The Composite Sensing of Affect

Gordon McIntyre[1] and Roland Göcke[1,2]

[1] Research School of Information Sciences and Engineering,
Australian National University, Canberra, Australia
[2] Seeing Machines, Canberra, Australia

Abstract. This paper describes some of the issues faced by typical emotion recognition systems and the need to be able to deal with emotions in a natural setting. Studies tend to ignore the dynamic, versatile and personalised nature of affective expression and the influence that social setting, context and culture have on its rules of display. Affective cues can be present in multiple modalities and they can manifest themselves in different temporal order. Thus, fusing the feature sets is challenging. We present a composite approach to affective sensing. The term composite is used to reflect the blending of information from multiple modalities with the available semantic evidence to enhance the emotion recognition process.

1 Introduction

Recognising emotions from the modulations in another person's voice and facial expressions is perhaps one of our most important human abilities. Such interaction is inherently multimodal and for computers to adapt and respond in a natural, yet robust, manner in real-world situations demands a similar capability. This is a great challenge. *Affective sensing* is the neologism used to describe recognition of emotional cues by machines. It is the process of mapping measurable physical responses to affective states. Several studies have successfully mapped strong responses to episodic emotions such as happiness, anger and surprise. However, few studies deal with the more subtle emotions such as anxiety and depression and most research takes place in a controlled environment, ignoring the importance that social settings, culture and context play in dictating the display rules of affect.

At present, reported examples of affective sensing systems tend to be very application specific [1,2,3,4,5,6,7,8,9,10,11]. However, in a natural setting, emotions can present themselves in many ways, and in different combinations of modalities. Thus it seems that some level of semantic incorporation is essential. For instance, during a diplomatic exchange, anger is more likely to be signaled through verbal content than, say, in an incident during a football game where a player remonstrates wildly with the referee. In this paper, a novel approach is presented which integrates semantic descriptions with standard speech recognition and computer vision feature sets.

C. Peter and R. Beale (Eds.): Affect and Emotion in HCI, LNCS 4868, pp. 104–115, 2008.

The remainder of the paper is structured as follows. Section 2 discusses the physiology of emotional display. Section 3 gives a brief overview of the recognition of emotions by machines. It also motivates the discussion of the limitations in current emotion recognition due to inheriting much of its techniques from automatic speech recognition (ASR) technology. Section 4 describes how we might add semantics to the emotion recognition process. Finally, Section 5 presents conclusions and future work.

2 The Physiology of Emotions in Speech

Age, gender, culture, social setting, personality and well-being all play their part in suffusing our communication apparatus even before we begin to speak. Darwin raised the issue of whether it was possible to inhibit emotional expression [12]. This is a pertinent question in human emotion recognition and in emotion recognition by computer systems. Intentional or not, the voice and face are used in everyday life to judge verisimilitude in speakers.

2.1 Vocal Speech

Speech carries a great deal more information than just the verbal message. It can tell us about the speaker, their background and their emotional state. Changes in brain patterns result in modulations in our major anatomical systems.

Stress tenses the laryngeal muscles, in turn, tightening the vocal folds. The result is that more pressure is required to produce sound. Consequently, the fundamental frequency and amplitude, particularly with regard to the ratio of the open to the closed phase of the cycle, varies the larynx wave. The harmonics of the larynx wave vary according to the specific balance of mass, length and tension that is set up to produce a given frequency [13].

	fear	**anger**	**sorrow**	**joy**	**disgust**	**surprise**
speech rate	much faster	slightly faster	slightly slower	faster or slower	very much slower	much faster
pitch average	very much higher	very much higher	slightly lower	much higher	very much lower	much higher
pitch range	much wider	much wider	slightly narrower	much wider	slightly wider	
intensity	normal	higher	lower	higher	lower	higher
voice quality	irregular voicing	breathy chest tone	resonant	Breathy, blaring	grumble chest tone	
pitch changes	normal	abrupt on stressed syllable	downward inflections	smooth upward inflections	wide downward terminal inflections	rising contour
articulation	precise	tense	slurring	normal	normal	

Fig. 1. The effect of emotion on the human voice [14]

Some affective states like anxiety can influence breathing resulting in variations in sub-glottal pressure. Drying of the mucus membrane causes shrinking of the voice. Rapid breath alters the tempo of the voice. Relaxation tends to deepen the breath and lowers the voice. Changes in facial expression can also alter the sound of the voice. Figure 1 represents the typical cues to the six most common emotion categories [14].

