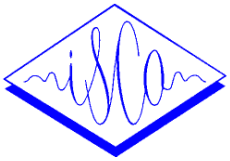

ETRW on Dialogue and Prosody
Veldhoven, The Netherlands
September 1-3, 1999

# CHANGING EMOTIONAL TONE IN DIALOGUE AND ITS PROSODIC CORRELATES

*R. Cowie, E. Douglas-Cowie and A. Romano*

The Queen's University of Belfast

## ABSTRACT

This paper examines changing emotional tone in dialogue and its prosodic correlates. 'Feeltrace' has been developed to trace emotional tone over time. It uses a simple but tractable representation of emotional tone based on psychological research. Listeners rated the emotional tone of arguments between friends (in real time) by positioning a pointer in a two-dimensional space whose axes are evaluation and activation. Feeltrace ratings were correlated with prosodic measures. The strongest correlations involve change in emotional tone, with change in activation level the most clearly and consistently marked. The common marker of change in evaluation involves pausing. However the correlation patterns are not the same for different speakers, and several features have variable significance.

## 1. INTRODUCTION

Structured change in emotional tone is an integral part of dialogue. Speakers rarely reach extremes of pure emotion, but they often become more or less animated and positive about each other and/or their subject matter. Intuitively there appear to be norms governing the management of these movements, and simulated dialogue that ignores these norms is likely to have unintended side effects. It also seems likely that prosodic variables are integral to both the expression and the perception of these movements. We have developed tools which make it possible to study these issues, and we report preliminary findings.

The literature on emotion and speech tells us rather little about how emotional tone fluctuates over time in dialogue or about how the fluctuation might relate to prosodic variation. That is probably because studies of speech and emotion tend to be rooted in a particular theoretical view of emotion, which we will call the classical view. The classical view reflects two main ideas about the essence of emotion. One is that emotion reduces to a few primary or pure emotions. The other is that emotion has particularly close connections with biology: human emotions are continuous with animals, and both have simple biological functions with a relatively direct relationship to survival. These functions are linked to homeostasis, and are controlled by a particular structure in the brain (the hypothalmus). The classical approach has exerted a powerful influence, particularly in the area of speech. Influential studies in the area of emotion and speech take it for granted that the material to be studied should express the primary emotions and the measures to be taken should relate as directly as possible to the physiological events associated with them. Hence most of the experimental studies on emotion and speech focus on the prosodic analysis of emotional extremes (regarded as 'pure' or 'primary' emotions) portrayed in monologue form by actors and sustained over short periods of time, e.g. [1]. However, research directly concerned with emotion has moved away from the classical preoccupation with pure or primary emotions. Instead, it has developed an emphasis which is more akin to natural history: it has focused on describing the various aspects of emotional life, and finding effective ways of organising them. It is in that context that we turn our attention to the study of less extreme emotions and the way they fluctuate over time in natural conversation. We report a study of the prosodic correlates of changing emotional tone in conversations between friends.

Taking that approach to emotion and speech raises a number of problems - most obviously, there is the issue of how to measure emotional tone over time and how to relate it to relevant prosodic measures. These are problems that emerge in the only other prosodic study that sets out to address the issue of changing emotional tone in dialogue (by Selting) [2]. Selting demonstrates the existence of what she calls an 'emphatic style' in which linguistic devices are used to signal heightened emotional involvement. The style is prosodically marked: the prosodic parameters that Selting shows to be relevant are dense accentuation and marked rhythm, variations in global pitch and loudness, and locally marked accent variants. However Selting does not separate the measurement of emotional tone in psychological terms from its prosodic correlates. In fact she uses prosodic features as indicators of emotional tone. This produces a circularity in approach.

The present study depends on tools that we have developed to study changing emotional tone in dialogue in an experimental and systematic way. Two tools are relevant. The first, called Feeltrace, provides measurement in real time of changing emotional tone in a dialogue. The output is numerical, not categorical. The numerical output allows the capture of gradual change and fluctuation in emotional tone in a way that discrete categorical categories do not. Basically subjects position a pointer controlled by a mouse in a two-dimensional space relevant to the description of emotion. Subjects move the pointer as they think the emotions change. The co-ordinates of the pointer are recorded at one sixtieth of a second intervals and these form the numerical output. A second tool, ASSESS [3,4] provides the prosodic measures. ASSESS generates a range of prosodic measures automatically which have been shown to be relevant to the expression of emotion [5,6]. The two sets of measures, emotional and prosodic are then correlated.

