Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Speech Communication 40 (2003) 33–60

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Emotional speech: Towards a new generation of databases

Ellen Douglas-Cowie [a,*], Nick Campbell [b], Roddy Cowie [a], Peter Roach [c]

[a] *Schools of English and Psychology, Queen's University, Belfast BT7 1NN, N. Ireland, UK*
[b] *ATR, Human Information Science Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan*
[c] *School of Linguistics and Applied Language Studies, University of Reading, Whiteknights, Reading RG6 6AA, UK*

Abstract

Research on speech and emotion is moving from a period of exploratory research into one where there is a prospect of substantial applications, notably in human–computer interaction. Progress in the area relies heavily on the development of appropriate databases. This paper addresses four main issues that need to be considered in developing databases of emotional speech: scope, naturalness, context and descriptors. The state of the art is reviewed. A good deal has been done to address the key issues, but there is still a long way to go. The paper shows how the challenge of developing appropriate databases is being addressed in three major recent projects—the Reading–Leeds project, the Belfast project and the CREST–ESP project. From these and other studies the paper draws together the tools and methods that have been developed, addresses the problems that arise and indicates the future directions for the development of emotional speech databases.

Résumé

L'étude de la parole et de l'émotion, partie du stade de la recherche exploratrice, en arrive maintenant au stade qui est celui d'applications importantes, notamment dans l'interaction homme–machine. Le progrès en ce domaine dépend étroitment du développement de bases de données appropriées. Cet article aborde quatre points principaux qui méritent notre attention à ce sujet: l'étendue, l'authenticité, le contexte et les termes de description. Il présente un compte-rendu de la situation actuelle dans ce domaine et évoque les avancées faites, et celles qui restent à faire. L'article montre comment trois récents projets importants (celui de Reading–Leeds, celui de Belfast, et celui de CREST–ESP) ont relevé le défi posé par la construction de bases de données appropriées. A partir de ces trois projets, ainsi que d'autres travaux, les auteurs présentment un bilan des outils et méthodes utilisés, identifient les problèmes qui y sont associés, et indiquent la direction dans laquelle devraient s'orienter les recherches à venir.

Keywords: Databases; Emotional speech; Scope; Naturalness; Context; Descriptors

1. The context

Research on speech and emotion is moving from a period of exploratory research into one where there is a prospect of substantial applications, notably in human–computer interaction.

* Corresponding author.
E-mail address: e.douglas-cowie@qub.ac.uk (E. Douglas-Cowie).

0167-6393/02/$ - see front matter
PII: S0167-6393(02)00070-5

A recent ISCA workshop discussed the changes that were needed to make that transition (Cowie et al., 2000a; www.qub.ac.uk/en/isca/index.htm). The most widely agreed priority was a change in the scale and quality of databases.

There are many problems surrounding database development, some of which may not become obvious until it is too late. This paper aims to articulate key lessons from existing projects, so that new database projects can learn from them. In addition to providing a broad overview of sources, the paper takes three major projects as case studies. They are the Reading–Leeds Emotion in Speech project, which collected the first large scale database of naturally occurring emotional speech and devised a speech labelling system (Greasley et al., 1995; Roach et al., 1998; www.linguistics.rdg.ac.uk/research/speechlab/emotion/); the Belfast project (Cowie et al., 2000a; Douglas-Cowie et al., 2000), which assembled the first large audio–visual database of emotion as part of the Principled Hybrid Systems and Their Application (PHYSTA) project (Cowie et al., 2001; www.image.ntua.gr/physta/); and the CREST–ESP project (www.isd.atr.co.jp/esp) which is currently developing a database of "expressive speech" in English, Japanese and Chinese, for the purpose of expressive speech synthesis. The co-authors of this paper have been involved in developing these databases.

The discussion is organised around four broad questions. First, what should the scope of speech and emotion databases be, both in terms of numbers of subjects and in terms of the range and numbers of emotions? Second, what should the nature of the material be—natural or acted, deliberately induced by the researcher or culled from existing sources? Third, what kind of context needs to be provided for episodes that carry vocal signs of emotion—considering both the time course of emotional episodes and the other modes of information (verbal, facial, etc.) that accompany vocal signs? Finally, what descriptors should we attach to the speech and to the emotional content of the databases?

Satisfactory answers depend on assimilating information from diverse sources. In the rest of this section we identify key sources. Four bodies of literature are relevant. These deal with existing emotional speech datasets and descriptions of them; the psychological literature on emotion; sources concerned with speech data collection in general; and applied research on speech synthesis and recognition. To these we add two informal sources—the debate and discussion among researchers that took place at the ISCA workshop, and our own practical experience in putting together databases of emotional speech. The next section of the paper then discusses each question in turn, in the light of the sources that have been outlined.

Most of the literature on emotion in speech is underpinned by sources that we call 'datasets' rather than 'databases'. They are comparatively small-scale collections of material, typically created to examine a single issue, and not widely available. These datasets yield both positive and negative lessons. Positively, they incorporate methodologies and descriptive tools that are potentially valuable for a new generation of databases. Negatively, they highlight problems, particularly problems to do with scale, validity, and generalisability.

The psychological literature on emotion might be expected to have a major influence on the selection and description of emotions for database research, but that has not always been the case, and when psychological ideas are invoked, they are often dated. The paper notes some of the ways that recent psychology impinges on databases: an extended discussion of the topic is provided by Cowie and Cornelius (2003).

Several efforts within the speech community are relevant to data collection. Socio-linguists have emphasised the importance of appropriate fieldwork techniques (Milroy, 1987). Corpus linguistics and speech recognition research illustrate the benefits of large shared databases (McEnery and Wilson, 1996; ten Bosch, 2000). Finally COCOSDA, The International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques, which promotes collaboration and information exchange in speech research, has recently adopted emotional speech as a future priority theme (www.slt.atr.co.jp/cocosda).

Our fourth set of sources relates to the growing emphasis on applications in the area—the syn-

thesis of emotionally coloured speech and recognition of emotional speech. That orientation brings to the fore questions that pure research might overlook—for instance, is a successful application likely to depend on considering speech as only one of several mutually supportive information sources, and therefore on the availability of multi-modal databases (i.e. necessitating the collection of visual as well as audio material)? We argue that rather than pure and applied emphases being in conflict, each stands to benefit from awareness of the other.

A final source of information is informal discussion, particularly at the ISCA workshop on Speech and Emotion (op. cit). Many of the participants at the workshop expressed concerns about the collection and description of emotional speech, and the discussions helped to identify common concerns.

From those sources, we draw an assessment of the issues to be addressed in database development. That assessment then provides a framework for discussing current datasets, and the developments that are needed.

2. Content and descriptors: needs and issues

Four main issues need to be considered in developing a database—the scope, naturalness and context of the content; and the kinds of descriptor it is appropriate to use. This section explains what the terms mean and why the issues are important. Recognition is also given to the practical issues of format and distribution.

2.1. Scope

We use the term 'scope' to cover several kinds of variation that a database or dataset may incorporate, notably number of different speakers; language spoken; type of dialect (e.g. standard or vernacular); gender of speakers; types of emotional state considered; tokens of a given state; social/functional setting. These kinds of variation are potentially important for any attempt to generalise. The issue would be less pressing if signs of emotion in speech were highly consistent across individuals and occasions. There do seem to be facial signs of emotion that are effectively universal (Ekman, 1994), and there are reasons to expect that the same is true of at least some vocal signs (Williams and Stevens, 1972). But a recent review of the literature (Cowie et al., 2001) shows that although some features are remarkably consistent across studies, others are quite variable. The findings for hot anger seem consistent, as do those for joy/elation. But there are inconsistencies for most other emotions and emotion-related states that have been studied at all frequently. Sadness generally seems to be marked by a decrease in mean F_0, but there are cases where there is no change, and a source not covered in the review shows an increase (Pereira, 2000a). It is often reported that fear is marked by an increase in F_0 range and speech rate; but there are contradictory findings for both for variables. Studies of disgust report both an increase in mean F_0 and a decrease in mean F_0. For boredom, speech rate has been reported both to increase and to decrease.

Some variations may simply reflect inconsistent procedure or interpretation of emotion categories, or differences between real and simulated data. Others, though, seem likely to reflect real differences in the vocal expression of emotion, from speaker to speaker, from culture to culture, and across genders and situations. Comparisons between languages and cultures are limited, but they suggest substantial differences. For example, in Japanese society, an open display of emotion may be considered anti-social or "selfish" behaviour, and it is considered normal to show a smile when angry or embarrassed. This may be partly responsible for the "inscrutable" image sometimes referred to by westerners. On the other hand, the implicit expression of emotion by use of pressed-voice is often used to show positive concern or to display sympathy for the listener (for general discussion of relevant Japanese cultural differences see (Marcus and Kitayama, 2001)). Gender is recognised as a key socio-linguistic variable (Trudgill, 1983). It would be surprising if it were not important in emotion too, and there are some indications that it is (Brend, 1975). The same is true of what we have called social setting, i.e. the

relationship that governs interaction between speaker and listener. At the most basic, the normal setting for vocal expression of emotion is dialogue. Monologues may be easier to control, but they cannot be expected to reveal all the features that will occur in dialogue. At a finer level, it seems quite likely that what is accepted as business-like in a task-oriented setting would convey irritation in a sociable one.

The importance of variation depends on research goals. For the purpose of synthesis, it may well be enough to study a single speaker, so that his or her methods of expressing emotion can be modelled. On the other hand, research aimed at recognising emotion needs databases that encompass as many as possible of the signs by which a given emotion may be expressed. Failure to address that issue may contribute to the notorious difficulty of recognising emotion from speech in anything approaching a naturalistic context (Batliner et al., 2000). Conversely, speech synthesis needs to understand how linguistic context can affect expression, whereas recognition may only need to identify the contexts in which emotion can be reliably inferred. Pure research has freedom to choose how it approaches those issues, but individual researchers should be explicit about the choices they make.

A second aspect of scope relates to the range of emotions considered. It seems fair to say that the default option is to consider a relatively small set of 'basic' emotions, the most obvious being fear, anger, sadness and happiness. Certainly many participants at the ISCA workshop appeared to regard that type of approach as self-evidently correct. It reflects the popular theory that a few universal types underlie the whole of emotional life. Sophisticated versions of the theory have substantial support (Ekman, 1999), but its application to speech cannot be taken for granted, for a number of reasons. Emotional life in general is modulated by strong cultural influences (Harré, 1986) and constrained by 'display rules' (Ekman and Friesen, 1969). Since speech is a cultural activity par excellence, signs of emotion in speech may well be particularly subject to cultural influences. Also, speech in daily life tends to express moderate emotional states rather than full-blown basic emotions. These issues are covered more fully by Cowie and Cornelius (2003).

