VOQUAL'03, Geneva, August 27-29, 2003

A study of perceived vocal features in emotional speech

Tanja Bänziger & Klaus R. Scherer

Department of Psychology, University of Geneva, Switzerland
Tanja.Banziger@pse.unige.ch

Abstract

Perceived vocal features of emotional speech have rarely been investigated. In this contribution, a procedure allowing to collect reliable judgments on the perception of voice characteristics of emotional speech is presented. Relations between acoustic parameters and perceived features of speech are described. Some benefits and potential drawbacks of studying perceived vocal features in emotional speech are introduced and discussed.

1. Introduction

"Studies using electromechanical methods of analyzing speech surely would be an important step toward defining the vocal cues of feeling. But in addition to such studies, it would also be imperative to investigate the auditory cues which can be discriminated by listeners, rather than by electronic devices, for in the final analysis, the cues heard by listeners must carry the emotional meanings involved in interpersonal, vocal communication." Davitz [1] 1964, p.26

Up to now, this proposition, formulated by Davitz in 1964, did not arouse much interest. Subsequent studies on vocal communication of emotions have investigated acoustic correlates of expressed emotions but have only seldom examined *"auditory cues which can be discriminated by listeners"*. Reviews of studies that have described acoustic profiles of emotions can be found elsewhere (see for instance Scherer [2] or Juslin & Laukka [3]). A noteworthy exception to the general lack of interest for perceived voice features in the field of emotional communication is the work done by van Bezooijen [4] who examined a large range of perceived vocal features assessed by semi-professional voice specialists who where briefly trained.

Davitz's suggestion has been followed by Scherer [5] who proposed to use a *Brunswikian lens* paradigm to study the non-verbal communication process. In this paradigm, the communication process is represented in 4 steps. The internal state of the speaker (step 1) is, more or less systematically, encoded in different external/measurable cues (acoustic parameters, step 2); lay observers (listeners) can perceive some of those cues (step 3); and, finally, some of the perceived cues will be used by the listeners to infer the internal state of the speaker (step 4).

The lack of interest for perceived vocal features in the study of emotional speech (step 3 in the *Brunswikian lens* paradigm) is probably due to multiple factors, however only two major problems will be briefly described here.

First a model of the different vocal dimensions that would have to be included in a comprehensive study of perceived features of emotional speech is not available. Systematic models for the description of voice quality have been proposed (e.g. Laver's production model [6] used by van Bezooijen [4]) but the dimensions they specify are not necessarily adapted to the study of emotional speech. Furthermore, most dimensions used by voice professionals are unfamiliar (and often incomprehensible) to lay persons and therefore cannot be used for obtaining judgments from listeners who have not been specifically trained.

Second, research on the assessment of perceived pathological voice quality showed that the inter-rater and the test-retest reliability of judgments on dimensions such as *roughness* or *breathiness* are low (see Kreiman & Gerratt [7] for more details on those issues). Hence, results from the field of pathological voice quality assessment strongly suggest that the standards for rating voice quality differ from listener to listener and also vary over time for the same listener.

In the present contribution, we present a first attempt to measure perceived voice features of emotional expressions using a method developed to address (and if possible, control) the reliability issues mentioned above. The relation of a set of perceived vocal features to a set of measured acoustic parameters is examined and two illustrations of the benefits of looking at perceived vocal features of emotional speech are presented.

2. Judgment procedure

Two major problems linked to the assessment of perceived vocal features appear to be: (a) inter-individual variation of the definitions/representations of voice features described by terms such as *breathiness* or *roughness* or, more generally, insufficient vocabulary to describe voice quality in a reliable way; (b) unstable comparison standards within raters (low test-retest reliability).

In an attempt to overcome both problems, we adapted a procedure proposed by Granqvist [8]. In a traditional approach, the recordings to be assessed are displayed in random order and are evaluated by listeners on a number of different scales immediately after the display. In the approach used here, a scale is presented to the listener on a computer screen (see Fig. 1). The task of the listener is to position the recordings on this scale. All recordings produced by one given speaker are represented on the screen in the form of identical icons (small blue "flags", see Fig. 1), the recordings are displayed when the listener double-clicks the icons and can be moved on the scale to any position selected by the listener. Lay participants can listen to the recordings and can modify/correct their answers as often as they wish. The answers are recorded on a continuous scale ranging from 0 (for recordings positioned on the extreme left of the scale) to 10 (for recordings positioned on the extreme right of the scale).

