A NEW EMOTION DATABASE: CONSIDERATIONS, SOURCES AND SCOPE

Ellen Douglas-Cowie, Roddy Cowie and Marc Schröder

School of English / School of Psychology, Queen's University Belfast

ABSTRACT

Research on the expression of emotion is underpinned by databases. Reviewing available resources persuaded us of the need to develop one that prioritised ecological validity. The basic unit of the database is a clip, which is an audiovisual recording of an episode that appears to be reasonably self-contained. Clips range from 10 – 60 secs, and are captured as MPEG files. They were drawn from two main sources. People were recorded discussing emotive subjects either with each other, or with one of the research team. We also recorded extracts from television programs where members of the public interact in a way that at least appears essentially spontaneous. Associated with each clip are two additional types of file. An audio file (.wav format) contains speech alone, edited to remove sounds other than the main speaker. An interpretation file describes the emotional state that observers attribute to the main speaker, using the FEELTRACE system to provide a continuous record of the perceived ebb and flow of emotion. Clips have been extracted for 100 speakers, with at least two for each speaker (one relatively neutral and others showing marked emotions of different kinds).

1. INTRODUCTION

Our interest in creating an emotion database arises from an EU project called PHYSTA [1]. The aim of the project is to develop a system that will recognise emotion from facial and vocal signs. The system is to be based on hybrid computing, that is to say, it will use a combination of neural net techniques and traditional symbolic computing. To train the neural net component, we need audiovisual emotional material.

At the start of the project, we assumed that databases containing suitable material were already available. After exploration, we concluded that existing collections were not sufficient for the project. That led us into two linked undertakings. One was a to articulate more clearly the criteria against which databases could be evaluated. The second was to assemble a database that met the most basic of our requirements.

1.1 Guiding Principles

Most pre-existing databases consisted of examples representing a few archetypal states. The rationale behind that approach is rarely spelled out, but the only obvious way to justify it is to postulate that the whole space of emotional signs can be reconstructed from information about a few cardinal types. We will call that the benign interpolation hypothesis.

Exploratory work convinced us that the benign interpolation hypothesis ought not to be taken for granted, for several reasons. One key point is that people usually aim to avoid socially unacceptable signs of emotion. As a result, emotion often seems to be signalled by flaws in an attempt to portray an acceptable emotional state – a smile that is too fixed, or a catch in a voice that is meant to be unconcerned. Observations like that suggest that even if everyday signs of emotion involve the same elements as archetypal examples, the rules governing the combinations that occur naturally, and their meanings, are not likely to be reconstructed by a priori interpolation.

It follows that at the very least, the benign interpolation hypothesis ought not to be assumed without evidence. It may be true; but if so, the point can only be established by assembling .a range of samples against which it can be tested, i.e. a database with some claim to ecological validity. If it is false, then an ecologically valid database is the only training corpus that is appropriate for a system that is to recognise emotion in everyday situations.

Note that the argument is more pressing for recognition than it would be for simulation. It seems likely that a given type of state may be signalled in various ways. Effective recognition depends on sensitivity to all the variants that are likely to be used. Simulation only needs to match one of them (at least as a first approximation).

The attempt to achieve ecological validity was guided by four more specific considerations:

1. Genuine emotion Our core decision was to use material generated by people experiencing genuine emotion. That contrasts with the widespread tendency to use records of actors simulating emotion. It is not obvious whether actors reproduce the genuine article or generate a stylised idealisation in which some features of the everyday reality are heightened, and others ignored. The doubt is underlined by the observation that it rarely takes long to recognise whether emotional material on television is part of a drama or a genuinely emotional interview. At the very least, acted emotion cannot be a sufficient basis for conclusions about the expression of emotion. Proposals based on it need to be tested against natural material.
2. Emotion in interaction We also chose to focus on examples derived from people engaged in human interactions. Some studies deal with emotion that is real, but produced in private – by watching a film or playing a computer game. There is a case for that approach in studies of facial signs of emotion. However, our core concern is with speech.

Speech as a medium is intrinsically oriented towards another person, and the natural contexts in which to study it are interpersonal.

3. Gradation Long standing research traditions direct attention towards archetypal examples of emotion – fullblown fear, happiness, etc. However, archetypal emotions form a very small part of naturally occurring emotional behaviour. Hence ecological validity entails sampling situations where emotion is mixed or controlled in the ways that typically occur in everyday life.