The states that seem most practically important are often emotion-related rather than pure emotions per se. In terms of speech synthesis, it is unclear why we should want to synthesise full-blown fear or anger or sadness. Milder forms of expression are more likely to be required, including what Scherer calls interpersonal stances, such as friendliness, interest, and pleasure (Scherer, 2000). Similarly, 'stress' is practically important, and has already attracted a good deal of research (Johannes et al., 2000; Fernandez and Picard, 2000).

Those observations suggest that the emotional scope of databases needs to be thought through carefully. Since standard lists contain more than a hundred words for (non-basic) emotions (Cowie et al., 2001), the scope may have to be very large. It is presumably possible to work with a smaller number of 'landmark' states and interpolate, but establishing a set of landmarks that is appropriate for speech research is an empirical task, which itself depends on access to data that spans the known range of emotional states.

2.2. Naturalness

The easiest way to collect emotional speech is to have actors simulate it. The difficulty with that approach is that strikingly little is known about the relationship between acted data and spontaneous, everyday emotional speech.

It is certainly true that good actors can generate speech that listeners classify reliably. Material studied by Banse and Scherer (1996), for example, produced recognition rates of 78% for hot anger, 76% for boredom and 75% for interest, though scores for other emotions were lower with an average recognition rate of 48% across 14 emotions. However, that kind of evidence does not establish how closely the speech mirrors spontaneous expression of emotion.

There are many reasons to suspect that there are systematic differences between acted and natural emotional speech. Acted speech is often 'read', not spoken, and read speech is well known to have distinctive characteristics (Johns-Lewis, 1986).

Neither the words nor the phrasing are typically chosen to simulate emotional speech. The typical form is a non-interactive monologue, and so interpersonal effects are not represented. The context is typically minimal, so the material does not indicate how vocal signs of emotion build and fade over time, or relate to other kinds of signal.

It would therefore be unsurprising if attempts to express emotion under these very atypical circumstances had very atypical features. At one extreme, it may amount to caricature (which would, of course, make for high recognition rates). Skilled actors who are engaged in an interpersonal drama may be a different matter. Our intuition is that even their performances would not usually be confused with truly natural behaviour. Once again, the only way to establish the point is by reference to databases of naturally occurring emotion.

The price of naturalness is lack of control. Emotion has an unpredictability that makes it difficult to collect samples of people in a target state, whether it is induced or spontaneous. Particularly if it is spontaneous, identifying the emotion that is being expressed becomes a substantial issue. Some applications (e.g. concatenative synthesis) need phonetically and prosodically balanced data sets, and it is difficult to imagine easily achieving that kind of balance with truly natural speech. The long-term solution to those problems may well be 'bootstrapping', i.e. using truly natural material to guide the production of material that is acted, but genuinely close to nature.

Again, research goals matter, and in some cases, naturalness may actually not be the relevant goal. For instance, a simulated newsreader should presumably produce the kind of modulation that a real newsreader does rather than simulating genuine fury or grief.

2.3. Context

There is direct evidence that listeners use context to determine the emotional significance of vocal features (Ladd et al., 1986; Cauldwell, 2000). Hence if research aims to understand human performance, or to match it, it needs databases that contain evidence on the way vocal signs relate to their context. One of the obvious doubts about acted speech is whether it captures subtler aspects of contextualisation in naturally emotional speech. Four broad types of context can be distinguished.

(a) Semantic context: Genuinely emotional speech is likely to contain emotionally marked words. There is a clear potential for interaction between content and vocal signs. Various kinds of relationship can be envisaged—such as trade-off calculated to control the overall level of emotionality conveyed, selective allocation of vocal signs of emotion to emotionally significant words, and tendency of vocal signs to follow emotive words.
(b) Structural context: It seems likely that many signs of emotion are defined relative to syntactic structures—stress patterns, default intonation patterns, etc. If so, misleading conclusions may be drawn if databases fail to allow for comparison across relevant syntactic forms, or if investigators ignore relevant distinctions. Less often noted is the possibility that emotion may be signalled by variations in style, which are expressed in structural characteristics of the utterances (long or short phrases, repetitions and interruptions, etc.).
(c) Intermodal context: The fact that we can communicate a wide range of emotions over the telephone shows that analysis concerned with speech alone is a reasonable undertaking. However, speech may often function as a supplement to other sources of information about emotion rather than as a stand-alone source. Normally we both hear and see a speaker, and the visual channel provides several kinds of emotion-related information (notably facial expression, gesture, and posture). There is reason to suspect that audio information could at least sometimes play a rather specific role within that context: it is known that in speech-reading, audio and visual channels are to a considerable extent complementary (Summerfield, 1983). Experiments have begun to consider whether the same is true of emotion (de Gelder and Vroomen, 2000a,b; Massaro and Cohen, 2000), but without access to audio–visual databases, it is difficult to know whether

the critical combinations have been addressed. Other modes may also be relevant in practical applications, e.g. a user's keyboard behaviour or temperature.

(d) Temporal context: Natural speech involves distinctive patterns of change as emotion ebbs and flows over time. Databases need to include material that reflects that linear sequential development if virtual agents are to reproduce it or to exploit it (e.g. by using nearby phrases to resolve local ambiguity in emotional tone). It also seems likely that at least sometimes, the emotional significance of a speech pattern may only be evident in the context of other pointers to an emotional build-up.

2.4. Descriptors

Constructing a database requires techniques for describing the linguistic and emotional content on one hand, and the speech on the other.

The requirements for accurate labelling of emotional content may interact with naturalness. Acted material may well be adequately described in terms of category labels such as sad, angry, happy, etc. Natural databases, though, are likely to involve gradation in and out of emotional peaks, coincidence of different emotions, and relatively subtle states (e.g. 'vengeful anger'). The result is a serious tension between faithful description and statistical tractability. The psychological literature offers alternative ways of describing emotion that may ease the problem, and we return to developments in that area in Section 4.

In terms of speech descriptors, two issues stand out. First, coding needs to acknowledge the full range of features involved in the vocal expression of emotion, including at least voice quality, prosody and non-linguistic features such as laughter, crying, etc. Second, it needs to describe the attributes that are relevant to emotion. A fundamental choice is between categorical descriptors (e.g. ToBI) and continuous variables. The relative merits of the two types remain to be resolved.

If databases are multi-modal, then additional types of label (e.g. facial and gestural) may also be needed. There are now well-established standards for describing relevant facial gestures, in particular, the FACS model (Ekman and Friesen, 1978), from which derives the ISO MPEG-4 standard (1996).

2.5. Accessibility

The value of a database increases enormously if it is available to the whole speech community, so that effort does not need to be duplicated, algorithms can be compared on the same data, and so on. Two main issues have a bearing on availability: format and ethics.

The format of the data files needs to be standard and/or transparent. This applies not only to formats for coding raw material (e.g., wav), but also to the coding of descriptors. Experience suggests that the temptation to adopt ad hoc conventions can be overwhelming. Format also needs to encode all relevant details. For instance, MPEG files have obvious advantages in terms of storage and transmission, but it is less clear whether they provide full enough information about the signal or the details of its collection.

More fundamental are problems of ethics and copyright, particularly with natural data. Natural emotional data is often very personal, and subjects may object to wide circulation. Radio and television provide rich sources, in chat shows, documentaries, etc., but accessing them raises serious copyright problems.

It is clear that there are challenges in assembling and describing databases of the type that meet the needs we identify. We turn to look at the state of the art in more detail.

3. Datasets: the status quo

This section attempts to set out the current state of the art in terms of datasets of emotional speech. It does so in the form of a table and accompanying text. The table is not intended to be an exhaustive description of every dataset—rather to indicate the kind of data that has been used to date in research on speech and emotion. On one hand, it aims to convey how limitations at that level currently limit the conclusions that can be drawn: on the other, it

draws attention to the range of techniques for collecting and describing data that have been explored. Three of the key points identified above—scope, naturalness and context—are addressed within Table 1. The issue of descriptors is discussed separately.

The table is designed to give key information briefly. The identifier for the dataset may be either a general name, a literature reference or a website. Scope covers number of subjects, emotions considered, and language involved (to indicate the cultural range of existing datasets). Under 'naturalness', we include several categories—simulated, semi-natural and natural; scripted or unscripted; and type of material (e.g. passages, sentences, numbers). 'Semi-natural' covers a variety of techniques that might be expected to generate something between outright simulation and total naturalness: examples are given as appropriate. Under context we note whether there is any attempt to address the issue of emotional development and change over time, and whether the data is audio or audio–visual.

The table is organised in terms of the simulated/semi-natural/natural distinction, beginning with sources that are unequivocally acted, and moving through various intermediate types to sources that are fully natural. A number of general points can be made about the material. They are summarised under the headings explained in the previous section. Some involve the limitations of available resources, but there are also indications that there is movement towards consensus on some key issues.

3.1. Scope

Historically, most studies have been limited in scope, in terms of number of speakers, range of languages, and emotions covered. However, there are exceptions, though rarely in all respects, and recent studies show increasing recognition of the need for scope at least in terms of numbers of speakers—including the studies considered more fully in Section 4.

The number of subjects studied has tended to be small, so that it is difficult to gauge the extent of inter-subject variability. The possibility of gender effects compounds the problem. Samples are sometimes balanced for gender as in the Berlin corpus (www.kgw.tu-berlin.de/) and the Hebrew corpus (Amir et al., 2000), but the total numbers are often not large enough for useful statistical comparisons on the basis of gender (Berlin corpus, 5 males, 5 females; Hebrew corpus, 16 males, 15 females; van Bezooijen (1984), 4 male, 4 female; Banse and Scherer 6 male, 6 female), though there are exceptions (Tolkmitt and Scherer, 1986; France et al., 2000).

With respect to languages, the picture divides into two parts. Most work has been done on the Germanic languages. Coverage for other language groups is sparse, though there are datasets for Spanish, Russian, Hebrew, Korean and Japanese. As a result, it is difficult to gauge how many of the relationships that the literature describes may be specific to single relatively homogeneous cultural milieu. Less obviously, it seems likely that even within Western Europe, most of the information available relates to educated individuals using standard variants of the languages involved. There could easily be considerable socio-linguistic variation in the expression of emotion within a single country; for example, it is commonly thought that non-standard speakers make more use of expletives to signal intense feelings. If that were so, the available data would not show it.