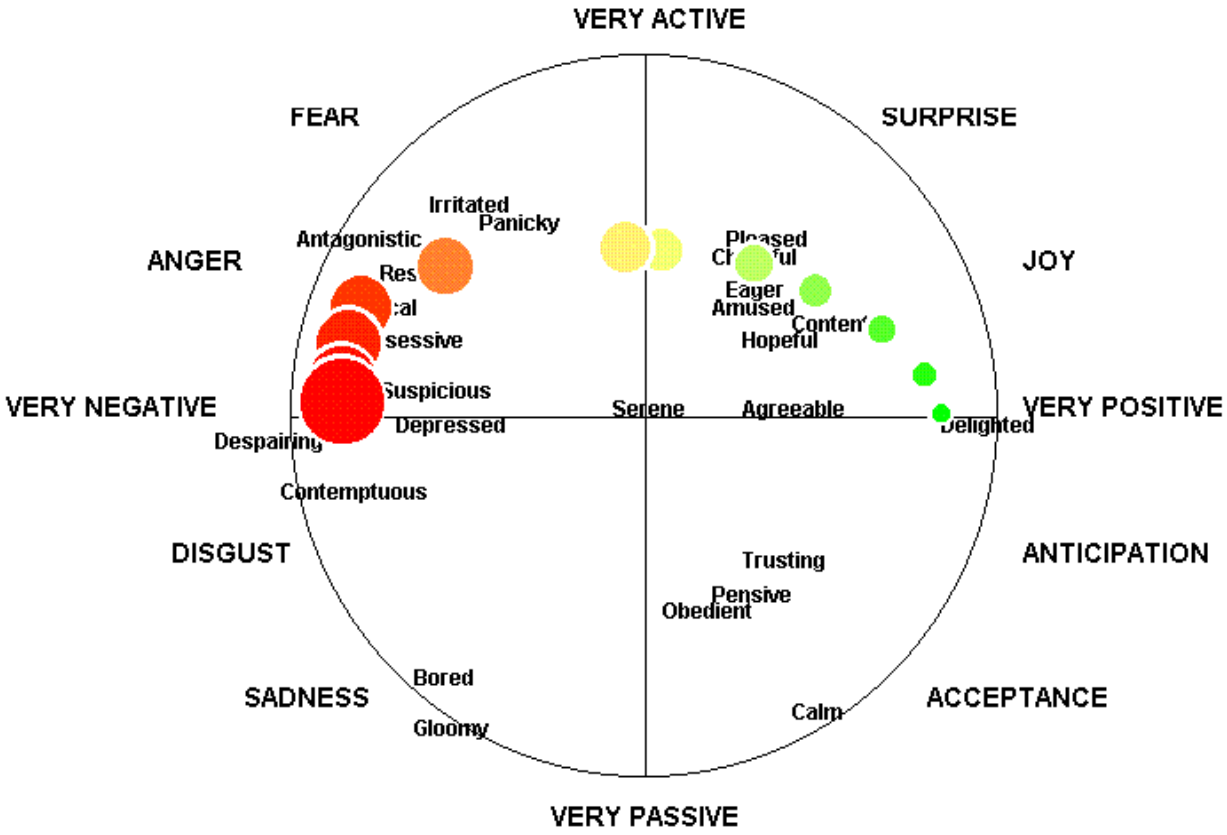


**Figure 1**: Example of the display that a subject using Feeltrace sees at a particular instant

## 2. METHOD

### 2.1. Data collection

To elicit dialogue with significant emotional content, we invited groups of three friends to argue about issues that they felt strongly about in a recording studio. They were asked to specify in advance the topics that they wanted to talk about. Three groups were recorded on video tape. Each discussion lasted for about an hour. At any one time, one participant acted as 'chair' while the other two carried the discussion. After the session, participants reviewed the discussion and reported how genuinely engaged they had been at each stage, so that passages which were unnatural (e.g. because speakers were aware of the studio context) could be discarded.

### 2.2. Measuring emotional tone

A program that we have called Feeltrace was used to measure emotional tone. Feeltrace is a continuous input device that we have developed for recordingjudgements of emotional content. It uses a simple but tractable representation of emotional tone based on psychological research [7,8]. In the literature, many authors agree that emotions can be organised roughly into a two-dimensional space whose axes are evaluation (i.e. how positive or negative the emotion is) and activation (i.e. the level of energy that a person experiencing the emotion is likely to display). Feeltrace provides users with an accessible way of assigning co-ordinates in evaluation-activation space to continuous expressions of emotion (conveyed through the face, voice, or music).

