AUTOMATIC STATISTICAL ANALYSIS OF THE SIGNAL AND PROSODIC SIGNS OF EMOTION IN SPEECH

Roddy Cowie & Ellen Douglas-Cowie

School of Psychology / School of English
The Queen's University of Belfast
Belfast BT7 1NN, Northern Ireland

ABSTRACT

We highlight two broader domains surrounding specific attributions of emotion and the specific features of speech that underlie them, and argue for caution over compartmentalising these broader domains. It seems to be a general rule that variations in what we call the augmented prosodic domain (APD) are emotive - perhaps because they signal departure from a reference point corresponding to a well-controlled, neutral state. Our studies show that various departures from that reference point are reflected in the APD, including central and sensory impairments (schizophrenia and deafness) as well as emotion. Intuitively it seems right to acknowledge that departures from well-controlled neutrality are highly confusable, and it is unclear that phonetics should to try draw those distinctions more sharply than listeners tend to. A system called ASSESS automatically measures properties in the APD, opening the way to explore it in an empirical spirit.

1. EMOTION: IN WHAT CONTEXT?

Any interesting topic is worth approaching from more than one angle. We have come to the study of emotion from an angle that is somewhat different from most people's. The result is a distinctive perspective which we try to convey here. We regard it as a natural counterbalance to other approaches, expressing ideas which ought to be kept in mind as a null hypothesis even if one chooses to follow an alternative approach.

So far as the symposium is concerned, the distinctive idea in our approach is that emotion is not a special problem: it is properly handled in the context of wider concerns. That idea applies at two levels. At the level of understanding listener responses, we have treated emotion as part of a wider domain concerned with evaluative judgements about speakers and reactions to them. At the level of understanding relevant speech variables, the variables that we have related to emotion form part of a domain that we will call the Augmented Prosodic Domain. We explain below what that entails.

We stress that we are not interested in doctrinal confrontations. We believe that our approach captures enough of the truth to be useful in a range of contexts. The same is almost certainly true of approaches which focus more selectively on emotion. We assume that in the longer term more of the truth will be captured when approaches are formulated that combine the merits of both. In the meantime both approaches have practical uses.

Practically, our approach rests on a system called ASSESS [1]. The core of ASSESS is a highly simplified representation of the speech signal based on a few features that we call landmarks. The main landmarks are peaks and troughs in the profiles of pitch and speech intensity and the boundaries of pauses and fricative bursts. These and a few other parameters define a representation of the speech signal which is equivalent to (and can be used to generate) a sketch of its main prosodic features.

The early stages of ASSESS generate this core representation automatically. There is nothing very profound about the methods that they use. They are chosen for robustness rather than quantitative precision.

The later stages of ASSESS derive summary statistics from the core representation. Most of them are straightforward, dealing with attributes like the range and midpoint of rises and falls in intensity or pitch. A few are more sophisticated, like the parameters of the quadratic functions that best fit 'tunes' in the sample (a tune being the portion of the pitch contour that lies between two pause boundaries). The statistics also provide information on average spectra and spectral change, using the landmarks to find (e.g.) subspectra associated with intensity peaks (which tend to correspond to vowel centres).

The point of the approach is to capture a subset of the speech signal which is intuitively natural and empirically significant. Intuitively, it summarises the output of channels concerned with prosody and some aspects of voice quality. For convenience, we will say that it deals with an augmented prosodic domain, or APD for short. Its empirical significance is that information about that domain seems to account for a considerable proportion of the judgements and evaluations that people make on the basis of speech.

2. DEAFENED PEOPLE'S SPEECH

The problem that led us into the augmented prosodic domain was describing what happens to people's speech when they lose their hearing. When we began work in that area, the literature tended to suggest that deafened people's speech was not a problem. It did not generally become unintelligible, and nothing further needed to be said or done. Goehl and Kaufman underlined that view in a memorable exchange [2, 3].

Our experience with deafened people led to a different view: even when intelligibility is not a problem - and sometimes it is - there certainly is a problem with the reactions and impressions that deafened people's speech evokes. Many of the reactions involve attributions of emotion.

We probed these reactions in a questionnaire study [4]. Factor analysis identified seven main themes in the response patterns. Of these three related to inferred emotions - warmth, social

poise (which included lack of anxiety or timidity), and stability. The others were competence, plus motor disability, intellectual disability, and aversiveness. Judgements on all of these variables related to level of hearing loss. This is to say that speech variables which actually reflected the limitations of control imposed by impaired hearing were misinterpreted by listeners as signals of the speakers' emotional makeup.