2.2 Visual Speech

The most widely used system for explaining the facial expression of emotion is that of Ekman's Facial Action Coding System (FACS) [12] [15,16,17,18]. Facial muscles are mapped to "Action Units" that produce movement. The combinations of "Action Units" are mapped to emotional states. The changes associated with emotional expression are usually brief, i.e. a few seconds.

McNeill [19] has shown how tightly integrated and important a role gesture plays in speech. It often precedes vocal expression, exposes our inner thoughts and can disambiguate utterances. Gestures can be expressed through various body parts (e.g. hands, arms, head) as well as the entire body.

3 Recognition of Emotions by Machines

Affective sensing is an attempt to map manifestations or measurable physical responses to affective states. Non-obtrusive sensing of affect from the voice and facial expressions is commonly based on ASR technology and computer vision. ASR is concerned with the analysis of sound patterns, phonemes, words, sentences, and dialogues. However, when extended to the detection of emotions in vocal speech, the focus tends to be on prosody and energy levels.

Computer vision techniques to detect emotions from facial expressions are often used in conjunction with some codebook of muscle movements such as Ekman's FACS. FACS is typically used in conjunction with probabilistic models, e.g. Hidden Markov Models [20]. Several studies have used computer vision to detect features and build evidence of FACS Action Units [21] [22].

Several researchers have reported improved recognition of emotions when sensory cues from multiple modalities are fused [23]. In [24] facial features, prosody and lexical content in speech are fused. In his dissertation, Polzin used a similar technique, using separate, composite hidden Markov models to model each emotion [25].

However, ASR and computer vision approaches are grounded in pattern matching and statistical machines learning techniques. Hence, the premise is that samples of real world data can be matched against samples of test data. One inherent weakness in this premise, for emotion recognition, is in the elicitation method of the sample data. The topic of the elicitation of emotional speech samples has been well covered by other reviews [26,27,28], so it is only briefly covered in the next section.

3.1 Eliciting Emotional Speech Samples

Naturally Occurring Speech. To date, call centre recordings [10,29], recordings of pilot conversations, and television reports [30] have provided sensible sources of data to research emotions in speech. These types of samples have the highest ecological validity. However, aside from the copyright and privacy issues, it is very difficult to construct a database of emotional speech from this sort of naturally occurring emotional data sources. In audio samples, there are the complications of background noise and overlapping utterances. In video, there are difficulties in detecting moving faces and facial expressions. A further complication is the suppression of emotional behaviour by the speaker who is aware of being recorded.

Induced Emotional Speech. One technique introduced by Velten [31], is to have subjects read emotive texts and passages which, in turn, induce emotional states in the speaker. Other techniques include the use of Wizard of Oz setups where, for example, a dialogue between a human and a computer is controlled without the knowledge of the human [32]. This method has the benefit of providing a degree of control over the dialogue and can simulate a natural setting. The principal shortcoming of these methods is that the response to stimuli may induce different emotional states in different people.

Acted Emotional Speech. By far the most popular approach is to engage actors to portray emotions [11,33,34]. This technique provides for a lot of experimental control over a range of emotions and like the previous method provides for a degree of control over the ambient conditions.

One problem with this approach is that acted speech elicits how emotions should be portrayed, not necessarily how they are portrayed. The other serious drawback is that acted emotions are unlikely to derive from emotions in the way that Scherer *et al.* [35] describe them, i.e. episodes of massive, synchronised recruitment of mental and somatic resources to adapt or cope with a stimulus event subjectively appraised as being highly pertinent to the needs, goals and values of the individual.

3.2 Discussion on the Elicitation Methods

We display emotions in an extemporaneous symphony of modalities and with insouciant ease. Some of us are Rembrandts in concealing and revealing our feelings. Cultural, social, physiological, and contextual factors dictate the display rules of emotions. Yet, as implied in the last section, few studies ever take these factors into account. In computer science, we like to hold certain variables constant in order to find ways of explaining the change in the others. In this case, the variables that are held constant are the most important ones that contribute to the selection and production of affect.