4. Richness Research tends to deal with the expression of emotion in one modality at a time – audio or facial. Presumably handling audiovisual inputs is expected to be a matter of straighforward interpolation. However emotional expression is typically extended both in time and in modality, in the sense that vocal expression is linked to facial expression, gross gestures ('body language'), and verbal content. Hence ecological validity entails collecting samples which make it possible to study whether those elements are effectively independent or interactive, and how they evolve and cohere in time.

Prioritising ecological validity has its costs. One of them is that elegant designs are harder to achieve – which is a serious issue for training. That point is considered more fully later. A particular aspect of the design problem should be mentioned here, though.

If database construction is not governed by a list of archetypal emotions, an alternative approach to ensuring coverage is needed. The approach that we have taken depends on the idea (derived from psychology) that emotions can be considered as points in a continuous space. Broadly speaking, the coverage that we have aimed for is one that ensures samples throughout standard representation of emotion space – with some qualifications, which are discussed later.

The use of emotional space is one example of a wider issue. To be useful, a database needs to contain information about examples as well as raw records. An integral part of the challenge in developing an ecologically valid database is to find descriptive tools that allow the material to be described appropriately. Section 3 takes up that issue.

1.2 Existing databases

This section reviews existing databases in light of the considerations above.

Genuineness and interaction Simulated expression is the norm in existing databases, although there are some important exceptions. Since separation of audio and visual modalities is almost universal, we survey audio databases first, then facial databases.

Audio databases of emotional expression typically consist of acted material read from a written text. The material tends to consist of isolated words or sentences, though there are some longer passages. The content is often deliberately neutral, so that actors can impose different emotions without conflict between expression and content. Examples are the Danish Emotional Speech Database [2], the Berlin corpus [3] and the Groningen ELRA corpus which is only partially oriented to emotional expression [4]. The Danish database contains recordings of four actors reading two words (yes and no), nine sentences (four of them questions) and two passages. Each is read to express five states – neutrality, surprise, happiness, sadness and anger. The Berlin database consists of ten utterances spoken by five male and five female German actors expressing hot anger, quiet sadness, joy, fear, disgust, boredom and neutrality. The Groningen corpus contains over 20 hours of read speech from 238 readers. The material read includes two texts containing passages that lend themselves to emotional expression (particularly passages of direct speech).

Audio databases of totally natural vocal emotional expression are rare. A database described by Amir [5] moves part of the way. It records 30 subjects recalling an emotional event in which they participated. That represents a move away from read text. Closer again to complete naturalism is the database collected by the Reading-Leeds project. It consists of extracts from radio broadcasts of material such as interviews, in which emotion arises spontaneously from the content or the interaction [6].

The norm in facial databases is also to use acted or staged material. Most of the material is static rather than kinetic. The classic printed collection of static images showing facial emotion is that by Ekman and Friesen [7]. Others are available online. For example, the Yale database with 15 subjects [8] has ten distinct images of each subject including some emotional states (happy, surprised, sleepy) and the ORL database with 40 subjects [9]contain 11 distinct images of each subject with some aspects of facial expression which are at least broadly relevant to emotion (open/closed eyes, smiling/not smiling). The PICS database at Stirling is bigger again, containing 689 face images with four expressions represented [10]. Kinetic samples of faces are less frequently encountered, and kinetic sequences which are emotionally characterised are even less common. Where they do occur, they are simulated, and involve short sequences from a few subjects. One example is a set of video sequences containing facial expressions which can be downloaded from the MIT Media Lab Perceptual Computing Group ftp server [11]. This contains video expressions of approximately 10 frames per expression. Expressions covered are smile, anger, disgust and surprise.

Again, there are some databases that show genuine rather than simulated facial expression of emotion. The main example comes from the Geneva group, who have used interactive computer games to elicit emotions [12]. They devised experimental computer games for the specific purpose of eliciting specified emotions. Subjects were videotaped while they were playing the experimental games. The approach does seem capable of eliciting genuine emotion in a reasonably controlled way. On the other hand, the context is a very specific one, and it remains to be seen how closely it relates to the way emotion is expressed in social contexts.