The program presents users with a particular representation of this two-dimensional space on a computer screen. The form it takes is associated with Plutchik [7] and many others. Possible emotions are arranged in a circle. Strong emotions lie at the periphery: the centre represents an emotion-free state of alert neutrality. The vertical axis of the circle represents activation level, the horizontal axis evaluation - positive emotions are on the right, negative on the left. Key emotion are arranged round the periphery to provide landmarks and help subjects to orient themselves within the space. Users are asked to specify the emotional tenor of what they are listening to by moving a

pointer on a computer screen (using a mouse) so that it shows where on those two dimensions the emotions they are perceiving fall at any given instant.

The Feeltrace display is designed to convey the basic idea of emotion as a point in a 2-D space. It incorporates several features which are meant to ensure that subjects understand what a pointer position means. The main axes are marked and described, one (activation) running from very active to very passive; the other (evaluation) running from very positive to very negative. The colour of the pointer is keyed to its position using a colour coding introduced by Plutchick [7], which subjects find reasonably intuitive. The cursor is green in positions corresponding to highly positive emotional states, and red in positions corresponding to highly negative emotional states; yellow in positions corresponding to highly active emotional states, and blue in positions corresponding to very inactive emotional states. Selected emotion words are presented at the point in the space where their reported co-ordinates [8] indicate that they lie. Each octant of the emotion space is labelled with a term describing the archetypal emotion associated with that region of the space. The dimension of time is represented indirectly, by keeping the circles associated with recent mouse positions on screen, but having them shrink gradually (as if the pointer left a trail of diminishing circles behind it). Figure 1 shows an example of the display that a subject using Feeltrace sees at a particular instant. The subject's rating of the episode that he or she is observing has moved from being active/ positive to active/negative. (Colours are not indicated).

In this study three subjects independently used Feeltrace to rate their perceptions (in continuous form) of the emotional tenor of all three conversations (using audio presentation). Subjects had been trained in the use of Feeltrace. However, the study was carried out early in the development of Feeltrace, and the training routine was less full than those we have developed since. Fuller training leads to close agreement among users of Feeltrace.

## 2.3. Prosodic analysis

Prosodic analysis was applied to a section of one of the conversations. Basic signal processing was carried out using Entropic WAVES[a], operating on a SUN SPARC II workstation. WAVES[a] was used to recover the basic attributes of the signal — estimates of F0 in 10ms intervals, confidence levels for the estimates, and signal intensity in each interval. F0 points and confidence levels were recovered using the formant routine. The only estimates of F0 used in later analysis were those with confidence greater than 0.9.

The ASSESS system [3] was then used to generate a range of prosodic measures from the WAVES output. These measures were generated for each conversational turn. ASSESS generates measures automatically with minimum hand interference. (Hand editing operates on a graphical display of F0 estimates. The user removes estimates which are clearly unreliable). ASSESS measures are of two general types: pointwise and piecewise. Examples of pointwise measures are mean, standard deviation and other descriptive statistics for the points that make up the F0 or the amplitude contour. The name reflects the fact that each number in the set summarised by a pointwise statistic describes a single point in the relevant contour. Piecewise measures are based on dividing the signal into 'pieces' that are structurally simple, but that nevertheless last for an appreciable time. Each number in the set summarised by a piecewise statistic describes a property of one of those 'pieces'. For example, F0 is broken down into continuous movements up or down (called rises and falls) and statistics are calculated for the duration, degree of pitch movement (magnitude), and rate of change in pitch (slope) for each type of movement. Signal level is also measured during each movement, giving an analogue of stress. A range of statistics for pauses is also given.

Point-wise measures were obtained for each conversational turn. They were mean F0, standard deviation of F0, and the 10th, 50th, and 90th percentile points (these were based on WAVES[a] points remaining after hand editing). A number of piecewise measures were also generated for each turn. They were derived by considering four types of 'piece' — F0 rises, F0 falls, F0 level stretches and pauses — and calculating for each one the mean and standard deviation of (a) their magnitude, (b) their duration, (c) their slope and (d) their amplitude (which reflects sound level within a piece). Statistics were also generated for rises and falls combined. Most of the measures are reasonably easy to understand. Pause magnitude needs a gloss: this refers to the pitch interval between the end of one vocalisation and the beginning of another. The number of rises, falls and pauses in each turn were also calculated. Numbers were normalised so that they did not simply reflect the length of a turn.