The possibility to directly compare a set of vocal expressions on a given dimension/scale and to assess those expressions relatively to one another addresses the problem of the modification of internal standards over time. To deal with the issue of the un/shared standards of comparison between listeners, two recordings are presented under each scale. They illustrate two extreme instances of each dimension (e.g. 'very low pitched' recording versus 'very high pitched' recording). Those recordings can be displayed at any time by double-

clicking on the icons represented under the scale. Participants were instructed to use those recordings as examples of extreme instances of the vocal dimension represented by the scale and not for direct comparison with the recordings to be assessed.

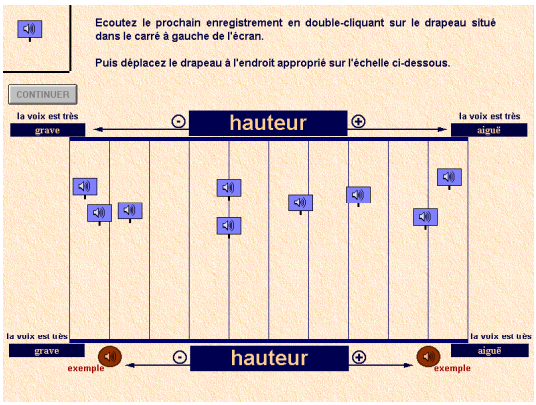

Figure 1: Illustration of the procedure. The bipolar scale 'pitch' as displayed for the participants

2.1. Recordings used

144 emotional expressions have been sampled from a larger set of emotional expressions described in detail by Banse & Scherer [9]. Expressions produced by 9 actors have been selected. All actors pronounced 2 sequences of 7 syllables (1. "hät san dig prong nju ven tsi", 2. "fi gött laich jean kill gos terr") and expressed 8 emotions : cold anger ('irrit') and hot anger ('rage'), anxiety ('anx') and panic fear ('paniq'), sadness ('sad') and despair ('desp'), happiness ('joy') and elation ('elat').

Another speaker was recruited to produce the audio-illustrations presented under each scale to establish a common definition of the vocal dimensions under examination. He pronounced the French sentence "je ne peux pas le croire".

2.2. Participants and scales/vocal dimensions

The procedure presented above being more time consuming than a traditional procedure, 4 groups of participants were recruited to assess 4 subsets of the 144 recordings. The groups were composed of 15 to 16 students in their first year at the University of Geneva. Each group assessed the recordings produced by 3 speakers. To check for a potential group effect on the ratings, the recordings produced by one of the 9 speakers were assessed by all the participants. Each participant rated 48 expressions (2 sequences of syllables x 8 expressed emotions x 3 speakers) on the 8 following dimensions. The directions of the scales are given in brackets and the French words available to the participants are represented in square brackets:

- 'pitch' (low ↔ high) ['hauteur', grave ↔ aiguë]
- 'intensity' (weak ↔ strong) ['volume', faible ↔ forte]
- 'intonation' (monotonous ↔ modulated) ['mélodie', monotone ↔ modulée]
- 'speed' (slow ↔ fast) ['vitesse', lente ↔ rapide]
- 'articulation' (bad ↔ good articulation) ['articulation', mal ↔ bien articulée]
- 'instability' (steady ↔ shaky) ['stabilité', ferme ↔ tremblante]
- 'roughness' (not rough ↔ rough) [qualité 'rauque']
- 'sharpness' (not sharp ↔ sharp) [qualité 'perçante']

3. Results

The reliability of the ratings was satisfactory. Table 1 shows the intraclass correlations (single and average measure) for each of the participant groups and for each vocal dimension. The reliability is generally high, but the ratings for some vocal dimensions (e.g. 'intensity') are substantially more reliable than others (e.g. ratings for 'roughness' or 'articulation').

Table 1: Reliability of the ratings - intraclass correlations (r = single measure, R = average measure)

dimension	group 1		group 2		group 3		group 4	
	r	R	r	R	r	R	r	R
articulation	0.27	0.85	0.31	0.87	0.22	0.81	0.47	0.93
intonation	0.41	0.92	0.34	0.89	0.37	0.90	0.47	0.93
intensity	0.85	0.99	0.84	0.99	0.84	0.99	0.88	0.99
pitch	0.58	0.96	0.49	0.93	0.51	0.94	0.50	0.94
roughness	0.29	0.87	0.19	0.78	0.30	0.87	0.44	0.92
speed	0.64	0.97	0.62	0.96	0.66	0.97	0.72	0.98
sharpness	0.39	0.91	0.61	0.96	0.63	0.96	0.72	0.97
instability	0.49	0.94	0.50	0.94	0.41	0.91	0.47	0.93

Systematic differences between the ratings given by the 4 groups (for the subset of expressions evaluated by all participants) were not found. Consequently, the ratings given by all participants (for the complete set of recordings) are considered simultaneously in the following result description. One fourth of the ratings given for the recordings that were assessed by all participants were randomly selected. The mean ratings for each of the 144 expressions are thereafter computed on the basis of 15 to 16 judgements.

