HOW EMOTION IS EXPRESSED IN SPEECH AND SINGING

Klaus R. Scherer
University of Geneva

ABSTRACT

This contribution focusses on some of the major issues in the study of emotional expression in the speaking and the singing voice. Adopting a sociopsychobiological approach, it is claimed that affect vocalizations have multiple determinants and serve multiple functions. Based on research examples, it is demonstrated how these determinants can be empirically distinguished. Furthermore, recent data on emotion differentiation via acoustical profiles is presented. Brief allusions are made concerning the appeal and the symbol functions of affect vocalization. Finally, an approach to study emotional expression in the singing voice is presented.

INTRODUCTION

In his influential manual of rhetorics, the *Orator*, Cicero remarks: "There are as many movements of the voice as there are movements of the soul, and the soul is strongly affected by the voice". While the terminology is no longer fashionable, this brief statement sums up much of what current research on the vocal expression of emotion in speech and music is empirically documenting. In this contribution, I will present some of the work of our research group, both with respect to theory and data.

MULTIPLE FUNCTIONS AND MULTIPLE DETERMINANTS

While it has been customary to consider vocal expression in animals primarily as indicative of underlying affective or motivational states [1], recent research indicates that the situation is more complex. Marler and his colleagues, studying the alarm calls of vervet monkeys, found that calls are not only indicative of the emitter's fear state but are also specific to certain types of predators [2]. Alarm calls produced for leopards, eagles or snakes, for example, have different sounds, energy levels and frequency ranges. Therefore, Marler and his colleagues reject the notion that animal communication is limited to indicating the animal's emotional or motivational state, arguing that most animal calls have a very strong referential symbolic component. Supporting this notion is the observation that alarm calls seem to be partially learned. Therefore, the alarm call system does not simply "push out" the underlying affect but reflects the outcome of predator classification, which reflects rudimentary cognitive processes.

An exclusive emphasis on either a motivation-affect expression or a symbolic function, neglects the fact that most vocal signals are multifunctional. The Organon model, developed by Bühler [3], can be used to analyse the functions of vocal affect signals. In this model, a sign has three functions: as a *symbol* in representing the object, event or fact it stands for, as a *symptom* of the state of the sign user and as an *appeal*, or signal, trying to elicit a response from the receiver. The vervet monkey alarm call, for example, serves all three: a) as a symbol of different predators, b) as a symptom of the fear state of the animal, and c) as an appeal to others to run away. Furthermore, these functions are mutually interdependent: if a call refers to an air predator, both the emotional reaction and the appeal may be very different from one made in reference to a ground predator - in the first case, both emitter and receiver might seek shelter under a bush and freeze; in the second, they might become highly activated and run up a tree.

Apart from multiple functions we also need to consider multiple determinants. We have suggested to distinguish between *push effects* in which physiological processes such as muscle tone push vocalisations in a certain direction and *pull effects* where external factors such as the expectations of the listener pull the affect vocalisation toward a particular acoustic model [1]. In the push effect, given that muscle tone is likely to be higher in sympathetic arousal, the fundamental frequency of the voice (F0) will also be higher. Pull effects, on the other hand, are governed by social conventions such as display rules. These cultural conventions influence the production of signs required in social situations by specifying a particular acoustic target pattern, as opposed to mental concepts or internal physiological processes which push out expression. This distinction is important in understanding the differences between vocal productions.

Thus, push factors are defined as producing changes in subsystem states in the organism which have a direct effect on vocalisation parameters. They work largely involuntarily; the effects on vocal organs and the resulting acoustic parameters are almost exclusively determined by the nature and force of physiological mechanisms. Pull factors on the other hand, although they are mediated through internal systems, are externally based - they operate toward the production of specific acoustic patterns or models, as in the case of detailed optimum signal transmission features or socially defined signal values.

Clearly, the push/pull concept of two major types of determinants of vocal signals is directly linked to the Bühler model of multiple functions. One can argue that the symptom aspect, i.e. the expression of an internal state, represents push, whereas the symbol and appeal aspects represent pull. Different factors might determine the nature of the expression in each case. And, it might be the antagonism between push and pull, e.g. high physiological arousal pushing voice fundamental frequency up and the conscious attempt to show "control" pulling it down, which can produce mixed or even contradictory messages. If this were so, it would be important to empirically isolate the two determinants. Future research and theorizing in this area will need to more clearly differentiate between these multiple determinants and multiple functions in order to avoid futile controversies about the "true nature" of affect vocalizations..