A predecessor of ASSESS suggested what the speech variables in question might be. Judged stability correlated with relatively slow change in the lower spectrum. Judged poise correlated with narrow variation in F0 accompanied by wide variation in intensity. Judged warmth correlated with a predominance of relatively simple tunes, a tendency for change to occur in the mid-spectrum rather than at the extremes, and a low level of consonant errors (the last was established by phoneticians, not by the ASSESS-like analysis). Competence was associated with the pattern of changes in the intensity contour.

These measurements were made in the context of understanding deafness and its effects, and published accounts describe them in that context. However, they can also be set in the context of emotion. In that context, they make the point that the enhanced prosodic domain carries information about a range of speaker attributes. They also make the point that listeners do not fully disentangle the various factors that actually impinge on the domain - in particular, they do not disentangle the effects of impaired control due to hearing loss and emotionality.

ASSESS developed the idea that the kinds of measures which those studies had considered form part of a natural domain with recurring links to issues of emotion and evaluation. A large study which is currently under way applies ASSESS to deafened speakers in order to evaluate the effects of cochlear implants [5, 6]. The analysis confirms that the augmented prosodic domain is highly sensitive to differences between deafened speakers and controls. However, we have also applied the approach to other domains. One of the first was emotion per se.

3. ASSESS APPLIED TO EMOTION

Our study of emotion used passages 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.

Recordings were digitised at 20kHz, after low pass filtering at 10kHz. ASSESS can estimate absolute intensity by using a calibration signal with a known dB level, but here no absolute referent was available, and level was normalised by treating the opening of a passage as a referent and setting its median intensity at 60dB. This seems unlikely to confound results. Technical information on statistics is given elsewhere [7].

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.

	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

Table 1: Selected intensity contrasts between groups.

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).

Pauses were longer in the intensity marked passages, most markedly so in happiness and sadness. 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.

	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

Table 2: Duration features and negative emotions (ms).

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.

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. That is not surprising: it parallels a well-known effect of tension on spectral balance [8]. 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. Separating maxima and minima shows a weak passage effect for minima and a much stronger one for maxima. In all three cases, range is widest for happiness and nearly as wide for anger, with the lowest range in the neutral passage.

	Afraid	**Angry**	**Happy**	**Sad**
Spectrum				
• midpt & slope		+		–
Pitch movement				
• range		+	+	
• timing			+	+
Intensity				
• marking	+	+	+	
• duration	+	+		+
Pausing				
• total			+	+
• variability				+

Table 3: Summary of distinctions among passages

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 also happens in 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.

Table 3 provides a compact outline of the findings. This shows that features in the augmented prosodic domain distinguish each of the passages from any other in several ways.

4. PROSODY IN 'FLATTENED AFFECT'

Our study of emotion is a module within a larger clinical project concerned with the attributes of speech that differentiate schizophrenics from normal speakers. Clinicians' judgements about speech are a key element in diagnosis [9,10], particularly in the context of 'flattened affect', which is linked to poor prognosis and hospitalisation [11,12]. It is obviously a matter of concern that diagnosis might be influenced by individual differences in clinicians' aptitude for impressionistic judgements about speech. We studied the possibility of providing objective measures for that reason.

Our schizophrenic sample consists of 72 subjects from Belfast who have a diagnosis of schizophrenia (DSM-III-R)[13] of more than one year's duration. At the time of testing 40 were outpatients who were attending a local psychiatric outpatient clinic and 32 were long-stay inpatients in a psychiatric hospital. The passages that they read were the same as the passages used in the study of emotion, and they were compared with the controls whose results have been outlined above.

Two broad types of comparison are worth drawing. The first involves contrasts between schizophrenic and control subjects in the delivery of an emotionally neutral passage. The second involves contrasts in the way that they differentiate between neutral passages and passages with strong emotional content. The approach that we have taken allows both to be made within a single framework, because it treats the expression of emotion as part of a broader descriptive task.

Comparing schizophrenics and controls on the neutral passage, several significant contrasts emerge. Schizophrenics showed higher mean F0, along with lower pitch variability in terms of both the amount by which pitch tended to rise between neighbouring minima and maxima, and the amount by which tunes tended to differ in their central pitch. The way they varied intensity was more stereotyped: both rises and falls in intensity showed a narrower range of variation than in controls, as did the duration of falls in intensity. Conversely, they showed increased variability in the duration of silences.

The variables involved in these contrasts are emotion-like. As the term 'flattened affect' suggests, they convey something like absence of any emotional colour, even the tints that we expect in so-called neutral speech. But it is not clear how accurate that is. The second comparison underlines uncertainty whether the issue is the existence of emotion or its expression.

Table 4 is parallel to Table 3, but it juxtaposes the contrasts between neutral and emotional passages in controls (C columns) and patients (P columns). Most of the entries are self-explanatory. The pattern of intensity marking is worth commenting on - the crescendo effect is absent in the cases where controls show it, but the opposite effect - a diminuendo - occurs in sadness, where the controls show level pitch.

