Unobtrusive Multimodal Emotion Detection in Adaptive Interfaces: Speech and Facial Expressions

Khiet P. Truong, David A. van Leeuwen, and Mark A. Neerincx

TNO Human Factors, Dept. of Human Interfaces, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
{khiet.truong, david.vanleeuwen, mark.neerincx}@tno.nl

Abstract. Two unobtrusive modalities for automatic emotion recognition are discussed: speech and facial expressions. First, an overview is given of emotion recognition studies based on a combination of speech and facial expressions. We will identify difficulties concerning data collection, data fusion, system evaluation and emotion annotation that one is most likely to encounter in emotion recognition research. Further, we identify some of the possible applications for emotion recognition such as health monitoring or e-learning systems. Finally, we will discuss the growing need for developing agreed standards in automatic emotion recognition research.

Keywords: emotion detection, emotion recognition, classification, speech, facial expression, emotion database.

1 Introduction

In human-human communication, we give emotional signals to each other all the time: when we talk and when we write (e.g., through emoticons), we want to create a feeling of mutual understanding and share our feelings and intentions with each other. Emotions are part of human nature. Machine's inability to feel is one of the main reasons why communications and interactions between humans and machines fail. Therefore, researchers have been trying to automatically detect emotions in order to improve human-machine interaction. In this way, interfaces can be designed to adapt to the user's emotions: for example, a computer assisted language learning system may sense the frustration in the student's tone of voice and facial expressions, and may decide to lower the level of difficultness or to switch to another exercise in order to keep the student motivated. When we talk to each other face to face, we can see emotions expressed in face, body gestures etc. and we can hear emotions expressed in vocal sounds. Facial expressions and speech are considered to be very accessible, visible and non-obtrusive modalities and therefore, we will focus on these two channels.

The term 'emotion' is a term that can have many senses and interpretations. Other terms that can be used to refer to 'emotion' are 'affective', 'expressive', 'emotional state' or 'mood'. We will use 'affective', 'expressive' and 'emotional state' interchangeably with 'emotion'. 'Mood' on the other hand is usually described as an emotion that can last

D.D. Schmorrow, L.M. Reeves (Eds.): Augmented Cognition, HCII 2007, LNAI 4565, pp. 354–363, 2007.

for a longer period of time. In short, we will continue to use the term ‘emotion’ in its broad sense, meaning that we will use ‘emotion’ to describe a broad range of feelings that humans can have and express and which can influence humans in their behavior [1].

This paper is structured as follows: Section 2 gives an overview of emotion recognition studies that have been carried out on speech and facial expressions. In Section 3, we will elaborate on some general difficulties in emotion research. Section 4 describes possible real-life emotion recognition applications. And finally, we conclude this paper with a discussion and general conclusions in Section 5.

2 Short Overview: State of the Art

Emotions can be measured through different modalities. Usually, physiological measures such as heart rate or skin conductivity (e.g., [2, 3, 4]) are considered obtrusive, while speech and facial expressions are relatively non-obtrusive measures. Therefore, the focus will be on emotional analyses of speech and facial expressions.

2.1 Automatic Emotion Recognition from Speech

By making variations in the melody of an utterance, by changing the speaking rate or by changing the loudness etc., humans can produce emotional speech which seems to be a prerequisite for effective human-human communication. In voice-based automatic emotion recognition, we are more interested in *how* words are spoken rather than *what* words are spoken (although knowing *what* words are spoken may also help emotion recognition, e.g., swear words). In the course of time, many studies have investigated voice-based emotion recognition (see Table 1). The acoustic-phonetic correlates of emotional speech have been exhaustively investigated (e.g., [5-9]). Based on previous studies we can observe that often prosody-related features are used such as statistics of F0 (fundamental frequency of speech), statistics of intensity, speech rate, F0/intensity contour and duration. Further, quality-related speech features such as Mel-Frequency Cepstrum Coefficients (MFCC), Hammarberg Index, centre of spectral gravity, the energy distribution in the spectrum, jitter and shimmer are also frequently used. The emotions are modeled through these features and a modeling technique; frequently used modeling techniques include Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Neural Networks (NN) and K-Nearest Neighbors (KNN). Further, note the small number of subjects used in most of these studies. In order to perform subject-independent classification experiments and to make reliable statements about the results, more subjects should be used.