The average ratings obtained for the 144 emotional expressions on the 8 vocal dimensions are correlated. A factor analysis (principal components with varimax rotation) yielded two components with eigenvalues bigger than one. Together, both components account for 73% of the total variance of the vocal dimensions. Table 2 shows the loadings of the 8 vocal dimensions on the 2 components (loadings smaller than 0.3 are not displayed). Dimensions related to "intonation" (sharpness, intensity, intonation, pitch and speed) load on the first component. Dimensions descriptive of good versus bad "voice quality" (instability, articulation, roughness) load on the second component.

Table 2: Factor loadings of a principal component analysis based on 8 perceived vocal dimensions

Vocal dimensions	Component 1	Component 2
sharpness	0.974	
intensity	0.926	
intonation	0.886	
pitch	0.851	
speed	0.740	
instability		0.838
articulation		-0.824
roughness	0.325	0.625

A number of acoustic parameters were computed for the 144 emotional expressions using the software PRAAT [10]. Parameters derived from F0 and intensity contours, duration and spectral distribution of energy were measured (the choice of parameters corresponds to Banse and Scherer's selection [9]). The acoustic parameters being highly correlated, a principal component analysis was used to select a subset of relatively independent parameters. Nine parameters with high loadings on one of nine components extracted by the factor analysis were selected: F0 minimum (F0.min) and F0 range (F0.range), intensity range (int.range), total duration (dur.tot), proportional duration of voiced segments (dur.v/art), proportion of energy in voiced segments between 300 and 500 Hz (v.300-500) and between 600 and 800 Hz (v.600-800), proportion of energy below 1 kHz in voiced and in unvoiced segments (v.0-1k and n.0-1k). Average intensity (int.mean) was added to those parameters out of theoretical considerations. The 8 perceived vocal dimensions were regressed on the 10 selected acoustic parameters in 8 stepwise regressions. Table 3 shows the parameters that significantly contributed to each of the 8 regressions. Directions of the relations between the parameter in the regressions and the vocal dimensions are reported in the form of (+) for positive relations and (–) for negative relations. The proportion of explained variance (R^2) for each scale is represented on the right in this table.

Table 3: Stepwise regressions of 8 perceived vocal dimensions on 10 acoustic parameters

vocal dimension	acoustic parameters	R^2
intensity	int.mean$^{(+)}$, int.range$^{(+)}$, v.0-1k$^{(-)}$	0.88
sharpness	int.mean$^{(+)}$, F0.range$^{(+)}$, F0.min$^{(+)}$, v.0-1k$^{(-)}$, int.range$^{(+)}$	0.87
speed	dur.tot$^{(-)}$, int.mean$^{(+)}$, v.0-1k$^{(-)}$, dur.v/art$^{(-)}$	0.79
intonation	F0.range$^{(+)}$, int.mean$^{(+)}$, int.range$^{(+)}$, F0.min$^{(+)}$ n.0-1k$^{(-)}$, dur.tot$^{(-)}$	0.67
pitch	F0.min$^{(+)}$, F0.range$^{(+)}$, int.mean$^{(+)}$	0.65
instability	F0.min$^{(+)}$, dur.tot$^{(+)}$, v.0-1k$^{(+)}$	0.35
articulation	F0.min$^{(-)}$, int.mean$^{(+)}$, int.range$^{(+)}$, F0.range$^{(-)}$, n.0-1k$^{(-)}$	0.32
roughness	n.0-1k$^{(+)}$, v.0-1k$^{(-)}$, int.mean$^{(+)}$, dur.tot$^{(+)}$	0.28

The variance of vocal dimensions related to "intonation" is largely accounted for by the acoustic parameters. Average intensity is the best predictor of perceived intensity and sharpness. Total duration is the best predictor of perceived speed; F0 range is the best predictor of perceived intonation and F0 minimum is the best predictor of perceived pitch. On the other hand, the acoustic parameters could account for only a small part of the variance of vocal dimensions related to "vocal quality", showing that those dimensions describe vocal aspects that were not captured by the relatively simple acoustic measures used in this study.

