Speech in the Process of Becoming Bored

R.Cowie, A.McGuiggan, E.McMahon †, and E.Douglas-Cowie‡
† Psychology, Queen's University, Belfast
‡ English, Queen's University, Belfast
E-mail: r.cowie@qub.ac.uk, e.m.mcmahon@qub.ac.uk, e.douglas-cowie@qub.ac.uk

ABSTRACT

We have developed paradigms for studying speech in everyday emotional states, including boredom. 12 subjects spent 30 mins each watching repetitive computer displays and describing them. We recorded their speech, and three indices of their state – error rate; time per display; and self rated boredom. There appeared to be three phases: 1) fresh; 2) when boredom ratings reached a ceiling; 3) towards the end, with similar ratings, but impaired performance. Several different patterns of change were seen in speech. Energy below 500Hz rose throughout, and F0 standard deviation fell throughout. Others changes mirrored subjective ratings: time spent speaking per screen, and number and duration of substantial pauses, fell from phase 1 to phase 3, then stabilized. Others mirrored error rate, with change mainly after phase 2 (number of short breaks and energy in the range 2-5kHz). Such patterns integrate well into recent accounts that view emotion in terms of loosely correlated changes in multiple variables.

1. INTRODUCTION

Speech carries self-descriptive information – i.e. information about the speaker – alongside information about the topics that are overtly being discussed. One of the functions of the self descriptive information is to convey the speaker's emotional state. Intuitively, it seems likely that speech interfaces between humans and machines will have to deal with emotion-related information in order to be fully acceptable to human users [1].

There is a considerable body of research on speech and emotion [2]. However, there has been growing unease with the dominant paradigm, which views emotions as states epitomised by a few archetypes – fear, anger, happiness, sadness, etc. Such a paradigm directs attention to speech produced in pure versions of these states, expecting to deal with weak or mixed states by interpolation. There is increasing consensus that the anticipated transfer to applied problems does not occur [3],[4].

We have moved to the view that a more fundamentally ecological approach is needed, at least to complement older approaches. To deal with the emotional states that matter in everyday interaction, one has to study the emotional states that matter in everyday interaction. That entails thinking not only about speech in these states, but also about the states themselves and ways of describing them [5].

This paper focuses on a state that is common enough to be practically important, but which does not feature in standard lists of emotions (see, eg [6],[7]). It is boredom. If people want to argue that boredom is not an emotion, we will concede the point, and call it an emotion-related state. It is a side issue where an academic community decides to draw semantic boundaries. What does matter is that ideas in the recent emotion literature seem to illuminate boredom, and vice versa.

Emotions as correlated and decorrelated processes

Descartes conceptualized archetypal emotions as pure states, analogous to primary colours (hence the term 'primary emotions', which is still in popular use, though the research literature on emotion has largely discarded it). That conception has been giving way to views implying that emotions are intrinsically complex: what defines them is precisely the way multiple elements come together.

Scherer [8] advocated a view of emotions as episodes in which multiple processes synchronise in distinctive ways. Reisenzein [9] took the approach a step further, proposing that the various strands associated with an archetypal emotion – feeling, somatic response, cognitive response, behavioural elements – were correlated, sometimes quite loosely, rather than tightly synchronized.

Implicit in views like Scherer's and Reisenzein's is a point that should perhaps be self-evident, but which it is nevertheless easy to overlook. It is that emotions are temporally patterned - not so much states as episodes.

Background to the study

We have developed a dual track approach to understanding the speech styles associated with everyday emotional states – collecting samples of (relatively) spontaneous speech, and developing paradigms for eliciting speech in relevant states. We report a study of the second type, designed to elicit bored speech.

The context of the study is the ORESTEIA project [10], which explores new technologies for monitoring states that threaten hazards to life or health if no action is taken. One of the main examples is vehicle control. Various states of a driver or a pilot are known or suspected to carry real risk to life – inebriation, distraction, overload, anger ('road rage'), and lowered arousal, which becomes drowsiness or sleep in the extreme case. Various technologies have been applied

 ISBN 1-876346-48-5

to the task of detecting these states. Some research has monitored the eyes to detect drowsiness, but it would be preferable to have a measure that detected problems before the eyes began to shut. Various physiological measures that correlate with emotionality.might provide earlier warning. Even more promising is evidence that voice may signal lowered arousal without the need for sensors that are both intrusive and error-prone [11].