The picture is quite complex with respect to the scope of emotions covered. A few core states are considered in a wide range of studies—anger, happiness, sadness, fear and neutrality (with disgust on the margin of the group). However, the table suggests a good deal of dissatisfaction with the default approach of collecting datasets that cover only that kind of range. Two main alternatives emerge. Some investigators have moved towards a fuller coverage of the range of emotions, using a larger number of emotion categories (often about a dozen), and often distinguishing between forms of some core emotions. It is increasingly recognised that hot and cold anger are distinct, and different forms of happiness (such as elation and contentment) are sometimes separated. A few also distinguish more and less intense forms of a single emotion. In the other direction, a number of investigators have chosen to study a relatively narrow range of emotional states in depth rather

Table 1
Examples of how datasets address the issues of scope, naturalness and context

Identifier	Scope			Naturalness			Context	
	Number subjects	Description given of emotions	Language	Simulated, semi-natural, natural	Scripted unscripted	Linguistic nature of material	Time sensitive	Mode
Danish emotional speech database (Engberg et al., 1997)	4	Anger, happiness, neutrality, sadness, surprise	Danish	Simulated	Scripted	Subjects read 2 words, 9 sentences and 2 passages in a range of emotions (material not emotionally coloured)	No	Audio
Groningen, 1996 ELRA corpus number S0020 (www.icp.inpg.fr/ELRA)	238	Database only partially oriented to emotion	Dutch	Simulated	Scripted	Subjects read 2 short texts with many quoted sentences to elicit emotional speech	No	Audio
Berlin database (Kienast and Sendlmeier, 2000; Paeschke and Sendlmeier, 2000)	10 (5 male, 5 female)	Anger–hot, boredom, disgust, fear–panic, happiness, neutrality, sadness–sorrow	German	Simulated	Scripted	10 sentences (material selected to be semantically neutral)	No	Audio
Pereira (Pereira, 2000a,b)	2	Anger (hot), anger (cold), happiness, neutrality, sadness	English	Simulated	Scripted	2 utterances (1 emotionally neutral sentence, 4 digit number) each repeated	No	Audio
van Bezooijen (van Bezooijen, 1984)	8 (4 male, 4 female)	Anger, contempt, disgust, fear, interest, joy, neutrality, sadness, shame, surprise	Dutch	Simulated	Scripted	4 semantically neutral phrases	No	Audio
Alter (Alter et al., 2000; also this journal)	1	Anger (cold), happiness, neutrality	German	Simulated	Scripted	3 sentences, 1 for each emotion (with appropriate content)	No	Audio
Abelin (Abelin and Allwood, 2000)	1	Anger, disgust, dominance, fear, joy, sadness, shyness, surprise	Swedish	Simulated	Scripted	1 semantically neutral phrase	No	Audio
Polzin (Polzin and Waibel, 2000)	Unspecified no of speakers. Segment numbers 1586 angry, 1076 sad, 2991 neutral	Anger, sadness, neutrality (other emotions as well, but in insufficient numbers to be used)	English	Simulated	Scripted	Sentence length segments taken from acted movies	No (segments chosen for consistent emotion)	Audio–visual (though only audio channel used)

Banse and Scherer (Banse and Scherer, 1996)	12 (6 male, 6 female)	Anger (hot), anger (cold), anxiety, boredom, contempt, disgust, elation, fear (panic), happiness, interest, pride, sadness, shame	German	Semi-natural. Actors were given scripted eliciting scenarios for each emotion, then asked to act out the scenario. (Each contained the same 2 semantically neutral sentences for acoustic comparison.)	Scripted	2 semantically neutral sentences (nonsense sentences composed of phonemes from Indo-European languages)	No	Audio–visual (visual info used to verify listener judgements of emotion)
Mozziconacci (Mozziconacci, 1998)[a]	3	Anger, boredom, fear, disgust, guilt, happiness, haughtiness, indignation, joy, neutrality, rage, sadness, worry	Dutch	Semi-natural. Actors asked to read semantically neutral sentences in range of emotions, but practised on emotionally loaded sentences beforehand to get in the right mood	Scripted	8 semantically neutral sentences (each repeated 3 times)	No	Audio
Iriondo et al. (Iriondo et al., 2000)	8	Desire, disgust, fury, fear, joy, surprise, sadness	Spanish	Semi-natural. Subjects asked to read passages written with appropriate emotional content	Scripted	Paragraph length passages (20–40 mm s each)	Unclear	Audio
McGilloway (McGilloway, 1997; Cowie and Douglas-Cowie, 1996)	40	Anger, fear, happiness, neutrality, sadness	English	Semi-natural. Subjects asked to read 5 passages written in appropriate emotional tone and content for each emotional state	Scripted	Paragraph length passages	No. Emotional tone intended to be fairly constant throughout passage	Audio
Belfast structured database. An extension of McGilloway database above (Douglas-Cowie et al., 2000)	50	Anger, fear, happiness, neutrality, sadness	English	Semi-natural. Subjects read 10 McGilloway-style passages AND 10 other passages—scripted versions of naturally occurring emotion in the Belfast natural database	Scripted	Paragraph length passages written in first person	Yes. The scripts span a period in which the emotion varies in intensity	Audio

Table 1 (*continued*)

Identifier	Scope			Naturalness			Context	
	Number subjects	Description given of emotions	Language	Simulated, semi-natural, natural	Scripted unscripted	Linguistic nature of material	Time sensitive	Mode
Amir et al. (Amir et al., 2000)	60 Hebrew speakers and 1 Russian speaker	Anger, disgust, fear, joy, neutrality, sadness	Hebrew and Russian	Semi-natural. Subjects asked to recall personal experiences involving each of the emotional states	Unscripted	Non-interactive discourse	Unclear (1 min allowed per emotional state)	Audio
Fernandez et al. (Fernandez and Picard, 2000; also this journal)	Data reported from 4 subjects	Stress	English	Semi-natural. Subjects give verbal responses to maths problems in simulated driving context	Unscripted	Numerical answers to mathematical questions	Yes. Recordings span period—stress level raised and lowered	Audio
Tolkmitt and Scherer (Tolkmitt and Scherer, 1986)	60 (33 male, 27 female)	Stress (both cognitive and emotional)	German	Semi-natural. Two types of stress (cognitive and emotional) were induced through slides. Cognitive stress induced through slides containing logical problems; emotional stress induced through slides of human bodies showing skin disease/accident injuries	Partially scripted	Subjects made 3 vocal responses to each slide within a 40 s presentation period—a numerical answer followed by 2 short statements. The start of each was scripted and subjects filled in the blank at the end, e.g. 'Die Antwort ist Alternative...'	No	Audio
Reading–Leeds database (Greasley et al., 1995; Roach et al., 1998; this paper)	Around 4.5 h material	Wide range	English	Natural. Unscripted interviews on radio/television in which speakers have been induced by interviewers to relive emotionally intense experiences	Unscripted	Unscripted interactive discourse	Yes	Audio

Belfast natural database (Douglas-Cowie et al., 2000; this paper)	125, 31 male, 94 female	Wide range (details in this paper)	English	Natural. 10–60 s long 'clips' taken from television chat shows, current affairs programmes and interviews conducted by research team	Unscripted	Unscripted interactive discourse	Yes. Each 'clip' shows the context in which the emotion occurs and its development over time	Audio–visual
Geneva Airport Lost Luggage Study (Scherer and Ceschi, 1997, 2000)	109	Anger, good humour, indifference, stress, sadness	mixed	Natural. Unobtrusive videotaping of passengers at lost luggage counter followed up by interviews with passengers	Unscripted	Unscripted interactive discourse	Yes	Audio–visual
Chung (Chung, 2000)	6, 1 Korean speaker, 6 Americans	Joy, neutrality, sadness (distress)	English and Korean	Natural. Television interviews in which speakers talk on a range of topics including sad and joyful moments in their lives	Unscripted	Interactive discourse	Yes. Interviews lasted 20–40 min. Speech fragments extracted at intervals to reflect emotional development through interview	Audio–visual
France et al. (France et al., 2000)	115, 48 females, 67 males. Female sample 10 controls (therapists) 17 dysthymic 21 major depressed. Male sample 24 controls (therapists) 21 major depressed 22 high risk suicidal	Depression, neutrality, suicidal state	English	Natural. Therapy sessions/phone conversations. Post-therapy evaluation sessions were also used to elicit speech for the control subjects	Unscripted	Interactive discourse	Yes. Samples selected from each therapy session substantial in time—2 min 30 s	Audio

[a] Note: Database recorded at IPO for SOBU project 92EA.

than attempting to cover the whole gamut. Several studies focus on stress (Fernandez and Picard, 2000; Tolkmitt and Scherer, 1986). A number of other significant studies, not in the table, are also stress oriented (for example Bonner, 1943; Karlsson et al., 1998; Roessler and Lester, 1979; Johannes et al., 2000). In addition, the Geneva group has reported a study of travellers who had lost their luggage in Geneva airport, where predominant emotions were stress and anger (Scherer and Ceschi, 1997; Scherer and Ceschi, 2000, see table). Several studies also focus on depression-related states including France et al. (see table), and three studies not included in the table (Hargreaves et al., 1965; Stassen et al., 1991; Frolov et al., 1999).

Either of the alternative strategies can be defended, but both clearly need development. The multi-emotion strategy needs agreement on a descriptive system that provides acceptable coverage of the range of commonplace emotional states. If different investigators choose different sets of categories or dimensions, it becomes frustratingly difficult to integrate data across studies. The selective strategy needs to establish ways of checking whether signs that are distinctive within the database are shared with states outside it. An observation by Cowie et al. (1999b) illustrates why the issue matters. Vocal features that are typical of depression seem to be shared with poor reading, raising the prospect of potentially serious confusion.

3.2. Naturalness

The table shows that research has relied relatively heavily on material that is acted and involves read, non-interactive material—typically consisting of non-emotional text.

There are datasets that include fully natural speech in emotion-related states. They are often quite large, but conversely, they tend to deal with a rather specific type of state, which is often not emotion in a strict sense. The Vanderbilt II database (used in France et al., 2000) covers depression and suicidal states. There are others not listed in the table. The SUSAS database (Hansen and Bou-Ghazale, 1997) covers stress. The database used by Slaney and McRoberts (1998) covers mother–child interactions, which it has been argued are vocally related to emotion (Trainor et al., 2000). More directly related to emotion, but still narrowly focussed, is the Geneva group's recording of travellers who had lost their luggage (see table). The projects covered in Section 4 stand out in contrast as studies that cover genuinely natural speech in a broad range of emotional states.

A considerable proportion of the work involves intermediate strategies—labelled semi-natural in the table. Examples in the table show that studies at the artificial end of this group use actors to read material that lends itself to the required emotion (McGilloway, 1997). The Belfast structured database (Douglas-Cowie et al., 2000) moves nearer naturalness in that the texts to be read are taken from the natural Belfast database, so that their vocabulary and phrasing are appropriate to the emotion. Beyond that, Amir's subjects (Amir et al., 2000) recall particularly emotional events in conducive contexts. The technique still falls short of naturalness, though. Recalling fear in a secure laboratory may generate real emotion, but it is very unlike facing a real and present threat to life and limb; Stemmler (1992) has shown that the distinction matters physiologically. More realistic in that respect are studies that involve laboratory-induced emotions. A range of techniques are used such as solving maths problems aloud under different levels of stress in a simulated environment (Fernandez and Picard, 2000) and responding to unpleasant pictures (Tolkmitt and Scherer, 1986). Note, though, that these techniques can force the verbal content into a very specific and atypical mode. Studies of a similar type, not described in the table, are (Scherer et al., 1985; Bachorowski and Owren, 1995; Karlsson et al., 1998).

