Principles and history

Roddy Cowie, Ellen Douglas-Cowie, Ian Sneddon, Anton Batliner, Catherine Pelachaud

Angaben zur Veröffentlichung / Publication details:

Cowie, Roddy, Ellen Douglas-Cowie, Ian Sneddon, Anton Batliner, and Catherine Pelachaud. 2011. "Principles and history." In *Emotion oriented systems*, edited by Roddy Cowie, Catherine Pelachaud, and Paolo Petta, 167–96. Berlin [u.a.]: Springer. https://doi.org/10.1007/978-3-642-15184-2_11.

Principles and History

Roddy Cowie, Ellen Douglas-Cowie, Ian Sneddon, Anton Batliner, and Catherine Pelachaud

Abstract Developing databases for emotion-oriented computing raises specific and complex issues at multiple levels, from the practicalities of recording to conceptual issues in psychology. Whether it is developing databases or using them, research in emotion-oriented computing needs to think about these issues rather than reflexively importing habits derived from other fields. Contemporary research identifies a number of principles that are relevant to making appropriate choices. They can be grouped under three broad headings – function; structure and scope; and relationship to psychological theory. These principles were not obvious when research in the area began. They have emerged gradually over a decade of relatively sustained work, and it is reviewed. Databases that have played a significant role in the process are listed, and selected case studies are examined in more depth. Lessons are drawn for future work at two levels, first at the level of an abstract overview and then at the level of practical issues that need to be addressed.

1 Introduction

Databases are central to the development of systems that use human-like channels of communication. As a result, they have been part of emotion-oriented computing since research in the area began. They have also been recognised as an area where there were major problems to be dealt with. Most obviously, there was a shortage of appropriate databases. Less obviously, there were unresolved questions about the kind of database that would count as appropriate. Broadly speaking, early research tended to transfer working habits from cognate fields, such as the recognition of words in speech or recognition of an individual from different photographs of his/her face. However, it gradually became apparent that some important working habits might not transfer to the new area. The ISCA Workshop on Speech

R. Cowie
Department of Psychology, Queen's University Belfast, Belfast, Northern Ireland, UK
e-mail: roddy.cowie@qub.ac.uk

and Emotion in 2000 (Douglas-Cowie et al., 2000) is a useful marker. There ten Bosch (2000) argued in principle that word recognition paradigms might not transfer; Batliner et al. (2003) argued that standard ways of constructing databases were not practically appropriate to the field; and groups from England, Ireland and Japan (Douglas-Cowie et al., 2003) described work on emotion databases collected according to principles very unlike traditional databases for speech or face recognition.

One of the key aims of this chapter is to convey why research that uses databases in emotion-oriented computing needs to think about database issues rather than reflexively importing habits derived from other fields and to identify the principles that contemporary research suggests are relevant to selecting databases in this particular area. In a decade, a handbook of emotion-oriented computing may not need to cover those issues. Imported habits may seem too obviously inappropriate to be worth mentioning. However, at the time of writing, it remains true that reviewers are continually faced with articles full of sophistication in the computational domain, but painfully naïve about the kinds of data that their sophistication should be applied to.

A second key aim is to inform people whose interest is in developing databases rather than using them (or as well as using them). In many areas, developing databases is not seen as central to the field. The general assumption is that databases ought to be there to be pulled off the shelf as needed. In the context of emotion, there are good reasons for thinking that the task of database development needs a higher profile. That is not only because the task is large but also because it demands very particular skills. If the development of emotion-related databases becomes a subdiscipline in its own right, as we believe it probably should, it needs its own theoretical framework.

2 Theoretical Foundations

A wide range of conceptual issues may be relevant to the design or the selection of a database involving emotion in a broad sense. This section tries to draw out the main principles, grouping them under three broad headings – function; structure and scope; and relationship to psychological theory. A later section considers another range of issues which are essentially practical.

2.1 Database Functions

There may be many different motives behind the construction or the selection of a database. That point was recognised early on in HUMAINE using a simple distinction between two extreme types of goal.

> At one extreme, databases can play a purely supportive role, allowing investigators working within a well-established paradigm to fill in or check details of processes that they assume are understood in general terms. At the other extreme, databases can play a provocative role, providing examples that help investigators to expand and restructure their thinking about the area (Douglas-Cowie et al., 2004).

The supportive/provocative distinction is useful because it is simple. However, it reflects a much more complex set of distinctions in the aims relevant to database research. This section sets out the issues more fully.

The archetypal function of a database is to provide material for training systems that will be used for recognition. Databases with that function need to be large and structured, with many tokens of each significant type. The class itself subdivides. At one extreme, the aim of learning may be to acquire something that is understood as a general competence (such as recognising a large vocabulary of words in clear speech). At the other, the aim may be to train a system for a particular application, and only for that application. Often both aims play a part. There is a target application, but there is also an aspiration to address it in a way that has some generality – not least because the unforeseen happens, and that is bad news for systems trained too narrowly to deal with the foreseen.

Synthesis makes different demands. For some applications (such as copy synthesis), a small number of cases, or even a single case, may be sufficient to extract parameters that will be used to generate a particular kind of output – a kind of expression, for instance, or a gait. On the other hand, other applications may require large numbers, because they need a complete set of components (samples of phonemes in all relevant contexts, for instance).

Although these are the applications that come to mind immediately, a high proportion of papers that use databases have rather a different kind objective: broadly speaking, to test a technique or to select among alternatives. For example, a speech database may be used to select a good set of features from the enormous range that could conceivably be used for recognition or to compare the effectiveness of different statistical learning techniques. As a special case, certain data sets become the basis for tasks which are accepted as a ‘gold standard’; systems are compared by their performance on them.

Databases also serve more traditional functions. There are areas where skilled observers – psychologists, linguists, ethologists, and so on – are central to the process of developing systems. Their intuition allows them to formulate hypotheses (particularly qualitative hypotheses) that machines cannot, for instance, about the kinds of emotion-related meaning that signs convey. That kind of intuition needs to be informed by data. If the examples available simply reflect a set of presuppositions which are orderly but naïve, they cannot give intuition the right kind of impetus.

A special version of that function comes to the fore in the particular case of emotion databases. It is bootstrapping the development of databases. It tends to be by engaging with examples in an existing database that people recognise new types of issue that they may need to consider. A single example in a preliminary collection may draw attention to a large domain that had been not considered a priori, but that once recognised clearly needs to be covered.

That kind of issue arises in several forms. The simplest is recognising that a particular kind of example is important, and that more of the same kind should be collected. Preliminary databases are also fundamental to the development of descriptive systems. Examples highlight the information that needs to be provided to make sense of the fact that emotion is sometimes displayed in one way, sometimes in another. Linked to both of those, deciding the form that databases should eventually take depends on understanding how the domain of emotion is structured; but that understanding has to come from databases. For example, it is possible that in the long run, databases will only need to contain examples of archetypal states, because interpolation will be enough to cover the rest of the domain. However, that can be tested only using a database that contains a sufficient variety of non-archetypal examples.

2.2 Structure and Scope

It follows from the last section that several kinds of requirement have a bearing on the amount and kind of material that a database should contain. Some can be formulated numerically, others are in practice intuitive. Their importance will depend on the particular application that the researcher has in mind.

2.3 Size and Formal Structure

When databases are used to train recognisers, the processes involved are generally grounded in statistics. There is a large statistical literature on the problem of identifying sample sizes sufficient to achieve robust distinctions between classes. The analyses translate into prescriptions for the sizes of the databases that are needed to train recognisers.

Classical analyses (e.g. Raudys and Jain, 1991) produce figures that are in line with reasonably common practice. For a simple classification rule, such as Fisher's linear discriminant analysis, if p features are used for classification, about $10p$ training samples per class are needed to obtain a reliable decision rule. Matters become more difficult with more complex rules, not least because the required number of samples per class tends to rise faster than p. For instance, with a quadratic discriminant analysis, for 8 features, the requirement is about $16p$; for 20, it is about $22p$; for 50 it is about $40p$. The rapid rise with number of features is linked to the so-called 'curse of dimensionality'.

These numbers are based on analytic procedures that make tacit assumptions about distributions. Where the relevant properties are not ideally distributed, empirical approaches suggest that sample sizes may need to be an order of magnitude bigger (Han et al., 2005).

Balance between categories is also critical. If a database covers several classes of phenomenon, and the frequencies are different, the size of the smallest class is generally used as a basis for deciding whether a particular statistical technique can safely be applied. It follows that a database of a given total size is much more useful for training if the phenomena that it covers are evenly distributed.

The issues are usually stated in terms of a classification paradigm. Similar considerations apply when the analysis is concerned with continuous relationships between variables. For the simplest analysis of that kind, multiple regression, a standard formulation is that 40 cases per variable are needed. Corresponding to the issue of equal class sizes, statistical effectiveness falls if the variables being considered are correlated (Tabachnick and Fidell, 2001).

Turning to synthesis, examples give a sense of the size and the structure that are needed to achieve respectable synthesis of speech. CMU Arctic databases are regarded as somewhere near the lower bound for acceptable synthesis. An Arctic database is a reading of the Arctic prompt set by a single speaker in a specified style of delivery (plus associated files). The prompt set contains 10,045 words and 39,153 phones – about 2 h of speech including pauses (Kominek and Black, 2004).

A minimal constraint on database size is that it should contain all the theoretically relevant kinds of unit – which for basic unit synthesis, techniques mean all phoneme pairs. In practice, though, phoneme pairs differ according to context. As a result, databases close to the theoretical minimum size are not ideal. Larger databases are more likely to contain not only the right units but also the right units in the right context. A neat illustration of the way that can be exploited is that some systems replay whole words or even phrases directly if they can be found in the corpus, rather than constructing them from simpler elements (sequences of a few phonemes).