The acquisition and the use of realistic emotional speech data in emotion recognition remain challenges. Most of the studies in Table 1 have used acted or semi-spontaneous speech data that is elicited through a Wizard-of-Oz experiment (subjects interacting with a system, not knowing that the system is actually being operated by a human being). For speech analysis, a clean speech signal is preferred, i.e., background noise, overlap or crosstalk in speech, clipping etc. should be avoided. However, the more realistic the setting is in which the data is acquired, the harder it is to avoid noisy data (see Fig. 1).

Table 1. Short overview of automatic emotion recognition studies based on speech (where data can be acted real or obtained via woz=Wizard Of Oz, SI=SubjectIndependent, SD=SubjectDependent)

Study	Data	SI/ SD	Speech features	Method+Accuracy
Banse 1996 [5]	14 emotions, 12 subjects (acted)	?	F0, energy, speech rate, long-term spectrum	LDA: 25-53%
Ang 2002 [6]	2 emotions (woz)	?	F0, energy, speech rate, duration, pauses, spectral tilt	Decision tree: 75%
Nwe 2003 [7]	6 emotions, 12 subjects (acted)	SD	LFPC (Log Frequency Power Coefficients)	HMM: 77%-89%
Vidrascu 2005 [8]	2 emotions, 404 subjects (real)	?	F0, energy, duration, spectral features, disfluency	SVM: 83%
Batliner 2005 [9]	4 emotions, 51 subjects (woz)	SI	F0, energy, duration	LDA: 78%

In addition to finding acoustic profiles for basic emotions, researchers have been investigating acoustic correlates of emotion dimensions. The two most frequently used dimensions in the emotional space are that of activation (or arousal, active-passive) and evaluation (valence, positive-negative). Acoustic correlates found on the activation scale are much stronger than the correlates found on the evaluation scale. It seems to be much more difficult to describe negativity or positivity in terms of acoustic features.

In summary, although the classification results in Table 1 show accuracies that are well above chance, they are based on artificial conditions and therefore, we must conclude that automatic emotion recognition in speech is still in its development phase.

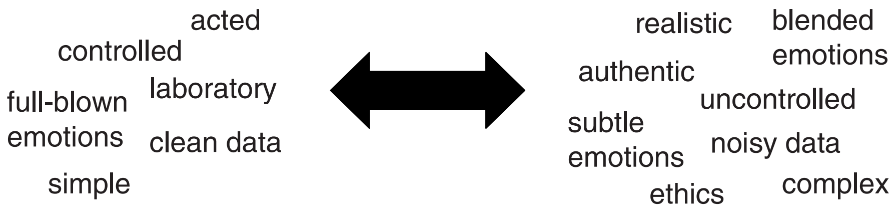

Fig. 1. Emotion research: from laboratory to real-life

2.2 Automatic Emotion Recognition from Facial Expressions

The movements of certain (combinations of) landmarks in the face can reveal much about the expressed emotions in the face: e.g., raised eyebrows typically indicate surprise, frowned eyebrows are usually used to express anger or dislike and smiles are usually characterized by an upward lip movement. The process of automatically recognizing a facial expression can be divided into three sub-processes: 1) detection of the face, 2) feature detection and extraction and 3) classification of emotions. The best known method for facial data extraction is the facial action coding system (FACS, see [11]). The FACS system describes facial movements in terms of Action Units (AU). This system consists of a taxonomy of 44 AUs with which facial expressions can be described, and has attracted many researchers from the field of computer vision to develop automatic facial expression analyzers based on AUs.

In Table 2, we show a short summary of more recent facial expression recognition studies; for a more exhaustive overview, readers are referred to [12]. In many aspects, the approach and challenges in facial emotion recognition studies resemble voice-based emotion recognition studies, e.g., small number of emotions, acted data etc. Difficulties with noisy data include e.g., bad illumination or background issues, different head poses and movements, facial hair and glasses.