We have developed a method of inducing boredom that generates substantial amounts of speech data. Subjects watch computer generated displays of simple geometric shapes (e.g. a blue square, a red triangle, and a green circle). Figure 1 shows an example. The task is to describe the shapes on a screen, press a key to bring up the next screen, describe the shapes on it, and so on. The process continues for tens of minutes, and subjects find it intensely boring. It is made more boring by the fact that up to 25 successive screens may show the same shapes.

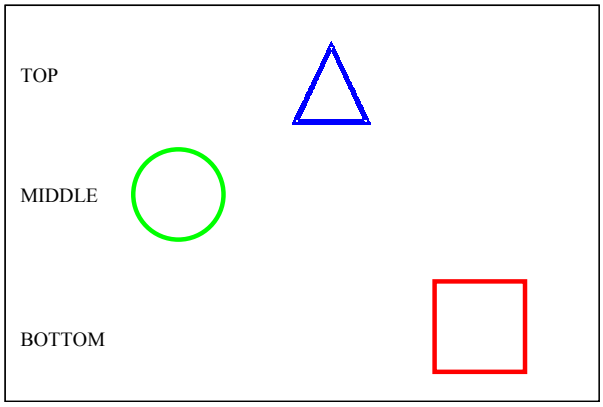

Figure 1: Example of a screen to be described. The response expected would be 'green circle in the middle, blue triangle at the top, red square at the bottom'.

The pilot study investigated a range of measures:

- Self report On every tenth trial, subjects rated their subjective level of boredom
- Speech was recorded and a range of measures was derived using the ASSESS system (see below)
- Speed of response was monitored
- Accuracy of subjects' descriptions was monitored
- Heart rate was recorded
- Movement was monitored by a detector strapped to the left hand

There may be trends in the physiological and gross movement measures, but they were not stable enough to be informative in small samples. The other measures suggested more robust patterns. In particular, the process of becoming bored seemed to involve at least three phases. In the first, subjects were fresh, but relatively slow at the task. In the second, subjective boredom ratings had reached a ceiling, but performance was faster and errors remained low. By the third phase, accuracy had dropped, and there were substantial numbers of errors. The timings of the phases varied from subject to subject.

Several speech variables appeared to change from phase to phase. However, the details of the task left doubts about the exact emotion that they reflected. The naming task was demanding enough to mean that subjects sometimes became confused and /or irritated rather than simply bored.

The main study followed up these observations. The less informative measures – heart rate and gross movement – were dropped; and the naming task was simplified.

2. METHOD

Induction

Twelve subjects spent about ½ hour each watching computer generated displays of simple geometric shapes. The shapes were as in figure 1, but they were always on the horizontal midline, and the responses were of the form 'green circle, blue triangle, red square'.

Information recorded

Subjects' speech was recorded, and three indices of their state were collected – error rate; rate at which trials were completed; and a self rating of interest or boredom, which was presented to them after every 10th display.

Speech analysis

For each subject, three speech samples were digitized for analysis, one in each of the phases identified by the performance and self rating data. Each sample consisted of responses to five screens of the type shown in Figure 1. Samples generally lasted for 0.5-1 min.

Samples were analysed using ASSESS [12], [13]. Measures were chosen in advance on the basis of exploratory work in the pilot study and earlier work on related problems [3],[14],[15]. They were

Intensity (excluding periods not classified as silences): mean (m1),standard deviation (m2), and 90th percentile point (m3).

F0: mean (m4), standard deviation (m5), and 90th percentile point (m6).

Spectral: spectral slope (m7), and mean energy in each of the following bands – 0-500Hz (m8); 0-2kHz (m9); 2-5kHz (m10); 5-8kHz (m11).

Time spent speaking excluding all silences (m12) and excluding only silences long enough to be classed as pauses, i.e. >154ms (m13).

Mean duration of continuous rises or falls in pitch (m14) and amplitude (m15).

Number and mean duration of silences long enough to be classed as pauses (m16, m17) and of 'breaks', i.e. shorter silences (m18, m 19)

Number and mean duration of 'bursts', i.e. phrase-like units of speech bounded by pauses (m20, m 21)

