THE ENCODING OF AFFECT: A REVIEW AND DIRECTIONS FOR FUTURE RESEARCH

J. Pittam* and K.R. Scherer†

*University of Queensland, Australia

†University of Geneva, Switzerland

This paper reviews the work conducted into the encoding of affect in synthesised speech signals and isolates a number of major problem areas that researchers need to consider in the future.

INTRODUCTION

It has long been recognised by those working in the various areas of speech science and technology that there is a need to come to terms with the non-linguistic elements of voice and speech if one of the fundamental goals of the discipline is to be realised: the construction of machines capable of speaking in a clear and natural sounding manner. Over and above the realisation of this goal, however, numerous current and projected applications require specific information about long-term speaker effects. Yet, even today, when details of the segmental elements of speech are relatively well understood, and even the suprasegmental areas are receiving attention, little work has been conducted on longer-term communicative dimensions such as emotion or the various aspects of social identity. Emotion, which is the focus of the present paper, has received more attention than the latter, although in absolute terms there has been little consideration given either to the encoding of discrete emotions and more general affective states into synthesised speech signals, or the computer recognition of affect.

ACOUSTIC BASES OF EMOTION COMMUNICATION

Changes in emotional state result in changes to respiration, phonation and articulation, which lead to changes in the acoustic signal. Listeners can then infer affective state from that signal. Unfortunately, there is little systematic knowledge about the details of the encoding process (involving externalization of underlying psychophysiological states and realization of culturally determined display characteristics) or the decoding process (with respect to the acoustic cues listeners use in inferring speaker state), although work has appeared in both areas over the last few years. For one recent example of synthesised affective states, see Cahn (1990).

Physical measures of human speech and vocal sounds are based on three perceptual dimensions: loudness, pitch, and time. The central concern is the selection and measurement of appropriate acoustic cues from these dimensions. One selection criterion that can be adopted is to use acoustic variables and measurement techniques that are relatively long-term or at least suprasegmental, since affect is expected to last longer than individual speech segments. Adoption of such a criterion does not mean we must ignore short-term measures related to individual speech sounds, but that we will often use them in ways that simulate longer time frames. Thus, formant frequencies can be averaged over many instances of the same vowel to give an overall frequency measure. Table 1 lists major physical measures that have been linked to emotion.

In addition to using these three dimensions individually, more complex variables based on more than one dimension have been calculated. One measure combining amplitude

Major Types of Physical Measure
FO mean; range; variability; perturbation; contour details
Formant means; bandwidth
Intensity mean; range; variability
Speech rate; pausing
Slurring of articulation
Spectral noise
Proportion of high frequency energy to low
LTS contour and frequency band
Short-term spectral envelope measures

Table 1. Major types of physical measure

and time - the 'vocal envelope' (Scherer & Oshinsky, 1977) - measures the time taken for an auditory signal to reach maximum amplitude, and the time it takes to decay to zero amplitude. Other variables combine amplitude and frequency. The higher formants, for example, vary not so much with vowel quality as an individual speaker's vocal tract. As carriers of affect, therefore, they are potentially useful, although this is lessened by the difficulty often experienced in locating them. Another long-term measure of voice that combines these dimensions is the long-term averaged spectrum (LTS). LTS have been calculated for the voiced and unvoiced sections of speech combined, and for the voiced sections alone. The one selected may depend on the type of affect studied (i.e., whether the affect is carried by the voiced parts alone).

When we hear a voice speaking, acoustic cues from all three dimensions enter our auditory system. It is not always clear, however, just what the relationship is between these physical cues and the way they are perceived. Even the relationship between FO and pitch is complex. The relationship between many other acoustic cues and their perceptual counterparts, particularly the more complex measures such as the LTS, remains largely unexplored. It is not clear, for example, how the various parts of the LTS relate to one another or how they might be used perceptually. Different emotions or affect states may be difficult to distinguish in the LTS unless the lower frequency band, covering the highest point that FO reaches, is excluded. Large energy shifts extraneous to emotion can occur across quite wide frequency bands in the lower regions. Even filtering out low frequency bands, however, will not remove all traces of FO. It will still be reflected in the remaining harmonic structure. Corrections for changes in FO that affect the harmonics may also be needed, therefore.

THE ENCODING OF DISCRETE EMOTIONS

A number of discrete emotions have been studied, although apparent contradictions occur on which acoustic correlates are important for which emotions and in which way. One possibility for this may be that different studies measure different types of emotion while using the same broad category label. Anger, for example, seems to cover a range of types from irritation to rage. We present here a brief overview of findings for five broadly labelled emotions that have been regularly studied. For more detailed accounts and more comprehensive reviews of the literature, see Pittam and Scherer (in press) and Scherer (1986).