Gradation Databases have focused almost exclusively on archetypal representations of the basic emotions - anger,

happiness, sadness, fear, disgust, and to a lesser extent surprise. There appear to be no systematic collections dealing with the milder and subtler kinds of emotional state that are commonplace in everyday life. An exception of a sort occurs in phonetic textbooks, often under the heading 'attitude'. These present sentences, usually constructed on the basis of the writer's intuition, that exemplify the kind of prosodic pattern that would be used to express various attitudes. Schubiger [13], for example, and O'Connor and Arnold [14] between them list up to 300 labels of attitudinal states and indicate their vocal correlates. Examples of labels they use are 'abrupt, accusing, affable, affected, agreeable, amused, apologetic, approving …'

Richness The term richness covers three main issues – modality, time and context. Existing databases that contain emotional material deal with one modality at a time – audio or facial. There are few databases of audiovisual material in general, and none of them focuses on emotional material. The issue of time is an interesting one. A large proportion of the material available features isolated words or short sentences, but some databases do contain fairly long read passages, e.g. [4] or spoken episodes [6]. There seems to be no collection dealing with material on the scale of a sustained conversation involving emotional ebb and flow. Context relates to both modality and time, and involves considering how face and speech and verbal content cohere or relate, and how they do so in time. Systematic study of those issues is limited by the constraints on modality and time that have already been noted.

In summary, existing databases have not generally been developed with ecological considerations in mind. There are probably two main reasons, one practical and one theoretical. In practical terms, there are real difficulties in collecting naturally occurring samples of emotion. They involve problems that are familiar from disciplines such as sociolinguistics that try to obtain naturalistic data, plus others associated with the need for audio as well as video recordings, and with people's wariness about displaying emotion. In theoretical terms, the classical emphasis on basic emotions has exerted a powerful influence. Researchers have focused on collecting archetypal emotions, and paid relatively little attention to situations where emotion is more controlled, weaker, or subtler, even though they are much commoner in everyday life.

Theory and practice perhaps reinforce each other. For example if one is dealing with only a few clearcut primary emotions, then it is natural to think of generating that kind of clearcut data in artificial contexts. Similarly, if tracking emotion over time is a real practical difficulty, it is easier to stick with a theory that treats emotion as a static quality that can be just as fully present in a short sample as in a long one.

2. CREATING THE BELFAST DATABASE

The development of our database was guided by the ecological approach that we have outlined. It was designed to sample genuine emotional states including archetypal and other states, to involve both modalities, and allow exploration of emotion over time. The particular demands of the PHYSTA project also set a number of constraints involving quality and comparator data (see 2.2).

2.1 Sources

We explored two main sources. First, we made our own studio recordings. Second, following the approach pioneered by the Leeds/Reading group, we recorded extracts from selected television programmes.

Studio recordings The prime attraction of making our own recordings was the prospect of controlling the material, both technically (e.g. in terms of camera position) and in terms of content (e.g. to obtain a balanced set of states for each speaker). The general approach was to record people who knew each other well talking about emotive issues. We tried two versions of the approach.

In the first version, we asked postgraduate students who knew each other well to decide on a few topics that provoked strong feelings, and then to come into a television studio and discuss them. This was expected to produce displays of negative emotions in an interactive context. They were debriefed afterwards to what their emotional state had been. The situation was set up for three students at a time – one acting effectively as a chair, with express instructions to get the other two 'going'. We recorded three groups of this type. The approach was not taken further because subjects' behaviour was generally very constrained. For example, some subjects indicated in debriefing that their attitudes had been very negative, and yet they smiled throughout.

In the second version, we made audiovisual recordings of one to one interactions involving a researcher with fieldwork experience and a series of colleagues and friends. Each session lasted about 1-2 hours. The aim was to cover topics that would elicit a range of emotional responses (i.e. active positive emotion, active negative emotion, passive positive emotion, passive negative emotion). Fieldwork techniques were based on standard procedures in sociolinguistics. In particular, care was taken over three issues. First, the physical setting was made as informal as possible (by use of unobtrusive wall mounted cameras, physical props such as coffee table, etc.). Second, recordings were long (sociolinguistic research shows that even in formal situations, subjects relax after an hour and speak more freely [15]). Third, the interviewer used prior knowledge of each subject to tailor the conversation. Each interview session followed the same broad pattern. The interviewer started with fairly neutral topics (enquiries about the family, description of job), moved to positive topics and finally to negative topics. Positive topics typically included holidays, children's successes, birth of children/grandchildren, reminiscing to happy times and events. Negative topics were typically political trouble in Northern Ireland, bereavement, problems at work.

Some of the material obtained from that approach was judged useful, but it was almost all mild. Even when subjects were well known to the interviewer, and discussing highly charged experiences, they rarely showed dramatic signs of emotion. In terms of coverage, the method was a useful source of material close to the centre of emotion space (which corresponds to alert

neutrality); but it seemed unlikely to allow coverage of extreme regions of emotion space.