Table 2. Short overview of emotion recognition studies based on facial expressions (SI=SubjectIndependent, SD=SubjectDependent)

Study	**# Classes of Emotions/Database**	**SI/ SD**	**Facial features/modeling**	**Method+Accuracy**
Pantic 2000 [13]	7 emotions (posed)	?	Action Units	Hybrid: Fuzzy+NN: 91%
Cohen 2003 [14]	7 emotions, 210 subjects (posed)	SI	Motion Units	Tree-Augmented-Naïve Bayes: 73%
Sebe 2004 [15]	4 emotions, 28 subjects (realistic)	SI	Motion Units	KNN: 95%
Den Uyl 2005 [16]	7 emotions (posed)	?	Active Appearance Model	NN: 85%

2.3 Multimodal Automatic Emotion Recognition

An increasing number of researchers believe that multi-modality is a key factor in automatic emotion recognition. In one of the first bimodal emotion recognition studies, De Silva et al. [17] found that some emotions were better recognized by humans through the auditory modality than the visual modality, and vice versa: anger, happiness, surprise and dislike were more visual dominant, and sadness and fear were more audio dominant. Since some modalities may carry complementary information and since humans also make use of multimodal information, it seems natural to fuse different modalities, which may lead to higher classification accuracies.

In Table 3, we can observe that indeed classification accuracies increase when audiovisual information (AV) is used instead of individual audio (A) or video channels (V). Usually, we make a distinction between fusion on feature-level and decision-level. On feature-level, features from different modalities can be concatenated to each other to form one large *N*-dimensional feature vector. Feature selection techniques may then be used to remove redundant features. Fusion on decision-level means that the features of the different modalities are processed separately, and are fused when the separate classifiers give outputs/scores which are usually in terms of posterior probabilities or likelihoods. These scores are then subsequently fused by summing, or taking the product of the scores etc. Fusing classifiers and data streams is not straightforward; we will discuss in Section 3.3 the difficulties that may arise in the fusion process.

Other studies have not only used speech and facial expressions, but also other physiological measures such as skin response, heart rate etc. [2-4]. However, physiological measures are usually measured with specialized hardware and sensors that are attached to the body which can be perceived as obtrusive. In general, the use of certain multimodal features depends on the application that one has in mind and the allowed degree of obtrusiveness.

Table 3. Short overview of bimodal emotion recognition studies based on speech and facial expressions (A=audio, V=video, AV=audiovisual, SI=SubjectIndependent, SD=Subject Dependent)

Study	**Data**	**SI/ SD**	**Fusion method**	**Accuracy**
Chen 1998 [18]	6 emotions, 2 subjects (acted)	??	Feature-level (concatenation)	75% (A), 69% (V), 97% (AV)
De Silva 2000 [19]	6 emotions, 2 subjects (acted)	SD	Dominant rule-based fusion	72% (AV)
Sebe 2006 [20]	11 emotions, 38 subjects (acted)	SD	Feature-level, Bayesian topology	45% (A), 56% (V), 89% (AV)

3 General Difficulties in Automatic Emotion Recognition

Apart from specific speech related and facial expressions related difficulties which were discussed in Section 2.1 and 2.2 respectively, there are also some more general difficulties to tackle in automatic emotion recognition research.

3.1 How Should We Annotate Emotion?

As the summaries of emotion recognition studies show (Table 1, 2 and 3), most of these studies still use categorical emotion labels. However, these labels are not always useful for real, genuine spontaneous emotion data since these labels tend to represent the extremes of each emotion category that are rarely encountered in spontaneous speech data. Further, humans can also express degrees of happiness or sadness. Taking this into account, a dimensional approach to emotion representation can offer an elegant solution. An advantage of this approach is that labels and categories of emotions have become redundant; we can express emotions now in terms of degrees of activation and evaluation. However, few studies have performed detection of degrees or shades of emotions in terms of emotional dimensions.

There remains discussion about how to obtain ground truth emotion annotations. On the one hand, we can define a ground truth emotion as an emotion that is perceived by people and that is agreed upon by most of the receivers. On the other hand, we can define a ground truth emotion as the experienced, true emotion as felt by the person her/himself. However, there can be a discrepancy between the perceived and experienced emotion: people may not always express their (true) emotions, especially when they are in conversation and obey the unwritten conversational rules. An option would be to let the subjects annotate their own emotional expressions (self-annotations) and compare these with the annotated emotions as perceived by other subjects.