EMPIRICAL ASSESSMENT OF MULTIPLE DETERMINANTS

We argue that the type of determinant will have a major effect on the *coding*, i.e. the relationship between the underlying referent and the sign features. If, in a push condition, muscle tension goes up under stress, producing an increase in the fundamental frequency of the voice, we would expect direct *covariation* between the amount of muscle tension increase measured by electromyography and the increase in fundamental frequency as measured by digital voice analysis. In this convariation model, we would expect a continuous (and probably linear) covariation between the two variable classes. The alternative is what we call the *configuration* model [4]. The configuration model is more "linguistic" than the covariation model, itself more psychological in nature. The configuration model argues that to achieve a certain effect in the listener, one uses a particular combination of intonation, accent, word and/or syntactic structure, e.g. a rising intonation contour in a WH-question, a falling one for in a Y/N question. There are no variable dimensions, no continua, in configuration effects: certain classes of phenomena have to co-occur to produce an effect. In terms of push and pull, push is likely to follow covariance rules, while pull, if anything, would follow configuration rules. In trying to understand how communication processes follow either a covariation or a configuration model, we need to gather insight into the determinants.

Scherer, Ladd and Silverman conducted two studies to distinguish covariation and configuration. The first [4] used a corpus of questions (from a large scale study on interactions between civil servants and other citizens) that were homogenous in structure, but that varied in terms of pragmatic force. Some questions were clearly reproaches, though phrased as WH questions, while others were factual information questions. We used three filtering or degrading techniques to systematically isolate particular acoustic cues: a) low pass filtering, b) random splicing, and c) reversing. (see [5] for a comparative list of the acoustic cues retained by the respective techniques).

The results show that even when the text of the questions used is rendered unintelligible, much of the affective meaning remains in the acoustive signal. This confirms the covariance model in its claim that nonverbal vocal cues convey affect in a direct and context-independent way. However, this is true

only for those masking conditions in which voice quality cues are audible, i.e. in the random splicing and reversing conditions. In both of these conditions, the intonation contour of the sentence is lost or destroyed. This would seem to imply that this feature of speech utterances plays no major role in the communication of affective meaning. Obviously, this is rather counter-intuitive and contradicts empirical evidence showing that this information is relevant for the communication of affect. Could it be that intonation follows configuration rather than covariance rules? In order to investigate this question with the speech material in this study, we divided the questions into WH and Yes/No questions and classified the intonation contours into final fall and final rise [4]. The results show that intonation contour obviously has a strong effect on the impression of speaker affect but it seems to be mediated by context effects, verbal or syntactic.

One might interpret these results as reflecting the traditional descriptions of "normal" or "unmarked" intonation for the two different question types. The supposedly "normal" combinations of intonation type (i.e. falling WH questions and rising yes/no questions) were judged as more polite and agreeable. "Marked" combinations on the other hand were rated rather more negatively. This clearly points to strong configuration effects. It is possible then, to presume that some acoustic cues, such as voice quality, operate according to covariance rules, whereas others, such as intonation contours, are used in accordance with configuration rules. This would make sense in terms of a psychobiological approach to communication. One could argue that those cues that show a remarkable degree of phylogenetic continuity - such as the differential nature of phonation which yields different voice qualities - are closer to direct covariance with physiological states. In contrast, cues that have been domesticated within a language system, such as intonation, should follow a configuration model.

In order to further test these notions, we used digital resynthesis of speech in order to be able to experimentally vary different acoustic cues in a factorial design. Such an approach obviously avoids the disadvantage of using a natural corpus, since it allows greater experimental control of the variables under study. In a series of studies [6,7], we used this technique to systematically vary intonation contour, F0 range, intensity, timing, accent, structure, and other parameters. The advantage of this procedure, as pointed out above, is that all of these acoustic features under study can be manipulated independently of each other in a factorial design while leaving all of the remaining acoustic cues constant. Three major types of findings will be highlighted. First, we did not find any interaction effects in the analysis of variance, suggesting that the acoustic variables we studied function largely independently of each other. Secondly, in those studies where we used several speakers and several utterances, we found virtually no interaction between these factors and the acoustic variables manipulated. This encourages one to think that the effects can be generalised over a wide range of speakers and utterances. Thirdly, out of the variables studied, F0 range had the most powerful effect by far on the judgment of the raters, particularly on the attributions of arousal. Furthermore, we were able to show that these effects seemed to be a continous function of changes in F0 range since arousal related ratings go up in a linear fashion with increasing range.