The overall situation is not so much a movement away from acted material as a recognition that it needs to be complemented by other sources. Acted material continues to be collected and used by sophisticated teams. However, increasing attention is being paid to methods of ensuring that the acted and semi-natural material is an adequate reflection of reality. Fully natural databases that can be used as a comparison or an aid to development are an integral part of that strategy.

3.3. Context

Much of the material in the table is singularly devoid of context—purely audio recordings of short utterances with neutral semantic content and a preset linguistic structure. However, several projects have addressed context-related issues directly or indirectly, and they help to highlight significant issues.

Several datasets do contain vocal signs in relevant semantic contexts. Not all of those contain appropriate structural context: for instance, the passages used by McGilloway (1997) are in a literary style that seems unlikely to reflect the phrasing of spontaneous emotional speech. Batliner et al. (2000) have outlined a proposal for using speech in the context of semantic and structural sources.

Related to semantic content is communicative intent. Fridlund (1994) in particular has questioned the usual assumption that signs of emotion essentially reflect inner states. Instead, he argues, they are expressive displays with social motives. If so, it is a serious problem that even the semi-natural datasets rarely locate expressions of emotion in anything approaching a credible communicative context.

The great majority of the datasets are purely audio, presumably because investigators have generally assumed that audio and visual channels function independently. However, several projects have considered audio–visual material—the Geneva group (Scherer and Ceschi, 1997; Scherer and Ceschi, 2000); Polzin and Waibel (2000); and the Belfast group (see Section 4).

The issue of temporal context is not often considered explicitly, but the way naturalistic studies select units for coding implies intuitions about the patterns that need to be considered. Several different strategies are represented. The units considered by Amir et al. (2000) are both relatively long and relatively homogeneous—monologues on a single emotive subject, lasting a minute each. Polzin and Waibel (2000) consider much shorter units, single sentences, but also require them to be emotionally uniform. The Reading–Leeds database considers 'turns' which average about 15 words, and may contain marked internal variation in emotional tone. The Belfast naturalistic database uses 'clips' that may contain several turns, and pairs emotional clips with a relatively neutral one from the same speaker. Section 4 gives more information. It is revealing that none of the groups consider more than about a minute of speech necessary to contextualise signs of emotion.

3.4. Descriptors

Section 3.1 above deals with the most salient group of issues surrounding description of emotion, hinging on the choice of a suitable set of category labels. Psychology offers a well-developed alternative in the form of dimensional approaches: for details, see (Cowie and Cornelius, 2003). It has been used in a few studies of speech, particularly the Belfast naturalistic study, which is presented more fully in Section 4.

Speech descriptors are not specified in the table. They varied enormously. In most studies, a few descriptors were selected, with relatively little comment on the choice. As a result, it is difficult to form a cohesive summary of the dimensions on which speech varies. However, there seems to be clear evidence that continuous acoustic measures can support automatic discrimination. Measures related to pitch, intensity, spectral shape, and timing all contribute (Banse and Scherer, 1996; Amir et al., 2000; Batliner et al., 2000; Polzin and Waibel, 2000; Cowie and Douglas-Cowie, 1996). A few groups have tried to develop a systematic framework that captures the full range of relevant properties—the Belfast and the Reading groups, whose approaches are considered in the next section, and the Geneva group.

The task of describing speech for emotion recognition clearly overlaps with more standard speech description tasks. Sophisticated tools have been developed for corpus annotation, and some use is made of them in connection with emotion (Polzin and Waibel, 2000; the Reading–Leeds project 1994–98). Recent developments in corpus technology are clearly of interest (Bird and Harrington, 2001). However, the case studies considered in the next section make the point that capturing the features relevant to emotion is a

distinctive task, and probably requires developments that are specific to it.

3.5. Overview

Reviewing the sources on which it is based, it is clear that contemporary knowledge about speech and emotion is likely to be incomplete in multiple respects. On the other hand, a good deal has been done to identify the kinds of development that are needed. The next section considers three projects that have responded to that challenge.

4. Towards a new generation of databases

Three databases—the Reading–Leeds database (www.linguistics.rdg.ac.uk/research/speechlab/emotion/; Greasley et al., 1995; Roach et al., 1998), the Belfast database (Douglas-Cowie et al., 2000; Cowie et al., 2001) and the CREST–ESP database (www.isd.atr.co.jp/esp)—are reviewed here in some depth. They are described in chronological order. The Reading–Leeds database pioneered large-scale naturalistic data collection. The Belfast database was influenced by it, but added an audio–visual dimension and explored more quantitative descriptive strategies. The CREST database is a third generation project, taking up ideas from the previous two, but applying them in a different context.

The databases represent some of the most sustained efforts to overcome core problems facing the field. In particular, they have set out to obtain genuinely natural data on a substantial range of emotions, and systematically addressed the problems that arise from that decision. Both the solutions that they have developed and the problems that remain are significant for the field as a whole.

4.1. The Reading–Leeds database

The project (ESRC grant no. R000235285) was begun in 1994 to meet the apparent need for a large, well-annotated body of natural or near-natural speech stored in an orderly way on computer. The project made advances in three broad areas. First, it identified types of natural material where phonetic marking of emotion was (and was not) evident. Second, it established some broad characteristics of that kind of material. Third, it developed principled techniques for annotating both the emotional content of the material and the features of speech that might carry emotional information. In the process, it revealed major difficulties in a number of areas.

Different aspects of the project were handled by the Speech Research Laboratory of the University of Reading, and the Department of Psychology at the University of Leeds, and the material is lodged in the ESRC Data Archive.

The essential aim of the project was to collect speech that was genuinely emotional rather than acted or simulated. The ideal recording was a passage of speech in which the speaker could also be heard speaking relatively normally so as to provide a baseline for comparison with the affected portion. Although the researchers were aware that this strategy would create serious problems in terms of the control of variables, and would severely limit the possibility of using statistical analyses on the results, it was felt that the value of having 'genuine' emotions being expressed would outweigh these disadvantages. It was also assumed that the speech should exhibit phonetic effects that could be causally ascribed to the effect of one or more emotions.

At the beginning of the project, a pre-existing body of data was expected to serve as the foundation of the corpus. It contained recordings of people being interviewed by a psychologist, who was asking questions about their emotional state in relation to particular stimuli. In fact, the information in that material turned out to be almost wholly negative. Most of the recordings turned out to be rich in verbal descriptions of emotional states, but very impoverished in terms of phonetic exponents of the states. The material was not analysed in detail, and other sources were explored. Nevertheless, it has been included in the corpus of recordings. It indicates that speech may convey an impression of emotionality without using phonetic marking to any great extent—reinforcing the point made earlier that there may be non-trivial interactions between verbal and pho-

netic markers of emotion (such as trade-off, i.e. speakers tend to avoid giving both).

The source which did provide emotional marking was broadcast material. Most of the recordings in the database consist of unscripted interviews in which speakers have been induced by interviewers to relive emotionally intense experiences. A typical example of such material is Esther Rantzen interviewing a man whose daughter was murdered. Laughter while speaking was recorded from speakers taking part in broadcast panel games. In addition, some other broadcast material was included where it was felt that the speaker was genuinely affected by emotion (e.g. the commentary of the Hindenberg Disaster).

The total amount of recorded material in the corpus is 264 min; of this, 78 min has been annotated with phonetic/prosodic information, while 72 min has been given psychological coding. In summary, the total amount of data that has received some useful annotation that could be valuable in research in emotional speech is 81 min. A further 183 min is included in the corpus but has not been analysed.

The psychological character of the material was explored in a series of three studies conducted by the Leeds group. The first study (Greasley et al., 1996, 2000) considered the complexity of emotions in naturally occurring speech. It compared the free-choice codings (by word or phrase) of stretches of emotional speech with fixed-choice codings (choosing from five basic emotion labels). Results indicated that the fixed-choice labelling was adequate to capture subjects' responses to some extracts, but not to others; hence it is probably necessary to include both levels in a database.

The second study (Greasley et al., 2000) highlighted the problem of studying the contribution of speech to emotional effect when the verbal content is also emotional. Respondents used Osgood et al.'s (1957) dimensions (Evaluation, Potency and Activity) to report perceived emotional content of individual words, both in their natural context (presented auditorily in stretches of emotional speech) and out of context (presented as typed lists). Codings in the two conditions differed significantly in 44% of the cases. The technique offers a first step towards analysing the way speech contributes to conveying emotion in a multi-modal context.

A third study (Sherrard and Greasley, 1996) extended the theme, again using Osgood et al.'s (1957) Evaluation and Activity dimensions. These time codings of emotionally relevant words were made either in the context of stretches of emotional speech or from decontextualised recordings of the same speech segments. The contextualised codings produced linear plots across utterances that were visually simpler, and more congruent with the basic-emotion codings of the speech segments that had been obtained in the first study.

Together, the findings underline the importance of considering ambiguity and context-dependence in spontaneous speech. Naturalistic databases are needed to explore these issues, but they can only serve that function if they are coded in appropriate ways.

The philosophy behind the coding systems developed for the project was to provide truly comprehensive systems of annotation, both for the features of speech that could be observed and for the emotional content that they conveyed.

The psychological coding, outlined in (Greasley et al., 1995; Waterman and Greasley, 1996), uses a range of ideas from contemporary psychology. There are four levels, each reflecting a different approach to the description of emotion, on the principle that the combination of descriptors should specify emotional content more or less uniquely. The first level uses everyday emotion labels. The second specifies emotion strength, together with a sign to indicate valence (i.e. whether the feeling is pleasant or unpleasant). The third is derived from an analysis due to Ortony et al. (1988), and categorises the emotion in terms of its object and the individual's appraisal of it ('reproach emotion', indicating disapproval of another person's actions). The fourth is essentially an expansion of the third, which specifies the presumed cognitive antecedents of the emotion.

In similar manner, the coding of speech used different techniques chosen to complement each other. Quantitative acoustic measurement was carried out with the xwaves+ package running on Unix workstations. The initial analysis was of fundamental frequency (F_0), using the F_0

extraction program built into xwaves+, which is generally regarded as one of the best available. To the F_0 trace was added an indication of the articulation rate (syllables per second excluding pauses) using a special program written for the purpose (detailed by Arnfield et al., 1995). This program displays a trace similar in appearance to an F_0 trace, in a separate xwaves+ window which has the time calibrated on the x-axis and syllables per second on the y-axis.

