong participant agreement that each category of affective sentences conveyed the intended emotion. The first filtering process weeded out affective sentences with an average ICC lower than .65, thereby leaving sentences with high average ICCs. Neutral sentences showed a low average ICC, significantly lower in comparison with that for all affective sentences. This is also a result of the first filtering process, which removed all “neutral” sentences with an ICC higher than .35. The resulting low average ICC reinforces our confidence that, for the sentences that were considered validly neutral, the raters could not agree that any one emotion was prominent.

Happy sentences yielded notably higher ICCs than did sad, angry, or fearful sentences. It seems that happiness was more easily identified and agreed upon by raters than was any other emotion that was intended by sentence content. Possibly happiness is more easily recognized because it is unique in being the only positive emotion of the universal emotions. Raters who judge sentence content can very simply differentiate between positive and negative connotations, and thus can much more readily identify a happy sentence. Parallel findings regarding the recognition of happiness in facial expressions have also been demonstrated (Gosselin, Kirouac, & Doré, 1995; Sackeim, Gur, & Saucy, 1978)

We found that female participants perceived slightly, but significantly, higher emotional content than did male participants when rating the same sentences. No interaction between sex and emotion was found, so male and female participants seem to distinguish between emotional content similarly. For affective sentences, emotional content is more apparent to female participants than to male participants, which should be of little concern when these

sentences are used as audio stimuli. There could be concern that some stimuli, with content that is intended to be neutral, might appear to be more emotional to female participants. The use of neutral sentences with ICCs lower than .35 alleviated this problem, because these sentences were largely devoid of emotional content according to both male and female raters.

A limitation of this study is the small sample size of the raters. We opted for a high quality of data from a small number of motivated, highly proficient, and closely supervised participants, and we expected 12 undergraduate students to provide fairly reasonable and conscientious judgments of the basic emotions. Furthermore, we selected the sentences that the raters agreed upon most strongly. Further validation with larger samples from more diverse cultures, however, could increase confidence in the objectivity and generalizability of the emotional sentence content. Another limitation is the subjectivity involved in rating the sentences on a scale of 1 to 10. Such a scale is inherently inconsistent among raters and is difficult to standardize. This does not pose a large problem, however, because the measurement of perceived emotion is merely the amount of agreement among listeners about what emotions are being expressed.

In summary, the method described is a procedure of validating affective and neutral sentence content, using the ratings of a group of participants. The results from our sentence list showed reasonable agreement among participants and, therefore, allowed us to further validate a subset of sentences against an objective algorithm for detecting the presence of affect in our sentences.

Acknowledgments

This research was supported by a University Scholars Program Grant at the University of Pennsylvania to J.B.R. and by NIMH Grant MH 60722 to R.C.G.

APPENDIX: Selected Sentences

Neutral
I’m on my way to the meeting
I wonder what that is about
Have you seen him?
The airplane is almost full
Can you hear me?
Maybe tomorrow it will be cold
I would like a new alarm clock
Can you call me tomorrow?
I think I have a doctor’s appointment
We’ll stop in a couple of minutes
How did he know that?
Don’t forget a jacket
I think I’ve seen this before

The surface is slick

Happy

I really enjoy our family vacations

We had so much fun last night

You look so excited

What could be better?

I love spending time with you

This chicken is excellent

The new version is the best

That show makes me laugh

That magazine is my favorite

I always enjoy when she visits

That was better than the first time

You look wonderful

What joke could be funnier than that?

That was a blast

I highly recommend that professor

Isn’t that beautiful?

He’s always pleasant to be around

I look forward to meeting you

The soup is delicious

The mountains are supposed to be nice this time of year

The atmosphere there is very nice

I can’t wait to see you

I would definitely like another slice

Have you ever tasted anything better?

Sad

I’m so sorry for hurting you

He never listens to me anymore

I miss the trips we used to take

I regret that we broke up

I miss the time we spent together

It’s terrible that such a thing could happen

I can’t seem to do well on my exams

Please forgive me

If only I could go back

If only I hadn’t said those things

My best friend moved away

That movie made me cry

That newspaper article was depressing

I wish I could please them

My brother is very sick

I felt so helpless

I didn't mean to hurt your feelings

My dog died yesterday

Angry

This microwave is useless

What makes you think you can yell at me?

Why would you say such a thing?

I hate when you ignore me

What do you want from me?

She never shows up on time

I think that article was ridiculous

Quit bothering me

That class is completely worthless

Don't raise your voice at me

I'm tired of her attitude

Those prices are way too high

He's so unhelpful

Don't ever speak to me that way

I'm going to write a complaint

Why are you always testing my patience?