3.2 Lack of Spontaneous Multimodal Data

One of the major obstacles in emotion research is the lack of annotated, spontaneous emotion data. Consequences are that most emotion recognition systems are trained on relatively small datasets containing a small number of subjects and that the classification results do not transfer very well to other data sets or real-life situations. However, we have seen that it is difficult to acquire analyzable signals in real-life situations (see Section 2.1, 2.2 and Fig. 1). Also note that for speech analysis, it is important that the content is independent of the expressed emotion to avoid confounding, e.g., [21].

One of the largest spontaneous audiovisual emotion databases (to date) is the Belfast Naturalistic database [22] which consists of TV clips. Other ways of collecting or eliciting (semi-) spontaneous emotions include showing movies or still pictures [23], listening to music [24], playing games [3] or interacting with virtual characters [25]. Playing (video) games seems to be particularly suitable for collecting emotion data: game developers are increasingly instructed to develop video games that in some way can trigger a range of emotions [26]. Also, manipulated games offer better control over the elicited emotion.

Finally, we should enable easier comparison and interpretation between studies by collecting a representative emotion database that can serve as a basis for benchmarking.

3.3 Fusion of Multimodal Measures

Table 3 showed that fusion of multimodal features could improve the performance of an emotion recognition system substantially. But, in most cases, fusion of multimodal features is not straightforward due to the different properties and behaviors of the features. For instance, the segmental units over which the features are measured are most likely to differ for many features which make it difficult to synchronize the features, e.g., different frame rates or lag times. Further, how should we deal with feature streams that have missing data while other streams have continuous output: e.g., speech features are usually measured over non-silent segments while heart rate can be measured continuously. And how should the system cope with conflicting outcomes of the classifiers as this can occur frequently in real-life data where blended emotions are not rare [27].

It seems that some researchers prefer a more human-like approach to fusion in emotion recognition [10, 20]: they prefer to fuse features on feature-level, which simulates humans who process multimodal information simultaneously and not separately. However, decision-level fusions are more informative and more explaining, e.g., the behavior of each modality during the recognition process can be more controlled and can be made more visible for feedback purposes. Further, decision-level fusions are somewhat easier to perform and have proven to be powerful in e.g., speaker recognition.

3.4 Evaluation of Emotion Recognition Systems in a Detection Framework

In emotion classification, the classifier's task is usually defined as "classify a given sound/image in one of these *N* emotion categories". An average accuracy percentage, based on calculations on the confusion matrix, is usually given as a single performance measure. However, this average accuracy measure depends on the number of emotion categories and the proportions of the used trials of each emotion category which is not very useful for making comparisons between studies. Instead of *classification*, we prefer to speak in terms of *detection* in which the classifier's task is defined as "does this given sound/image sound/look like emotion X, yes or no?". In this case, we can adopt the detection framework and evaluate the discrimination performance with a single measure Equal Error Rate (EER, which is defined as the point where the false alarm rate is equal to the miss rate). Further, it should also be clear whether the detection/classification experiment was performed subject-independently so we can interpret the results better. At this moment, comparing performances between emotion recognition systems is difficult which is partly due to the lack of shared datasets and shared evaluation standards.

4 Emotion Recognition in Adaptive Interfaces

It is clear that many efforts are taken to investigate automatic emotion recognition, but what exactly drives researchers in pursuing an automatic emotion recognition system? From a scientific point of view, we would like to put our knowledge about human emotion to the test and build a machine with human-like traits that enables an improved human-machine interaction: typically, this involves adapting interfaces to the user via emotion recognition. We are also interested in improving automatic speech recognition (ASR) systems by employing emotion-sensitive acoustic models (since it is generally known that ASR performances decrease when affective speech is uttered). Emotion recognition can also be used in computer-mediated communication (i.e., video conferencing, audio chat) in e.g., the e-health domain: monitoring a patient's emotional state from a distance during a medical consult or therapy can be very useful for a doctor. Other environments that may benefit from automatic emotion recognition include call centers, meetings (meeting browsers [28]), crisis management and surveillance.

Finally, adaptive interfaces employing emotion recognition can adjust their interfaces in games or e-learning systems to the player's or student's emotional state in order to increase his/her motivation. Grootjen et al. [29] have been investigating automatic assessment of stress and task load in order to develop an emotion-sensitive adaptive interface that can adapt to the operator's stress level so that tasks of a stressed operator can be allocated to another operator. They have been collecting multimodal measures of operators performing tasks on a navy ship. Working on that data, they have experienced many of the similar issues discussed above: speech analysis is difficult due to background noises consisting of loud beeps, facial expression analysis is difficult due to the pose of the head and background issues (see Fig. 2), fusion of multimodal measures is difficult because of their different time scales etc. With the development of advancing emotion recognition technology that can cope with these problems, opportunities for future interesting and useful applications increase.

