ACOUSTIC PROFILES IN PROTOTYPICAL VOCAL EXPRESSIONS OF EMOTION

Tom Johnstone, Rainer Banse, and Klaus R. Scherer
Department of Psychology, University of Geneva

ABSTRACT

Acoustic data from a large study on actor portrayals of vocally expressed emotions are reanalysed on the basis of the differences in the accuracy of recognition of the voice samples used. The results show differentiated patterns with respect to the similarity and variability of the mean acoustic profiles for well versus poorly recognized portrayals.

INTRODUCTION

Banse & Scherer [1] report a large-scale study on the expression of emotion in multiple communication modalities, in which 12 professional actors were asked to portray 14 emotions varying in intensity and valence or quality. The results on decoding replicate and extend earlier findings demonstrating the ability of judges to infer vocally expressed emotions with much better than chance accuracy for a large number of emotions. Consistently found differences in the recognizability of different emotions are also replicated. A total of 224 different portrayals, 16 per emotion category, were subjected to digital acoustic analysis to obtain profiles of vocal parameters for different emotions, using a large set of acoustic variables. The data provide first indications that vocal parameters not only index the degree of intensity typical for different emotions but also differentiate valence or quality aspects. In particular, the data are used to test theoretical predictions on vocal patterning based on Scherer's component process model of emotion [2]. While most hypotheses are supported, others need to be revised on the basis of the empirical evidence.

Discriminant analysis and jack-knifing were used to determine how well the 14 emotions can be differentiated on the basis of the vocal parameters measured. The results show remarkably high hit rates and patterns of confusion that closely mirror those found for listener-judges. One of the major results of this study was the identification of typical acoustic profiles for 14 major emotions. However, the portrayals used to compute these profiles varied substantially in the extent to which their emotional content was recognized by listener-judges, despite the fact that they had been preselected for clarity of emotional expression. In this study we report a new, secondary analysis of the earlier data set in order to examine potential differences between acoustic profiles for portrayals that are and that are not well recognized by listener judges. One can argue that portrayals that are well recognized on the basis of vocal expression alone represent prototypical examples of vocal emotion communication. In consequence, their acoustic profiles should represent more closely the acoustic parameters which index different emotional speaker states in natural speech. In contrast poorly recognized portrayals should show greater parameter variation and a less pronounced profile.

METHOD

The mean accuracy of the judgments (computed on the basis of recognition scores ranging from 0 to 12, corresponding to the number of judges who correctly categorized each of the intended emotions as portrayed by the actors) was used to split the vocal utterances into two groups: well recognized vs. poorly recognized (yielding an average score of 8.5 for the well recognized stimuli as compared to 3.2 for the poorly recognized). The respective profiles over 29 acoustic parameters reported previously [1] were computed separately for the two groups of stimuli.

RESULTS

Splitting the utterances produced two groups of utterances for each emotion with substantially different mean recognition scores except in the cases of disgust (difference in means = 2.6) and shame (difference in means = 3.2). These two emotions were badly recognized overall (with overall mean recognition scores of 1.5 and 3.2 respectively) and thus the small difference between the well and poorly recognized groups might be explained as a floor effect. It is likely that the actors found it very difficult to express these emotions and in groping for ways of expressing the requested emotion, either did not produce systematic changes to the vocal signal or consistently produced utterances which were confused with another emotion.