I never want to see you again

He always acts like he's better than everyone

I won't shop there again

Do you think you can push me around?

That was insulting

I have been waiting in line for too long

That noise is getting really annoying

What he said was very offensive

Do you know how unjust that is?

Fearful

She never returned from the campout

I was terrified that night

That man looks suspicious

I have no idea if I'm ready for the exam

She shouldn't walk there at night

These woods are creepy

I wish we didn't have to walk through the graveyard

Are you sure everything's alright?

The forest is eerie when no one's around

I think someone's in here besides us

That noise made me jump

Is someone in here?

That abandoned warehouse doesn't seem safe

I don't like the looks of that house

They say he was a murderer

It's getting dark and I can't find my way back

He said he wanted to meet but it doesn't sound good

REFERENCES

Edwards J, Jackson HJ, & Pattison PE (2002). Emotion recognition via facial expression and affective prosody in schizophrenia: A methodological review. Clinical Psychology Review, 22, 789–832. [PubMed: 12214327]

Ekman P (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. Psychological Bulletin, 115, 268–287. [PubMed: 8165272]

Ekman P, Friesen WV, O'Sullivan M, Chan A, Diacoyanni-Tarlatzis I, Heider K, et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. Journal of Personality & Social Psychology, 53, 712–717. [PubMed: 3681648]

Gomez KA, & Gomez AA (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Gosselin P, Kirouac G, & Doré FY (1995). Components and recognition of facial expression in the communication of emotion by actors. Journal of Personality & Social Psychology, 68, 83–96. [PubMed: 7861316]

Leitman DI, Foxe JJ, Butler PD, Saperstein A, Revheim N, & Javitt DC (2005). Sensory contributions to impaired prosodic processing in schizophrenia. Biological Psychiatry, 58, 56–61. [PubMed: 15992523]

Mitchell RLC, Elliott R, Barry M, Cruttenden A, & Woodruff PWR (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. Neuropsychologia, 41, 1410–1421. [PubMed: 12757912]

Monrad-Krohn GH (1947). The prosodic quality of speech and its disorders (a brief survey from a neurologist's point of view). Acta Psychiatrica et Neurologica Scandinavica, 22, 255–269.

Sackeim HA, Gur RC, & Saucy MC (1978). Emotions are expressed more intensely on the left side of the face. Science, 202, 434–436. [PubMed: 705335]

Table 1

Means and Standard Deviations for Ratings

		Emotional Context							
		Happy		Sad		Angry		Fearful	
Sentence Emotion	***n***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Happy	25	8.93	0.68	1.88	0.66	1.72	0.69	1.31	0.30
Sad	24	1.93	1.09	8.87	0.73	4.41	2.01	4.50	1.89
Angry	26	1.49	0.42	4.81	1.46	9.10	0.51	2.70	1.30
Fearful	25	2.15	1.24	4.03	1.61	3.49	1.08	8.54	0.96
Neutral	30	4.62	1.66	4.61	1.43	5.86	1.52	4.27	2.19

Table 2

Interclass Correlations (ICCs) for Each Sentence Category

Sentence Emotion	*n*	Avg. ICC	Variance
Neutral	30	.32	.06
Happy	25	.87	.01
Sad	24	.72	.02
Angry	26	.80	.00
Fearful	25	.68	.03

Note—$p < .001$ for neutral versus average ICCs of happy, sad, angry, and fearful sentences, and for happy versus average ICCs of sad, angry, and fearful sentences.

Table 3

Coefficients of Variation (CVs) for Each Sentence Category in Each Context

Sentence Emotion	*n*	Emotional Context	Average CV	Variance	*p*
Happy	24	Happy	.15	.01	<.001
		Avg. SAF	.64	.14	
Sad	18	Sad	.15	.01	<.001
		Avg. HAF	.59	.07	
Angry	25	Angry	.12	.00	<.001
		Avg. HSF	.67	.07	
Fearful	17	Fearful	.15	.01	<.001
		Avg. HSA	.65	.05	
Neutral	14	Happy	.58	.04	.31
		Sad	.67	.03	
		Angry	.59	.02	
		Fearful	.54	.03	

Note—H, happy; S, sad; A, angry; F, fearful.

Table 4

Sex Differences in Mean Ratings by Emotion

	Happy		Sad		Angry		Fearful	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Male	3.477	0.395	4.081	0.693	4.417	0.609	3.826	0.629
Female	4.227	0.936	5.522	1.226	5.559	0.946	4.678	0.981