# The Influence of Context Knowledge for Multi-modal Affective Annotation

Ingo Siegert, Ronald Böck, and Andreas Wendemuth

Cognitive Systems Group, Otto von Guericke University Magdeburg, Germany
ingo.siegert@ovgu.de

**Abstract.** To provide successful human-computer interaction, automatic emotion recognition from speech experienced greater attention, also increasing the demand for valid data material. Additionally, the difficulty to find appropriate labels is increasing.

Therefore, labels, which are manageable by evaluators and cover nearly all occurring emotions, have to be found. An important question is how context influences the annotators' decisions. In this paper, we present our investigations of emotional affective labelling on natural multi-modal data investigating different contextual aspects. We will explore different types of contextual information and their influence on the annotation process.

In this paper we investigate two specific contextual factors, observable channels and knowledge about the interaction course. We discover, that the knowledge about the previous interaction course is needed to assess the affective state, but that the presence of acoustic and video channel can partially replace the lack of discourse knowledge.

**Keywords:** emotion comparison, affective state, labelling, context influence.

## 1 Introduction

In future, technical systems should provide more human-like interaction abilities. Therefore, these systems have to be adaptable to the user's individual skills, preferences and current emotional states [20]. To enable such systems to determine a user's affective state, the recognition needs to rely on all signals humans use in the interaction, like speech, facial expressions and gestures. To provide successful human computer interaction, automatic emotion recognition from speech experienced greater attention, also increasing the demand for valid data material.

The recognition of the user's affective state is still a challenging task. Many years the focus was set on acted data e.g. [3], also due the lack of available datasets. While, in acted data, the label (ground-truth) is clearly instructed to the actor, resulting in clear and high expressive emotional recordings, see [21] the application of classifiers trained on acted data material within a realistic or naturalistic human computer interaction shows that these databases do not

M. Kurosu (Ed.): Human-Computer Interaction, Part V, HCII 2013, LNCS 8008, pp. 381–390, 2013.

include the variety of natural occurring affective states [9]. So, within research community, the focus changed from acted emotions to more realistic emotional expressions like [7,8,18], because emotion can induce changes in speech that cannot be controlled by the speaker. This recordings mostly have a lower expressiveness of the affective states and raises the problem of a reliable ground-truth generation [6]. So the difficulty, to find appropriate labels is increasing. Whereas in acted emotional data, the *label* is clearly instructed and can assessed via perception tests in realistic recordings the expressions are uncontrolled and not as obvious. The generation of a *label* is persuaded by an annotation process where a large number of annotators is used to assess the observed affective state, by choosing a suitable label.

Therefore, labels, which are manageable by evaluators and cover nearly all occurring emotions, have to be found. But besides the utilized label also the design of the labelling process is important. An important question is how context influences the annotators' decisions [4]. The authors in [5] investigated the influence of the context onto the perception of anger. They argue that traditional associations between tones and attitudes are misleading and that the contextual factor can neutralise the anger perception. The authors perform two studies, where they could show, that neutral uttered wh-words are perceived as anger, when heard without surrounding context. A further study is performed in [13], where the authors investigate the role of channel information onto aggression detection utilizing three different settings: audio only, video only and audio plus video. They stated, that for 46% of their material the annotation of all three sets differs for the same samples.

This study supports our hypothesis, that the context plays an important role within affect recognition. In our study, we want to combine both investigations of surrounding information and channel influence. Furthermore, we investigated both context influences within a realistic human-computer interaction.

The remainder of the paper is structured as follows: In Section 2 the utilized dataset is described in detail. Afterwards,we introduce our research methods in Section 3. The results of our study are presented and discussed in Section 4. Finally, in Section 5, an outlook for further research is given.

## 2 Dataset

The conducted study utilizes the LAST MINUTE corpus [15]. It contains multimodal recordings of 130 native German subjects collected in a Wizard-of-Oz experiment. The technical recordings and first classification results are described in detail in [10]. As background information the subjects are told to test a new natural language communication interface. The setup revolves around a journey to an unknown place "Waiuku" that they have won. Using voice commands the subjects have to prepare the journey, assembling the baggage and select clothing. The task is designed to generate affective enriched material from a naturalistic human computer interaction [17]. First results on multi-modal affect recognition can be found in [12,14]. The utilized TTS uses an artificial and mechanical voice, providing a lot of explanation, which leads to long monologues by the system.

During the dialogue critical events provoking negative emotions and could may leading to a break off of the dialogue, are induced. We focus on three key events, where the user should be set in a certain condition: Baseline (BL), Challenge (CH), and Waiuku (WA). All three events are designed in such a way, that the system gives a specific information whereupon the user shall show a specific reaction. At the BL event, occurring after 5-10 minutes, the test person has been adapted to the experimental situation and the first excitement is gone. The person starts packing the luggage. The system only confirms the action requests. The CH event happens when the system creates mental stress by suddenly claiming to reach a previously luggage limit. This event arises after 15-20 minutes of the experiment. In the WA event a second strategy change has to be performed, when claiming a different voyage destination. It is winter instead of summer at the destination. At this point the subject notice, that it has to re-arrange its complete baggage. This event occurs at about 20-25 minutes of the experiment. Neither we can be sure about the real stress factor for the particular subject, nor can we assure the real duration of any higher stress level.