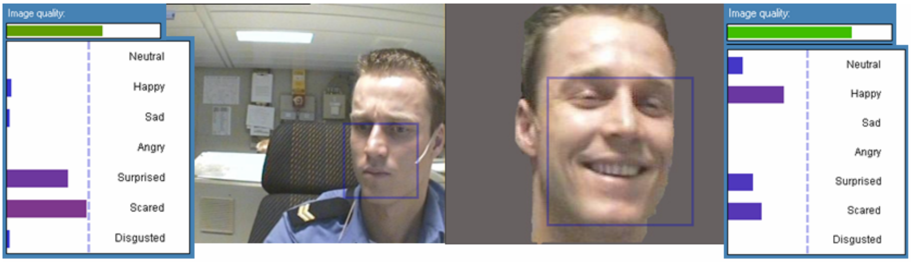

Fig. 2. The FaceReader trying to classify data from Grootjen et al. [29]. Left: ambiguous output. Right: the FaceReader is sensitive to background which was manually removed to improve classification output.

5 Discussion and Conclusions

We have discussed some of the problems that one can encounter in automatic emotion recognition research with the focus on speech and facial expressions analysis. Some of the difficulties can be partly solved by agreeing upon standards for the emotion community, and some of the difficulties can be solved by developing advanced technologies that can deal with noisy data. By describing these difficulties and problems for the development of an emotion recognition system, we do not want to discourage researchers. Rather, our intention is to encourage researchers in the emotion community to elaborate on the problems and to agree upon standards. A lot of work is being carried out by a European project HUMAINE [30] that aims at laying foundations for the development of ‘emotion-oriented’ systems. Consequently, if we agree upon standards such as the definition of “emotion”, evaluation measures and annotation issues, we can further develop data sets that can be used for benchmarking emotion recognition systems. One of the reasons why it is difficult to agree upon emotion standards may be related to the subjectivity and dependency of emotion phenomena. Further research should indicate to what extent emotion recognition and production is subject, culture or context dependent so we can take this into account in our research.

It is clear that building an automatic emotion recognition system can be very complex, especially if we want to incorporate an accurate and complete model or theory of emotion. However, is it always necessary or realistic to pursue such an ideal system that is based on a complete and complex model of emotion? Researchers must keep in mind that detection of ‘simple’ striking emotions in context (e.g., ‘panic’) can also be of high practical value for adaptive interfaces.

One of the conclusions that we can draw from the short overviews of uni and multimodal emotion recognition studies is that we have arrived at a point where we should bridge the gap between working and training emotion models with simulated emotion data and applying these models to real-life emotion data. Furthermore, we should enable easier comparisons between emotion recognition studies by developing some standards for an automatic emotion recognition framework. Researchers are now increasingly working with authentic emotion data and the results of these emotion recognition systems are promising but still a lot of improvements can be made. It is rather incredible that humans are able to make judgments about someone's emotions based on a bulk of multimodal information. For now, we think it is fair to say that humans still 'have the best feel for emotions'.

Acknowledgements. This study is supported by the Dutch BSIK project MultimediaN (http://www.multimedian.nl).