Qualitative phonetic coding was divided into two parts. The ToBI transcription system (Beckman and Ayers, 1994; Roach, 1994) was used to specify a prosodic "skeleton" showing the major points at which pitch-accents and international phrase boundaries occurred. ToBI requires five tiers or windows, one showing the F_0 trace, one giving the time-aligned orthography, one with pitch-accent marking, one showing the Break Indices, and one for the "Miscellaneous" tier. All of these were present on-screen during the transcription and analysis.

The second type of phonetic coding reflected the judgement that descriptions based on fundamental frequency alone could not be adequate. A well-known and tested transcription system for a full range of prosodic and paralinguistic features was presented by Crystal and Quirk (1964) and Crystal (1969), and a menu-driven transcription system based on Crystal's work was incorporated in the annotation conventions for the corpus. Table 2 provides a brief summary of the features coded.

Table 2
Summary of prosodic and paralinguistic features coded in Reading–Leeds database

Feature type	Specific codings
Pause	ToBI break index tier
Pitch range	High/low, wide/narrow
Loudness	Loud/quiet, crescendo/diminuendo
Tempo	Fast/slow, accelerating/decelerating, clipped/drawled
Voice quality	Falsetto, creak, whisper, rough, breathy, ventricular, ingressive, glottal attack
Reflex behaviours	clearing the throat, sniffing, gulping, audible breathing, yawning
Voice qualifications	Laugh, cry, tremulous voice

The system is described by Roach et al. (1998), with further commentary by Roach (2000).

The project as a whole must be judged to have been over-ambitious in its attempt to produce a large-scale fully annotated database of emotional speech. Nevertheless, it establishes a reference point for future databases, in a number of respects.

The coding systems are a systematic implementation of one of the natural approaches to encoding in emotion-related databases, that is, coding that describes events in terms of qualitative labels. The systems are not perfect, and the coding scheme for prosodic and paralinguistic transcription in particular continues to be worked on and developed. However, they are based on sound theoretical principles, and capable of being applied reliably in practice. As such, they provide a natural point of departure for future work with qualitative codings.

One of the major difficulties to emerge from the project relates to the qualitative coding strategy. Because of the number of categories, the number of occurrences in a given category tends to be small. For example, inspection of terminal tone contours shows that most types occur less than ten times even in a gross emotion category (anger, disgust, fear, happiness, sadness, neutrality), let alone in a more precisely specified state (Stibbard, 2000). With numbers of that order, it may be possible to derive useful hypotheses, but there is little prospect of finding statistically robust effects. A working estimate might be that something of the order of ten times as much material might be needed, even without considering finer emotion categories.

The selection of material provides both a major success and arguably the most serious problem to have emerged. The source that was expected to provide vocal signs of emotion, interviews with psychologists, turned out not to. In contrast, broadcast material proved a rich source, but the copyright problem restricted its value dramatically. The great majority of the recordings were made off-air, and it transpires that making them generally available could result in legal action from the broadcasting companies or the speakers involved. Other groups' experience confirms that the problem is not easily resolved (see next section).

Strange as it may seem, finding an acceptable solution is a substantial issue for research on speech and emotion.

4.2. The Belfast database

The Belfast database was developed as part of an EC project called PHYSTA (Principled Hybrid Systems and Their Application; www.image.ntua.gr/physta/; Cowie et al., 2001). The aim of the project was to develop a system capable of recognising emotion from facial and vocal signs. The system was to be based on hybrid computing, i.e. a combination of neural net techniques and traditional symbolic computing. The core function of the data was to train the neural net component.

It was assumed that the system was unlikely to achieve real-world applications unless the training material was naturalistic. Hence, collection was guided by four principles.

(i) The material should be spoken by people who at least appeared to be experiencing genuine emotion.
(ii) The material should be derived from interactions rather than from reading authored texts, even in a genuinely emotional state.
(iii) The primary concern was to represent emotional states of the type that occur in everyday interactions rather than archetypal examples of emotion (such as full-blown fear or anger).
(iv) The material collected was audio–visual as opposed to audio alone. The decision was partly driven by the specific needs of the PHYSTA project, but they converge with general ecological principles in this respect.

The ideal goal was that the system should form the same emotional judgements as people would. Hence objective knowledge about a speaker's true emotional state was not considered critical.

Two main sources were used—television programmes, and studio recordings carried out by the Belfast team. The use of broadcasts followed the approach pioneered by the Reading–Leeds group. Television was the main source of material. A few programme types reliably presented real interactions with a degree of emotional content. The most useful were chat shows and religious programmes, though use was also made of programmes tracing individuals' lives over time and current affairs programmes. Shows that seemed to include an element of 'staging' were excluded. Chat shows provided strongly emotional material, but with a bias towards negative emotions. They typically dealt with an emotive issue, such as divorce, death or drugs, with an audience composed of people who were personally affected by it. Interviews from religious programmes yielded a higher proportion of positive emotions.

Studio recordings were based on one to one interactions between a researcher with fieldwork experience and close colleagues or friends. Standard socio-linguistic fieldwork procedures were used, with care taken over informality of setting, length of recording and prior knowledge (Milroy, 1987). The aim was to cover topics that would elicit a range of emotional responses. The interviewer started with fairly neutral topics (mainly work or families), then 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. The interactions were carried out in a University television studio, and each lasted about 1–2 h.

A selection was made from both types of source, following the principles outlined earlier. The basic aim was to extract material that showed an individual departing from emotional neutrality in a reasonably consistent way for an appreciable period. The emotional states were not required to be particularly extreme, so long as clear signs of emotion were present. Mixed emotional states were included when the signs were strong enough to signal departure from neutrality despite a degree of conflict or instability. Emotional material was only included if it was also possible to identify a passage of relatively neutral material from the same individual. As in the Reading–Leeds project, broadcast material contained far stronger signs of emotion than other sources. Since the studio interviews included discussions between people who had known each other for 15 years, about episodes

such as being assaulted and robbed by a gunman, the finding is not trivial. It underlines the urgent need to clarify the contexts in which people show vocal signs of emotion.

Following exploratory work, material was extracted in units which will be called 'clips'. These are episodes which appear to provide within themselves at least most of the context necessary to understand a local peak in the display of emotion and to show how it develops over time. For example, a typical clip from a chat show might start with an interviewer posing the question which led to an emotional response, and conclude with the interviewer drawing a conclusion or moving onto another topic or person. Clips ranged from 10–60 s in length. Selection was made by the first author.

The database currently contains 298 audiovisual clips from 125 speakers, 31 male, 94 female. For each speaker there is one clip showing him or her in a state that the selector judged relatively neutral, and at least one in a state that she judged relatively emotional. Clips from the first 100 speakers, totalling 86 min of speech, have been labelled psychologically and acoustically (additional rating is under way). The clips are stored as MPEG files, with audio data extracted into .wav files.

The techniques used to describe speech and emotional content overlap with the Reading–Leeds schemes, but develop in a different direction. Broadly speaking, the Belfast project focused on developing quantitative descriptions.

The psychological coding included elements comparable to the Reading–Leeds approach. There were two levels of description based on everyday verbal categories, one using a 'basic emotion vocabulary' of 16 terms (shown in Table 3), and the other allowing choices (up to 2) from a larger vocabulary of 40 emotion terms. The vocabularies were chosen on the basis of preliminary studies reported by Cowie et al. (1999a). As in the Reading–Leeds database, each term was associated with a rating of the intensity of the state. Category labels were attached to the clip as a whole.

The coding strategies diverged mainly because the Belfast team concluded that uncertainty and gradation were intrinsic features of the data, and looked for ways of reflecting them. To reflect uncertainty about the emotion displayed in a particular clip, the database included ratings from individual subjects rather than trying to establish a consensus. To reflect gradation, the project exploited another of the options offered by contem-

Table 3
Main emotion categories used in the Belfast natural database and their frequency of use (as first choice)

Label	Frequency of use	Frequency of full agreement	Broad group	Numerical coding
Neutral	273	31	Not strongly emotional	7
Angry	114	19	Strong negative	2
Sad	94	12	Strong negative	1
Pleased	44	3	Unoriented positive	15
Happy	37	0	Unoriented positive	16
Amused	26	6	Unoriented positive	17
Worried	19	0	Strong negative	4
Disappointed	17	0	Not strongly emotional	6
Excited	17	0	Oriented positive	12
Afraid	13	0	Strong negative	3
Confident	13	0	Not strongly emotional	8
Interested	12	0	Not strongly emotional	9
Affectionate	10	0	Oriented positive	14
Content	4	0	Not strongly emotional	10
Loving	3	0	Oriented positive	13
Bored	3	0	Unassigned	5
Relaxed	3	0	Unassigned	11

porary psychology, the dimensional approach associated with Osgood. Two dimensions, activation and evaluation, are known to capture a relatively large proportion of emotional variation. A computer program called Feeltrace was written to let users describe perceived emotional content in terms of those dimensions. The space was represented by a circle on a computer screen, alongside a window where a clip was presented. The vertical axis represented activation, the horizontal axis evaluation. Raters used a mouse to move a cursor inside the circle, adjusting its position continuously to reflect the impression of emotion that they derived from the clip. Cowie and Cornelius (2003) give more information about the system.

The database is not representative in any strict sense, but it provides some guidance on the kinds of emotion that tend to occur in natural speech. The summary below is based on ratings of the first 100 speakers by three trained raters. The second column of Table 3 shows how often the raters used each emotion category as their first choice, and the third shows how many clips were assigned the same label by all three raters. Generally, the distribution indicates that genuine interactions present a considerable number of emotional states, many of them relatively subtle. There are not many examples of states that are positive but inactive, but that is mainly because a high proportion of these clips were drawn from TV programmes which tended to be highly charged. The variety and subtlety of the emotions is linked to the low rates of complete agreement, underlining the point that uncertainty is a major issue in naturalistic data.

The task of measuring inter-rater agreement highlights some of the difficulties associated with category labels as descriptors. Table 4 illustrates several strategies. Simple calculation of agreement among categories gives the kappa values shown in the second column. The results clearly underestimate real consensus, because they ignore the fact that labels may be similar even if they are not identical. The simplest response is to aggregate categories that behave similarly. Inspection suggested that responses fell naturally into four broad groups, which are indicated in the fourth column of Table 3. The column headed 'grouped categorical' in Table 4 shows the kappa coefficients derived by considering those groups. They are higher than the second column, as one might expect, but still moderate. An alternative convenient response is to replace category labels with numerical equivalents, chosen so that labels are assigned similar numbers if they are similar in meaning and tend to be applied to the same clips. The last column of Table 3 shows the best numbering of that kind that the Belfast team could construct. The fourth column of Table 4 shows correlations based on it. They confirm that there is more consensus than the simpler techniques suggests; but because the approach is fundamentally ad hoc, it is difficult to draw stronger conclusions.