Television programmes Television was the main source of material involving relatively strong emotion. We began by watching a range of programmes over a period of several months, and eventually identified a few programme types that were potentially useful. All of them dealt with real interactions rather than acted material. The programme types were (i) chat shows, (ii) religious programmes (iii) programmes tracing the life of real people over time (iv) current affairs programmes. We excluded programmes where we believed there was an element of 'staging'.

Chat shows provided the most obviously emotional material, though the emotional range tended to be limited to negative emotions. They typically dealt with an emotive situation or issue, such as divorce, death, or drugs, with an audience composed of people who had direct experience of the particular type of situation or issue. The two programmes that eventually provided most material were Trisha (Independent Television Network) and Kilroy (British Broadcasting Corporation).

Religious programmes were often a source of positive emotion. They were used as a counterbalance to the negative emotions expressed in the chat show data. Our best source here was the BBC 1 programme, 'Songs of Praise'. This weekly programme moves around the country, and people in different regions are recorded coming together to sing hymns. Usually between the hymns there is an interview with the presenter in which a member of the local community is interviewed. The member of the community has often had some special emotionally charged experience. It is usually a positive experience often attributed to religious faith. The tone of these interviews is often positive, sometimes exhuberant and sometimes tranquil.

Two other television sources, programmes tracing the life of real people over time and current affairs programmes were also used. BBC Panorama was a useful current affairs programme which gave some quite intense material related to deaths resulting from food poisoning. The BBC series 'The Village' traced the daily life of villagers in a particular area over a long period of time, and occasionally this gave emotional material of both positive and negative nature, e.g. moving to a new house, losing a job.

2.2 Selection

A selection was made from both types of source. The selection was made on the basis of the ecological approach outlined earlier and constraints concerned with the practicalities of analysis.

The central target consisted of episodes where an individual appeared to depart from emotional neutrality in a reasonably consistent way for an appreciable period. That included emotional states which were not particularly extreme, so long as the signs of emotion seemed strong and stable enough to be detectable to an observer. Mixed emotional states were included when the signs were strong enough to signal departure from neutrality despite a degree of conflict or instability. A secondary target, required for comparison, consisted of episodes where an individual who showed emotion elsewhere appeared to be effectively neutral.

From a total of 20 studio recordings, 9 were identified as containing usable material. Each on average contained 3 or 4 episodes that were regarded as being at all emotionally marked Out of a four-month period, 45 television broadcasts were identified as containing usable material. Within each of these broadcasts, there were on average 2 episodes that could be described as strongly emotionally marked.

The basic material of the database consists of 'clips' extracted from the selected recordings. Clips ranged from 10 – 60 secs in length. Several issues governed their selection. Each clip focussed on a selected individual. It was required to contain both audio and visual material for the selected person, including at least some shots where the two modes co-occurred. The basic aim was that the material should be perceived as a relatively self-contained episode. Where the episode included a display of emotion, the clip was long enough to contextualise the display and to show how the emotional state developed over time. Each emotional clip was paired with a comparatively neutral clip for the same person. Where possible, more than one emotional clip was selected for the same person.

Clips were captured as MPEG files, using a Broadway card for capture. The soundtrack of each clip was copied into an audio file in .wav format.

For the purpose of acoustic analysis, .wav files were edited. Voices other than the selected individual's were cut by setting signal level to zero during the periods when they were present. The resulting files retained the true time course of the interaction, but presented only one voice.

2.3 Overview of recordings

The database currently contains material from a total of 100 people, with at least one emotional and one comparatively neutral clip for each, giving a total of 239 clips. Of the clips, 209 are from the TV programmes, and 30 from the interview recordings.

That is a small return from a large amount of raw material. The reason relates to a point that has been earlier, without evidence. It is that archetypal emotions are a rare phenomenon. There are two related observations here.

First, displays of intense emotion are rare. The clips that make up the database are highly selected, starting from recordings chosen for their emotional content and then extracting emotional highlights from it. Even so, they include episodes that observers are reluctant to consider 'truly' emotional, partly because fullblown emotionality was so limited even in the selected sources.

Second, clear examples of 'pure' primary emotions are even rarer. Even where there was strong emotion, we found that anger and sadness, for example, often seemed to combine. Those are significant points if we are interested in systems that

can recognise naturally occurring emotion. They should not be designed on the assumption that emotion will generally consist of archetypal extremes, or even approach them.