Evidence shows that increasing database size improves quality. For instance, a study by Sak (2000) compared performance using two databases, one containing 3 h of speech and the other containing 19 h. At the upper end, the XIMERA system was developed using corpora from three speakers, of 20, 60 and 111 h, respectively (Kawai et al., 2004). These figures are all for the synthesis of essentially neutral styles. Introducing convincing emotional expression would require correspondingly larger corpora.

2.3.1 Units of Analysis

Databases need to be divided into appropriate units of analysis. In the speech material mentioned above, the primary units of analysis are relatively clear – groups of two or three adjacent phonemes and words. In emotion databases, choice of units is a vexed issue. There is a strong tradition of using individual frames as a basis for emotion recognition from faces, but gestures extend over much longer times. Speech-oriented work has divided data into words (Batliner et al., 2006) and phrase-like units (Fragopanagos and Taylor, 2005). Divisions based on emotion include some designed to extract episodes of sustained emotion (Douglas-Cowie et al., 2003) and some to include build-up from ‘rest’ levels to a peak and return to ‘rest’ levels (see the chapter on the HUMAINE database in Part III).

A widely used pattern in HUMAINE is to use a primary division into ‘clips’, where a clip is an episode chosen so that emotion-related signals within it are generally understood in the same way as they would be in a much larger context. There are typically many nested and overlapping units within a clip, often with fuzzy boundaries. A good formalisation of these issues would be very useful.

2.3.2 Modalities

There is increasing agreement that emotion is multimodal. Signals in different channels are likely to complement and illuminate each other rather than duplicating the same information. The term modality is commonly used not just to refer to gross distinctions between visible, audible and physiological sources but also to distinguish within them. So facial, gestural and postural modalities would normally be distinguished within the broad category of optical sources.

It is also increasingly clear that signs need to be related to the context in which they occur. Some very successful procedures interpret facial signals in the context of ongoing attempts to solve particular problems. It may be stretching semantics to call context a modality, but it certainly needs to be considered alongside modalities like gesture and prosody.

For those reasons, emphasis has shifted away from unimodal databases. They remain important, not least because some applications are inherently unimodal – such as detecting emotion in phone calls (to help desks, clinics, etc.). But there is a clear need to collect material that clarifies how modalities interact.

A complicating factor is that the modalities seem not to combine independently. It is hard to find situations that give rich emotional signs in vocal, facial and gestural modalities simultaneously. As a result, databases drawn from a relatively uniform kind of source cannot practically be expected to be equally rich in all modalities.

2.3.3 Realism

There is a standard method of generating resources like the Arctic or XIMERA databases mentioned above. A ‘talent’ is given a body of material to speak, and it is recorded in a soundproof room. Translating that method directly to the domain of emotion would involve asking ‘talents’ to simulate the emotion on demand – i.e. acting. One of the most contentious issues in the area is how useful acted renditions of emotion are.

It is important to be clear that the issue is not simple. Using acted data has great potential advantages. It is likely to be much more economical than collecting spontaneous samples of emotion as it appears in everyday life; it is much easier to ensure that the data is well structured; the recordings are likely to be more tractable; and it is usually clearer what the underlying emotional state is (so long as one sets aside awkward general questions about what the underlying state of an actor simulating an emotion might be).

It is harder to articulate the advantage of naturalistic sampling. It often seems to be implied that the issue is whether something essential is present in real data but not in acted data, such as physiological arousal which gives rise to telltale changes in muscle tone or reactivity. But it is probably at least as important that naturalistic sampling has the potential to reflect the kind of connectivity that is outlined in Chapter ‘Editorial: ‘Theories and Models’ of Emotion’. Real emotion has connections at its core, and it characteristically affects a multitude of connections around it. It impacts the way people think, what they look at, their choice of words and discourse structure, the actions they take and how they execute them; and how all that depends on the context, social and physical. Recognising emotion in these things is rather like recognising water in coffee, or spilled on a floor, or in a damp cloth, or in a balloon, or in processed meat, or in a carrot, or in hydrated crystals, or in aquiferous rock, or in a cloud, or in a rainbow. Pictures of pure water in laboratory beakers would not be an ideal basis for recognising water in these contexts. The same problem arises with pictures of idealised and decontextualised emotion, however true they may be to what they depict.

These points are intimately related to the earlier point about bootstrapping database development. It is only by building up large collections that it can become clear how much variety needs to be represented in a functional set of databases and how much of the load can be carried by a few features which recur in virtually any context.

2.4 Databases and Psychological Theory

Psychological theory is indirectly relevant to a wide range of questions about the design or the selection of a database, including several that have already been addressed. However, there are areas where the psychology is absolutely central. They often raise quite difficult issues.

2.4.1 Respecting Psychological Semantics

Emotion-oriented computing has an obligation not to deceive. Unfortunately, it is extremely easy to slip into deception by misusing the words that describe emotion in both expert and 'naïve' psychology. Databases are central to avoiding that kind of deception.

The problem arises when, for instance, it is claimed that a system recognises anger, and in fact it only recognises activation. That is very likely to happen if a system is trained on a database that consists of audio samples that are either angry or neutral. A system trained on that basis is likely to identify anger when it is presented with samples of fear, surprise, amusement, happiness, stress and many other states. The reason is that they all involve elevated activation. Level of activation is not the only difference between anger and neutrality, but in the auditory modality, it is much the easiest difference to detect; and therefore, it is the dimension that is likely to dominate the choices of a system whose training rests on samples of anger and neutrality.

To avoid being party to deception, teams that develop databases have to consider how the descriptive terms that they use relate to the coverage of the database. It is inviting misrepresentation for a database to use the label 'anger' unless it contains a sufficient range of other states to ensure that the term is properly contrasted with the alternatives that people have in mind when they say 'anger'.

In practice, that means databases always have to emotional space quite widely. The granularity is another matter. Some areas may be so easily discriminated that they do not need to be heavily represented. But there should be real concern about databases that do not allow at least one state to be contrasted with all the others that a human being would contrast with the target state.

2.4.2 Theory-Driven Sampling

There are well-known lists or taxonomies that aim to cover the whole domain of emotion, the simplest being the 'big six' emotions proposed by Ekman (1992). It is natural to expect that a database will be complete if it includes samples of all the states in such a list.

It is critical that theories in psychology, in this area and others, need to be understood as hypotheses. A theory corresponds to a framework that is recognised as conceptually attractive. It is another matter whether the framework deals satisfactorily with the empirical evidence. That is the test against which a framework has to be judged. Hence, for the medium term, theories need to be measured against databases; and the theories need to be called into question if the range of phenomena that they cover does not match up to the range that the database allows us to see occurs in real life.

It cannot be predicted in advance how the question will be answered. The conclusion may be that the theory is sound and the database is skewed, or that theory is too narrow. Deciding which of those is true is part of the bootstrapping process that has been mentioned already.

2.4.3 Prototypes and Ecological Validity

It is clear that in practice, people who are collecting material for a database often look for cases that they regard as 'good examples'. It is less clear what status the concept of a 'good example' has. The most straightforward idea is that it corresponds to something that has a special status in people's mental representation of emotion – what psychological theories like Rosch's (1978) call a prototype (see also chapter 'Editorial: 'Theories and Models' of Emotion').

There may be value in going out to look for prototypical examples, but it is clearly not the same as finding examples that represent what is likely to happen in a relevant range of situations – that is, looking for ecological validity. There are good psychological reasons to suspect that people will tend to underestimate the difference – events that confirm stereotypes tend to have a salience in memory that their objective frequency does not warrant and memories of events tend to become more like prototypes than the original events actually were (Sutherland, 2007). The more sophisticated that investigators are about the way they approach these issues, the less likely they are to misunderstand what they have actually collected.

2.4.4 Annotation

The point of annotation is to distinguish states that are functionally different and group states that are functionally similar. It is reasonable to assume that that is effectively a psychological task, and therefore psychology should provide the theoretical basis of annotation. In practice, the link has not always been particularly direct.

In some cases, it seems fair to say that the reason is a kind of naïve realism. Researchers sometimes invoke everyday categories in a way that suggests unquestioning belief that those categories correspond exactly to the natural kinds of emotion. The first chapter of this handbook indicates why that is hard to defend.

On the other hand, there are reasons why descriptions derived from psychological theory may not transfer easily to annotation tasks. Plain familiarity is an issue, particularly for applications that are likely to involve people who are not experts in the theory of emotion. Complexity is also an issue. A key early example is the Leeds–Reading database, which is discussed later. It used sophisticated descriptions rooted in psychological theory on one side and linguistic theory on the other; but the result was a fine-grained subdivision showing too few instances of any individual category for statistical relationships to be established.

2.4.5 Forms of Validation

It is clear that a particular type of label should not be used in a database unless it is in some sense trustworthy. However, care is needed about the sense in which it needs to be trustworthy.

Two kinds of tests are commonly invoked. One is that the labels should represent ‘ground truth’ – i.e. the label anger should be applied only if it can be shown that the person is genuinely angry when it is applied. The other is that the label should not be used unless there is a high degree of agreement associated with it, i.e. the labelling is statistically reliable. Neither of those criteria should be taken for granted. The point is related to the distinction between cause-and-effect-type descriptions drawn by Cowie et al. (2001).

It is appropriate to ask for ground truth when the aim is to give a cause-type description – that is, to establish what really behind the signs that can be seen in a recording. However, it is often more relevant to give an effect-type description – that is, to specify how a person can be expected to interpret the signs, rightly or wrongly. It often makes sense to assume that the interpretation observers put on the signs another person gives is probably the most reliable indicator of the real situation that is available. It certainly should not be assumed that brain scans or psychophysiological records would be more reliable.