The Feeltrace measures of evaluation and activation avoid comparable problems. The 'Feeltrace co-ordinates' columns in Table 4 show that raters agreed quite closely on both dimensions, particularly evaluation. Agreement on categorisation can be measured using a related technique. Each category can be represented by two numbers, i.e. the co-ordinates of the mean Feeltrace cursor position associated with trials where that category was selected. The procedure yields arrays which can be correlated, with the results shown in the last two columns of Table 4. It is reassuring that the pattern of correlations is similar to the pattern for

Table 4
Measures of agreement among three Belfast database raters on categorical and numerical descriptors of emotion

Raters being compared	Simple categorical (kappa)	Grouped categorical (kappa)	Category numbers (rho)	Feeltrace co-ordinates: evaluation	Feeltrace co-ordinates: activation	Category co-ordinates: evaluation	Category co-ordinates: activation
R3 vs. R1	0.38	0.46	0.71	0.84	0.69	0.71	0.60
R3 vs. R2	0.50	0.60	0.67	0.82	0.56	0.68	0.54
R2 vs. R1	0.29	0.42	0.66	0.85	0.58	0.68	0.45

Feeltrace ratings as such: it suggests that differences between raters related mostly to raters their' underlying judgements about emotion, rather than to one or both of the response modes. Since categorical description is a more familiar medium, one might expect it to be less subject to individual differences: but comparing the correlations based on Feeltrace as such with those based on categories, in whatever form, it seems that if anything, the opposite is true. It is also of interest that agreement was closer on the evaluation dimension even when the co-ordinates were recovered from categorical responses. It suggests that the dimensions capture factors that affect subjects' judgements even when they are making categorical responses.

Description of speech is based on a system called Automatic Statistical Summary of Elementary Speech Structures (ASSESS, Cowie et al., 1995). The philosophy behind ASSESS is to extract a comprehensive set of summary statistics from the speech signal, so that it possible to explore a wide variety of potentially interesting relationships. For each clip, ASSESS constructs a stylised description specifying straight-line approximations to intensity and F_0 contours, pause boundaries, high frequency bursts, and basic spectral properties. Pause boundaries are used to divide the passage into 'tunes' (episodes of speech between substantial pauses). Statistics are then derived for each tune, and for the passage as a whole, to describe its components at various levels—'slices' (25.6 ms samples), rises and falls in intensity and F_0, pauses, high frequency bursts, and trends across the whole unit. The result is a battery of 352 measures per unit, covering properties related to its spectrum, intensity profile, and F_0 profile.

For large corpora, manual coding is prohibitively slow and expensive, and so it is critical that ASSESS is automatic—or rather semi-automatic, because during analysis it displays spectra and profiles of intensity and F_0 in a way that allows users to adjust global settings if automatic decisions about issues such as pause boundaries are not credible. Related to that, it is critical that ASSESS is robust, because it is difficult to guarantee that naturalistic recordings will be acoustically impeccable. Developing analysis systems with those properties is important for progress in understanding emotion as it naturally occurs. Note that some limitations are intractable, though—for instance, gain control tends to be varied during broadcasting, so that intensity measures have to be treated with caution (particularly in comparisons between clips).

In contrast to the Leeds–Reading approach, the Belfast database embodies ways of implementing quantitative encoding in emotion-related databases. The quantitative format lends itself to exploring continuous relationships of the kinds that several investigators have described, for instance between parameters of pitch and activation (Pereira, 2000b). Surface comparison suggests that such an approach reveals statistically significant relationships between speech and emotion descriptors more readily than qualitative coding. However, the only way to establish whether one approach has a real advantage over the other is to compare them on a single body of data, larger than either of the databases described so far. That remains to be done.

It was intended that the database would also include descriptions of emotion-relevant parameters for faces, describing the positions of key points on the face in each frame. In fact, automatic identification of the relevant points has proved difficult, and information is only available for a limited number of frames. The role of different information sources has been probed psychologically, though, by collecting Feeltrace codings for visual, audio and audio–visual modes of presentation. Preliminary results indicate that some kinds of emotion judgement are relatively independent of visual input, but others are not. Visual input seems to play a particular role in conveying that emotion is strongly positive.

PHYSTA was a conspicuously ambitious project, on a considerable scale. The main database contains information on over 2000 tunes. Exploratory studies suggest that for statistical purposes, that is too small by an order of magnitude. The Belfast structured database was begun as a way of amplifying the data set. Actors were given transcripts of selected passages, and asked to reproduce them with the appropriate emotional

colouring. If they are successful, the result includes much more emotion-relevant context than traditional methods provide. An adaptation of Feeltrace (called Validtrace) allows raters to indicate how convincingly the reading appears to approximate genuine emotion. The structured database is also relevant to access issues. Considerable fees for acquiring broadcast material do not extend to distributing it, and so access to the naturalistic database remains restricted. The structured database is not subject to the same restrictions, and wider distribution has already begun.

4.3. CREST: the expressive speech database

The expressive speech processing (ESP) project started in Spring 2000 and will run for five years. It is part of the JST/CREST (Core Research for Evolutional Science and Technology) initiative, funded by the Japanese Science and Technology Agency. Its research goals are (a) collecting a database of spontaneous, expressive speech that meets the requirements of speech technology (particularly concatenative synthesis); (b) statistical modelling and parameterisation of paralinguistic speech data; (c) developing mappings between the acoustic characteristics of speaking-style and speaker-intention or speaker-state; and (d) the implementation of prototypes and testing of the software algorithms developed in (b) and (c) in real-world applications.

The focus on applications means that the states of most interest to ESP are those that are likely to occur during interactions between people and information-providing or service-providing devices. These certainly include emotional states (such as amusement) and emotion-related attitudes, such as doubt, annoyance, surprise. It is not clear how relevant the classical basic emotions are. Since the 'expressive speech' associated with these states may be specific to a language community, material is being collected in three languages (Japanese (60%), Chinese (20%) and English (20%)). The target is to collect and annotate a total of 1000 h of speech data over 5 years. To date, 250 h of natural-speech data have been collected and about 10% transcribed. The data for the corpus have primarily been collected from non-professional, volunteer subjects in various everyday conversational situations, but samples for analysis also include emotional speech recorded from television broadcasts, DVD and video.

4.3.1. Levels of data

The key design problem is to balance between the demands of automatic speech processing on the one hand, and paralinguistic investigation on the other. That entails developing methods for obtaining speech samples which are clear enough to be processed by automatic techniques and yet which are not stilted, acted, prompted, or otherwise less than natural. The problem has been addressed by collecting several levels of data. The speech samples range from highly structured studio readings of phonemically and prosodically balanced sentences for use in waveform-concatenation speech synthesis, to completely unstructured recordings of casual conversational speech. Similarly, a range of microphone arrangements and recording devices has been tested in order to balance recording quality with freedom of expression and naturalness in each case.

For truly natural speech, a "Pirelli-Calendar" approach is being taken (named for the fact that photographers once took 1000 rolls of film on location in order to produce a calendar containing only 12 photographs). Volunteers are fitted with long-term recorders, and samples of their day-to-day, throughout the day vocal interactions are recorded. Perhaps only by over-collecting speech data in this way can adequate and representative coverage be guaranteed. However, the task of annotating the resulting data is enormous, and automatic transcription using speech recognition is very difficult.

An effective way of collecting clear but naturally spontaneous conversational speech is to record one side of a telephone conversation using a studio-quality microphone in a sound-treated room. The speaker soon becomes immersed in the telephone conversation to the extent that the surroundings are forgotten and the speech is very natural. Interactions with different interlocutors result in very different speaking styles and attitudes. By recording both sides of the conversation separately, high-quality recordings can be

obtained without cross-talk or overlapping speech causing problems for the analysis. At the other extreme, samples of highly dysfluent speech from autistic and physically handicapped subjects are also being collected and annotated with prosodic and phonation-type labels for an analysis of their expressive and speaking-style characteristics.

While some of the speech is being recorded in audio–visual conditions, the intrusion of the camera into the situation has a detrimental effect on spontaneity, and although a visual check would be useful to confirm indications of a given emotion, the cost in terms of reliability of the speech produced is high. Volunteers who may be prepared to wear a sound recorder (for the sake of science) even while having a tantrum (for example) are much more wary of being filmed doing it.

The issue of recording quality is also handled by including several different levels. While studio-recorded speech is optimal for acoustic analysis, speakers tend to act less naturally when confined in a studio. It was decided to accept four levels of recording quality, which are inversely correlated with naturalness of expression in the speech.

The highest quality level is that of studio-recorded speech with, optionally, Laryngograph signals for separate recording of the fundamental frequency at the glottis. This quality of speech will be used primarily for speech-synthesis databases, which have to include phonetically and prosodically balanced speech samples for unit concatenation.

The second level of quality is from recordings on DAT tape (i.e., 48 kHz 16 bit stereo), using studio-quality microphones with different speakers on different tracks, but in a quiet and relaxed conversational setting outside the studio or laboratory. The portable DAT recorders are considerably heavier and more bulky than the newer Minidisc ('MD') recorders, but they have an advantage in recording quality.

The size of the MD recorders means that they can be worn for extended periods while the subjects go about the everyday tasks of life. Small head-mounted studio-quality microphones allow accurate reproduction of the speaker's voice, while at the same time preserving the confidentiality of third parties, whose voices may be inadvertently included in the recordings, by the high drop-off with distance from the mouth. MDs (on monaural setting) enable 160 min of continuous high-quality recording (at 44.1 kHz) on a single 2.5-inch disc and are light enough to be worn comfortably in a shirt or skirt pocket. The limitation is that the recording is filtered by Adaptive Transform Acoustic Coding (www.minidisc.org/aes_atrac.htm) using a masking algorithm to remove (allegedly) imperceptible information from three separate filter bands across the spectrum for data compression of the speech signal. It has been confirmed that the quality of the MD-recorded speech is superior to that of MPEG encoding, and that it is suitable for signal processing and prosodic analysis (Campbell, 2002).

The lowest level of speech quality (from the point-of-view of acoustic processing) is that obtained by use of a single far-field microphone placed between subjects talking in an informal environment; e.g., on a table during family meals, or on a desk in conversations between colleagues. These conversations provide useful material for perceptual or subjective analyses, and can provide exceptional samples of genuinely emotional speech, but it can be very difficult to extract reliable pitch contours or to perform automatic phonemic segmentation of the data because of the high degree of speech overlap and background noise.

4.3.2. Tools for preliminary analysis

A key contrast between the ESP Project and those considered earlier is the effort devoted to ensuring that the preliminary analysis of the data can be done by automatic methods. There are two main reasons for that emphasis. First, the project is expected to generate much more data than the earlier projects. Second, the synthesis application highlights questions about the role of linguistic structures, from phonemes to syntactically significant contours. Answering those questions depends access to information about linguistic structures. Extracting that information depends on tools and software that are described in this section.