Results for intonation contours and voice quality were complex and seem to require further study. In the case of intonation contours, this may well be due to the important role of the configuration model for this variable. In consequence, we feel that the distinction between configuration and covariance rules may be very useful in understanding the communication of affect in vocal utterances and it would seem useful to continue this type of research with the aid of modern digital signal manipulation techniques.

ACOUSTIC EMOTION PROFILES

The research reported above dealt with affective states of relatively low intensity as one is likely to encounter in normal social interactions. Full-blown, intensive emotions are difficult, if not impossible, to study in an experimental fashion. Therefore, much of the work on the acoustic concomitants of emotion has used actor portrayals of different emotional states to obtain vocal expression samples that could then be analyzed acoustically. Pittam & Scherer [8] have summarized the state of the literature to date as follows:

Anger: Anger generally seems to be characterized by an increase in mean F0 and mean energy. Some studies, which may have been measuring "hot" anger (most studies do not explicitly define whether they studied hot or cold anger), also show increases in F0 variability and in the range of F0 across the utterances encoded. Studies in which these characteristics were not found may have been measuring cold anger. Further anger effects include increases in high frequency energy and downward directed F0 contours. The rate of articulation usually increases.

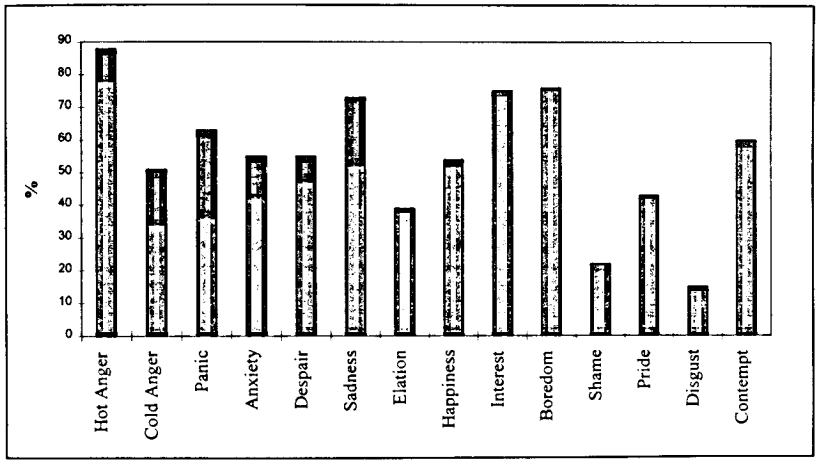

Figure 2. Accuracy of emotion recognition in vocal actor portrayals (based on data from[10]). The shaded parts of the bars represent confusions between members of the same emotion family.

Fear: There is considerable agreement on the acoustic cues associated with fear. High arousal levels would be expected with this emotion, and this is supported by evidence showing increases in mean F0, in F0 range, and high frequency energy. Rate of articulation is reported to be higher. An increase in mean F0 has also been found for milder forms of the emotion such as worry or anxiety.

Sadness: As with fear, the findings converge across the studies that have included this emotion. A decrease in mean F0, F0 range, and mean energy is usually found, as are downward directed F0 contours. There is evidence that high frequency energy and rate of articulation decrease. Most studies have investigated the quieter, subdued forms of this emotion rather than the more highly aroused forms such as desperation. The latter variant might be characterized by an increase of F0 and energy.

Joy: This is one of the few positive emotions studied, most often in the form of elation rather than more subdued forms such as enjoyment or happiness. Consistent with the high arousal level that one might expect, we find a strong convergence of findings on increases in mean F0, F0 range, F0 variability and mean energy. There is some evidence for an increase in high frequency energy and rate of articulation.

Disgust: The results for disgust tend to be inconsistent across studies. The few that have included this emotion vary in their encoding procedures from measuring disgust (or possibly displeasure) at unpleasant films to actor simulation of the emotion. The studies using

the former found an increase in mean F0, whereas those using the latter found the reverse - a lowering of mean F0. This inconsistency is echoed in the decoding literature.

Even though these results seem to indicate a rather clear acoustic differentiation of the major basic emotions, it cannot be excluded that many of the differences are due to a simple arousal factor - high sympathetic arousal, which is typical for several emotions, driving up F0, energy, and high-frequency spectral energy. The issue of whether vocal expression only indexes sympathetic arousal, rather than qualitative emotion differences (as found in prototypical facial expressions) has been one of the major concerns in this area [9].