References

1. Cowie, R., Schröder, M.: Piecing together the emotion jigsaw. In: Machine Learning for Multimodal Interaction, pp. 305–317 (2005)
2. Kapoor, A., Picard, R.W.: Multimodal affect recognition in learning environments. In: Proceedings of the ACM International Conference on Multimedia, pp. 677–682 (2005)
3. Kim, J., André, E., Rehm, M., Vogt, T., Wagner, J.: Integrating information from speech and physiological signals to achieve emotional sensitivity. In: Proceedings of Interspeech, pp. 809–812 (2005)
4. Zhai, J., Barreto, A.: Stress Recognition Using Non-invasive Technology. In: Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference FLAIRS, pp. 395–400 (2006)
5. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, 614–636 (1996)
6. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in Human-Computer Dialog. In: Proceedings of the ICSLP International Conference on Spoken Language Processing, pp. 2037–2040 (2002)
7. Nwe, T.L., Foo, S.W, De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Communication 41, 603–623 (2003)
8. Vidrascu, L., Devillers, L.: Detection of real-life emotions in call centers. In: Proceedings of Interspeech, pp. 1841–1844 (2005)
9. Batliner, A., Steidl, S., Hacker, C., Nöth, E., Niemann, H.: Tales of tuning - prototyping for automatic classification of emotional user states. In: Proceedings of Interspeech, pp. 489–492 (2005)
10. Pantic, M., Rothkrantz, L.J.M.: Towards an Affect-Sensitive Multimodal Human-Computer Interaction. Proceedings of the IEEE 91, 1370–1390 (2003)
11. Ekman, P., Friesen, W.V.: Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto (1978)
12. Pantic, M., Rothkrantz, L.J.M.: Automatic Analysis of Facial Expressions: The State of the Art. IEEE Transaction on Pattern Analysis and Machine Intelligence 22, 1424–1445 (2000)
13. Pantic, M., Rothkrantz, L.J.M.: Expert system for automatic analysis of facial expression. Image and Vision Computing Journal 18, 881–905 (2000)
14. Cohen, I., Sebe, N., Chen, L., Garg, A., Huang, T.S.: Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. Computer Vision and Image Understanding 91, 160–187 (2003)

15. Sebe, N., Sun, Y., Bakker, E., Lew, M.S., Cohen, I., Huang, T.S.: Towards Authentic Emotion Recognition. In: IEEE SMC International Conference on Systems, Man, and Cybernetics, pp. 623–628 (2004)
16. Den Uyl, M., Van Kuilenburg, H.: FaceReader: an online facial expression recognition system. In: Proceedings of 5th International Conference on Methods and Techniques in Behavorial Research, pp. 589–590 (2005)
17. De Silva, L.C., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multi-modal information. In: Proceedings of the ICICS International Conference on Information, Communications and Signal Processing, pp. 397–401 (1997)
18. Chen, L.S., Tao, H., Huang, T.S., Miyasato, T., Nakatsu, R.: Emotion recognition from audiovisual information. In: Proceedings of the IEEE Workshop on Multimedia Signal Processing, pp. 83–88 (1998)
19. De Silva, L.C., Ng, P.C.: Bimodal Emotion Recognition. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, pp. 332–335 (2000)
20. Sebe, N., Cohen, I., Gevers, T., Huang, T.: Emotion Recognition Based on Joint Visual and Audio Cues. In: Proceedings of the ICPR International Conference on Pattern Recognition, pp. 1136–1139 (2006)
21. Chen, L.S.: Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. Phd thesis (2000)
22. Douglas-Cowie, E., Cowie, R., Schröder, M.: A new emotion database: Considerations, sources and scope. In: Proceedings of ISCA ITRW Workshop on Speech and Emotion, pp. 39–44 (2000)
23. Lang, P.J.: The emotion probe - studies of motivation and attention. American Psychologist 50, 371–385 (1995)
24. Wagner, J., Kim, J., André, E.: From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: Proceedings of the IEEE ICME International Conference on Multimedia & Expo, pp. 940–943 (2005)
25. Cox, C.: SALAS Sensitive Artificial Listener induction techniques. Paper presented to HUMAINE Network of Excellence Summer School (2004), Retrieved 5 Februay 2007 from http://emotion-research.net/ws/summerschool1/SALAS.ppt#259
26. Lazarro, N.: Why we play games: 4 keys to more emotion. Paper retrieved February 5, 2007 from http://www.xeodesign.com/xeodesign_whyweplaygames.pdf
27. Douglas-Cowie, E., Devillers, L., Martin, J., Cowie, R., Savvidou, S., Abrilian, S., Cox, C.: Multimodal databases of everyday emotion: facing up to complexity. In: Proceedings of Interspeech, pp. 813–816 (2005)
28. AMI project, http://www.amiproject.org
29. Grootjen, M., Neerincx, M.A., Weert, J.C.M., Truong, K.P.: Measuring Cognitive Task Load on a Naval Ship: Implications of a Real World Environment. In: Proceedings of ACI (this volume) (2007)
30. HUMAINE project, http://emotion-research.net/