# 3. RESULTS

Close analysis focuses on the conversation of one of the three groups. This was the group where the participants reported most genuine engagement. A section from the conversation was selected for detailed study. The section selected lasted 7 minutes and consisted of 77 conversational turns. A number of predetermined criteria were applied in making the selection - (i) The section had to occur at least half way through the one-hour conversation and had to be rated as fairly natural by the participants afterwards (this minimised the effects of being recorded in a television studio), (ii) The section had to have a definite topic focus with a reasonably clear break point marking the beginning and end. The topic under discussion in the 7 minute section was 'Do you need religion?' It involved two speakers, one male, one female.

Figure 2 shows Feeltrace ratings for this section. The top panel shows ratings for Speaker R (male) and the bottom panel shows ratings for Speaker J (female). Each point denotes ratings for a conversational turn. The y axis shows co ordinates for ratings in the Feeltrace circle. There are two sets of values represented on the y axis. The first set (which are usually positive) represent activation level. The second set (which are usually negative) represent evaluation. Circles denote Feeltrace Rater 1 (EDC), Triangles denote Feeltrace Rater 2 (DAC) and

crosses denote Feeltrace Rater 3 (AR). Time (in seconds) is on the x axis. A number of points can be made from the figure.

First, there is broad agreement among raters on the type of emotion being conveyed in the conversation. In terms of the dimensions on which the conversation was rated, the emotions are broadly active and negative. This can be seen from the fact that the filled circles, triangles and xs (representing the activation dimension) mostly have positive values, while the blank circles, triangles and crosses (representing the evaluation dimension) mostly show negative values.

Second, emotional tone appears to form orderly peaks, falls and level stretches. The top panel for Speaker R and bottom panel for Speaker J show similar patterns: this means that both subjects shift emotional tone in tandem. The peaking pattern shows that emotional intensity rises and falls across the passage. Closer inspection indicates that adjacent turns tend to be similar in tone. Increase in emotional intensity usually involves a gradual crescendo while reduction in emotional tone can sometimes be deliberately engineered rather sharply.