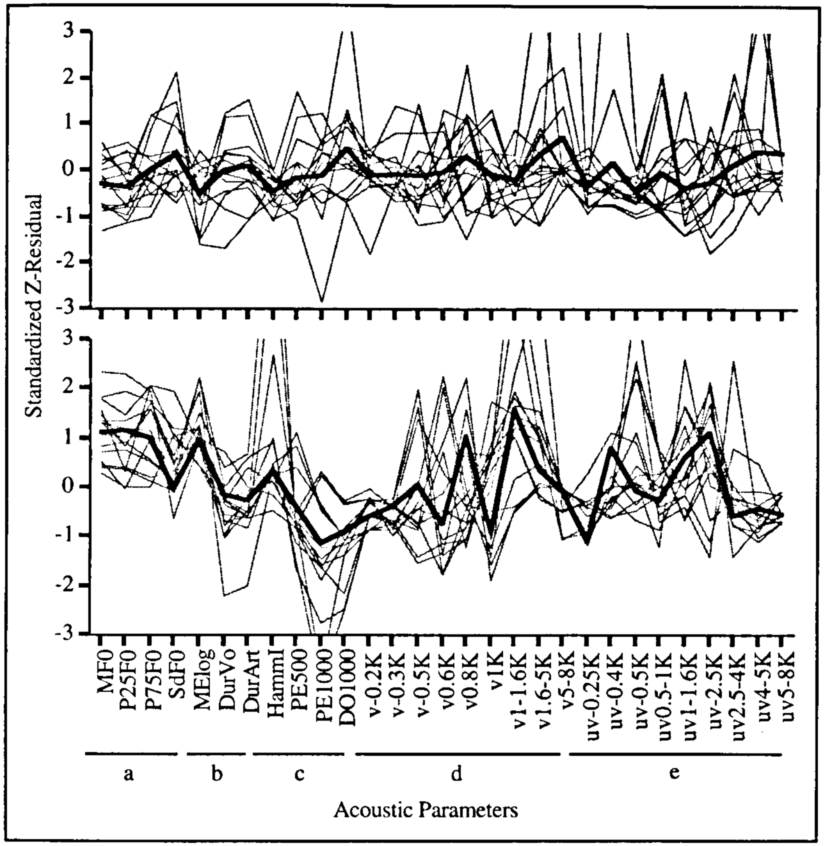

Figure 1. Acoustic profiles of all disgust (top) and hot anger (bottom) utterances, with the mean profiles shown as the dark lines (Acoustic parameters: a) F0 measures b) energy and duration measures, c) high-low frequency ratio, d) voiced spectral parameters, e) unvoiced spectral parameters: see Banse and Scherer [1], Table 6, for a full explanation of the acoustic variables).

The correlation between the mean profiles for well recognized utterances and those for poorly recognized utterances for each emotion was calculated to provide a measure of profile similarity. The emotions can be divided into three classes; those with low, medium and high correlations between the well versus poorly recognized sample profiles respectively.

The utterances expressing disgust (r=0.02) and interest (r=0.12) fall into the *low correlation* class. As mentioned previously, disgust had a poor overall recognition score. This can be attributed to the lack of any consistent acoustic profile, as shown in Figure 1, and is consistent with previous studies of disgust which show the emotion to be difficult to recognize in speech [3, p.190]. Possibly, the expression of disgust typically involves the use of affect bursts

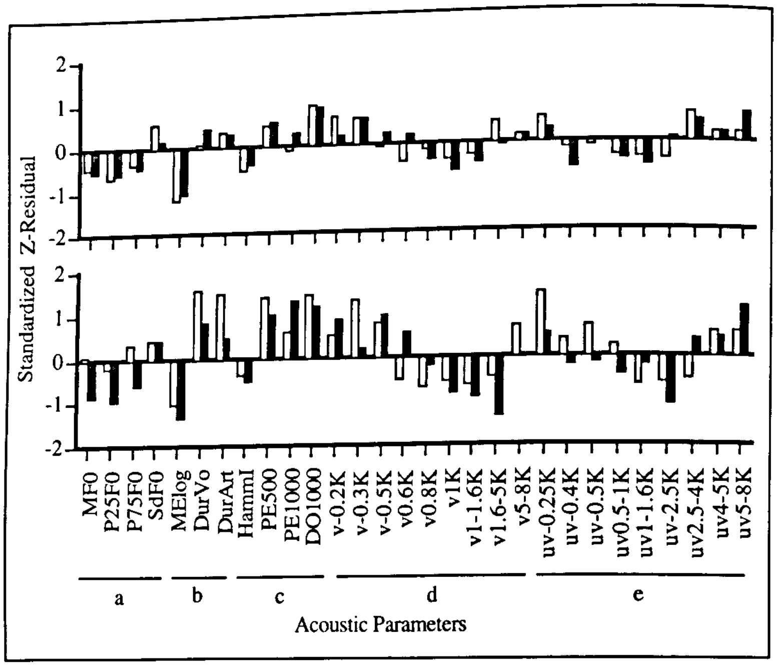

Figure 2. Acoustic profiles of shame (top) and sadness (bottom). White columns are for poorly recognized and shaded columns for well recognized utterances.

rather than the nonverbal modulation of fluent speech [4]. In contrast, interest had a high overall mean recognition score of 11. It is possible that, of the 29 acoustic parameters making up each profile, only a few are used in the expression of interest. Other parameters not measured in the study, such as the type of F0 contour, could play an important part in the expression of interest. Thus the profiles measured in this analysis would not be very well defined despite the high recognition of the utterances.