Relatively little research into affect has been based on natural speech. In many cases, the approach to affect recognition has simply been an extension of ASR, i.e. acquiring a corpus of acted speech, then annotating sequences containing affect within the corpus. In the case of automatic recognition of episodic emotions, this approach is plausible, based on the assumption that clear-cut bursts of episodic emotion will look and sound somewhat similar in most contexts [28]. However, recognition of pervasive emotions present a much greater challenge and, intuitively, one would think that awareness of personal and contextual information needs to be integrated into the recognition process.

Fernandez and Picard [36] used eighty-seven features and concluded that the recognition rate was still below human performance. One would have to question how much extrapolation it would take to extend the ASR approach to affective sensing in a natural setting. Studies by Koike *et al.* [37] and Shigeno [38] have shown that it is difficult to identify the emotion of a speaker from a different culture and that people will predominantly use visual information to identify emotion. The implications are that the number of feature sets and the amount of training samples required to take into account natural, social, physiological, and contextual factors would be infeasible.

Richard Stibbard [39] who undertook the, somewhat difficult, Leeds Emotion in Speech Project reported,

"The use of genuine spoken data has revealed that the type of data commonly used gives an oversimplified picture of emotional expression. It is recommended that future work cease looking for stable phonetic correlates of emotions and look instead at dynamic speech features, that the classification of the emotions be reconsidered, and that more account be taken of the complex relationship between eliciting event, emotion, and expression."

In keeping with speech recognition, much of the effort to date in emotion recognition has been concerned with finding the low-level, symbolic representation and interpretation of the speech signal features. Only a handful of reports involve real-time facial feature extraction in the emotion recognition process [40] [30]. Similar points about the need to recognise emotions in natural settings, and the difficulties of doing so, were made by [41]. To address this deficiency, some level of semantic reasoning seems essential.

4 Adding Semantics to the Emotion Recognition Process

There have been some attempts at representing real-life emotions in audio-video data with non-basic emotional patterns and context features [42] [43]. [44] have shown that recognition of speech can be improved by combining a dictionary of affect with the standard ASR dictionary. [45] have developed a rule-based system for interpreting facial expressions. This recent activity in the field suggests that the incorporation of some level of semantic reasoning in the recognition process is now seen by many as a necessary evolution.

Some systems have incorporated elaborate syntax checking rules but there are fewer examples where semantics within a domain of interest has been used. Speech processing and computer vision techniques were discussed previously. An important distinction between the two is that visual information is inherently

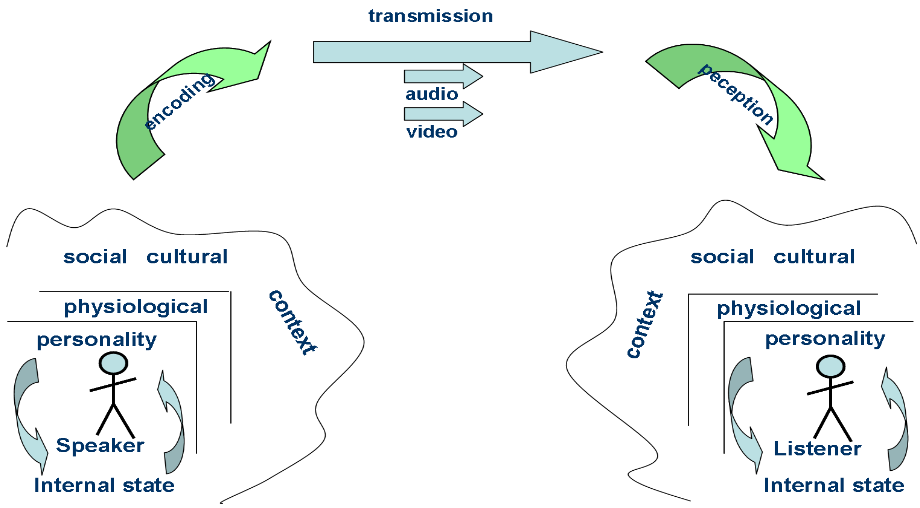

Fig. 2. A generic model of affective communication

more ambiguous and semantically impoverished [46]. The currently available computer vision techniques are still no match with human interpretation of images. However, by combining modalities with other available semantic evidence it could be possible to enhance not only the emotion recognition process but the recognition of speech.