Anger: The frequency domain seems to be particularly important for the encoding of anger, although intensity also has been found to play a vital role. Whether it is cold or hot anger, this emotion seems to be characterised by an increase in mean FO and mean intensity. Some studies, which may have been measuring hot anger, show increases in FO variability and in the range of FO across the utterances encoded. Other studies, however, have not found these characteristics, although they may have been measuring cold anger. Other anger effects include increases in high frequency energy and downward directed FO contours. The rate of articulation usually goes up.

Fear: There is considerable agreement on the type of acoustic cue associated with fear. High arousal levels would be expected with this emotion, and this is supported by evidence showing increases in mean FO, in FO range, and high frequency energy. Rate of articulation is reported to be speeded up, while an increase in mean FO 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 FO, FO range, and mean intensity is usually found. There is also evidence for downward directed FO contours. There is evidence that high frequency energy and the rate of articulation decreases. Most studies reported in the literature have studied the quieter, resigned forms of this emotion rather than the more highly aroused forms such as grief or desperation.

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

Disgust: The results for disgust tend not to be consistent across the encoding studies. The few that have included this emotion vary in their induction 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 FO, whereas those using the latter found the reverse - a lowering of mean FO. This inconsistency is echoed in the decoding literature.

From this brief review it becomes evident that where there is considerable consistency in the findings, it is usually related to the dimension of arousal, and in particular the high arousal emotions such as anger, fear and elation. Given the relationship between such emotions and the sympathetic nervous system, this is perhaps not too surprising. It should not be taken as evidence that discrete emotions are not differentiated by vocal cues, however. There are several reasons why there is as yet little evidence for vocal differentiation of individual emotion states. The range of acoustic cues used, for example, has been very limited in the majority of studies. Many others could be used, including the more complex measures derived from short-term and long-term spectra. Also, there is a definite need to differentiate emotions more precisely. Even at the level of affective dimensions more differentiation is needed. Arousal is only one dimension, albeit an important one. There has been little effort made to systematically study other emotional dimensions such as control and valence - although see Pittam, Gallois and Callan (1990) who found evidence that these dimensions were reflected in different frequency bands of the LTS.

RESEARCH PROBLEMS AND PERSPECTIVES FOR THE FUTURE

Emotion induction and research design

The central problem shared with most other approaches to the study of emotion, is the difficulty of studying real and strong emotions in situ or producing such states in the laboratory. Very few studies have used naturally occurring emotions. Consequently, researchers have either studied emotion portrayals by actors (see Scherer, Banse, Wallbott, & Goldbeck, 1991, for a detailed review of this type of approach) or have used ethically acceptable induction techniques, resulting in relatively weak affect states of the subjects studied. Both approaches entail advantages and disadvantages. While actor portrayed emotion states are generally clearly differentiated and of sufficient intensity, it is conceivable that the actors portray, at least in part, socially shared expression prototypes. Induction approaches face the difficulty of ascertaining that the elicitation procedure has actually produced the same emotion in different subjects. This is both methodologically and practically difficult, especially if the intensity is low. Data stemming from these two approaches needs to be compared systematically in an effort to determine how much of the acoustic variations in actor emotion portrayals are due to theatre conventions or cultural display rules, and how much of what is not found in induction studies might be due to lack of intensity or clear differentiation of different states.

To begin with, the emotion categories studied should not be limited to the basic ones, but be more differentiated (e.g., cold vs. hot anger as well as compound emotions such as contempt). In addition, given the importance of the arousal and valence dimensions to emotions, it might be useful to obtain appropriate judgments on these dimensions, in order to calibrate the emotion labels used. If one could compare the respective locations in this two-dimensional space of emotion states studied by two different researchers, one might be able to get clues that help to explain differences in findings and interpretations. Such a procedure might also be useful to better understand the link between physiological processes underlying arousal and valence and the voice production process.

Recent studies have shown the importance of personal and social identity in emotion encoding in the voice particularly in terms of gender differences and interindividual variation. Much of what looks like lack of replication may well be due to inadequate sampling, therefore. The same is true with respect to contextual cues (e.g., the nature of the respective social situation, audience effects, task demands, etc.) which have all been shown to affect speech and voice production. Unless these variables are tightly controlled, it is difficult to expect replication between studies.

One possible solution to the difficult task of bringing about greater comparability between studies in this area might be a stronger move toward applied settings. In particular, the development of a knowledge base to help computers understand emotionally toned speech and to produce affectively variable synthetic utterances might provide the constraint needed to keep different researchers focussed on a similar set of variables and procedures. The automatic inclusion of evaluation procedures to measure the performance of the system is likely to provide a set of comparable measures.

