PROSODIC SIGNS OF EMOTION IN SPEECH: PRELIMINARY RESULTS FROM A NEW TECHNIQUE FOR AUTOMATIC STATISTICAL ANALYSIS

S. McGilloway, R. Cowie & E. Douglas-Cowie
School of Psychology / School of English, Queen's University, Belfast, UK

ABSTRACT

Emotional expression has been studied using a technique in which automatic preprocessing extracts key speech features and statistical descriptions of them are generated. Five broad types of marker are found to distinguish among passages - spectral balance, range of pitch movement, timing of pitch movement, timing of intensity changes, and intensity distribution.

INTRODUCTION

There are many reasons for trying to analyse the vocal expression of emotion in objective and quantitative terms. It is a natural extension of research on machine speech to explore methods of recognising and generating speech which is not emotionally neutral [1]. Social and behavioural studies could benefit from reliable methods of measuring vocal signs related to emotion. Clinical applications also exist. Our own interest in the area stems from one of these. Lack of emotional expression in speech is diagnostically important in schizophrenia [2], but the absence of formal measures in the area impedes refinement and evaluation of diagnostic practices, as well as leaving critical decisions to depend on the ear of someone who need have no particular aptitude for phonetics. We studied markers of emotion in normal speech as a necessary part of a project concerned with that clinical problem.

The study uses a system called ASSESS which is described more fully elsewhere in these Proceedings [3]. It extends earlier work on the statistical description of speech [4]. The main innovation is that statistical description is preceded by preprocessing which extracts key features of the speech signal and simple units associated with them. ASSESS then generates an array of approximately 400 statistics specifying the attributes and variations of these features.

METHOD

Speech sample

Passages were constructed to suggest four emotions - fear, anger, sadness, and happiness. A fifth, emotionally neutral passage was used as a baseline. The passages were of comparable lengths, taking about 25-30 seconds each to read.

Speakers were 40 volunteers from the Belfast area, 20 male and 20 female, aged between 18 and 69. There was a broad distribution of social status, and accents represented a range of local types.

Subjects familiarised themselves with the passages first and then read them aloud using the emotional expression they felt was appropriate. They were presented in computer-generated random orders.

Recordings were digitised using a CED 1401 signal capture system. Sampling was at 20kHz, after low pass filtering at 10kHz. System limitations meant that files had to be entered in sections of 7 seconds or less and rejoined at a later stage of processing. Splits were placed by hand within substantial pauses.

Acoustic analysis

ASSESS is based on standard descriptors: the speech spectrum; the intensity contour; and the pitch contour. It breaks these up into significant units using techniques chosen for robustness rather than elegance or precision, otherwise hand correction would be essential. Contours are smoothed before finding inflections, i.e. points at which volume or pitch stops rising and starts falling, or vice versa. Rises, falls, and plateaux (periods of relatively flat pitch or intensity flanking an inflection) are then found. Spectral information is used to identify transitions which mark at least roughly natural units. Four types are considered - silences, sound blocks, tunes, and fricative bursts. Sound blocks are defined by the way intensity rises after a silence, peaks, and falls to the next silence. Tunes are defined by the way pitch rises and falls between silences long enough to be considered pauses. Fricative bursts are defined by the distribution and amount of energy high in the spectrum. Subspectra are formed for special types of episode, fricative busts and peaks in the intensity contour.

Properties and relationships of these units are summarised in a battery of statistics, primarily measures of midpoint and spread, generally in parametric and nonparametric forms (the latter are less sensitive to occasional erratic data points). Lines and curves are also fitted to specify the shape of tunes and spectra.

ASSESS can estimate absolute intensity by using a calibration signal with a known dB level. However in this study no absolute referent was available, and level was normalised by treating the first file in a passage as a referent and setting its median intensity at 60dB. This seems unlikely to confound the results.

Statistical analysis

The basic statistical procedure was analysis of variance followed by post hoc comparison of group means using Duncan's range test. Characteristics were considered distinctive only if the overall analysis of variance was significant with $p<0.05$ and the emotional passage in question contrasted with the neutral passage with $p<0.05$. Passage was treated as a between groups variable for both tests. This is conservative, i.e. it is likely to conceal real differences rather than generate spurious ones. Sex was considered as a third variable in analyses involving pitch height, because otherwise variance is inflated by sex differences and effects of emotion are swamped.

RESULTS

There was wide range of differences between passages - over 1/3 of the measures considered yielded significant differences. The challenge is to reduce these to a manageable set.