The proposed approach consists of a generic model of affective communication and a domain *ontology* of affective communication. The model and ontology are intended to be used in conjunction as a standardised way to describe the content.

4.1 A Model for Affective Communication

Figure 2 presents a model of emotions in spoken language. Firstly, note that it includes speaker and listener, in keeping with the Brunswikian lens model as proposed by Scherer [26]. The reason for modelling attributes of both speaker and listener is that the listener's cultural and social presentation vis-ã-vis the speaker may also influence judgement of emotional content. Secondly, note that it includes a number of factors that influence the expression of affect in spoken language. A brief description of the components of the model follows.

Context is linked to modality and emotion is strongly multimodal in the way that certain emotions manifest themselves favouring one modality over the other [28]. **Physiological** measurements change depending on whether a subject is sedentary or mobile. A stressful context such as an emergency hot-line, air-traffic control, or a war zone is likely to yield more examples of affect than everyday conversation.

Agent characteristics such as facial hair, whether a person wears spectacles, and their head and eye movements all affect the ability to visually detect and interpret emotions. As Scherer [26] points out, most studies are either speaker oriented or listener oriented, with most being the former. This is significant when you consider that the emotion of someone labelling affective content in a corpus could impact the label that is ascribed to a speaker's message.

Culture-specific display rules influence the display of affect [28]. Gender and age are established as important factors in shaping conversation style and content in many societies.

It might be stating the obvious but there are marked differences in speech signals and facial expressions between people of different **physiological** make up, e.g. age, gender and health. The habitual settings of facial features and vocal organs determine the speaker's range of possible visual appearances and sounds produced. The configuration of facial features, such as chin, lips, nose, and eyes, provide the visual cues, whereas the vocal tract length and internal muscle tone guide the interpretation of acoustic output [47].

Social factors temper spoken language to the demands of civil discourse [28]. For example, affective bursts are likely to be constrained in the case of a minor relating to an adult, yet totally unconstrained in a scenario of sibling rivalry. Similarly, a social setting in a library is less likely to yield loud and extroverted displays of affect than a family setting.

Internal state has been included in the model for completeness. At the core of affective states is the person and their experiences. Recent events such as winning the lottery or losing a job are likely to influence emotions.

4.2 An Application Ontology for Affective Communication

An ontology is a statement of concepts which facilitates the specification of an agreed vocabulary within a domain of interest. Creating an ontology introduces a common way of laying down the knowledge and facilitates intelligent searching and reuse of knowledge within the domain. Ontologies have been used for some time in the annotation of web pages and in the medical fields. In its simplest form it is a hierarchical database of definitions. In a more complex setup, it is a sophisticated knowledge base with embedded logic and semantic constraints.

Figure 3 shows an example application ontology for affective communication in a context of investigating dialogues. During the dialogue, various events can occur, triggered by one of the dialogue participants and recorded by the sensor system. These are recorded as time stamped instances of events, so that they can be easily identified and distinguished. In this ontology, we distinguish between two roles for each interlocutor: sender and receiver, respectively. At various points in time, each interlocutor can take on different roles. On the sensory side, we distinguish between facial, gestural, textual, speech, physiological and verbal[1] cues. This list, and the ontology, could be easily extended for other cues

[1] The difference between speech and verbal cues here being spoken language versus other verbal utterings.