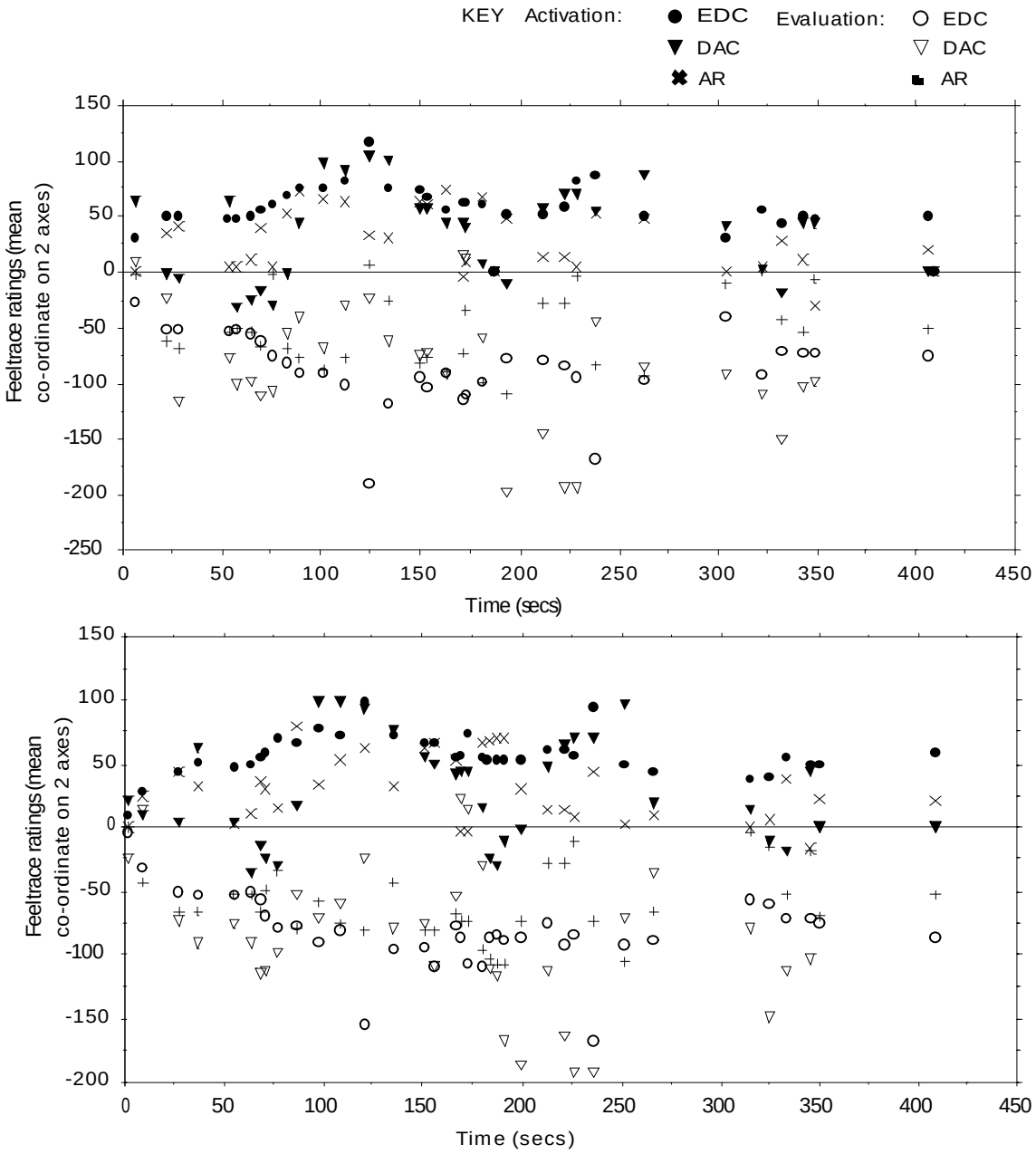


**Figure 2**: Feeltrace ratings by 3 raters (EDC, DAC, AR) for section of conversation involving 2 speakers, speaker R (top panel) and speaker J (bottom panel).

| Feeltrace measure | Direct-ion of relation | Gross measures | Amplitude associated with F0 features | Pitch movement features | Pause features |
|---|---|---|---|---|---|
| **SPEAKER R** | | | | | |
| Evaluation mean | (+ve)<br><br>(-ve) | <br><br>min F0 | stand dev of amplitude for falls | stand dev of length for rises& rise/falls | |
| Evaluation standard deviation | (+ve) | | stand dev of amplitude for levels | | stand dev of pause length |
| Activation mean | | | | | |
| Activation standard deviation | (+ve)<br><br>(-ve) | min F0;<br><br>rise & fall nos | <br><br>rise amplitude mean; stand dev of amplitude for falls & levels | <br><br>stand dev of length for rises, falls & rise/falls | |
| **SPEAKER J** | | | | | |
| Evaluation mean | | | | | |
| Evaluation standard deviation | (+ve)<br><br><br><br><br>(-ve) | F0 stand dev;<br><br><br><br><br>min F0 | stand dev of amplitude for rises, falls & rise/falls | stand dev of magnit & slope of falls | pause nos;<br>mean & stand dev of pause magnitude;<br>stand dev of pause length |
| Activation mean | (+ve) | F0 stand dev; 90% F0 point | stand dev of amplitude for rises | | pause nos, mean & stand dev of pause magnitude; |
| Activation standard deviation | (+ve)<br><br><br><br><br>(-ve) | F0 stand dev;<br>90% F0 point;<br><br><br><br>fall nos & rise/fall nos | mean amplitude of falls;<br><br><br><br>stand dev of amplitude for falls | mean length of rises,falls & rise/falls;<br>mean magnit of falls & rise/falls;<br><br>mean slope of falls | |

**Table 1**: Prosodic measures which correlated consistently with Feeltrace measures (mean and standard deviation of Activation and Evaluation for each turn) for each speaker (+ve indicates a positive correlation, -ve indicates a negative correlation)

The general patterns observed in Figure 2 informed the next stage of the study, which was prosodic analysis of emotional tone. On the basis of the plots in Figure 2, 24 turns (12 from each participant) were selected for analysis of relationships between Feeltrace ratings and prosody. The turns were selected to provide reasonably consistent inter rater judgements, but contrasting ratings. In general, that meant that they represented the main peaks and troughs in intensity seen in Figure 2.

The relationship between prosody and emotional tone was studied by correlating measures derived from Feeltrace and measures derived from ASSESS. Specifically, for each turn, the mean and standard deviation was calculated for each Feeltrace dimension; and the resulting measures were correlated with ASSESS measures describing the prosodic characteristics of each turn.

Table 1 summarises the cases where measures derived from all three raters showed a reasonable correlation, in the same direction, with a prosodic measure.Several points points of interest emerge from the table and from associated exploration of the correlational data.