3. ASSOCIATED DESCRIPTIONS

Work is in progress on providing three main types of descriptor for each clip. They will describe (i) perceived emotional content, (ii) significant auditory and visual features and (iii) acoustic properties recovered by the ASSESS system.

3.1 Descriptors of emotional content

The database incorporates two types of description for the emotional content of each clip – dimensional and categorical. Both types of representation have uses in particular contexts.

Dimensional Files containing dimensional descriptions are associated with each clip. Each one specifies the clip's emotional content as perceived by a particular subject in terms of activation-evaluation space. The techniques are summarised here: more information is given elsewhere in these proceedings[16]. Activation and evaluation are dimensions that are known to discriminate effectively between emotional states. Activation values indicate how dynamic a state is – e.g. they are high in excitement, low in boredom. Evaluation values give a global indication of the positive or negative feeling associated with the emotional state – e.g. they are positive in happiness, negative in despair. To a first approximation, emotion terms correspond to points in a space defined by those two axes. The space is naturally circular: alert neutrality lies at the centre, and states which are at the limit of emotional intensity define a circle around it.

A computer program called Feeltrace based on that representation allows users to generate time-varying descriptions of emotional content as they perceive it. Activation-evaluation space is represented by a circle on a computer screen, and observers describe perceived emotional state by moving a pointer to the appropriate point in the circle using a mouse. The output records the position of the pointer on the two axes at intervals of a few milliseconds.

FEELTRACE is well suited to an approach concerned with gradation and richness in emotion. Three particular reasons stand out. First, Feeltrace locates emotional states in terms of continuous scales rather than discrete categories. Second, it allows emotional content to be tracked over time. Third, because the techniques apply equally to clips presented in different modalities, it is possible to explore the relationship between time and modality in the expression of emotion.

Categorical In addition to Feeltrace files, we also provide categorical labels for the emotional content of each clip. (e.g. angry, happy etc).

Categorical labels are given by raters immediately after Feeltracing a clip. The procedure has two parts. The first part is designed to ensure access to a description that is coarse but tractable (because it uses a small number of categories). The second allows a description that is finer grained but uses a relatively large number of categories.

Raters are given two lists of emotion labels based on preparatory experiments [17]. The first contains the 16 words that subjects choose most often when they are asked to select a minimal vocabulary of emotion words. The second list contains the next most commonly selected 24 words. Raters choose the label from the first list that best describes the dominant emotion in the clip. If that is not a satisfactory description, they may also choose up to two more labels from either list. Order of choice is recorded. For each label, subjects also report the intensity of the emotion on a scale from 1-3.

3.2 Other descriptors

For each clip, the ASSESS system automatically summarises acoustic properties that are potentially relevant to its emotional impact. It is described elsewhere in these proceedings[18]. Work is also in progress on a rating form designed to record visual and auditory features that trained observers judge may contribute to the emotional impact of a clip.

4. DATA AND DEVELOPMENTS

Work in progress is outlined here because it illustrates the kind of use to which the database can be put.

To date, three raters have rated the emotional content of all 239 clips in the database, presented in audiovisual mode. All received systematic training in the use of Feeltrace and passed an initial test designed to exclude individuals who rate eccentrically.

Figure 1 summarises Feeltrace results from the first two raters. Each point represents the mean score on the two axes (evaluation and activation) for a clip. One of the effects of the displays is to indicate the kind of coverage that has been achieved. The database appears to cover the upper part of activation-emotion space reasonably well. The lower part is a problem, presumably because the main source (television) tends not to dwell on low-activation emotions – drowsy happiness, boredom, etc. In mitigation, it is not obvious how important those emotions are likely to be for speech research.

It is also noticeable that the two distributions are not identical. Because category labels are available as well, it is possible to check whether the patterns reflect different uses of Feeltrace or deeper lying differences in the perception of emotion. The second is reasonably likely because people do differ quite markedly in the way they interpret signs of emotion. A good database should probably reflect that by including examples of contrasting assessments.

A second stage of Feeltracing has been completed but not analysed. Raters have been presented with material from selected clips in different modalities – audio only, video only, audio with verbal content removed by filtering and audiovisual. The clips were selected to be widely spaced in activation / evaluation space. Preliminary indications are that vocal signals

of emotion have quite a limited the role in ecologically valid contexts: discrimination is quite robust without sound, and poor when only voice is available. If so, it is all the more interesting to tease out exactly what role voice plays.