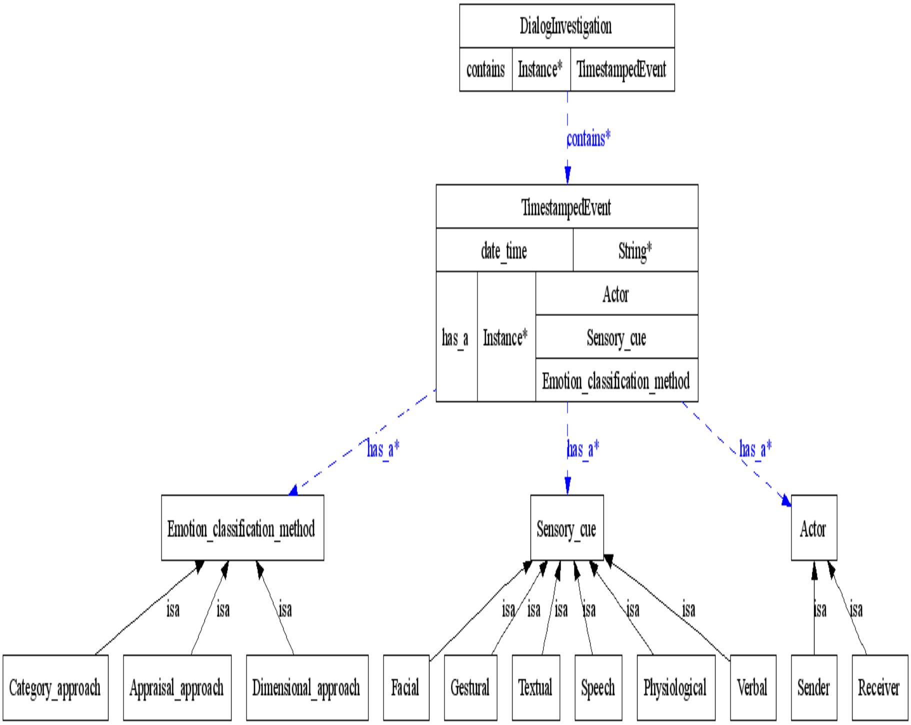

Fig. 3. An application ontology for affective sensing

and is meant to serve as an example here, rather than a complete list of affective cues. Finally, the emotion classification method used in the investigation of a particular dialogue is also recorded.

We use this ontology to describe our affective sensing research in a formal, yet flexible and extendible way. In the following section, a brief description of the facial expression recognition system developed in our group is given as an example of using the ontologies in practice.

4.3 Describing Semantics

One of the issues in emotion recognition, is that of reuse and verification of results. However, there is no universally accepted system of describing emotional content. The HUMAINE project is trying to remedy this through the definition of the Emotion Annotation and Representation Language (EARL) which is currently under design [48,49].

Another direction is that of the Moving Picture Experts Group (MPEG) who have developed the MPEG-7 standard for audio, audio-video and multimedia description [50]. MPEG-7 uses metadata structures or Multimedia Description Schemes (MDS) for describing and annotating audio-video content. These are provided as a standardised way of describing the important concepts in content description and content management in order to facilitate searching, indexing, filtering, and access. They are defined using the MPEG-7 Description Definition Language (DDL), which is XML Schema-based. The output is a description expressed in XML which can be used for editing, searching, filtering. The standard also provides a description scheme for compressed binary form for storage or transmission [51] [52] [53]. Examples in the use of MPEG-7 exist in the video surveillance industry where streams of video are matched against descriptions of training data [54]. The standard also caters for the description of affective content. Although it is a fairly modest offering, however, the standards are made to be extensible.

5 Conclusions and Future Work

The incorporation of the semantics of affective communication within a machine-processable ontology is expected to enhance the effectiveness of affective sensing systems. We have presented some of the issues in collecting emotional samples and the need for emotion recognition systems to be able to deal with genuine spoken data.

We have presented a framework for fusing background information (context, social, culture, agent characteristics, physiology, internal state), with the more traditional feature that describe an individual's emotional state. The framework consists of a generic model of affective communication to be used in conjunction with a domain ontology.

In future work, we intend to demonstrate the composite sensing of affect from multimodal cues and plan to include physiological sensors as another cue for determining the affective state of a user.