In broad terms, prosodic measures correlated strongly with emotional tone, but not simply. The number of correlations in the table show that prosodic features at this level do reflect the emotional tone of an argument. Equally obvious, however, is the fact that the correlation patterns are not the same for the two speakers.

Quite unexpectedly, the strongest correlations involved variation in emotional tone, as indicated by the standard deviation of a feeltrace measure. Conversely, steady emotional states (captured by mean activation and evaluation) were not

strongly marked - each speaker had one steady state which lacked any straightforward distinguishing feature (see blanks in mean activation column for Speaker R and mean evaluation column for Speaker J). It appears that prosodic features tend to signal turns where the speaker is shifting ground emotionally.

Change in activation level (as captured by the standard deviaton for activation) was the most extensively marked in both speakers, and also the most consistently marked. The number of pitch movements was low (see negative correlations for both speakers involving numbers of rises, falls and rise/falls), and so was variation in the amplitude of key segments (see negative correlations for both speakers for standard deviation of falls and levels). The common marker of change in evaluation (as captured by the standard deviaton for evaluation) involved pausing. Variable pause length (as measured by the standard deviation of pause length) was associated with change in evaluation in both speakers, and one (Speaker J) varied pausing in a range of other distinctive ways.

Several features seem to have variable significance. For speaker R, high standard deviation for the amplitude of falls and low minimum F0 signalled relatively positive evaluation. The same variables were associated in speaker J, but for her they signalled change in evaluation. Speaker R signalled changing activation by raising the lower end of the F0 distribution (minimum F0), speaker J by raising the upper end (90th percentile) - of which the associated rise in standard deviation is a natural consequence. Speaker J seems to use standard deviation as a generic marker of significant episodes, whereas it has no simple relation to emotional tone in speaker R. Speaker J also uses pausing to signal activation, whereas speaker R does not.

## 4. DISCUSSION AND CONCLUSIONS

Our main claim for this study is that it highlights intuitions which should inform a more extended program of research.

First and foremost, dialogue has an emotional structure. It is not normal or acceptable to exhibit an emotional tone which bears no relation to preceding discourse. Nevertheless, changes in emotional state are brought about during dialogue (and sometimes that is one of its main functions). It follows that there must be signs which allow participants to track each other's emotions and to display their own.

Everyday experience indicates that registering emotional tone is genuinely difficult – people do get it wrong, often with uncomfortable consequences. Hence should not go into analysis expecting to find clear cut, unequivocal signs.

One of the sources of difficulty, which is reflected in this study, is that there are different prosodic styles. Knowing the two participants, it rings true that speaker R gives fewer straightforward signs of emotion than speaker J. The difference may well be gender-related (R is male, J is female).

On the other hand, ASSESS measures do not capture all the available signals of emotional tone. Subtler prosodic variables may well be relevant. Certainly there are non-prosodic signals at work – voice quality possibly, visual signals probably – and an adequate discussion of the role of prosody should consider how it works with those other signals.

Similarly, it may be that what prosody signals is not simply related to the Feeltrace dimensions. The cases where prosodic patterns are complementary suggests one possibility that Feeltrace would not reveal. It may well be that when one is signalling offence, the other is signalling retreat.

We conclude that understanding the way people signal and detect gradations of emotional tone is a challenging problem. The techniques that we have developed may at least help to open these issues to systematic investigation.

## 5. REFERENCES


1. Banse, R., and Scherer, K. "Acoustic profiles in vocal emotion expression," *Journal of Personal. and Soc.Psychol*. 70 (3): 614-636, 1996.

2. Selting, M. "Emphatic speech style - with special focus on the prosodic signalling of heightened emotional involvement in conversation," J*ournal of Pragmatics* 22: 375-408, 1994.

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

4. Douglas-Cowie, E., and Cowie, R. "Intonational settings as markers of discourse units in telephone conversations," *Lang.and Speech* 41 (3 & 4): 347-370, 1998.

5. 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.XIIIth ICPhS* (1): 250-253, Stockholm, 1995.

6. Cowie, R. and Douglas-Cowie, E. "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," *Proc..Fourth ICSLP* (3): 1989-1992, Delaware, 1996.

7. Plutchik, R. *The psychology and biology of emotion*, Harper Collins, New York,1994.

8. Whissell, C. "The dictionary of affect in language." In R. Plutchik and H. Kellerman (Eds.), *Emotion, theory, research and experience,vol 4: the measurement of emotions*, Academic Press, New York, 1989.