In the remainder of this result section two illustrations of the relations that could be observed between perceived vocal dimensions and expressed or perceived emotions will be presented.

For emotional expressions with low activation (calm joy, anxiety, sadness and cold anger), a positive correlation between perceived speed of speech and perceived quality of articulation ($r = 0.38$, $p<.001$, $N = 72$) reflects a tendency for expressions perceived as faster to be also perceived as produced with a better articulation. Examination of the results for the 4 expressed emotions with low activation reveals that this unexpected positive correlation can be largely attributed to different "speaking styles" associated to different expressed emotions. Specifically, expressions of sadness are perceived as slow and not well articulated whereas expressions of cold anger are perceived as relatively fast and well articulated. Average ratings of speed and quality of articulation for 8 expressed emotions (with low and high activation) are represented on Fig. 2 and Fig. 3.

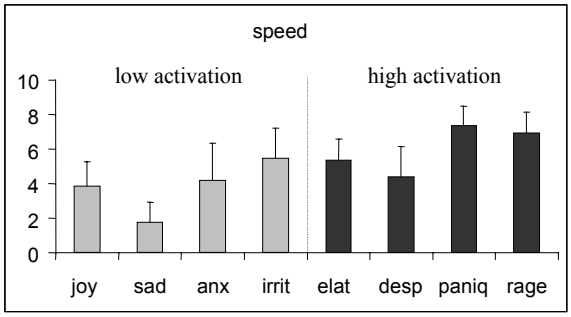

Figure 2: Average perceived speed for 8 expressed emotions (N = 18)

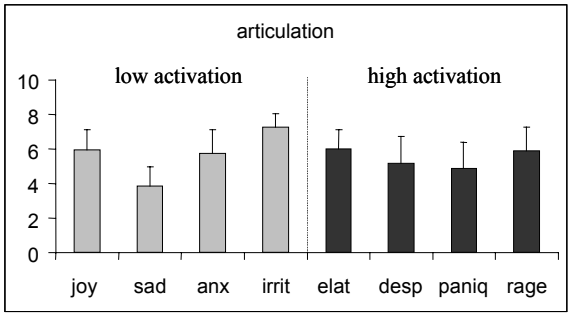

Figure 3: Average perceived quality of the articulation for 8 expressed emotions (N = 18)

Juslin [11] used a Brunswikian lens approach to study emotional communication in music performance. Following his approach we used the *lens model equation* (LME) (1) to compare the contribution of acoustic cues and the contribution of perceived vocal cues to the vocal communication of emotion. The LME splits the *communication achievement* (r_a, i.e. the correlation between expressed emotion and perceived emotion) into 2 multiplicative components: the *linear component* (i.e. the component of the correlation derived from the linear contributions of the variables entered in the model) and the *unmodeled component* (which includes systematic and unsystematic variance not accounted for by the linear models). The linear component is a function of *speaker consistency* (R_e, i.e. the multiple correlation of expressed emotion on the variables in the model), *listener consistency* (R_s, i.e. the multiple correlation of perceived emotion on the variables in the model) and *matching* (G, i.e. the correlation between the predicted values of the expressed emotion model and the predicted values of the perceived emotion model).

$$r_a = G\, R_e\, R_s + C \sqrt{(1-R^2_e)} \sqrt{(1-R^2_s)} \quad (1)$$

Judgements of perceived emotions were obtained in the form of perceived intensity of joy, fear, sadness and anger; using the same procedure as for the judgments on perceived vocal dimensions (separate groups of listeners were used). Fig. 5 shows an example of a LME decomposition of the correlation between expressed anger (dummy coded variable opposing cold and hot anger to other expressed emotions) and perceived anger. Communication achievement (r_a = .780) is split into a multiplicative *linear component* (G R_e R_s =.557) and an *unmodeled component* (C $\sqrt{(1-R^2_e)}$ $\sqrt{(1-R^2_s)}$ = .223). The model is computed with the 8 perceived vocal dimensions as predictors. The direction of significant predictors on both sides of the model are indicated by (+) and (-).