Multifactor determination of vocal characteristics

Vocal cues studied in this area are determined by biological, linguistic, and sociocultural factors. For example, intensity, F0 and temporal factors have all been shown to serve linguistic functions, to communicate personality characteristics and index individual differences, in addition to expressing physiologically mediated affect states and their culturally based display norms. There is also some evidence suggesting that prosody and emotion are interrelated in the acoustic signal. Liberman (1978) has pointed to the slight variations in F0 contour that provide the listener with information about subtle shifts in affect and attitude, while Cahn (1990) has shown a relationship between prosodic stress and emotion. Evidence is beginning to emerge, however (Hermansky and Cox, 1991), that linguistic and speaker dependent information can be separated during speech analysis, and merged during resynthesis. If so, it would be a significant step forward.

Three factors likely to be essential in trying to untangle biological, linguistic and sociocultural determinants are the universality of the expression of emotion, redundancy in the speech signal, and the variability of the various features in the speech signal. Some aspects of the vocal expression of emotion are thought to be universal and even phylogenetically continuous, and there is a considerable amount of redundancy in the speech signal, although we do not yet understand how much there is or what it signifies. Also, Ellman and McClelland (1986) point out that it is precisely the variability in the speech signal that permits listeners to understand speech in a variety of contexts and spoken by many speakers. While this may seem paradoxical, it does indicate that the study of vocal emotion communication could benefit from psycholinguistic work on speech production and perception (and vice versa). Most likely, the close interaction between linguistic and socio-affective information in speech can only be disentangled by close collaboration of the various disciplines concerned.

It is increasingly recognized that affectively toned vocalizations are jointly determined by an externalization of internal states and the requirements of species- or culture-specific normative models for affect signals or displays. Any model for generating affect in synthesised signals, therefore, must take account of both. It seems clear that we will not make any headway on this without conducting more cross-cultural and cross-species comparison studies.

One of the most urgent needs for change in research policy, however, concerns the theoretical basis for studies on emotion encoding and decoding with respect to vocalization. Given the large number of factors that affect voice production, a cumulation of empirical studies without clear guidance by theoretically based hypotheses is unlikely to advance our understanding of the phenomenon. Scherer (1986) has suggested an extensive set of theoretical predictions based on both emotion theory and recent work on voice production. Some of these predictions have been supported in recent studies using actor portrayals of emotions (Scherer, Banse, Wallbott, & Goldbeck, 1991). Further work in this direction, using more naturalistic emotional expressions and the integration of hypotheses concerning socio-cultural patterning of the expressions would greatly strengthen the case. Finally, it seems most useful to link the study of encoding and decoding in an overall research design (as discussed by Scherer, 1986, for example). Such an approach not only provides an overall analysis of the complete vocal emotion expression and communication process, it can also help disentangle the different factors that determine accuracy of recognition or lack of it. Table 2 summarises the issues raised in this paper that need to be considered the goal of successfully encoding affect into synthesised signals is to achieved.

Issues Needing Consideration in the Encoding of Affect
The relationship between acoustic cues and perception
Range of acoustic cues used
Greater differentiation of emotions and affective dimensions
Calibration of emotion labels
Systematic comparison of emotion induction techniques
Exploration of identity and contextual factors
Relationship between 'push' and 'pull' effects
Development of knowledge base of emotion production
Multifactor determination of vocal features
Relationship between emotion and prosody
Theoretically based hypotheses
Model incorporating encoding and decoding of affect

Table 2. Issues needing consideration in the encoding of affect

REFERENCES

Cahn, J.E. (1990) *The generation of affect in synthesized speech*, Journal of the American Voice I/O Society, 8, 1-19.

Ellman, J.L., & McClelland, J.L. (1986) *Exploiting lawful variability in the speechwave.* In J.S. Perkell & D.H. Klatt (Eds.) *Invariance and variability in speech processes* (pp. 360-380), (Hillsdale, NJ: Lawrence Erlbaum).

Hermansky, & Cox, (1991) *Perceptual linear predictive (PLP) analysis-resynthesis technique*, Proceedings of the Second European Conference on Speech Communication and Technology (pp. 329-332), (Genova, Italy, September).

Liberman, M.V. (1978) *The intonational system of English*, (Bloomington: Indiana University Linguistics Club).

Pittam, J., Gallois, C., & Callan, V.J. (1990) *The long-term spectrum and perceived emotion*, Speech Communication, 9, 177-187.

Pittam, J., & Scherer, K.R. (in press) *Vocal expression and communication of emotion.* In M. Lewis, & J. Haviland (Eds.). *The Handbook of Emotion*, (New York: Guilford Press).

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

Scherer, K.R., Banse, R., Wallbott, H.G., & Goldbeck, T. (1991) *Vocal cues in emotion encoding and decoding*, Motivation and Emotion, 15, 123-148.

Scherer, K.R., & Oshinsky, J.S. (1977) *Cue utilization in emotion attribution from auditory stimuli*, Motivation and Emotion, 1, 331-346.