A study recently conducted by our research group allows an advance with respect to this issue. 12 professional actors were asked to portray 14 emotions varying in intensity and valence or quality [10]. A total of 224 different portrayals, 16 per emotion category, were presented to judges who were asked to decode or infer the emotion category intended by the sender or encoder. 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. Figure 1 presents the differences in recognition accuracy across the 14 emotions. Consistently found differences in the recognizability of different emotions are also replicated.

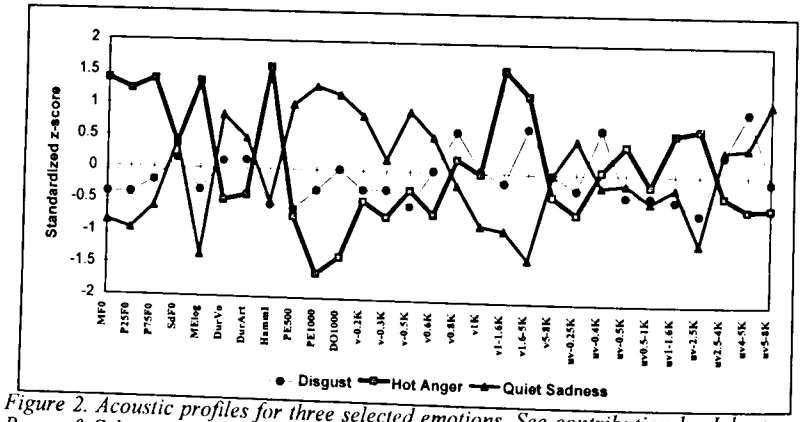

Figure 2. Acoustic profiles for three selected emotions. See contribution by Johnstone, Banse, & Scherer, this volume, for the legend of acoustic variables.

All 224 portrayals were subjected to digital acoustic analysis to obtain profiles of vocal parameters for different emotions, using a large set of acoustic variables. Figure 2 presents the acoustic profiles for some of the most interesting emotions. 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. They also suggest that with further refinement in acoustic measurement it might be possible to determine stable acoustic profiles for most emotions - provided that appropriately differentiated emotional states are used [9]. 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. It could also be shown that much of the variance in the judges' inferences could be predicted on the basis of the acoustic measurements, allowing a more detailed assessment of the acoustic cues that listeners use in inferring emotion from the voice.

APPEAL FUNCTIONS OF AFFECT VOCALIZATION

Research on the signalling or appeal aspect of vocal affect expression is particularly underdeveloped. However, it is possible to make a case for an important appeal function of vocal emotion expression in social influence settings, particularly persuasion. For example, it can be argued that appropriate emotional expression by a persuader will tend to increase the effectiveness of the persuasive message because of a) the attribution of greater credibility and trustworthiness to the sender, and b) the production of appropriate emotions in the audience which may induce the desired attitudes or behaviors or make the cognitive processing more amenable to accepting the message emitted by the persuader [11].

SYMBOLIC, REFERENTIAL FUNCTIONS OF VOCAL AFFECT EXPRESSION

I ventured a rather speculative proposal on how one might conceive of the symbolic function of vocal affect signs by arguing that the acoustic charateristics of an emotional vocalisation reflect the complete pattern of the cognitive appraisal process that produced the emotional state in the sender. This information about the criteria used in the emotion-antecedent evaluation should allow the listener to reconstruct the major features of the emotion producing event and its effect on the speaker [12]. In order to explain this postulate I have to expose some recent theorising on emotion. Many theorists in the field of psychology of emotion seem convinced that most human emotions are preceded by cognitive evaluation of events and situations (although the type of cognitive process can be relatively low level, automatic and unconscious). If this is the case, then knowing an organism's emotional state should allow us to infer the emotion eliciting cognitive processes, and thus, the approximate nature of the emotion eliciting event. If listeners are able to identify a particular emotional state of the sender from the acoustic features of the vocalisation, thus inferring the nature of the emotion producing event, then one might claim a symbolic function for emotional vocalisations. One could go even further. We are not only able to identify emotional states on the basis of acoustic cues, we may even have direct access to the results of the cognitive appraisals that have produced a particular emotional state. It is possible to elaborate predictions on how we would expect the major phonation characteristics to vary as a result of the major emotion antecedent evaluation criteria [1,9]. (The data from the actor portrayal study reported above were used to test these theoretical predictions on vocal patterning based on the component process model of emotion. While most hypotheses are supported, some need to be revised on the basis of the empirical evidence.)