References

1. McCann, J., Peppe, S.: PEPS-C: A new speech science software programme for assessing prosody. In: The Fifth Annual Parliamentary Reception for Younger Researchers in Science, Engineering, Medicine and Technology (SET for Britain. Taking science to parliament: The 2003 great British research and R&D show), the House of Commons, London (2003)
2. Devillers, L., Vasilescu, I., Vidrascu, L.: F0 and pause features analysis for anger and fear detection in real-life spoken dialogs. Speech Prosody (2004)
3. Jones, C.M., Jonsson, I.: Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. Technical report, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK and Department of Communication. Stanford University, California, USA (2005)
4. Jones, C.M., Jonsson, I.: Using paralinguistic cues in speech to recognise emotions in older car drivers. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868. Springer, Heidelberg (2008)
5. Breazeal, C.: Emotion and sociable humanoid robots. Int. J. Human-Computer Studies 59, 119–155 (2003)

6. Reilly, R., Moran, R., Lacy, P.: Voice pathology assessment based on a dialogue system and speech analysis. Technical report, Department of Electronic and Electrical Engineering, University College Dublin, Ireland and St James's Hospital, Dublin 8, Ireland (2000)
7. Picard, R.: Helping addicts: A scenario from 2021. Technical report (2005)
8. Kaliouby, R., Robinson, P.: Therapeutic versus prosthetic assistive technologies: The case of autism. Technical report, Computer Laboratory, University of Cambridge (2005)
9. Kaliouby, R., Robinson, P.: The emotional hearing aid: An assistive tool for children with asperger's syndrome. Technical report, Computer Laboratory, University of Cambridge (2003)
10. Petrushin, V.A.: Emotion in speech: Recognition and application to call centres. In: Artificial Neural Networks in Engineering (1999)
11. Yacoub, S., Simske, S.: X.Lin, Burns, J.: Recognition of emotions in interactive voice response systems. Technical report, HP Laboratories Palo Alto (2003)
12. Ekman, P.: Darwin, deception, and facial expression. Annals New York Academy of Sciences, 205–221 (2003)
13. Fry, D.B.: The Physics of Speech. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge (1979)
14. Murray, I., Arnott, L.: Toward the simulation of emotion in synthetic speech. Journal Acoustical Society of America 93(2), 1097–1108 (1993)
15. Ekman, P., Friesen, W.: Unmasking the Face. Prentice Hall, Englewood Cliffs (1975)
16. Ekman, P., Oster, H.: Emotion in the human face, 2nd edn. Cambridge University Press, New York (1982)
17. Ekman, P., Rosenberg, E.L.: What the Face Reveals. Series in Affective Science. Oxford University Press, Oxford (1997)
18. Ekman, P.: Facial Expressions. In: The Handbook of Cognition and Emotion, pp. 301–320. John Wiley and Sons, Ltd., Sussex (1999)
19. McNeill, D.: Gesture and language dialectic. Technical report, Department of Psychology. University of Chicago (2002)
20. Lien, J., Kanade, T., Cohn, J., Li, C.: Automated Facial Expression Recognition Based on FACS Action Units. In: International Conference on Automatic Face and Gesture Recognition, pp. 390–395 (1998)
21. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models - their training and applications. Computer Vision and Image Understanding 61(1), 38–59 (1995)
22. Nixon, M., Aguado, A.: Feature Extraction and Image Processing. MPG Books Ltd., Brodmin, Cornwall (2001)
23. Castellano, G., Kessous, L., Caridakis, G.: Emotion recognition through multiple modalities: face, body gesture, speech. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction, vol. 4868. Springer, Heidelberg (2007)
24. Fragopanagos, N., Taylor, J.: Emotion recognition in human-computer interaction. Neural Networks 18, 389–405 (2005)
25. Polzin, T.: Detecting verbal and non-verbal cues in the communication of emotions. PhD thesis, School of Computer Science. Carnegie Mellon University (2000)
26. Scherer, K.R.: Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256 (2003)
27. Cowie, R., Cornelius, R.: Describing the emotional states that are expressed in speech. Speech Communication 40, 5–32 (2003)
28. Cowie, R., Douglas-Cowie, E., Cox, C.: Beyond emotion archetypes: Databases for emotion modelling using neural networks. Neural Networks 18, 371–388 (2005)