Statistical reliability is an appropriate test if raters are supposed to be describing objective matters. That is typically the case when they are rating signs of emotion. However, it is a different issue when they are giving effect-type descriptions of their own reactions. It may well be the case that certain signs evoke reactions in different people, and discarding evidence that they do is distortion, not precision.

A special case of the distortion comes when databases are restricted to items that are identified with high reliability. The fact that a display is very reliably recognised does not mean that it is natural. On the contrary, the best way to ensure high recognition is to exaggerate.

In all these areas, the key rule is to consider what relevant demands are, rather than applying a rule unthinkingly. The demands that it is appropriate to make, depend on recognising that emotion is a complex psychological phenomenon.

3 The Development of the Field: Illustrative Case Studies

The principles that have been outlined were not obvious when research in the area began. They have emerged gradually over a decade of relatively sustained work. This section traces the way understanding of the issues has deepened through case studies of key research efforts. The area has progressed by registering the problems that arose in seemingly reasonable efforts to develop databases and looking for ways to address them.

3.1 The Berlin Database

This database represents the archetypal database in the 1990s – unimodal (speech only), acted and focused on a few emotions that were considered primary. Actors were asked to read the same set of sentences in each of five different emotions. Labelling simply consisted of identifying the emotion that the actor simulated.

Databases of that general kind formed the backbone of early work on emotion recognition. But it gradually became apparent that systems trained on this kind of data transfer poorly to real-life situations. The point was strongly made in an influential paper by Batliner et al. (2003).

3.2 The Leeds–Reading Database

The first notable departure from the style described above was the Leeds–Reading database (Roach, 2000). It too was unimodal, but it was not acted. It consisted of speech drawn from real-life emotional situations. The focus was on selecting the most intense emotional recordings and situations that could be found, including material from commentaries on major disasters.

The labelling of emotion was of very different order from the Berlin database. Labels had four elements. The first was an everyday emotion label (freely chosen), and the second specified the strength of the emotion. The third and fourth provided descriptors based on the appraisal theory of Ortony et al. (1988), setting the emotion in a generic category and describing its antecedents.

Despite the emphasis on intense material, labellers used a wide range of emotion words. This was an early indication that emotion in real life did not fit neatly into the ‘big six’ emotions identified in the theories most familiar to most research teams entering the area (happiness, anger, sadness, fear, contempt, disgust). Coding of speech was also sophisticated so that segments were classified into multiple types. The result was a multiplicity of labels, which meant that when data came to be analysed, there was no straightforward way to find patterns (Stibbard, 2001). Other problems were also realised in hindsight. For example, the researchers had focused on only the extreme peaks of emotion in the data and had thrown away the lead up and movement away from peaks of emotion. The result was that the emotional peaks were decontextualised, and there was nothing to compare them with. Particularly galling, copyright problems meant that the material could not be released.

The Leeds–Reading database was outstandingly ambitious for its time. Because of that, it gave an early indication of the complexities involved in selecting, labelling, analysing and distributing real data. These were key lessons for subsequent researchers in the area.

3.3 Belfast Naturalistic and EmoTV Databases

These databases drew on the Leeds–Reading experience. They used naturalistic data of the type that had been explored for the Leeds–Reading database – media broadcasts – but they used audiovisual data. Both selected emotional episodes to show the wider emotional context in which the emotional episode took place. Both also set out deliberately to address the nuances and complexities of emotion in real life and still avoid the statistical difficulties that had beset the Leeds–Reading database. The BND developed ‘trace’ techniques, which allowed raters to report their perceptions of the speaker’s emotions in quantitative terms, using the classical affect dimensions – positive to negative and active to passive. It also developed quantitative methods of describing speech so that straightforward statistical techniques could be used. The EmoTV database used a free choice of emotion words assigned over time but then found ways to group them into larger ‘cover classes’. It also looked for appropriate ways to label signs of emotion – not only speech (using an approach broadly similar to the BND) but also gesture.

Although these databases addressed known problems, they in their turn exposed a range of others. Although the recording quality was acceptable to humans, it did not allow the kind of machine processing that affective computing teams needed. It became obvious that both databases were very much rooted in specific contexts (mainly TV chat shows, news reporting and interviews) and that their relationship to other issues of concern to affective computing, such as persuasion, was not straightforward. Questions about the relationship between modalities were also highlighted. Representation of gesture in particular was strikingly uneven, and it was unclear whether this represented a failure to capture a significant modality or an indication that signs are often not given in all modalities. Both also uncovered problematic issues with labelling. EmoTV, for example, showed that much data consisted of mixed or blended emotions and that a labelling system needed to take this into account. The BND indicated that averaging across raters may mask the fact that individuals may ‘read’ the same recording in systematically different ways. Both showed widespread masking of emotion, and exposed the need for labelling techniques that could address the issue. Both databases also drew attention to issues of match and mismatch between signs of emotion in the face and in speech. For example, people who seemed to be fundamentally sad nevertheless smiled in a way that can easily be taken to convey happiness. Perhaps most acute of all, neither database could be released to the wider community because of copyright reasons.

These issues led the teams involved in both databases to explore techniques that provided more control over the data they would use. They also joined forces to develop a common labelling scheme informed by the two naturalistic databases.

3.4 AIBO, SAL and EmoTaboo

Experience with broadcast material has led to growing acceptance that, in effect, research teams need to become film-makers. Investigators have identified situations that induce interactions which are genuinely emotionally coloured but where they control inducing factors, recording and distribution. That has produced databases that support quite deep analysis. The AIBO database, recorded at Erlangen, pioneered the style. Children were recorded interacting with the AIBO robot, and its behaviour was manipulated to induce different kinds of emotionally coloured reaction. The speech has been thoroughly annotated, and techniques based on soft vectors have been developed to reflect the fact that the emotional content does not fall into sharply defined classes. The database forms the basis of the CEICES project, which is reported elsewhere in this handbook. The Belfast team moved from the BND to develop the SAL database, using an induction technique which simulated conversation with ‘artificial listeners’. Both audio and visual components have been coded using techniques similar to the BND, and high recognition rates have been reported (Fragopanagos and Taylor, 2005). The SAL technique does not produce much gesture, and the EmoTV team developed the EmoTaboo database to address that gap. Their induction technique was a game designed to elicit emotional gestures, in a reasonably naturalistic way. Labelling the gesture called for new techniques, and the work is still ongoing.

These examples have been selected to illustrate the general development of the field. Later chapters provide more detail on different aspects of them. The next section provides a wider ranging review, designed to let readers identify most of the significant databases that have been described in the literature.

4 An Overview of Existing Databases

This section considers emotion databases in three broad categories – those that are multimodal, those that are speech alone and face databases. Tables 1, 2 and 3 summarise the state of the art with regard to databases. Table 1 lists multimodal databases, Table 2 lists speech databases and Table 3 lists face databases. Data on

Table 1 Multimodal databases relevant to emotion

Belfast naturalistic database (Douglas-Cowie et al., 2000, 2003) *Modalities*: Audiovisual *Emotions*: Wide range *Elicitation*: Natural: 10–60 s long ‘clips’ taken from television chat shows, current affairs programmes and interviews conducted by research team *Size*: 125 subjects; 31 male, 94 female *Kind*: Interactive unscripted discourse *Language*: English

Geneva airport lost luggage study (Scherer and Ceschi, 1997, 2000) *Modalities*: Audiovisual *Emotions*: Anger, good humour, indifference, stress sadness *Elicitation*: Natural: unobtrusive videotaping of passengers at Geneva airport lost luggage counter followed up by interviews with passengers *Size*: 109 subjects *Kind*: Interactive unscripted discourse *Language*: French

Chung (2000) *Modalities*: Audiovisual *Emotions*: Joy, neutrality, sadness (distress) *Elicitation*: Natural: television interviews in which speakers talk on a range of topics including sad and joyful moments in their lives *Size*: 77 subjects; 61 Korean speakers, 6 Americans *Kind*: Interactive unscripted discourse *Language*: English and Korean

SMARTKOM www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom *Modalities*: Audiovisual, (+gestures) *Emotions*: Joy, gratification, anger, irritation, helplessness, pondering, reflecting, surprise, neutral *Elicitation*: Human machine in WOZ scenario: solving tasks with system *Size*: 224 speakers; 4/5-min sessions *Kind*: Interactive discourse *Language*: German

Amir et al. (2000) *Modalities*: Audio + physiological (EMG,GSR, heart rate, temperature, speech) *Emotions*: Anger, disgust, fear, joy, neutrality, sadness *Elicitation*: Induced: subjects asked to recall personal experiences involving each of the emotional states *Size*: 140 subjects 60 Hebrew speakers 1 Russian speaker *Kind*: Non-interactive, unscripted discourse *Language*: Hebrew, Russian

SAL database (http://semaine-project.eu/, D09) *Modalities*: Audiovisual *Emotions*: Wide range of Emotions/emotion-related states but not very intense *Elicitation*: Induced: subjects talk to artificial listener and emotional states are changed by interaction with different personalities of the listener *Size*: Study of 20 subjects in 2 SAL scenarios – Powerpoint SAL and Soild SAL, 24 sessions per scenario, 8 hours material per scenario *Language*: English

ORESTEIA database (McMahon et al., 2003) *Modalities*: Audio + physiological (some visual data too) *Emotions*: Stress, irritation, shock *Elicitation*: Induced: subjects encounter various problems while driving (deliberately positioned obstructions, dangers, annoyances ‘on the road’) *Size*: 29 subjects, 90 min sessions per subject *Kind*: Non-interactive speech: giving directions, giving answers to mental arithmetic, etc *Language*: English

Belfast boredom database (Cowie et al., 2003) *Modalities*: Audiovisual *Emotions*: Boredom *Elicitation*: Induced *Size*: 12 subjects: 30 min each *Kind*: Non-interactive speech: naming objects on computer screen *Language*: English