Utterances expressing the emotions of happiness, cold anger, boredom, pride and panic have *medium sized* correlations between well and poorly recognized group profiles (ranging from r =.37 for pride to .58 for cold anger). These emotions have medium overall recognition scores, implying that the actors were able to express the emotions reasonably well but that there was still considerable variability in the utterances. An examination of profiles indicates that the mean profiles for the poorly recognized utterances are quite similar in shape to those of the well recognized utterances, but usually involve smaller magnitudes. It is possible then that, in these cases, the poorly recognized utterances do not contain sufficient modulation of the relevant acoustic parameters to be identified accurately.

With the exception of shame, all the emotions with *high correlations* between well and poorly recognized utterances had medium to high overall recognition scores. These utterances are generally characterized by well defined acoustic profiles (e.g. the hot anger profile in Figure 1), which would presumably be responsible for the correct recognition of the intended emotion. It is possible that for those emotions which only had medium recognition scores, one or two acoustic parameters which are essential for the expression of the emotion are inconsistently used by the actors. Such idiosyncratic modulation of only a few parameters would not greatly affect the profile correlations. Thus whilst the profiles are consistent and highly correlated, some single important acoustic parameters may vary between actors, leading to poorer recognition of some utterances. It is also possible that in some cases, a number of poorly recognized utterances were not characterized by consistent profiles, due to high variability between speakers. In the cases of sadness and despair, there were significantly higher between-utterance variances for poorly recognized as opposed to well recognized utterances (t=3.1, p<0.05 and t=2.7, p<0.05 respectively). Thus the poorly recognized sets of utterances for these emotions did not represent prototypical emotion profiles.

Although utterances expressing shame had well defined profiles, they were very poorly recognized. Comparison of the acoustic profiles of sadness and shame indicates that actors may have been using the sadness prototype when trying to express shame. It is conceivable that, faced with difficulties expressing shame, actors reverted to the more familiar expression of sadness. This is supported not only by the similarity of the profiles for shame and sadness (Figure 2), but also by the large percentage of times shame utterances were falsely categorized as sadness by the judges in the study of Banse and Scherer [1].

CONCLUSIONS

It is apparent that studying actor portrayals of vocal emotion expression can reveal much about the nature of the acoustic parameters involved in the identification of emotion by a listener. At the same time certain emotions either do not seem amenable to consistent portrayal by actors or are not readily recognized by listeners. Certain emotions, such as hot anger and boredom, are portrayed using highly prototypical acoustic profiles which are easily produced by actors and accurately decoded by listeners. Others, while also characterized by quite consistent profiles, are not as well recognized, possibly due to the inability of actors to completely control all the aspects of voice or speech relevant to that emotion. This might be explained in terms of involuntary physiological changes to the vocal apparatus during real emotional episodes, which are inaccessible to voluntary production by actors. Still other emotions, such as interest, although well recognized by listeners, seem be communicated by suprasegmental features other than long term average modulation of the speech signal. Temporal changes in speech parameters such as F0 might be the primary method of encoding such emotions. Finally, disgust would seem to be universally badly encoded and recognized in speech. This could be due to the fact that disgust is more often expressed by brief affect bursts or interjections rather than by modulation of continuous speech.

The secondary analysis of the data set in [1] has shown the utility of using decoding data (i.e. contrasting well versus poorly recognized portrayals) to better understand the role of the encoding of vocally portrayed emotions (as measured by the variation of acoustic profiles). The results of the comparison yield a number of hypotheses which are amenable to further empirical research.

ACKNOWLEDGMENT

The preparation of this paper has been funded by a grant of the Swiss National Fund for Scientific Research (FNRS No. 21-32648.91) in the context of the ESPRIT-BRA VOX workshop programme.

REFERENCES

[1] Banse, R. and Scherer, K. R. (in press), "Acoustic profiles in vocal emotion expression.", *Journal of Personality and Social Psychology.*

[2] Scherer, K. R. (1986), "Vocal affect expression: A review and a model for future research.", *Psychological Bulletin,* vol. 99, pp. 143-165.

[3] 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 Press.

[4] Scherer, K. R. (1994). "Affect bursts." In S. H. M. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 161-196). Hillsdale, NJ: Erlbaum.