	Afraid		**Angry**		**Happy**		**Sad**	
	C	P	C	P	C	P	C	P
Spectrum								
• midpt & slope			+	+			–	:
Pitch movement								
• range			+	:	+	+		
• timing					+	:	+	:
Intensity								
• marking	+	:	+	:	+	:	:	-
• duration	+	+	+	+			+	+
Pausing								
• total					+	:	+	+
• variability	:	+	:	+	:	+	+	+

Table 4: Patient and control distinctions among passages. Colons mark effects that are absent in one group but not both.

The nature of schizophrenia is controversial, but whatever the underlying disorder is, it generates disturbance in the APD. That conveys something about the speaker's state to listeners. Conversely when these patients try to express emotion, there are changes in the APD domain, but the balance of markers is not a normal one.

5. CONCLUSION

We have drawn attention to two broader domains surrounding specific attributions of emotion and the specific features of speech that underlie them, and highlighted reasons for caution about compartmentalising these broader domains.

It seems to be a general rule that variations in the augmented prosodic domain are emotive. That may be because they signal departure from a reference point corresponding to a well-controlled, neutral state. Various departures from that reference point are reflected in the augmented prosodic domain, including central and sensory impairments as well as emotion. We have some evidence that stylistic and dialect variations, which intuitively also seem to be emotive, are also reflected in the augmented prosodic domain [4,14].

Intuitively it seems right to acknowledge that departures from well-controlled neutrality are multidimensional and highly confusable. It is often genuinely difficult to know whether somebody is angry or depressed or preoccupied or simply hoarse. On one hand, it is asking too much to expect phonetics to draw those distinctions when listeners cannot: on the other, it is failing to represent the fluidity of the categories in which representations of the speaker are couched. A cynic might add that the notorious difficulty of finding pure and natural samples of emotional speech reinforces the case.

That kind of view suggests a programme of inquiry which is thoroughly empirical, concerned with documenting the broad patterns of variation that occur in the augmented prosodic domain and the ways in which they are received. ASSESS reflects the fact that contemporary technology makes that kind of programme practical.

6. REFERENCES

1. Cowie, R., Sawey, M., and Douglas-Cowie, E. " A new speech analysis system: ASSESS (Automatic Statistical Summary of Elementary Speech Structures)," *Proc ICPhS* 3: 278-281, Stockholm, 1995.

2. Goehl, H., and Kaufman, D. "Do the effects of adventitious deafness include disordered speech?", *Journ. Speech and Hearing Disorders* 49: 58-64,1984.

3. Goehl, H., and Kaufman, D. " The real thing: a reply to Cowie, Douglas-Cowie and Stewart, " *Journ. Speech and Hearing Disorders* 51: 185-187,1986.

4. Cowie, R., and Douglas-Cowie, E. *Postlingually acquired deafness: Speech deterioration and the wider consequences*, Mouton de Gruyter, Berlin, 1992.

5. Summerfield, A.Q., and Marshall, D. *Cochlear Implantation in the UK* , HMSO, London, 1995.

6. Cowie, R., Douglas-Cowie, E., Sawey, M., and Mulhern, G. "The effects of cochlear implants on speech production in postlingually acquired deafness," *Proc ICPhS* 3: 198-201, Stockholm, 1995.

7. McGilloway, S., Cowie, R., and Douglas-Cowie, E. "Prosodic signs of emotion in speech: preliminary results from a new technique for automatic statistical analysis," *Proc ICPhS* 1: 250-253, Stockholm, 1995.

8. Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L. " Perceptual and acoustic correlates of voice qualities," *Acta Otolaryng*. 90: 441-451, 1980.

9. Andreasen, N. "Negative symptoms in schizophrenia: definition and reliability, " *Arch. Gen. Psychiatry* 38: 784-788, 1982.

10. Alphert, M. "The signs and symptoms of schizophrenia," *Comprehens. Psychiatry* 26 (2): 103-112, 1985.

11. Harris, A., and Metcalfe, M. " Inappropriate affect," *Journ. Neurology and Neurological Psychiatry* 19: 308-313, 1956.

12. Owens, D., and Johnstone, E. " The disabilities of chronic schizophrenia - their nature and the factors contributing to their development," *Brit. Journ. Psychiatry* 136: 384-395, 1980.

13. DSM III - Diagnostic and Statistical Manual of Mental Disorders, III, American Psychiatric Association, Washington, DC, 1987.

14. Douglas-Cowie, E., Cowie, R., and Rahilly, J. "The social distribution of intonation patterns in Belfast," in J. Windsor -Lewis (ed.), *Studies in General and English phonetics*, Routledge, London, 1995, pp. 180-186.