The largest set of differences reflect an effect that distinguishes two broad groups of passages: afraid, angry and happy on one hand, sad and neutral on the other. They involve intensity contrasts. It seems apt to call the groups intensity marked and intensity unmarked respectively.

Table 1 shows the main features of the effect. Measures are in bold face if they are significantly different from the neutral passage. The first two columns show intensity measures for all points outside pauses. These global measures are higher for fear, anger and happiness than for sad and neutral passages. However, intensity marking is not a simple matter of loudness. ASSESS reveals two types of structure in it.

Table 1 : Selected intensity contrasts between groups (normalised scale).

	mean	median	peaks	troughs
Anger	**64.11**	**61.57**	**66.87**	**59.97**
Fear	**63.64**	**61.51**	**66.45**	59.57
Happiness	**63.38**	**61.59**	**66.07**	59.52
Sadness	62.42	60.32	65.10	59.12
Neutral	62.33	60.73	64.87	58.83
p	0.000	0.003	0.000	0.095

First, note that intensity is normalised. Hence the first two columns do not mean that the first three emotions are associated with louder speech: it means that intensity rises after the first few phrases. This may be called a crescendo effect.

Second, note that the effect is more marked with means than with medians. That suggests it involves stretching in the top end of the intensity distribution rather than just a global upward shift. The inference is confirmed by the last two columns. The contrast in the level of peaks in the intensity contour is even more marked than the contrast in overall mean. However, there is much less contrast in the level of troughs (that is, minima). It is not significant overall, and the Duncan test shows only anger differs significantly from the neutral passage.

There is a trend for silences to be longer in the intensity marked passages, which is just short of significance ($p=0.051$) and most marked in anger. This is consistent with the general pattern of heightened dynamic contrast in the intensity marked passages.

Several other features distinguish intensity marked passages from the neutral passage, and to a greater or lesser extent distinguish them from each other.

Properties involving the duration of intensity features may tend to signal negative emotions: they do not affect happiness, and they may affect sadness. Table 2 summarises the data.

The durations of amplitude movements distinguish fear and anger from the neutral passage again. Both have longer median durations for both falls and rises. But in contrast to the crescendo and intensity stretching effects, this effect is stronger in fear than anger. Protracted intensity falls also characterise sadness. The durations of tunes show a similar pattern. Also broadly similar is a property of intensity plateaux. The interquartile range of their duration increases markedly in fear and sadness, and less so in anger.

Table 2: Aspects of duration associated with negative emotions (times in ms).

	rises median	falls median	tunes mean	plateau IQR
Fear	**82.35**	**84.8**	**1265**	**10.8**
Anger	**81.66**	**80.5**	**1252**	10.2
Happiness	78.03	77.4	1404	8.2
Neutral	78.50	77.2	1452	8.4
Sadness	77.28	**81.4**	**1179**	**11.0**
p	0.000	0.000	0.001	0.006

The passages differ in the distribution of energy across the spectrum, but few of the effects are easy to interpret.

Most straightforwardly, all the emotions are characterised by greater variability in the duration of fricative bursts (as measured by the standard deviation) than the neutral passage.

A second clear effect involves anger. Here the average spectrum for non-fricative portions of speech has a high midpoint. This is not surprising: it parallels a well-known effect of tension on spectral balance [5]. Conversely, the sad passage gives a significantly lower spectral midpoint than any of the intensity marked passages - it is lower even than the neutral passage.

Fricative bursts are associated with a number of effects which seem paradoxical at first sight. Anger is associated with high average energy in fricative bursts, but the average spectrum for slices classed as fricative has a low mean and a markedly negative slope. The implication appears to be that the intensity associated with frication is not rising as fast as the intensity associated with the lower spectrum. Fear and happiness are distinctive in terms of the subspectrum which shows variability in slices classed as fricative. These too show markedly negative slopes, indicating relatively low variability in the regions associated with frication. The effects may be less to do with frication than with raised variability in the lower spectrum.

Two aspects of the pitch contour show differences - the distribution of pitch height and the timing of pitch movement.

Passages do not differ significantly in pitch height per se. However, they do differ in its distribution. Again, the differences which are clearly significant fall into an orderly pattern. All of them involve interquartile intervals, which can be thought of as measures of the range a measure usually occupies. When all pitch inflections are considered together, the passage difference in interquartile interval just reaches significance ($F\ 4, 185 = 2.91$, $p=0.023$). Separating maxima and minima shows a weak passage effect for minima ($F4, 185=2.74$, $p=0.03$) and a much stronger one for maxima ($F4, 185 = 3.76$, $p=0.006$). In all three cases, range is widest for happiness and nearly as wide for anger, with the lowest range in the neutral passage.