XM2VTSDB multimodal face database http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/ *Modalities*: Audiovisual *Emotions*: None *Elicitation*: n/a *Size*: 295 subjects Video *Kind*: High-quality colour images, 32 kHz, 16-bit sound files, video sequences and a 3D model + profiles (left-profile and one right-profile image per person, per session, a total of 2,360 images), scripted four sentences *Language*: English

ISLE project corpora (http://nats-www.informatik.uni-hamburg.de/~isle/speech.html, IST project IST-1999-10647) *Modalities*: Audiovisual + gesture *Emotions*: None *Elicitation*: n/a *Size*, *Kind*, *Language*, unclear

Table 1 (continued)

Polzin and Waibel (2000) *Modalities*: Audiovisual (though only audio channel used) *Emotions*: Anger, sadness, neutrality (other *Emotions* as well, but in insufficient numbers to be used) *Elicitation*: Acted: sentence length segments taken from acted movies *Size*: Unspecified no. of speakers. Segment numbers 1,586 angry, 1,076 sad, 2,991 neutral *Kind*: Scripted *Language*: English

Banse and Scherer (1996) *Modalities*: Audiovisual (visual info used to verify listener judgements of emotion) *Emotions*: Anger (hot), anger (cold), anxiety, boredom, contempt, disgust, elation, fear (panic), happiness, interest, pride, sadness, shame *Elicitation*: Acted: actors were given scripted eliciting scenarios for each emotion, then asked to act out the scenario *Size*: 12 (6 male, 6 female) *Kind*: Scripted: Two semantically neutral sentences (nonsense sentences composed of phonemes from Indo-European languages) *Language*: German

Table 2 Databases of speech relevant to emotion

TALKAPILLAR (Beller et al., 2005) *Emotions*: Neutral, happiness, question, positive and negative surprised, angry, fear, disgust, indignation, sad, bore *Elicitation*: Contextualised acting: actors asked to read semantically neutral sentences in range of *Emotions*, but practised on emotionally loaded sentences beforehand to get in the right mood *Size*: One actor reading 26 semantically neutral sentences for each emotion (each repeated three times in different activation level: low, middle, high) *Kind*: Non-interactive and scripted *Language*: French

Leeds–Reading database (Greasley et al., 1995; Roach et al., 1998; Stibbard, 2001) *Emotions*: Range of full-blown *Emotions Elicitation*: Natural: unscripted interviews on radio/television in which speakers are asked by interviewers to relive emotionally intense experiences *Size*: Around 4 ½ h material *Kind*: Interactive unscripted discourse *Language*: English

France et al. (2000) *Emotions*: Depression, suicidal state, neutrality *Elicitation*: Natural: therapy sessions and phone conversations. Post-therapy evaluation sessions were also used to elicit speech for the control subjects *Size*: 115 subjects: 48 females and 67 males. Female sample: 10 controls (therapists), 17 dysthymic, 21 major depressed Male sample: 24 controls (therapists), 21 major depressed, 22 high-risk suicidal *Kind*: Interactive unscripted discourse *Language*: English

Campbell CREST database, ongoing (Campbell, 2002; see also Douglas-Cowie et al., 2003) *Emotions*: Wide range of emotional states and emotion-related attitudes *Elicitation*: Natural: volunteers record their domestic and social spoken interactions for extended periods throughout the day *Size*: Target – 1,000 h over 5 years *Kind*: Interactive unscripted discourse *Language*: English, Japanese, Chinese

Capital Bank Service and Stock Exchange Customer Service (as used by Devillers and Vasilescu, 2004) *Emotions*: Mainly negative – fear, anger, stress *Elicitation*: Natural: call centre human–human interactions *Size*: Unspecified (still being labelled) *Kind*: Interactive unscripted discourse *Language*: English

SYMPAFLY (as used by Batliner et al., 2003) *Emotions*: Joyful, neutral, emphatic, surprised, ironic, helpless, touchy, angry, panic *Elicitation*: Human machine dialogue system *Size*: 110 dialogues, 29.200 words (i.e. tokens, not vocabulary) *Kind*: Naïve users book flights using machine dialogue system *Language*: German

Table 2 (continued)

DARPA Communicator Corpus (as used by Ang et al., 2002) See Walker et al. (2001) *Emotions*: Frustration, annoyance *Elicitation*: Human machine dialogue system *Size*: Extracts from recordings of simulated interactions with a call centre, average length about 2.75 words 13,187 utterances in total of which 1,750 are emotional: 35 unequivocally frustrated, 125 predominantly frustrated, 405 unequivocally frustrated or annoyed, 1,185 predominantly frustrated or annoyed *Kind*: Users called systems built by various sites and made air travel arrangements over the phone *Language*: English

AIBO (Erlangen database) (Batliner et al., 2004) *Emotions*: Joyful, surprised, emphatic, helpless, touchy (irritated), angry, motherese, bored, reprimanding, neutral *Elicitation*: Human machine: interaction with robot *Size*: 51 German children, 51.393 words (i.e. tokens, not vocabulary) English (Birmingham): 30 children, 5.822 words (i.e. tokens, not vocabulary) *Kind*: Task directions to robot *Language*: German

Fernandez and Picard (2003) *Emotions*: Stress *Elicitation*: Induced: subjects give verbal responses to maths problems in simulated driving context *Size*: Data reported from four subjects *Kind*: Unscripted numerical answers to mathematical questions *Language*: English

Tolkmitt and Scherer (1986) *Emotions*: Stress (both cognitive and emotional) *Elicitation*: Induced: 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 *Size*: 60 (33 male, 27 female) *Kind*: Partially scripted: subjects made three vocal responses to each slide within a 40-s presentation period – a numerical answer followed by two short statements. The start of each was scripted and subjects filled in the blank at the end, e.g. ‘Die Antwort ist Alternative . . .’ *Language*: German

Iriondo et al. (2000) *Emotions*: Desire, disgust, fury, fear, joy, surprise, sadness *Elicitation*: Contextualised acting: subjects asked to read passages written with appropriate emotional content *Size*: Eight subjects reading paragraph length passages *Kind*: Non-interactive and scripted *Language*: Spanish

Mozziconacci (1998) Note: database recorded at IPO for SOBUproject 92EA. *Emotions*: Anger, boredom, fear, disgust, guilt, happiness, haughtiness, indignation, joy, rage, sadness, worry, neutrality *Elicitation*: Contextualised acting: actors asked to read semantically neutral sentences in range of *Emotions* but practised on emotionally loaded sentences beforehand to get in the right mood *Size*: Three subjects reading eight semantically neutral sentences (each repeated three times) *Kind*: Non-interactive and scripted *Language*: Dutch

McGilloway (1997) and Cowie and Douglas-Cowie (1996) *Emotions*: Anger, fear, happiness, sadness, neutrality *Elicitation*: Contextualised acting: subjects asked to read passages written in appropriate emotional tone and content for each emotional state *Size*: 40 subjects reading five passages each *Kind*: Non-interactive and scripted *Language*: English

Belfast structured database An extension of McGilloway database above (Douglas-Cowie et al. 2000) *Emotions*: Anger, fear, happiness, sadness, neutrality *Elicitation*: Contextualised acting: subjects read 10 McGilloway-style passages and 10 other passages – scripted versions of naturally occurring emotion in the Belfast naturalistic database *Size*: 50 subjects reading 20 passages *Kind*: Non-interactive and scripted *Language*: English

Danish emotional speech database (Engberg et al., 1997) *Emotions*: Anger, happiness sadness, surprise neutrality *Elicitation*: Acted *Size*: Four subjects read two words, nine sentences and two passages in a range of *Emotions* *Kind*: Scripted (material not emotionally coloured) *Language*: Danish

Table 2 (continued)

Groningen ELRA corpus number S0020 (www.icp.inpg.fr/ELRA) this new link is working (date: 16/01/2005):(www.elda.org/catalogue/en/speech/S0020.html) *Emotions*: Database only partially oriented to emotion *Elicitation*: Acted *Size*: 238 subjects reading two short texts *Kind*: Scripted *Language*: Dutch

Berlin database (Kienast and Sendlmeier, 2000; Paeschke and Sendlmeier, 2000) http://www.expressive-speech.net/ *Emotions*: Anger (hot), boredom, disgust, fear (panic), happiness, sadness (sorrow), neutrality *Elicitation*: Acted *Size*: 10 subjects (5 male, 5 female) reading 10 sentences each *Kind*: Scripted (material selected to be semantically neutral) *Language*: German

Pereira (2000) *Emotions*: Anger (hot), anger (cold), happiness, sadness, neutrality *Elicitation*: Acted *Size*: Two subjects reading two utterances each *Kind*: Scripted (one emotionally neutral sentence, four digit number) each repeated *Language*: English

van Bezooijen (1984) *Emotions*: Anger, contempt disgust, fear, interest joy, sadness shame, surprise, neutrality *Elicitation*: Acted *Size*: Eight (four male, four female) reading four phrases *Kind*: Scripted (semantically neutral phrases) *Language*: Dutch

Abelin and Allwood (2000) *Emotions*: Anger, disgust, dominance, fear, joy, sadness, shyness, surprise *Elicitation*: Acted *Size*: one subject *Kind*: Scripted (semantically neutral phrase) *Language*: Swedish

Yacoub et al. (2003) (data from LDC, www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28) *Emotions*: 15 *Emotions* Neutral, hot anger, cold anger, happy, sadness, disgust, panic, anxiety, despair, elation, interest, shame, boredom, pride, contempt *Elicitation*: Acted *Size*: 2,433 utterances from eight actors *Kind*: Scripted *Language*: English