The acoustic analysis of the data uses a three-pronged approach to extract phonemic, prosodic

and phonation-style information for annotation of the speech. Phonemic labelling (or segmental alignment) is a well-established technology, with free public-domain phonemic alignment software available (e.g., HTK, www.mbrola.org, Festival, etc.), but it still requires significant additional research in order to improve the detection of mislabelled speech segments. Prosodic labelling is currently being actively researched in several laboratories throughout the world, and will soon reach a usable level of maturity (see Sprosig, ongoing www.isca-speech.org/sprosig). Phonation style analysis has in the past been performed analytically, with manual intervention, and has not yet been fully automated; so this element of the analysis requires most effort for the development of data and tools. By combining these three levels of annotation, in conjunction with a semantic and syntactic analysis of the corpus text, data is being produced for a modelling of the mapping between acoustic events and different levels of intended "meaning" in the speech.

The Entropic software used in the Leeds–Reading project is no longer on the market, but satisfactory alternatives have been identified. ESP is currently using the Tcl/Tk extension "Snack" speech-processing libraries in conjunction with the "Wavesurfer" software developed and released in the public domain by the KTH laboratory in Stockholm (www.speech.kth.se/snack). For speech segmentation, the "Julius" software (released in the public domain by the Japanese IPA Project), and the "HTK-3.0" Hidden–Markov modelling software toolkit (released by Cambridge University) are being tested. Software for the analysis of voice quality is currently being implemented in conjunction with the above tools (Mokhtari and Campbell, 2002).

For the analysis and annotation of speaking style, the ESP team tested its own implementation of Feeltrace (see Section 4.2). Training is critical for the use of the original technique (Cowie et al., 2000a,b), but preliminary tests indicate that non-expert users may find it easier to track speaking-style or para-linguistic information in similar dimensions but separately. Hence, a one-dimensional three-track version of the Feeltrace software is being tested.

For pitch-contour labelling, MOMEL/IntSint software package from the University of Aix-en-Provence is being tested. This may eliminate the need for manual ToBI (Tones and Break-Index) labelling by the fitting of a quadratic spline to the fundamental-frequency signal and abstracting from the derived target points to produce a series of abstract specifiers of the underlying contour shape. It is not yet confirmed that the symbols thus derived can be satisfactorily predicted from text, but this investigation is now under way.

Initial experiments (performed under separate ATR-ICP, ATR-NAIST and ATR-Keio collaborations) have shown that both the recognition of attitude by machine processing and the control of emotion in synthetic speech can to a certain extent be achieved by signal analysis implemented as automatic labelling (Campbell and Marumoto, 2000; Iida et al., 2000). Thus the efficient labelling of speaking-style and voice-quality characteristics on a speech signal will form a key aspect of research. Particular interest is being paid to pressed-voice and breathy voice, both of which appear to signal information-bearing and meaning-related speaker-attitude differences independently of prosody.

4.3.3. Application

The final test of the corpus and of the software created for its development will be in real-world applications; notably automatic recognition of speaking style as an add-on to present speech recognition technology; and techniques for the more flexible expression of speaker attitudes in information-providing devices and services.

The tools developed by ESP for the automatic design and recording of prosodically balanced speech databases are designed to be compatible with waveform-concatenation speech synthesis software, so that recordings can be converted into CHATR-specific source-unit speaker databases (Campbell and Black, 1996). The methods are suitable in principle for the automatic creation of a speaker-database or of a speaking-style database, as long as a close transcription of the speech data is available. The main bottleneck that remains is that the transcription of unprompted speech has

not yet been automated, and still requires extensive (and expensive) manual labour.

5. Overview and way forward

Much of what we have said is common sense. Nevertheless, the issues need to be articulated, not least because they may not become obvious until effort has been invested in work which fails to take account of them.

The first and most fundamental conclusion of the review is that the development of emotion-rich databases needs to be recognised as a task in itself. The inherent variability in the area means that to support sound conclusions, databases need to be large. For databases that aim to span the range of everyday emotions, a low estimate is about 10 times the size of the Reading–Leeds database, i.e. around 12 h of emotional material. The ESP project aims much higher. Systematic comparison across cultures requires databases of a scale that has only begun to be contemplated.

Naturalness is at the centre of the problem. Real-world applications depend on coming to terms with the ways people express emotion, and they are complex and variable. It should be clear by this stage that naïve simulations are too far from that reality for techniques based on them to transfer (particularly in the context of recognition). That does not mean that simulation should be jettisoned. It means that simulation needs to become sophisticated. The literature now provides information about various techniques worth considering. It also warns that some techniques provide surprisingly little vocal expression of emotion, notably the interviews about emotional topics of the kind considered by both the Leeds–Reading and the Belfast teams.

A useful response to natural complexity is to retreat from the ideal of covering the whole domain of emotion, and to focus on a specific subregion. That approach has most often been taken in applied contexts, focusing on an emotion-related state rather than emotion per se, but it would make sense to extend it.

It is a matter of judgement whether speech is considered alone or with other modalities. Two substantial arguments favour a multi-modal approach. From a purely theoretical point of view, it seems clear that vocal signs of emotion form part of a multi-modal signalling system. Telephone conversation is an exception that proves the rule, because people adjust to the loss of other modes by adopting a distinctive 'phone voice' (Douglas-Cowie and Cowie, 1998). It makes sense to study emotion in telephone conversations as a purely vocal phenomenon, but perhaps not to treat speech extracted from face to face encounters in the same way. From a practical point of view, evidence is accumulating that high rates of emotion recognition are unlikely to be achieved from speech alone in applied settings, and it makes sense to consider it as one of many inputs.

Considerable progress has been made with regard to the description of emotion. The community is no longer wedded to description in terms of a small number of primaries, and there seems a degree of consensus that a replacement should include both a larger number of categories and dimensional description. Feeltrace provides a way of generating dimensional representations that is well suited to describing speech. Convergence on an agreed set of categories is clearly needed.

For the description of speech, there also appears to be consensus on the need to consider two levels. For linguistic descriptions, standard labelling systems exist and are increasingly becoming automated. The Reading–Leeds system provides a rational extension to deal with emotion-specific issues. Continuous acoustic measures are clearly relevant to recognition, and should ideally be incorporated in a database. No clear set of preferred candidates has yet emerged, and again, convergence is clearly needed.

Several very practical issues are critical. One is the release and sharing of resources. A good deal can be done via the web. For instance, Feeltrace has now been validated and will be released via the internet in 2001. However, databases of emotional material are surrounded by ethical and copyright issues. Neither the Reading–Leeds nor the Belfast database can be released because the material is held under a limited agreement with the broadcasters. Various other attractive sources—phone

calls and interviews with medical or psychological clients—also present legal problems.

It is natural to wonder whether the effort of developing adequate databases for emotion is worthwhile. In reality, it seems inevitable that the development will take place. Voice is becoming a key medium for human–computer interaction, and will be used increasingly to buy products or to retrieve information. Machines will also speak on behalf of people, as communication aids. Because the expression of emotion and feeling is such a characteristic feature of human speech, people will inevitably expect machines that use speech to register the affective content as well as the textual content of an utterance, and to follow basic conventions in their own use of affective colouring.

The only way to ensure that those interactions are satisfactory is to collect data that support sophisticated training of speech systems. It is apparent that research has begun to work towards developing genuinely large-scale databases of emotion. Arrangements for sharing data and tools relevant to database construction are also beginning to emerge, and a website has been established to facilitate the process (www.interactivesys.com/emotions). These developments can be expected to continue over the next decade as a solid body of emotional speech is accumulated.

Acknowledgements

We acknowledge the contributions made to development of the Belfast database, and the tools associated with it, by Martin Sawey, Susie Savvidou, Marc Schröder and Edelle McMahon.

References

Abelin, A., Allwood, J., 2000. Cross linguistic interpretation of emotional prosody. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 110–113.

Alter, K., Rank, E., Kotz, S., Toepel, U., Besson, M., Schirmer, A., Friederici, A.D., 2000. Accentuation and emotions two different systems? In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 138–142.

Amir, N., Ron, S., Laor, N., 2000. Analysis of an emotional speech corpus in Hebrew based on objective criteria. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 29–33.

Arnfield, S., Roach, P., Setter, J., Greasley, P., Horton, D., 1995. Emotional stress and speech tempo variation. In: Proceedings of the ESCA-NATO Tutorial and Research Workshop on Speech Under Stress, Lisbon, pp. 13–15.

Bachorowski, J., Owren, M., 1995. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. Psychol. Sci. 6 (4), 219–224.

Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Social Psychol. 70 (3), 614–636.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Noeth, E., 2000. Desparately seeking emotions or: actors, wizards and human beings. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 195–200.

Beckman, M., Ayers, G., 1994. Guidelines for ToBI Labeling, Version 2.0. Ohio State University, Department of Linguistics, Columbus.

Bird, S., Harrington, J. (Eds.), 2001. Special Issue on Speech Annotation and Corpus Tools Speech Communication 33.

Bonner, M.R., 1943. Changes in the speech pattern under emotional tension. Amer. J. Psychol. 56, 262–273.

Brend, R.M., 1975. Male–female intonation patterns in American English. In: Thorne, B., Henley, N. (Eds.), Language and Sex: Difference and Dominance. Newbury House, Rowley, Mass.

Campbell, W.N., 2002. Recording techniques for capturing natural everyday speech. In: Proceedings of the LREC-2002, Las Palmas.

Campbell, W.N., Black, A.W., 1996. CHATR – a multi-lingual speech re-sequencing synthesis system. Technical Report of IEICE SP96-7, pp. 45–52.

Campbell, W.N., Marumoto, T., 2000. Automatic labelling of voice quality in speech databases for synthesis. In: Proceedings of the ICSLP-2000, Beijing, Vol. IV, pp. 468–471.

Cauldwell, R., 2000. Where did the anger go? The role of context in interpreting emotion in speech. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 127–131.

Chung, S., 2000. Expression and Perception of Emotion extracted from the Spontaneous Speech in Korean and English, ILPGA, Sorbonne Nouvelle University, Paris, France. Available from <http://people.ne.mediaone.net/sangikoh/soojinchung.htm>.

Cowie, R., Cornelius, R., 2003. Describing the emotional states that are expressed in speech. Speech Communication 40 (1–2), this issue. PII: S0167-6393(02)00071-7.

Cowie, R., Douglas-Cowie, E., 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: Bunnell, T., Idsardi, W. (Eds.), Proceedings of the Fourth ICSLP, 3–6 October 1996, Philadelphia, pp. 1989–1992.

Cowie, R., Douglas-Cowie, E., Sawey, M., 1995. A new speech analysis system: ASSESS (Automatic Statistical Summary

of Elementary Speech Structures). In: Proceedings of the ICPhS 1995, Stockholm, Vol. 3, pp. 278–281.

Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A., Fellenz, W., 1999a. What a neural net needs to know about emotion words. In: Mastorakis, N. (Ed.), Computational Intelligence and Applications. World Scientific and Engineering Society Press, pp. 109–114.