If this line of reasoning is correct, one might conclude that by appropriate inferences from particular acoustic cues, receivers should be able to judge not only the nature of the emotional state of the speaker but also, and maybe even more directly, the outcomes of the pattern of cognitive appraisals which have produced the respective emotional states. In consequence, listeners should also be able to infer the approximate nature of the emotion producing event or situation as well as information about the speaker's ego involvement and coping potential. If this were the case, and if this effect were to be powerful enough to transcend individual ideosyncracies and the influence of contextual clues, then one would be justified in claiming a symbolic representational function for nonverbal vocal affect expression.

EMOTIONAL EXPRESSION IN SINGING

One can argue that emotion vocalizations might be at the root of all of human speech and singing [13]. It is not surprising, then, that much of what has been said above about multiple functions and multiple determinants is also true for singing. The acoustic signal produced by a singer reflects his or her emotional state, produces affect in the listeners, and often symbolizes abstract notions about emotionality (as shown.for example, in the *Affectenlehre* of Baroque opera). Reviews of the literature on all three of these aspects can be found in [14,15].

Unfortunately, empirical work is scarce in this area. I will conclude this contribution with an illustration of a recent study of our group on emotional expression in operatic singing [16]. Two

excerpts from the cadenza in Ardi gli incensi from Donizetti's opera Lucia di Lammermoor were acoustically analyzed for five recorded versions of the air by Toti dal Monte, Maria Callas, Renata Scotti, Joan Sutherland, and Edita Gruberova. The measured acoustic parameters of the singing voices were correlated with preference and emotional expression judgments, based on pairwise comparisons, made by a group of experienced listener-judges. In addition to showing major differences in the voice quality of the five dive studied, the acoustic parameters permit one to determine which vocal cues affect listener judgments. Furthermore, two component scores, based on a factorial-dimensional analysis of the acoustic parameters, allow the prediction of 84% of the variance in the preference ratings. Thus, we were able to show 1) that the different interpretations elicited significantly different listener ratings of emotional expressiveness, 2) that the voice samples of the five singers differ quite substantially with respect to objective acoustic variables, and 3) that we can quite successfully predict listener attributions on the basis of the objective acoustic characteristics.

REFERENCES

[1] 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.

[2] Marler, P. (1984), Animal communication: Affect or cognition? In K.R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 345-368), Hillsdale, N.J.: Erlbaum.

[3] Bühler, K. (1934), *Sprachtheorie*, Jena: Fischer (new edition 1984).

[4] Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984), Vocal cues to speaker affect: Testing two models, *Journal of the Acoustical Society of America*, vol 76, pp. 1346-1356.

[5] Scherer, K.R., Feldstein, S., Bond, R.N., & Rosenthal, R. (1985). Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research*, vol. 14, pp. 409-425.

[6]Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985), Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect, *Journal of the Acoustical Society of America*, vol. 78, pp. 435-444.

[7] Bergmann, G., Goldbeck, T., & Scherer, K.R. (1988), Emotionale Eindruckswirkung von prosodischen Sprechmerkmalen, *Zeitschrift für experimentelle und angewandte Psychologie*, vol. 35, pp. 167-200.

[8] 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.

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

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

[11]Scherer, K. R. (1993), Interpersonal expectations, social influence, and emotion transfer. In P. D. Blanck (Eds.), *Interpersonal expectations: Theory, research, and application* (pp. 316-336), Cambridge and New York: Cambridge University Press.

[12] Scherer, K. R. (1988), On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology*, vol. 7, pp. 79-100.

[13] Scherer, K.R. (1991), Emotion expression in speech and music. In J. Sundberg, L. Nord, & R. Carlson (Eds.). *Music, Language, Speech, and Brain.* Wenner-Gren International Symposium Series: vol. 59. (pp. 146-157), London: Macmillan.

[14] Scherer, K.R. (in press), Expression of emotion in voice and music. *Journal of Voice*.

[15] Sundberg, J. (1987), *The science of the singing voice*, DeKalb, IL: Northern Illinois University Press.

[16] Siegwart, H. & Scherer, K.R. (in press), Acoustic concomitants of emotional expression in operatic singing: The case of Lucia in Ardi gli incensi, *Journal of Voice*.