ISBN 1-876346-48-5

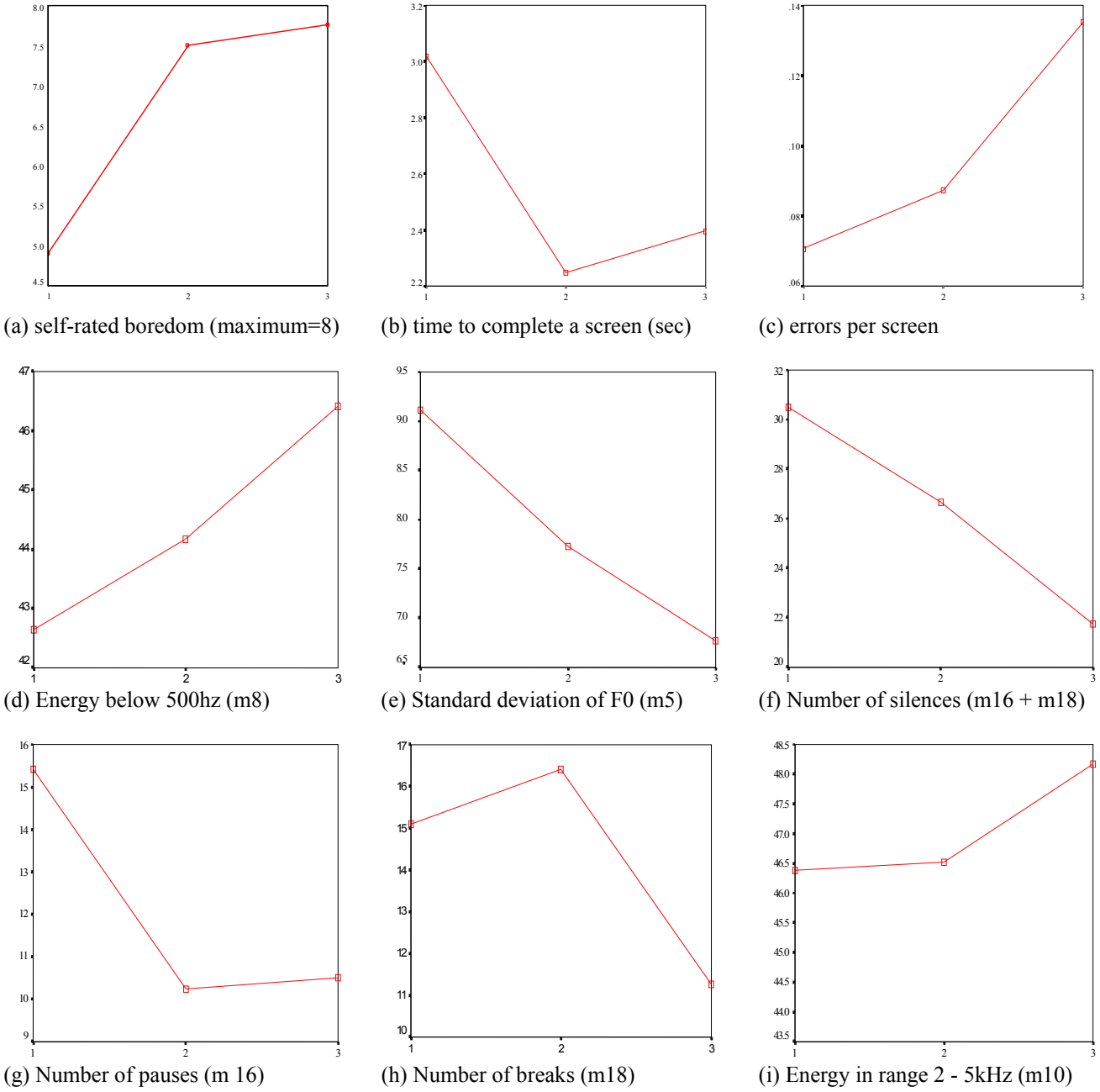

Figure 2 Main patterns observed in subjective state (panel a), performance (panels b & c), and speech variables (panels (d-i).

3. RESULTS

Figure 2 shows the main patterns observed. The horizontal axis is phase, using the divisions suggested by the pilot study. Linear trend relative to phase was used as the basic statistic. It is significant with $p<0.05$ in all panels except (i), where $F(91,11)=4.41$, $p=0.060$). All other cases where linear trend was significant are mentioned in the text.

The subjective and performance measures show broadly the expected pattern. In particular, it is in phase 3, where the task continues beyond the point of extreme subjective boredom, that marked performance deficits appear.

Speech variables relate to that pattern in multiple ways. Panels c-e in Figure 2 show the measures that followed the simplest pattern, continuous change from beginning to end of the experiment. The band 0-2kHz (m9) behaves like band 0-500Hz which is shown in Fig 2.