Table 3 Face databases

The AR face database (http://www.isbe.man.ac.uk~bim/data/tarfd_markup/tarfd_markup.html) *Emotions*: Smile, anger, scream neutral *Elicitation*: Posed *Size*: 154 subjects (82 male, 74 female) 26 pictures per person *Kind*: 1: Neutral, 2 smile, 3: anger, 4: scream, 5: left light on, 6: right light on, 7: all side lights on, 8: wearing sun glasses, 9: wearing sun glasses and left light on, 10: wearing sun glasses and right light on, 11: wearing scarf, 12: wearing scarf and left light on, 13: wearing scarf and right light on, 14–26: second session (same conditions as 1–13)

CVL face database (http://lrv.fri.uni-lj.si/facedb.html) *Emotions*: Smile *Elicitation*: Posed *Size*: 114 subjects (108 male, 6 female) seven pictures per person *Kind*: Different angles, under uniform illumination, no flash and with projection screen in the background

The Psychological Image Collection at Stirling (http://pics.psych.stir.ac.uk/) *Emotions*: Smile, surprise, disgust *Elicitation*: Posed *Size*: Aberdeen: 116 subjects Nottingham scans: 100 Nott-faces-original: 100 Stirling faces: 36 *Kind*: Contains seven face databases of which four largest are Aberdeen, Nottingham scans, Nott-faces-original, Stirling faces mainly frontal views, some profile, some differences in lighting and expression variation

The Japanese female facial expression (JAFFE) database (http://www.kasrl.org/jaffe.html) *Emotions*: Sadness, happiness, surprise, anger, disgust, fear, neutral *Elicitation*: Posed *Size*: 10 subjects seven pictures per subject *Kind*: Six emotion expressions + one neutral posed by 10 Japanese female models

Table 3 (continued)

CMU PIE database [CMU Pose, Illumination, and Expression (PIE) database] (http://www.ri.cmu.edu/projects/project_418.html) *Emotions*: Neutral, smile, blinking and talking *Elicitation*: Posed for neutral, smile and blinking 2 s video capture of talking per person *Size*: 68 subjects *Kind*: 13 different poses, 43 different illumination conditions, and with four different expressions

Indian Institute of Technology Kanpur database (http://vis-www.cs.umass.edu~vidit/IndianFaceDatabase/) *Emotions*: Sad, scream, anger, expanded cheeks and exclamation, eyes open–closed, wink *Elicitation*: Posed *Size*: 20 subjects *Kind*: Varying facial expressions, orientation and occlusions; degree of orientation is from 00 to 200 in both right and left directions, the similar angle variation are considered in the case of head tilting; and also head rotations both in top and bottom are taken into account. All of these images are taken with and without glasses in constant background; for occlusions some portion of face is kept hidden and lightning variations are considered

The Yale face database (http://cvc.yale.edu/projects/yalefaces/yalefaces.html) *Emotions*: Sad, sleepy, surprised *Elicitation*: Posed *Size*: 15 subjects *Kind*: One picture per different facial expression or configuration: centre-light, w/glasses, happy, left light, w/no glasses, normal, right light, sad, sleepy, surprised and wink

CMU facial expression database (Cohn–Kanade) (http://vasc.ri.cmu.edu//idb/html/face/facial_expression/index.html) *Emotions*: Six of the displays were based on descriptions of prototypic *Emotions* (i.e., joy, surprise, anger, fear, disgust and sadness). *Elicitation*: Posed *Size*: 200 subjects *Kind*: Subjects were instructed by an experimenter to perform a series of 23 facial displays that included single action units (e.g. AU 12 or lip corners pulled obliquely) and combinations of action units (e.g. AU 1+2, or inner and outer brows raised). Subjects began and ended each display from a neutral face

Caltech Frontal Face DB (http://www.vision.caltech.edu/html-files/archive.html) *Emotions*: Unclear *Elicitation*: *Size*: 27 subjects 450 images in total *Kind*: Different lighting, expressions, backgrounds

HumanScan BioID Face DB (http://www.bioid.com/support/downloads/software/bioid-face-database.html) *Emotions*: None *Elicitation*: n/a *Size*: 23 subjects *Kind*: Contains 19 manual markup points: 0 = right eye pupil; 1 = left eye pupil; 2 = right mouth corner; 3 = left mouth corner; 4 = outer end of right eyebrow; 5 = inner end of right eyebrow; 6 = inner end of left eyebrow; 7 = outer end of left eyebrow; 8 = right temple; 9 = outer corner of right eye; 10 = inner corner of right eye; 11 = inner corner of left eye; 12 = outer corner of left eye; 13 = left temple; 14 = tip of nose; 15 = right nostril; 16 = left nostril; 17 = centre point on outer edge of upper lip; 18 = centre point on outer edge of lower lip; 19 = tip of chin

Oulu University physics-based face database (www.ee.oulu.fi/research/imag/color/pbfd.html) *Emotions*: None *Elicitation*: n/a *Size*: 125 subjects *Kind*: All frontal images: 16 different camera calibration and illuminations

UMIST (http://www.sheffield.ac.uk/eee/research/iel/research/face.html) *Emotions*: None *Elicitation*: n/a *Size*: 20 subjects, 19–36 pictures per person *Kind*: Range of poses from profile to frontal views

Table 3 (continued)

Olivetti research (www.mambo.ucsc.edu/psl/olivetti.html) *Emotions*: None *Elicitation*: n/a *Size*: 40 subjects, 10 pictures per person *Kind*: All frontal and slight tilt of head

The Yale face database B (http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html) *Emotions*: None *Elicitation*: n/a *Size*: 10 subjects *Kind*: 9 poses × 64 illumination conditions

AT&T (formerly called ORL database) (http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html) *Emotions*: Smiling/not smiling *Elicitation*: Posed *Size*: 40 subjects *Kind*: 10 images for each subject which vary lighting, glasses/no glasses, and aspects of facial expression broadly relevant to emotion – open/closed eyes, smiling/not smiling

gestures and physiological signals is contained within Table 1, as this type of data (relatively rare) tends to occur within a multimodal context. The tables include material that consists of records with a limited amount of information rather than a fully marked-up and annotated corpus.

The tables are not intended to be a comprehensive list. The aim is to identify key databases and indicate the type of data that is available. Each table is arranged from left to right according to the same format – identifier for the database; modalities recorded (where there is more than one); description of how the data was elicited; indicator of size; further information regarding the type of data; and finally any information on cultural/linguistic range.

The information on each item is presented as a list rather than a geometric table. We have used table format in earlier summaries, but as the field develops, tables become too large for comfort. The format that we have used here seems to be a reasonable alternative.

4.1 Multimodal Databases

The dates of databases in this domain indicate that work on multimodality and emotion is relatively recent. Databases of emotion in multimodal contexts were unusual until the HUMAINE project, and even now large, structured and labelled databases are unusual. However, the area is fast gaining ground.

Perhaps the single most characteristic concern in recent work has been naturalness. Some of the work still uses actors, as, for example, the recent audiovisual database collected in Geneva (Baenziger and Scherer, 2007), but it has used professional actors working with a director. But most of the multimodal work has worked with more naturalistic settings or induction techniques to induce emotion. Some databases use real-life situations such as lost luggage offices (Scherer and Ceschi, 2000) or television chat shows (Chung, 2000; Douglas-Cowie et al., 2003; Devillers et al., 2006). Others use various induction techniques; key examples are mentioned in Douglas-Cowie et al. (2007), and fuller coverage is given in Part III, in the chapter “Issues in Data Collection”.

Because of the emphasis on naturalness, the range of emotions covered tends in the direction of everyday emotional behaviour rather than full-blown emotions. For example, the SMARTKOM database (Schiel et al., 2002; Steininger et al., 2002a, 2002b) like the SAL database is built from listeners' responses to a 'machine'. In fact the machine is actually two humans in another room operating in a Wizard of Oz-type situation. Users are asked to solve a range of tasks. The emotional states and related states recorded include joy/gratification, anger/irritation, helplessness, pondering/reflecting, surprise and neutral. Other multimodal databases reflect the same trend with the Geneva lost luggage database containing examples of good humour and indifference among the emotion-related states listed and the Belfast naturalistic database covering a wide range of emotional states with a wide spread on a two-dimensional representation of emotion (based on the dimensions of evaluation and activation). The French EmoTV database provides striking examples of the way even full-blown emotions tend to be mixed and of the way intensity shifts in the build-up to and away from an emotional peak. The recent Green data set recorded at Belfast shows attempts to persuade people into adopting a more 'sustainable lifestyle'. It is rich in what Baron-Cohen (2007) calls epistemic states – doubt, questioning, rejection and so on.

Coding material of that kind is a challenge, and the techniques that have been adopted are very varied. The main techniques are described in the chapter on labelling, and the synthesis developed for the HUMAINE database is described in the chapter on that database.

Some modalities are still not very fully covered. It has already been noted that gesture was not extensively covered, and the EmoTaboo database was developed to provide more material in that area. Several of the sources that there are, use driving as a context. The ORESTEIA database (McMahon et al., 2003, see also http://manolito.image.ece.ntua.gr/oresteia/) records physiological measurements (with some audio) from subjects on a driving simulator. The subjects encounter various problems while driving (deliberately positioned obstructions, dangers, annoyances 'on the road'). These are intended to induce emotional responses. An ongoing development of the approach (see the later chapter on issues in data collection) involves subjects being induced into various emotional states using techniques derived from the psychological literature and then drive through scenarios designed to test how the emotion affects their action. The recently constructed DRIVAWORK database provides audiovisual and physiological records of subjects using a driving simulator (Hönig, 2007).

The emphasis on naturalism has brought various practical problems with it. Genuinely natural data tends to involve noise, camera angles, lighting and various other factors that pose problems for machine analysis, and as noted above, copyright issues mean that significant material may not be freely available for researchers. The sources use a variety of contexts, but there is no systematic understanding of the range of contexts that data should aim to document. Multicultural data is also rare, though see Cube-G (Lipi et al., 2008). Finally different teams have used a wide variety of labelling conventions for emotion, not all of which are informed by contemporary understandings of emotion.