All the distinctive pitch duration features are associated with happiness. Pitch plateaux are shorter in the happy passages than elsewhere, and their durations generally lie within a narrower range (as measured by the inter quartile range). Conversely, pitch falls last longer in the happy passage than in the neutral one. This feature is shared with the sad passage. Pitch rises are also significantly faster in the happy passage than in the neutral passage. The overall picture is that happiness involves pitch movement which is not only wide, but constant.

The outline of the findings can be summarised in a table. This shows that each of the passages is distinguished in several ways from any other.

Table 3: Summary of distinctions among passages

	Afraid	Angry	Happy	Sad
Intensity				
• marking	+	+	+	
• duration	+	+		+
Spectrum				-
• midpt & slope		+		
Pitch movement				
• range		+	+	
• timing			+	+

DISCUSSION

It has been pointed out that the corpus of research on emotion in speech is not large, and studies tend to agree only partially among themselves [6]. Our main claim for this study is that it demonstrates the potential of an approach which may be relevant to those problems.

One reason for not drawing stronger conclusions lies in our speech sample. The passages do generally convey the emotions that they are meant to, but they are of simulations of emotion, made by people who have no expertise in simulating it. An obvious need is to obtain samples of genuinely emotional speech. That presents both practical and ethical difficulties, and it would also aggravate a problem which is present in this study, which is to distinguish effects due to linguistic content from effects due to emotion. It seems unlikely that the strongest features we have noted are due to linguistic content, but the possibility should be acknowledged.

With that qualification, our data make a simple point. Statistics which can be extracted automatically and conveniently do seem to distinguish emotional speech episodes. They include statistics of a higher order than the global measures of mean and range which have been in use for half a century [7],[8], and it seems likely that distinctions can be sharpened by using these higher order measures.

A significant theoretical attraction of reducing description at this level to numbers routinely extracted is that it frees us to explore pattern at a different level. We have noted that our passages are distinguished by feature combinations, and share features with each other. It is a natural extension to conjecture that different expressions of the same emotion bear similar relationships, with some features in common, but not all. This suggests a geometric picture which is familiar in research on automatic classification: emotions may be thought of as regions in a multidimensional space where points (corresponding to episodes of speech) are positioned by the strength of various attributes.

Theory apart, our approach has an obvious practical attraction. It points quite directly towards automatic methods of recognising emotion in speech. It seems clear that there are rather complex linguistic cues to emotion in speech [9], and capturing them automatically remains a long term project. However, using essentially simple statistical techniques seems a reasonably immediate prospect.

ACKNOWLEDGEMENT is due to Drs D Sykes and C Cooper for statistical advice.

REFERENCES

[1] Murray, I., Arnott, J. and Newell, F. (1988), "Hamlet: simulating emotion in synthetic speech", *Proc. 7th FASE Symposium*, Edinburgh, pp. 1217-1223.

[2] Andreasen, N., Alphert, M. and merrill, J. (1981), "Acoustic analysis: an objective measure of affective flattening", *Arch. Gen. Psychiatry*, vol 38, 281-285.

[3] Cowie, R., Sawey, M. and Douglas-Cowie, E. (1995), "A new speech analysis system: ASSESS (Automatic Statistical Summary of Elementary Speech Structures)", *Proc.13th International Congress of Phonetic Sciences*, Stockholm.

[4] Cowie, R., Douglas-Cowie, E. and Rahilly, J. (1991), "Instrumental measures of abnormalities in deafened speech", *Proc 12th International Congress of Phonetic Sciences*, Aix-en-Provence, pp. 350-353.

[5] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. "Perceptual and acoustic correlates of voice qualities" *Acta Otolaryngologica* vol. 90, pp. 441-451.

[6] Murray, I. and Arnott, J. (1993), "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *J. Acoust. Soc. Am.*, vol. 93 (2), pp. 1097-1108.

[7] Skinner, E. (1935), " A calibrated recording and analysis of the pitch force and quality of vocal tones expressing happiness and sadness", *Speech Monog.*, vol. 2, pp. 81-137.

[8] Fairbanks, G. and Pronovost, W. (1939), An experimental study of the pitch characteristics of the voice during the expression of emotion", *Speech Monog.*, vol. 6, pp. 87-104.

[9] Ladd, R., Silverman, K., Tolkmitt, F., Bergmann, G. and Scherer, K. (1985), "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect", *J. Acoust. Soc. Am.*, vol. 78 (2), pp. 435-443.