A final development relates back to the core task of training networks. For training purposes, the number of examples in the database remains low. However, it is possible to multiply them by extracting examples of genuine, spontaneous emotional utterances, and asking actors to deliver the same words with appropriate emotion. That approach directs them towards aspects of genuine emotional expression that they might not spontaneously register (notably choice of words and phrasing), but allows some of the advantages of a designed dataset.

It may be clear that our work has hardened our scepticism about the benign interpolation hypothesis. Real examples constantly underline the complexity of the interactions involved – between signs in different modalities, and between signs and a perceiver. From that perspective, database development becomes a much more substantial task, and a more interesting one, than has usually been assumed – a large, integral part of the process of understanding emotion, not an irksome preliminary to be bypassed as soon as possible.

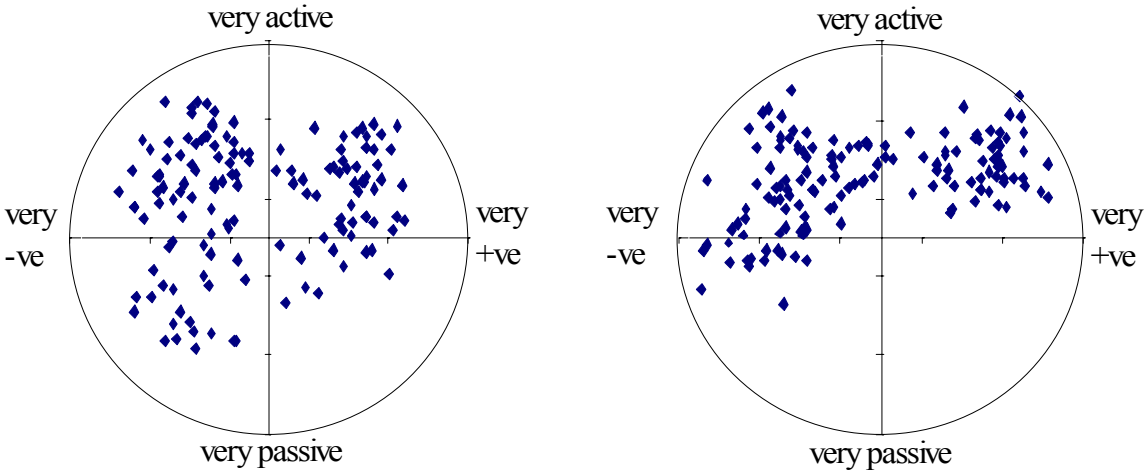

Figure 1: Feeltrace ratings of the whole database from two observers. Each point represents average rating for a clip.

5. REFERENCES

1. http://www.image.ntua.gr/physta/
2. Engberg, I. S., Hansen, A. V. et al. (1997) Design, recording and verification of a Danish Emotional Speech Database. *Proc. EuroSpeech*, Rhodes, 1997
3. Paeschke, A. &.Sendlmeier, W. (2000) Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements *This volume*
4. http://www.icp.inpg.fr/ELRA
5. Amir, N. et al. (2000) Analysis of an emotional speech corpus in Hebrew based on objective criteria *This volume*.
6. Greasley, P. et al. (1995) Representation of prosodic and emotional features in a spoken language database. *Proc XIII ICPhS*, Stockholm.
7. Ekman, P. & Friesen, W. (1975*) Pictures of Facial Affect*. Palo Alto CA: Consulting Psychologists Press.
8. http://cvc.yale.edu/projects/yalefaces/yalefaces.html
9. ftp://ftp.orl.co.uk/pub/data/
10. http://pics.psych.stir.ac.uk/index.html
11. ftp://whitechapel.media.mit.edu/pub/
12. http://www.unige.ch/fapse/emotion/
13. Schubiger, M. (1958*) English Intonation. Its Form and Function*. Tubingen: Niemeyer.
14. O'Connor, J. D. & Arnold, G. (1973) *Intonation of colloquial English*. London: Longman.
15. Douglas-Cowie, E. (1978) Linguistic code-switching in a Northern Irish village: social interaction and social ambition In P. Trudgill (ed*.) Sociolinguistic patterns in British English*. London: Edward Arnold, pp. 37-51.
16. Cowie, R. et al. (2000) Feeltrace: An instrument for recording perceived emotion in real time. *This volume*
17. Cowie, R. et al. (1999) What a neural net needs to know about emotion words. In N. Mastorakis (ed) *Computational Intelligence and Applications*. World Scientific Engineering Society
18. McGilloway, S. et al. (2000) Approaching automatic recognition of emotion from voice: a rough benchmark *This volume*.