Cowie, R., Douglas-Cowie, E., Wichmann, A., Hartley, P., Smith, C., 1999b. The prosodic correlates of expressive reading. In: Proceedings of the 14th ICPhS, San Francisco, 1–7 August, 1999, Berkeley, University of California, pp. 2327–2330.

Cowie R., Douglas-Cowie, E., Schroeder, M., 2000a. Proceedings of the ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework, Newcastle, N. Ireland, 5–7 September 2000. Textflow, Belfast. Available from www.qub.ac.uk/en/isca/index.htm.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schroeder, M., 2000b. 'FEELTRACE': an instrument for recording perceived emotion in real time. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human–computer interaction. IEEE Signal Process. Mag. 18 (1), 32–80.

Crystal, D., 1969. Prosodic Systems and Intonation in English. Cambridge University Press, Cambridge.

Crystal, D., Quirk, R., 1964. Systems of Prosodic and Paralinguistic Features in English. Mouton, The Hague.

de Gelder, B., Vroomen, J., 2000a. The perception of emotions by ear and eye. Cognition and Emotion 14, 289–311.

de Gelder, B., Vroomen, J., 2000b. Bimodal emotion perception: integration across separate modalities, cross-modal perceptual grouping, or perception of multimodal events? Cognition and Emotion 14, 321–324.

Douglas-Cowie, E., Cowie, R., 1998. International settings as markers of discourse units in telephone conversations. In: Special Issue: Prosody and ConversationLanguage and Speech 41 (3–4), 351–374.

Douglas-Cowie, E., Cowie, R., Schroeder, M., 2000. A new emotion database: considerations, sources and scope. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 39–44.

Ekman, P., 1994. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. Psychol. Bull. 115, 268–287.

Ekman, P., 1999. Basic emotions. In: Dalgleish, T., Power, M. (Eds.), Handbook of Cognition and Emotion. John Wiley, New York, pp. 45–60.

Ekman, P., Friesen, W., 1969. The repertoire of non verbal behavior: categories, origins, usage and coding. Semiotica 1, 49–98.

Ekman, P., Friesen, W., 1978. The Facial Action Coding System. Consulting Psychologists' Press, San Francisco, CA.

Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P., 1997. Design, recording and verification of a Danish Emotional Speech Database. In: Proceedings of the Eurospeech '97, Rhodes, Greece.

Fernandez, R., Picard, R., 2000. Modeling drivers speech under stress. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 219–224.

France, D., Shiavi, R., Silverman, S., Silverman, M., Wilkes, D., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Engng. 47 (7), 829–837.

Fridlund, A.J., 1994. Human Facial Expression. An Evolutionary View. Academic Press, San Diego.

Frolov, M., Milovanova, G., Lazarev, N., Mekhedova, A., 1999. Speech as an indicator of the mental status of operators and depressed patients. Human Physiol. 25 (1), 42–47.

Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., Horton, D., 1995. Representation of prosodic and emotional features in a spoken language database. In: Proceedings of the XIIIth ICPhS, Stockholm, Vol. 1, pp. 242–245.

Greasley, P., Sherrard, C., Waterman, M., Setter, J., Roach, P., Arnfield, S., Horton, D., 1996. The perception of emotion in speech. Abs. Int. J. Psychol. 31 (3/4), 406.

Greasley, P., Sherrard, C., Waterman, M., 2000. Emotion in language and speech: Methodological issues in naturalistic approaches. Language and Speech 43, 355–375.

Groningen corpus S0020 ELRA, 1996. Available from www.icp.inpg.fr/ELRA.

Hansen, J., Bou-Ghazale, S., 1997. Getting started with SUSAS: A speech under simulated and actual stress database. In: Proceedings of the Eurospeech 1997, Rhodes, Greece, Vol. 5, pp. 2387–2390.

Harré, R. (Ed.), 1986. The Social Construction of Emotions. Blackwell, Oxford, UK.

Hargreaves, W., Starkweather, J., Blacker, K., 1965. Voice quality in depression. J. Abnormal Psychol. 70, 218–220.

Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M., 2000. A speech synthesis system with emotion for assisting communication. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 167–172.

Iriondo, I., Guaus, R., Rodriguez, A., Lazaro, P., Montoya, N., Blanco, J., Beradas, D., Oliver, J., Tena, D., Longhi, L., 2000. Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 161–166.

Johannes, B., Salnitski, V., Gunga, H.-C., Kirsch, K., Voice stress monitoring in space – possibilities and limits. Aviation, Space and Environmental Medicine 71, 9, section II, A58–A65.

Johns-Lewis, C., 1986. Prosodic differentiation of discourse modes. In: Johns-Lewis, C. (Ed.), Intonation in Discourse. College-Hill Press, San Diego, pp. 199–220.

Karlsson, I., Banziger, T., Dankovicova, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K., 1998. Within speaker variation due to induced stress. In: Branderud, P., Traunmuller, H. (Eds.), Proceedings of Fonetik – 98 The Swedish Phonetics Conference, Stockholm University, 27–29 May 1998, pp. 150–153.

Kienast, M., Sendlmeier, W.F., 2000. Acoustical analysis of spectral and temporal changes in emotional speech. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 92–97.

Ladd, D.R., Scherer, K., Silverman, K., 1986. An integrated approach to studying intonation and attitude. In: Johns-Lewis, C. (Ed.), Intonation in Discourse. College-Hill Press, San Diego, pp. 125–138.

Marcus, H., Kitayama, S., 2001. The cultural construction of self and emotion: implications for social behavior. In: Parrot, W.G. (Ed.), Emotions in Social Psychology: Essential Readings. Taylor and Francis: Psychology Press, UK and USA.

Massaro, D., Cohen, M.M., 2000. Fuzzy logical model of bimodal emotion perception: Comment on 'The perception of emotions by ear and eye' by de Gelder and Vroomen. Cognition and Emotion 14, 313–320.

McEnery, T., Wilson, A., 1996. Corpus Linguistics. Edinburgh University Press, Edinburgh.

McGilloway, S., 1997. Negative symptoms and speech parameters in schizophrenia. Ph.D. thesis, Queen's University, Belfast.

Milroy, L., 1987. Observing and Analysing Natural Language. Blackwell, Oxford.

Mokhtari, P., Campbell, W.N., 2002. An evaluation of the Glottal AQ parameter automatically measured in expressive speech. In: Proceedings of LREC-2002.

Mozziconacci, S., 1998. Speech Variability and Emotion: Production and Perception. Technical University of Eindhoven, Proefschrift.

MPEG4 SNHC: Face and Body Definition and Animation Parameters., ISO/IEC JTCI/SC29/WG11 MPEG96/N1365, 1996.

Ortony, A., Clore, G., Collins, A., 1988. The Cognitive Structure of Emotions. Cambridge University Press, Cambridge.

Osgood, C., Suci, G., Tannenbaum, P., 1957. The Measurement of Meaning. University of Illinois Press, Urbana.

Paeschke, A., Sendlmeier, W.F., 2000. Prosodic characteristics of emotional speech: measurements of fundamental frequency movements. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 75–80.

Pereira, C., 2000a. Perception and expression of emotion in speech. Unpublished Ph.D. Thesis, Macquarie University, Australia.

Pereira, C., 2000b. Dimensions of emotional meaning in speech. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 25–28.

Polzin, T.S., Waibel, A., 2000. Emotion-sensitive human–computer interfaces. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 201–206.

Roach, P., 1994. Conversion between prosodic transcription systems: "Standard British" and 'ToBI'. Speech Communication 15, 91–99.

Roach, P., 2000. Techniques for the phonetic description of emotional speech. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 53–59.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., Setter, J., 1998. Transcription of prosodic and paralinguistic features of emotional speech. J. Int. Phonetic Assoc. 28, 83–94.

Roessler, R., Lester, J., 1979. Vocal pattern in anxiety. In: Fann, W., Pokorny, A., Koracau, I., Williams, R. (Eds.), Phenomenology and Treatment of Anxiety. Spectrum, New York.

Scherer, K., 2000. Emotion effects on voice and speech: paradigms and approaches to evaluation. In: ISCA Workshop on Speech and Emotion, Newcastle, N. Ireland, 5–7 September 2000 (oral presentation). Available from www.qub.ac.uk/en/isca/index.htm.

Scherer, K., Ceschi, G., 1997. Lost luggage emotion: a field study of emotion – antecedent appraisal. Motivation and Emotion 21, 211–235.

Scherer, K., Ceschi, G., 2000. Studying affective communication in the airport: The case of lost baggage claims. Pers. Social Psychol. Bull. 26 (93), 327–339.

Scherer, K., Feldstein, S., Bond, R., Rosenthal, R., 1985. Vocal cues to deception: A comparative channel approach. J. Psycholinguist. Res. 14, 409–425.

Sherrard, C., Greasley, P., 1996. Lexical valency in emotional speech. Int. J. Psychol. 31 (3–4), 4762.

Slaney, M., McRoberts, G., 1998. Baby ears. In: Proceedings of ICASSP, Seattle, WA, USA.

Sprosig: the special-interest-group for speech prosody, ongoing. Available from www.isca-speech.org/sprosig.

Stassen, H., Bomben, G., Gunther, E., 1991. Speech characteristics in depression. Psychopathology 24, 88–105.

Stemmler, G., 1992. The vagueness of specificity: Models of peripheral physiological emotion specificity in emotion theories and their experimental discriminability. J. Psychophysiol. 6, 17–28.

Stibbard, R.M., 2000. Automated extraction of ToBI annotation data from the Reading/Leeds Emotional Speech Corpus. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 60–65.

Summerfield, A.Q., 1983. Audio–visual speech perception, lipreading and artificial stimulation. In: Lutman, M., Haggard, M. (Eds.), Hearing Science and Hearing Disorders. Academic Press, London, pp. 132–182.

ten Bosch, L., 2000. Emotions: What is possible in the ASR framework? In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, 5–7 September 2000, Belfast, Textflow, pp. 189–194.

Tolkmitt, F., Scherer, K., 1986. Effect of experimentally induced stress on vocal parameters. J. Exp. Psychol.: Human Perception Perform. 12 (3), 302–333.

Trainor, L., Austin, C., Desjardins, 2000. Is infant-directed speech prosody a result of the vocal expression of emotion? Psychol. Sci. 11 (3), 188–195.

Trudgill, P., 1983. Sociolinguistics: An Introduction to Language and Society. Penguin, London (revised edition).

van Bezooijen, R., 1984. Characteristics and Recognizability of Vocal Expressions of Emotion. Foris Publications, Dordrecht.

Waterman, M., Greasley, P., 1996. Development of a qualitative instrument for coding cognitive antecedents of emotional responses. Int. J. Psychol. 31 (3–4), 4761.

Williams, C., Stevens, K., 1972. Emotions and speech: Some acoustical correlates. J. Acoust. Soc. Amer. 52 (4, part 2), 1238–1250.