One the other hand, monotonic change was by no means the only pattern. Figure 3 shows key examples of others. Number of pauses showed a pattern akin to subjective boredom, i.e. rapid change between phase 1 and phase 1, followed by stabilization. Pause duration (m 17) behaves

ISBN 1-876346-48-5

similarly, as does time spent speaking excluding either pauses (m13) or any kind of silence (m12), and the number of phrase-like units (m20). Together, these seem to reflect a global rate setting. In contrast, the number of silences too short to be considered as pauses (m 18) – mostly associated with marking stops and boundaries between words – changes only after phase 2. Energy in the upper spectrum (m10) follows a similar pattern.

Variables that have not been mentioned above showed no robust change with boredom. It is worth noting that these include basic measures such as mean intensity, pitch, and spectral slope, and measures which are sensitive to expressiveness, such as the duration of rises and falls in pitch.

4. CONCLUSIONS

Boredom is far from being a trivial topic in its own right. There are not many states that an artificial educator more obviously needs the ability to detect, and in air traffic control, reliable ways of detecting it could be the difference between life and death.

Our concern, though, is less with a particular state than with a general approach. We came by trial and error on the idea that it was possible to shed the image of boredom as a unitary state, and to think instead of becoming bored as a process with multiple elements, to which multiple aspects of speech might reasonably be thought to relate. We have found that a productive way of thinking, and we commend it to people working on any kind of emotion-related phenomenon.

Part of the attraction of the approach is that it invites theoretical reflection. It makes a great deal of sense that many changes in speech should be aspects of a more general speeding or slowing. It also makes sense that other changes should be related to decline in accuracy, particulary when the changes in question seem to be3 linked at least loosely to articulation.

Observations like that point the way towards a multicomponent analysis of speech and its links to emotion-related fluctuation. That seems a worthwhile extension of the project associated with Stevens [16] and Scherer [17] which tries not simply to document emotion-related changes in speech, but to understand them.

REFERENCES

[1] Cowie, R (2001) Non-verbal information processing and HCI *Proc International Workshop on Very Low Bitrate Encoding* Athens Oct 2001 pp. 12-14.

[2] Scherer, K R (2003) Vocal communication of emotion: A review of research paradigms *Speech Communication* **40**,*227-256*

[3] Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., Taylor J. (2001). Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine January 2001*. 32-80

[4] Batliner A., Fischer K., Huber R., Spilker J. and Nöth E. (2003) How to find trouble in communication *SpeechCommunication* **40**, 117-143

[5] Cowie R and Cornelius R. (2003) Describing the Emotional States that are Expressed in Speech *Speech Communication* 40, 5-32

[6] Ekman P (1999) Basic Emotions. In: Dalgleish, T., Power, M. (Eds.), *Handbook of Cognition and Emotion*. John Wiley,New York,pp.301 -320.

[7] Lazarus R S.(1999) *Stress and Emotion: A New Synthesis* Springer, New York.

[8] Scherer, K R (1984) On the nature and function of emotion: a component process approach. In: Scherer, K.R., Ekman, P. (Eds.), *Approaches to Emotion*. Erlbaum, Hillsdale, NJ, pp . 293–317

[9] Reisenzein, R (2000) Exploring the strength of association between the components of emotion syndromes:the case of surprise.*Cognition and Emotion* 14,1 -38.

[10] http://www.image.ntua.gr/oresteia/

[11] Hadfield, P & Marks P.(2000) This is your captain dozing. *New Scientist* 1682267, 21.

[12]Cowie R.., Sawey M., & Douglas-Cowie E. (1995). A new speech analysis system: ASSESS (Automatic Statistical Summary of Elementary Speech Structures). *Proc. International Congress of Phonetic Sciences,* Stockholm, 1995, vol. 3, pp. 278-281

[13] Douglas-Cowie E., Campbell N, Cowie R. & Roach P (2003) Emotional Speech: towards a new generation of databases *Speech Communication* 40, 33-60.

[14] Douglas-Cowie E. & Cowie R. (1998). Intonational settings as markers of discourse units in telephone conversations. *Language & Speech Special issue Prosody & conversation,* vol 41 3-4, 347-370.

[15] R. Cowie, E. Douglas-Cowie & A.Wichmann.(2002). Prosodic correlates of skilled reading: Fluency and expressiveness in 8- 10 year old readers. *Language and Speech* 45(1), 47-82.

[16]Williams, C.E., Stevens, K.N.,1972.Emotions and speech: some acoustical correlates. *Journal of the Acoustical Society of America* 52 (2),1238 –1250

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