The HUMAINE database was designed to reflect the main kinds of material that recent progress has made available and the labelling techniques that can be applied to it. It consists of 50 emotional episodes taken from a variety of induced and naturalistic settings and covers a wide range of emotional behaviour and signs of emotion (speech, face, gesture and physiological signs). The database is drawn from a number of sources that contain many more records less fully labelled. These include the Belfast naturalistic database, the SAL database and the EmoTaboo database, which are described above. The other main data sets used are the Reality Castaway TV database, the Belfast adventure and Spaghetti databases, the Erlangen driving database and the SAL Hebrew database. Chapter "Biological and Computational Constraints to Psychological Modelling of Emotion" Part I provides a full description of the HUMAINE database, and Chapter "Emotion: Concepts and Definitions" gives further information about individual sources from which the HUMAINE database has been compiled.

The emphasis in this section has been on multimodal emotion-related material. Deciding what to include is not always straightforward. Some databases contain audiovisual recordings, but the focus has been on the speech (Polzin and Waibel, 2000; Banse and Scherer, 1996). There are also major multimodal databases that have very little emotional content. Examples that appear to be in that category include XM2VTSDB (www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/) and ISLE (http://nats-www.informatik.uni-hamburg.de/~isle/speech.html, IST project IST-1999-10647), but ISLE in particular is relevant because of the labelling systems it has developed for multimodal work.

A particularly interesting case is the large corpus of meeting data collected by the AMI project. The investigators considered explicitly whether there was emotion present (Heylen et al., 2006). The terms that raters used to describe the mental states that they observed were (starting from the commonest) bored, confident, interested, attentive, serious, joking, friendly, curious, cheerful, at-ease, amused, relaxed, nervous, frustrated, decisive, uninterested, impatient, confused, agreeable, annoyed. It appears that the recordings involve mainly what Baron-Cohen (2003) calls epistemic states with an emotional dimension. It may well be that as emotion-oriented computing matures, it will become more concerned with states like these, but at present, they are not what most people in the field consider it their business to study.

4.2 Speech Databases

There has been a considerable body of audio data collected for speech and emotion studies, as reflected in the fact that more corpora appear in Table 2 than either of the other tables.

Much of the data has three characteristics: the emotion in it is simulated by an actor (not necessarily trained); the actor is reading preset material and he/she is aiming to simulate full-blown emotion (Yacoub et al., 2003; Kienast and Sendlmeier, 2000). Other examples in the general vein include (1997) Leinonen and Hiltunen

(1997), Nakatsu et al. (1999), Juslin and Laukka (2002), Nogueiras et al. (2001), Murphy (2002) and Oudeyer (2003). There are sometimes attempts to make the data more natural by contextualising the emotion, for example, using material to read that is inherently emotional in content [examples in Table 2 are from Mozziconacci (1998) and McGilloway (1997)].

At the other extreme, there are a few speech databases which are focused on naturalistic data. The Leeds–Reading database has already been described. Campbell' s CREST database (Campbell, 2006; see also Douglas-Cowie et al., 2003) has acquired a unique body of truly natural data with a wide range of everyday emotions from volunteers who were recorded for long periods as they went about their ordinary daily social interactions.

Also fairly natural, but narrower in emotional range, are data sets that use material recorded during specific types of events, such as game shows, emergency flight situations for pilots, affectively loaded therapy sessions and journalists' reports of emotion-eliciting events (e.g. France et al., 2000 in Table 2 but also Johannes et al., 2000; Frolov et al., 1999; Huttar, 1968; Kuroda et al., 1979; Roessler and Lester, 1976, 1979; Sulc, 1977; Williams and Stevens, 1969, 1972).

A good deal of work has also used data elicited by techniques designed to induce states that are both genuinely emotional and likely to involve speech. Early examples include a task where subjects are introduced to unpleasant images (Tollkmitt and Scherer, 1986) and some parts of the SUSAS database (speech under simulation and actual stress database) which use speech elicited in a range of stressful situations (Hansen and Bou-Ghazale, 1997, http://www.ldc.upenn.edu/Catalog/). More recent examples include simulations of call centres designed to elicit irritation (Mitchell et al., 2000; Batliner et al., 2003); a stressful driving task (Fernandez and Picard, 2003) and a spelling task designed to elicit embarrassment (Bachorowski and Owren, 1995). The Erlangen AIBO database (Batliner et al., 2004) produces a wider range of emotional states such as neutral (default), joyful, surprised, emphatic, helpless, touchy (irritated), angry, bored and reprimanding. It also generates a substantial amount of speech that the investigators describe as 'motherese', which makes a great deal of sense intuitively, and offers a useful reminder that standard lists of states do not necessarily fit easily onto behaviour patterns that are observed in naturalistic settings.

Data from call centres has been used extensively, and it epitomises both the advances that have been made and some of the key difficulties. Some involve human–machine dialogue, such as SYMPAFLY (Batliner et al., 2003) and the DARPA Communicator Corpus used by Ang et al. (2002) (see Walker et al., 2001). Other teams have used human–human call centre data (Devillers and Vasilescu, 2004).

These databases have several attractions. Firstly, the emotion is presumably genuine, not acted. Secondly, they deal with dialogue, which exposes issues missing from the monologue type data often produced in acted or elicited emotion. Thirdly, they are very directly related to a foreseeable application of emotion recognition.

But with these advantages come limitations. Access tends to be limited, both for commercial reasons and because of privacy issues. The frequency with which emotion is expressed tends to be low. To illustrate the scale of the problem, Ang et al. (2002) used material from the DARPA Communicator Corpus totalling 14 h 36 min of speech. The commonest strong emotion was frustration, of which he obtained 42 unequivocal instances. The nature of the interaction imposes constraints of the forms of utterances and probably on the way emotion may be expressed within those forms, raising major questions about generalisability.

Not least, the emotions tend to be from a narrow range, generally negative. Recent studies illustrate the point. The study by Ang et al., cited above, is a case in point. Similarly, Lee and Narayanan (2003) detected negative versus non-negative emotion using a corpus of utterances obtained from a commercially deployed human machine-spoken dialogue application; most dialogue turns had one utterance. Boozer et al. (2003) have reported work on neutral, frustrated and happy states using human–computer dialogues generated by a phone-based airline flight-planning system. The SYMPAFLY system offers a broader range, including states like helpless, panic and touchy.

In summary, speech databases have developed very rapidly over the last decade, with a strong movement towards naturalistic data. Researchers have been quite creative in this area in experimenting with different methods of collecting data, ranging from opportunistic use of pre-recorded naturalistic emotional situations to laboratory-based induction techniques. There is also a growth in the range of language and cultures covered, with work on Western European languages but also on Hebrew (Amir et al., 2000) and on Japanese and Chinese (Campbell, 2006). Nevertheless, core problems remain.

4.3 Face Databases

Table 3 shows a selection of key databases of facial expressions. As can be seen from the descriptions in the last column, these show faces under systematically varied conditions of illumination, scale and head orientation. Rather few consider emotional variables systematically, and the range of emotional expressions considered is quite limited and tends to focus on the ‘primary’ emotions. The data is also generally acted or posed and consist of static images. The term ‘staged’ is perhaps appropriate.

The seminal database of this type is the classic Ekman and Friesen collection of photographs showing facial emotion (Ekman and Friesen, 1975); this can be bought in electronic form. Others in the same mould are the Yale database which contains 11 images for each of 15 individuals, one per different facial expression or configuration – centre light, with glasses, happy, left light, without glasses, normal, right light, sad, sleepy, surprised and wink, and the ORL database of faces which contains 10 different images for each of 40 subjects. The images vary the lighting and aspects of facial expression which are at least broadly relevant to emotion – open/closed eyes, smiling/not smiling. Rather few databases contain samples of faces moving, and moving sequences which are emotionally characterised are even less common.

A good deal of material consists of images produced by research software, e.g. for facial animation, rather than the original video sequences used for the analysis or training. Examples can be found at www.cs.cmu.edu/~face/. Databases that combine speech and video are still rare and the few examples that there are have already been mentioned in Table 1, in particular SMARTKOM and the Belfast naturalistic database. The XM2VTSDB multimodal face database is also audiovisual but does not contain emotion.

The cultural range is dominated by the West, although there is a database of Japanese faces (see Table 3). In summary, the data is limited in emotional range and level of naturalness. However, the field is developing and genuinely natural data (of moving faces) is emerging with a much wider range of emotional expression. However, getting appropriate facial images is not straightforward, and researchers report practical problems along the way. Genuinely natural data involves quite a lot of jerky movement, frequent occlusion of the face, and particular angles that make its use for facial analysis limited. And audiovisual images create real problems in terms of the interference of speech with facial movement. Clearly appropriate balances need to be found.

5 Lessons for Future Research

The previous sections have shown that developing satisfactory emotion databases involves challenges at multiple levels, from the practicalities of recording to conceptual issues in psychology. This section reflects that range by drawing lessons for future work at two levels, first at the level of an abstract overview and then at the level of practical issues that need to be addressed.

5.1 Ideal Specifications and Obstacles to Achieving Them

One of the results of research is a gradually clarifying picture of the kind of data resource that the community might ideally hope for. At least the following properties are clearly desirable:

- It would be fully naturalistic (except insofar as some parts deliberately captured the behaviour of actors, newsreaders, etc.).
- It would sample the whole domain of emotion and emotion-related states.
- It would represent all the types of action through which emotion and emotion-related states can be expressed.
- It would sample the whole range of cultural and individual differences that are important for the expression of emotion.
- The recordings would be of high technical quality.
- There would be recordings in all relevant modalities.
- The data would be comprehensively labelled.
- The labelling would follow a consistent, standard pattern.