29. Lee, C.M., Narayanan, S., Pieraccini, R.: Recognition of negative emotions from the speech signal. Automatic Speech Recognition and Understanding (2001)
30. Devillers, L., Abrilian, S., Martin, J.: Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. Technical report, LIMSI, Centre national de la recherche scientifique, France (2005)
31. Velten, E.: A laboratory task for induction of mood states. Behaviour Research and Therapy 6, 473–482 (1968)
32. Schiel, F., Steininger, S., Türk, U.: The SmartKom Multimodal Corpus at BAS. Technical report, Ludwig Maximilans Universität München (2003)
33. Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotion in speech. Technical report, School of Computer Science. Carnegie Mellon University (1995)
34. Lin, Y.L., Wei, G.: Speech emotion recognition based on hmm and svm. In: Proceedings (2005)
35. Scherer, K.R.: Humaine Deliverable D3c: Preliminary plans for exemplars: theory (2004). Retrieved October 26, 2006, `http://emotion-research.net/publicnews/d3c/`
36. Fernandez, R., Picard, R.: Classical and novel discriminant features for affect recognition from speech. In: Interspeech, Lisbon, Portugal, pp. 473–476 (2005)
37. Koike, K., Suzuki, H., Saito, H.: Prosodic parameters in emotional speech. In: International Conference on Spoken Language Processing, pp. 679–682 (1998)
38. Shigeno, S.: Cultural similarities and differences in the recognition of audiovisual speech stimuli. In: International Conference on Spoken Language Processing, 1057th edn., pp. 281–284 (1998)
39. Stibbard, R.: Vocal expression of emotions in non-laboratory speech: An investigation of the Reading/Leeds Emotion in Speech Project annotation data. PhD thesis, University of Reading, UK (2001)
40. Silva, L.D., Hui, S.: Real-time facial feature extraction and emotion recognition. In: ICICS-PCM. IEEE, Singapore (2003)
41. Ward, R., Marsden, P.: Affective computing: Problems, reactions and intentions. Interacting with Computers 16(4), 707–713 (2004)
42. Liscombe, J., Riccardi, G., Hakkani-Tür, D.: Using context to improve emotion detection in spoken dialog systems. In: EUROSPEECH 2005, 9th European Conference on Speech Communication and Technology, pp. 1845–1848 (2005)
43. Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422 (2005)
44. Athanaselisa, T., Bakamidisa, S., Dologloua, I., Cowieb, R., Douglas-Cowie, E., Cox, C.: Asr for emotional speech: Clarifying the issues and enhancing performance. Neural Networks 18, 437–444 (2005)
45. Cowie, R., Douglas-Cowie, E., Taylor, J., Ioannou, S., Wallace, M., Kollias, S.: An intelligent system for facial emotion recognition. IEEE, Los Alamitos (2005)
46. Town, C., Sinclair, D.: A self-referential perceptual inference framework for video interpretation. In: Crowley, J.L., Piater, J.H., Vincze, M., Paletta, L. (eds.) ICVS 2003. LNCS, vol. 2626, pp. 54–67. Springer, Heidelberg (2003)
47. Millar, J.B., Wagner, M., Göcke, R.: Aspects of speaking-face data corpus design methodology. In: International Conference on Spoken Language Processing 2004, Jeju, Korea, vol. II, pp. 1157–1160 (2004)
48. Schröder, M., Devillers, L., Karpouzis, K., Martin, J.C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., Wilson, I.: What should a generic emotion markup language be able to represent? In: Proc. 2nd International Conference on Affective Computing and Intelligent Interaction (ACII 2007), Lisbon, Portugal (2007)

49. Schröder, M., Zovato, E., Pirker, H., Peter, C., Burkhardt, F.: W3c emotion incubator group final report. Technical report, W3C (2007)
50. MPEG-7 Committee: Retrieved June 2, 2007, http://www.m4if.org/m4if/
51. Chiariglione, L.: Introduction to MPEG-7: Multimedia Content Description Interface. Technical report, Telecom Italia Lab, Italy (2001)
52. Salembier, P., Smith, J.: MPEG-7 Multimedia Description Schemes. IEEE Transactions on Circuits and Systems for Video Technology 11, 748–759 (2001)
53. Rege, M., Dong, M., Fotouhi, F., Siadat, M., Zamorano, L.: Using MPEG-7 to build a Human Brain Image Database for Image-guided Neurosurgery. Medical Imaging 2005: Visualization, Image-Guided Procedures, and Display, 512–519 (2005)
54. Annesley, J., Orwell, J.: On the Use of MPEG-7 for Visual Surveillance. Technical report, Digital Imaging Research Center, Kingston University, Kingston-upon-Thames, Surrey, UK (2005)