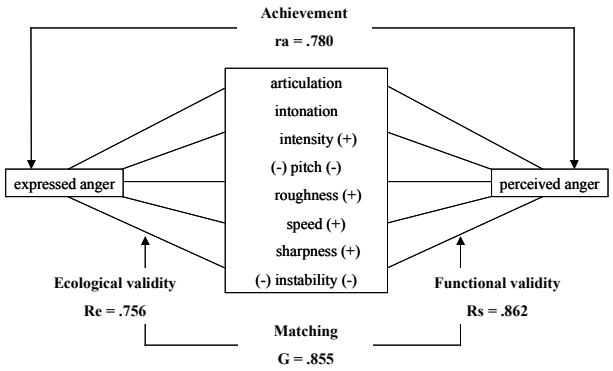

Figure 5: Lens model for communication of 'anger'

Table 4 summarizes 8 such models obtained for the communication of joy, fear, anger and sadness, using either 8 perceived vocal dimensions (as in Fig. 5) or 8 acoustic parameters (int.mean, int.range, F0.min, F0.range, dur.tot, dur.v/art, v.600-800, v.0-1k) as predictors. Best fitting models with acoustic parameters and perceived vocal features were obtained for anger. Acoustic parameters could not account for the correlation between expressed and perceived joy, whereas perceived vocal dimensions account for one third of the same correlation.

Table 4: Linear component of LME using acoustic or perceived vocal predictors

	joy		fear	
	acoust.	perceiv.	acoust.	perceiv.
r_a	0.754	0.754	0.677	0.677
$G*R_e*R_s$	0.026	0.265	0.277	0.357
$(G*R_e*R_s) / r_a$	0.03	0.35	0.41	0.53
	anger		sadness	
	acoust.	perceiv.	acoust.	perceiv.
r_a	0.780	0.780	0.796	0.796
$G*R_e*R_s$	0.456	0.557	0.331	0.538
$(G*R_e*R_s) / r_a$	0.58	0.71	0.42	0.68

4. Conclusion

The results presented above show that the assessment of perceived vocal features can provide interesting insights into the vocal features involved in the communication of emotion.

They allowed for instance to derive a tentative conclusion that sad expressions might be characterized by a "slurred" speaking style (slow speech with bad articulation) whereas cold anger would be characterized by a "controlled" speaking style (fast speech combined with good articulation). Perceived voice features could furthermore partially account for the communication of joy while simple acoustic parameters (mainly F0, intensity and duration derived parameters) failed to account for the relationship between expressed and perceived joy. On a more general level, perceived descriptions of vocal features provided a better explanation for the relationship between expressed and perceived emotions than the acoustic parameters did.

On the other hand, the use of perceived vocal feature to study emotional speech entails a number of potential shortcomings. The high inter-rater reliability indices obtained in this study probably reflect very big vocal differences in the expressions under examination. In other terms, the reliability of the ratings might be much lower for less extreme vocal expressions. Another crucial problem is the influence of perceived emotions on the assessment of vocal features. It cannot be excluded that the ratings of listeners on a given vocal scale are influenced by the emotions they perceive in the recordings. Those aspects should be more closely examined and controlled in further studies.

5. References

[1] Davitz, J. R., *The communication of emotional meaning,* McGraw-Hill, New York, 1964.

[2] Scherer, K. R., "Vocal communication of emotion: a review of research paradigms." *Speech communication*, *vol. 40*, pp. 227-256, 2003.

[3] Juslin, P. and Laukka, P., "Communication of emotion in vocal expression and music performance: different channels, same code?" *Psychological Bulletin*, in press.

[4] van Bezooijen, R., *Characteristics and recognisability of vocal expressions of emotion,* Foris Publications, Dordrecht, 1984.

[5] Scherer, K. R., "Personality inference from voice quality: The loud voice of extroversion", *European Journal of Social Psychology, vol. 8*, pp. 467-487, 1978.

[6] Laver, J., *The phonetic description of voice quality*. Cambridge University Press, Cambridge, 1980.

[7] Kreiman, J. and Gerratt, B. R., "Validity of rating scale measures of voice quality", *Journal of the Acoustical Society of America, Vol. 104*, pp. 1598-1608, 1998.

[8] Granqvist, S., "Enhancements to the Visual Analogue Scale", *TMH-QPSR KTH, Vol. 4/1996*, pp. 61-65, 1996.

[9] Banse, R. and Scherer, K. R., "Acoustic profiles in vocal emotion expression", *Journal of Personality and Social Psychology, vol. 70,* pp. 614-636, 1996.

[10] Boersma, P. and Weenink, D. J. M., *Praat, a system for doing phonetics by computer, version 3.4*, Institute of Phonetic Sciences of the University of Amsterdam, Report 132, 1996.

[11] Juslin, P., "A functionalist Perspective on Emotional Communication in Music Performance", Acta Universitatis Upsaliensis, Uppsala, 1998.

Note: More details about the study and the results presented in this contribution will be published in a forthcoming article. Readers with specific interests can contact the authors for more information.