- The labelling would include objective verification of all the emotional states involved.
- The material would be structured to facilitate statistically learning (for instance, it would include balanced samples, samples constructed to match in all respects but one key emotional contrasts, etc.).
- The number of instances would be large enough for statistically learning techniques to be applied.
- The material would be freely available.
- The process of obtaining, storing and distributing samples would be ethically sound.

The combined resources currently available fall far short of that ideal. That is partly because the total effort that has been invested is still small relative to the scale of the task, but there are also problems of principle at every turn. For instance,

- The demand for high-quality recordings with naturalness. Genuinely natural data tends to involve quite a lot of jerky movement, frequent occlusion of the face, and particular angles that make its use for facial analysis limited. And audiovisual images create real problems in terms of the interference of speech with facial movement.
- Attempting to achieve multimodality conflicts with naturalness, because it requires more and more intrusive types of recording. It is difficult to imagine recording more than two or possibly three modalities (recorded in ways that can usefully be analysed) without material loss of naturalness.
- The wish for objective verification conflicts with naturalness, because objective verification tends to demand either intrusive methods or tight control over the situation and with the subtler emotional states that make up a large part of emotional life, it is difficult to see how objective verification could be achieved at all.
- There is a balance to be struck between demand for quantity and comprehensiveness of labelling. The more detailed a labelling scheme is, the more labour intensive implementing it is likely to be, and the more skilled the people involved need to be.
- There is a balance to be struck between demand for statistical tractability and comprehensiveness of labelling. The 'curse of dimensionality' means that extracting meaningful relationships from highly detailed labelling schemes requires unrealistic quantities of data, unless radical alternatives to current statistical techniques become available.
- There is an almost unlimited range of types of action through which emotion and emotion-related states can be expressed, and it is difficult to imagine any way of combining them factorially (for instance, it is difficult to imagine a situation where emotional driving behaviour co-occurs with expression of emotion through large amplitude hand gestures and balletic movements).

The list could easily be extended. Two things are essential to deal with a domain that presents problems like these. One is a clear understanding of the need to reach intelligent compromises. It is a major barrier to progress if groups become wedded to one or two ideals, dismiss work that does not fulfil them completely and embrace work that does, even though it is far short on other criteria. The other essential is development of theory that provides a sound motivation for sampling. The traditional emphasis on basic or primary emotions was attractive partly because it seemed to meet that need. A more soundly based alternative is badly needed.

5.2 *Practical Issues*

It is a painful truth that excellent work invested in a database can be rendered useless by failure to engage with any one of a great many practical requirements. This section touches on the most important of these.

5.2.1 Ethics and Privacy

It is essential for the field to be well versed in the ethical issues involved in collecting emotionally coloured data. Most institutions now routinely require ethical approval for any work with humans, and it is not routine to give approval for procedures that involve eliciting negative emotions or deception, both of which are very common in emotion elicitation scenarios. Informed consent is usually *a sine qua non* and that makes it very difficult to record without the participants' knowledge, despite the obvious advantages in terms of naturalness. Retaining recordings requires another level of clearance and again must be backed by informed consent of the people involved, and release requires yet another level.

These issues should be handled through an appropriate ethics committee. Later chapters in this handbook give more detailed information.

5.2.2 Access

Ensuring access has proved a fatal obstacle in the past. It involves several sub-issues. The ethical dimension has been noted above. Release has to be backed by appropriate clearances. Teams will often want to retain control of material so that (for instance) they are not exploited for profit or used for inappropriate purposes. That can be achieved by ensuring that release is governed by conditions of use set out by the groups who constructed the databases. The governing 'CEICES' rubric proposed by FAU (Batliner et al., 2006) is a good example.

The fact that files are likely to be large raises another kind of problem. When files are not too large, the most convenient transmission medium is FTP (with a suitable password). It has been the norm to send larger bodies of data on DVDs, but that becomes onerous if there is a substantial demand. Recently, the HUMAINE Association has been exploring ways of making emotion-related data available via its server.

A third access issue is simply knowing that data exist, since it is not the norm for journals to publish articles that simply give the community details of a new database.

HUMAINE has made lists available on its website, and the HUMAINE Association will continue the practice. The tables in this chapter are based on the HUMAINE website tables.

5.2.3 Format of Recordings

It is all too easy to underestimate how important various technical criteria are if recordings are to be used for machine extraction. Low background noise and echo, fixed distance from the microphone, microphone response characteristics, and interference from nearby cables or other sources all have a major impact of the usefulness of audio recordings. Machine analysis of video can be seriously disrupted by highlights, colour balances, and lack of contrast with the background that a human being would not notice. Physiological data is highly sensitive to misplaced or poorly connected electrodes. Multimodal data needs to include well-defined markers to allow synchronisation.

Similar issues surround level of compression. Compressed files are much easier to store and transmit, but they may not be adequate for some purposes. For instance, in the video modality, MPEG format is suitable for many uses, but raw video files (.avi) may be needed to give the resolution required for accurate FAP extraction. In the context of physiological data, sampling rate raises similar issues. Quite low sampling rates are adequate for most emotion-related measures, but some information about cardiac activity depends on ECG records being sampled at 200 Hz.

In general, database creators need to understand the kind of format that potential users need.

5.2.4 Standardisation

It is a truism that databases are unlikely to be used unless they observe the standards of the field. Very little in the field of emotion-oriented computing even approaches formal standardisation. Exceptions are FACs coding, some MPEG standards, and the EARL described by Schroder et al. in Part IV. However, there are increasing terminologies, tools and practices that are shared by quite large groups, and it is advisable to respect them. Part I sets out terms and concepts that have become common currency within HUMAINE. The next two chapters describe technical resources that have a similar status.

6 Conclusions

This chapter reflects long and sometimes painful experience in the development of databases for emotion-oriented computing. As such chapters tend to do, it has tried to provide the information that the authors wish they had been able to access when they began to work in the area. Key areas are expanded in the chapters that follow; this chapter provides the context for them.

The single most important lesson of a decade and a half of research in the area is that the task is bigger than a beginner tends to assume. It involves understanding a considerable range of background concepts, addressing many different kinds of practicality and knowing what is there already. The chapter has attempted to provide grounding in those areas. The measure of success will be a new generation of databases that avoid repeating the errors of previous generations.

References

Abelin Å, Allwood J (2000) Cross linguistic interpretation of emotional prosody. ISCA workshop on speech and emotion. Newcastle, Northern Ireland, pp 110–113

Amir N, Ron S, Laor N (2000) Analysis of an emotional speech corpus in Hebrew based on objective criteria. In: Douglas-Cowie E, Cowie R, Schroeder M (eds) Proceedings of the ISCA workshop on speech and emotion, Belfast, Textflow, pp 29–33

Ang J, Dhillon R, Krupski A, Shriberg E, Stolcke A (2002) Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: Proceedings ICSLP, Denver, CO, Sept 2002

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 emotion expression. J Pers Social Psychol 70(3):614–636

Bänziger T, Scherer K (2007 Sept) Using actor portrayal to systematically study multimodal emotional expression: the GEMEP corpus. In: Paiva A, Pra-da R, Picard R (eds) Affective computing and intelligent interaction, Lisbon. Springer LNCS, Berlin, pp 476–487

Baron-Cohen S (2003) The essential difference: men, women and the extreme male brain. Penguin/Basic Books

Baron-Cohen S (2007) Mind reading: the interactive guide to emotions – version 1.3 . Jessica Kingsley, London

Batliner A, Fischer K, Huber R, Spilker J, Noeth E (2003) How to find trouble in communication. Speech Commun 40:117–143

Batliner A, Hacker C, Steidl S, Noth E, D'Arcy S, Russell M et al (2004) You stupid tin box—children interacting with the AIBO robot: a cross-linguistic emotional speech corpus. In: Proceedings LREC, Lisbon, 2004

Batliner A, Steidl S, Schuller B, Seppi D, Laskowski K, Vogt T, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V (2006) Combining efforts for improving automatic classification of emotional user states. In: Erjavec T, Gros J (eds) Language technologies, IS-LTC 2006. Infornacijska Druzba (Information Society), Ljubljana, Slovenia, pp 240–245

Beller G, Schwarz D, Hueber T, Rodet X (2005) Hybrid concatenative synthesis in the intersection of speech and music. JIM2005 Paris, CICM, 41–45

Boozer A, Seneff S, Spina M (2003) Using Prosodic features for emotion classification and recognition MIT Spoken Language Systems Group Summary of Research, Jul 2003, pp 51–54. http://groups.csail.mit.edu/sls//archives/root/publications/2003/ResSum2003.pdf. Accessed on 7/11/2010

Campbell N (2002) Recording and storing of speech data. Proceedings LREC 2002, Las Palmas, Canary Islands. http://www.mpi.nl/lrec/2002/papers/lrec-pap-06-nick-speech.pdf

Campbell N (2006) A language-resources approach to emotion: corpora for the analysis of expressive speech. In: Proceedings of the LREC Workshop on Corpora for Research on Emotion and Affect, Genoa, pp 1–5

Chung S-J (2000) L'expression et la perception de l.éḿotion extraite de la parole spontaneé: eví́dences du coreéń et de l'anglais. Unpublished doctoral dissertation, Universite ´de la Sorbonne Nouvelle, Paris III

Cowie R, Douglas-Cowie E (1996). Automatic statistical analysis of the signal and prosodic signs of emotion in speech. Proceedings of the international conference on spoken language processing, Philadelphia, 1989–1992

Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human–computer interaction. IEEE Signal Process Mag 18(1):33–80

Cowie R, McGuiggan A, McMahon E, Douglas-Cowie E (2003) Speech in the process of becoming bored. Proceedings of 15th international congress of phonetic sciences, Barcelona

Devillers L, Cowie R, Martin J-C, Douglas-Cowie E, Abrilian S, McRorie M (2006) Real-life emotions in French and English TV video corpus clips: an integrated annotation protocol combining continuous and discrete approaches. In: Proceedings LREC 2006, Genoa

Devillers L, Vasilescu I (2004) Reliability of lexical and prosodic cues in two real-life spoken dialog corpora. Proceedings LREC 2004, Las Palmas, Canary Islands

Douglas-Cowie E et al (2004) HUMAINE D5d, 2004 p7. http://emotion-research.net/projects/humaine/deliverables/D5d%20potential%20exemplars%20databases.pdf

Douglas-Cowie E, Campbell N, Cowie R, Roach P (2003) Emotional speech: towards a new generation of databases. Speech Commun 40(1–2):33–60

Douglas-Cowie E, Cowie R, Schroeder M (eds) (2000) In: Proceedings of the ISCA workshop on speech and emotion, Belfast

Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, McRorie M, Martin J-C, De-villers L, Abrilian S, Batliner A, Amir N, Karpouzis K (2007) The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In: Proceedings of the ACII 2007, Lisbon, pp 488–500

Ekman P (1992) An argument for basic emotions. Cogn Emot 6:169–200

Ekman P, Friesen W (1975) Pictures of facial affect. Consulting Psychologists' Press, Palo Alto, CA

Engberg IS, Hansen AV, Andersen O, Dalsgaard P (1997) Design, recording and verification of a Danish emotional speech database. EUROSPEECH-1997, 1695–1698

Fernandez R, Picard R (2003) Modeling drivers' speech under stress. Speech Commun 40:145–159

Fragopanagos N, Taylor J (2005) Emotion recognition in human–computer interaction. Neural Netw 18:389–405

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 Eng 47:7

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

Greasley P, Setter J, Waterman M, Sherrard C, Roach P, Arnfield S et al (1995) Representation of prosodic and emotional features in a spoken language database Proceedings XIIIth international congress of phonetic sciences, vol. 1. Stockholm, pp 242–245

Han JH, Ward AJI, Lavine BK (2005) The problem of adequate sample size in pattern recognition studies: the multivariate normal case. J Chemom 4(1):91–96

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

Heylen D, Nijholt A, Reidsma D (2006) Determining what people feel and think when interacting with humans and machines: notes on corpus collection and annotation. In: Kreiner J, Putcha C (eds) Proceedings 1st California conference on recent advances in engineering mechanics. California State University, Fullerton, January 12–14, pp 1–6

Hönig F (2007) DRIVAWORK – driving under varying workload. A multi-modal stress database in the automotive context. Vortrag: HUMAINE Plenary Meeting, Paris, 6 Jun 2007

Huttar GL (1968) Relations between prosodic variables and emotions in normal American English utterances. J Speech Hear Res 11:481–487

Iriondo I, et al (2000) Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. Proceedings of the ISCA workshop on speech and emotion, Belfast 2000, pp 161–166

Johannes B, Salnitski V, Gunga H-C, Kirsch K (2000) Voice stress monitoring in space – possibilities and limits. Aviat Space Environ Med 71(9):A58–A65 (section II)

Juslin P, Laukka P (2002) Communication of emotions in vocal expression and music performance. Psychol Bull 129(5):770–814

Kawai H, Toda T, Ni J, Tsuzaki M, Tokuda K (2004) XIMERA: a new TTS from ATR based on corpus-based technologies. In: Proceedings of the 5th ISCA ITRW on speech synthesis, Pittsburgh, PA, pp 179–184

Kienast M, Sendlmeier WF (2000) Acoustical analysis of spectral and temporal changes in emotional speech. In: Cowie R, Douglas-Cowie E, Schroeder M (eds) Speech and emotion: proceedings of the ISCA workshop. Newcastle, County Down, Sept 2000, Belfast, Textflow, pp 92–97

Kominek J, Black AW (2004) The CMU ARCTIC speech databases. In: Proceedings of the 5th ISCA ITRW on speech synthesis, Pittsburgh, PA, pp 223–224

Kuroda I, Fujiwara O, Okamura N, Utusuki N (1979) Method for determining pilot stress through analysis of voice communication. Aviat Space Environ Med 47:528–533

Lee C, Narayanan S (2003) Emotion recognition using a data-driven fuzzy inference system. In: Proceedings Eurospeech 2003, Geneva

Leinonen L, Hiltunen T (1997) Expression of emotional–motivational connotations with a one-word utterance. J Acoust Soc Am 102(3):1853–1863

Lipi AA, Yamaoka Y, Rehm M, Nakano Y (2008) Enculturating conversational agents based on a comparative corpus study. In: Prendinger H, Lester J, Ishizuka M (eds) Intelligent virtual agents, Springer

McGilloway S (1997) Negative symptoms and speech parameters in schizophrenia. Unpublished doctoral thesis, Queen’s University, Belfast

McMahon E, Cowie R, Kasderidis S, Taylor J, Kollias S (2003) What chance that a DC could recognise hazardous mental states from sensor outputs? In: Proceedings of the DC Tales conference, Sanotrini, Jun 2003

Mitchell CJ, Menezes C, Williams JC, Pardo B, Erickson D, Fujimura O (2000) Changes in syllable and boundary strengths due to irritation. In: Douglas-Cowie E, Cowie R, Schroder M (eds) Proceedings of the ISCA Workshop on Speech and Emotion, Belfast, Textflow, pp 98–103.

Mozziconacci S (1998) Speech variability and emotion: production and perception. Unpublished doctoral thesis, Technical University Eindhoven, Eindhoven

Murphy C (2002) Automatic recognition of spoken emotion using audio signal processing. Unpublished undergraduate thesis, Department of Electrical and Electronic Engineering, University College, Dublin

Nakatsu R, Tosa N, Nicholson J (1999) Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In: Proceedings of the IEEE workshop on multimedia signal processing, Copenhagen, pp 439–444

Nogueiras A, Moreno A, Bonafonte A, Marinõ J (2001) Speech emotion recognition using hidden Markov models. In: Proceedings of the Eurospeech 2001, Aalborg, Denmark

Ortony A, Clore G, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge, MA

Oudeyer P-Y (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum Comput Interact 59(1–2):157–183

Paeschke A, Sendlmeier WF (2000) Prosodic characteristics of emotional speech: measurements of fundamental frequency movements. In: Proceedings of the ISCA workshop on speech and emotion, Textflow, Belfast, 5–7 Sept 2000, pp 75–80

Pereira C (2000) Dimensions of emotional meaning in speech. Proceedings of the ISCA workshop on speech and emotion, Newcastle, Co. Down. Belfast, Textflow, pp 25–28

Polzin TS, Waibel A (2000) Emotion-sensitive human–computer interfaces. In: Douglas-Cowie E, Cowie R, Schroeder M (eds) Proceedings of the ISCA workshop on speech and emotion. Belfast, Textflow, pp 201–206

Raudys SJ, Jain AK (1991) Sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 13(3):252–264

Roach P (2000) Techniques for the phonetic description of emotional speech. In: Douglas-Cowie E, Cowie R, Schroeder M (eds) Proceedings of the ISCA workshop on speech and emotion, 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 JW (1976) Voice predicts affect during psychotherapy. J Nerv Ment Dis 163(3):166–176, Sep 1976

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, NY

Rosch E (1978) Principles of categorization. In: Rosch E, Lloyd, BB (eds) Cognition and categorization. Lawrence Erlbaum Associate, Hillsdale, NJ

Sak H (2000) A corpus-based concatenative speech synthesis system for Turkish. B.S thesis in computer engineering and information science, Bilkent University

Scherer KR, Ceschi G (1997) Lost luggage emotion: a field study of emotion-antecedent appraisal. Motivation and Emotion 21(3):211–235

Scherer KR, Ceschi G (2000) Studying affective communication in the airport: the case of lost baggage claims. Pers Soc Psychol Bull 26(3):327–339

Schiel F, Steininger S, Türk U (2002) The SmartKom multimodal corpus at BSA. In: Proceedings of the LREC 2002, Las Palmas, Gran Canaria, pp 200–206

Steininger S, Schiel F, Dioubina O, Raubold S (2002a) Development of user-state conventions for the multimodal corpus in SmartKom. In: Proceedings of the workshop ‘Multimodal Resources and Multimodal Systems Evaluation’ 2002, Las Palmas, Gran Canaria, pp 33–37

Steininger S, Schiel F, Glesner A (2002) Labeling procedures for the multi-modal data collection of SmartKom. In: Proceedings of the LREC 2002, Las Palmas, Gran Canaria

Stibbard R (2001) Vocal expression of emotions in non-laboratory speech. Unpublished doctoral dissertation, University of Reading, Reading, UK

Sulc J (1977) To the problem of emotional changes in the human voice. Act Nerv Super 19:215–216

Sutherland NS (2007) Irrationality: why we don’t think straight reissued. Pinter & Martin, London

Tabachnick BG, Fidell LS (2001) Using multivariate statistics. Allyn & Bacon, Boston, MA

ten Bosch L (2000) Emotions: what is possible in the ASR framework. In: Proceedings ISCA workshop on speech and emotion, Newcastle

Tolkmitt F, Scherer KR (1986) Effect of experimentally induced stress on vocal parameters. Exp Psychol Hum Percept Perform 12(3):302–313

van Bezooijen R (1984) The characteristics and recognizability of vocal expression of emotions. Foris, Dordrecht

Walker MA, Passonneau R, Boland JE (2001) Quantitative and qualitative evaluation of Darpa Communicator spoken dialogue systems. In: Proceedings of the 39th annual meeting on association for computational linguistics, Toulouse, France, pp 515–522

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

Yacoub S, Simske S, Lin X, Burns J (2003) Recognition of emotions in interactive voice response systems. In: Proceedings of the Eurospeech 2003, Geneva