A the expected time effort for labelling is up to four times higher than the material to be labelled, we selected 4 subjects from the whole corpus. This results in a subset of approx 23 minutes. Furthermore, we splitted the events into separate utterances, as it was already mentioned, that the surrounding words are important for a proper assessment [5]. So, we end up with 135 snippets with a length of 3 seconds to 50 seconds. The mean is 11 seconds. An overview about the number of snippets for each event and subject is given in Table 1.

**Table 1.** Overview of utilized snippets and their distribution for the selected subjects and experimental events

| subject | BL | CH | WA |
|---|---|---|---|
| 20101006aFM | 16 | 9 | 9 |
| 20101117bMT | 15 | 5 | 8 |
| 20101206bEG | 19 | 7 | 9 |
| 20110126aFW | 21 | 8 | 8 |
| Total | 71 | 29 | 34 |

## 3 Study Design

As stated in the introduction, we rely on the related work of [4,5,13], all claiming, that surrounding information and observable channels are important to receive a useful annotation. To verify our hypotheses, we design different labelling tasks, where we varied either the different observable channels or the interaction course. Hereby we will support the following hypotheses: labels can only be gathered, when both acoustic and visual information are present, information about interaction development supports the labelling process. Preliminary results onto the influence of system responses where presented in [17].

### 3.1 Define Dependent Variables

Based on this, we can now define the following two dependent variables and their expressions. The observable channel consists of the values “audio only”, “video only” and “audio plus video”. The interaction course can be either random or ordered. Therefore, we conducted six different sets, where both variables assume all its defined values. The resulting sets are presented in Table 2.

**Table 2.** Overview of the labelling sets, generated by the two dependent variables with their expressions

| interaction course | observable channels | label set |
|---|---|---|
| random | audio only | Set 1 |
| | video only | Set 2 |
| | audio + video | Set 3 |
| ordered | audio only | Set 4 |
| | video only | Set 5 |
| | audio + video | Set 6 |

### 3.2 Design the Annotation Process

The design of the annotation process is similar to [19]. Vidrascu proposes several phases to decide the list of labels, annotation scheme, and segment length and afterwards to start with the actual annotation process. To define the segment length, we rely on experience of [1], where an assessment based on a speech chunks is proposed. For our utilized database, these chunks are identical to the subjects utterances, as they are quite short in our material. A shorter length of single words will mislead the labellers, as investigated by [5]. A longer segment length including a complete dialogue turn, consisting of several human utterances and system responses, can be composed of several affective states.

To get the proper affective labels we utilized results of our study presented in [16]. There, we investigated the differences between three labelling methods and the observed emotional labels for a similar human-computer interaction utilizing the NIMITEK Corpus, see [11]. Therefore, we investigated the used list of labels and the annotation method usable for labelling human-computer interaction. The following affects proved to be useful, see Investigation I Table 3. As the NIMITEK corpus was designed to provoke negative emotions [11], whereas the LAST MINUTE corpus tries to investigate possible dialogue break offs [15], we conducted an experiment to gather more suitable emotional labels for the domain of the LAST MINUTE corpus. Therefore, we presented the utilized snippets to six labellers, all of them where psychologist students, utilizing the labels from [16], with the explicit task, to add emotional terms, they need to describe the affective state of the subject. This investigation results in additional labels, see Investigation II in Table 3. The affective labels found in both investigations and are combined into one word list, for the persuaded study, see Table 3 for an overview. Additionally, the labellers could asses (o) *no emotion*, if they assess, that no emotion was observed and we gave them the opportunity to leave a comment.

**Table 3.** Overview of the utilized affective word list

| | |
|---|---|
| Investigation I (see [16]) | |
| (a) sadness (b) contempt (c) helplessness (d) interest (e) hope (f) relief | |
| (g) joy (h) surprise (i) confusion (j) anger | |
| Investigation II | Additional |
| (k) shame (l) stress (m) concentration (n) impatientness | (o) no emotion |

We utilized 10 labellers, all of them with psychological background. To support the labellers during their annotation process, we used a variant of ikannotate [2]. The labellers could see or hear the actual snippet and could chose one or several words from the presented word list.The order of presented snippets could not be influenced by the labellers. The programme forces them watch the complete snippet and assess it afterwards, a repeated view of the actual one is possible.

## 4 Results

We evaluated our results on the basis of each set, as described in Table 2. To investigate the influence of the defined variables, we compared the assessed affective states resulting from a majority voting. Only the assesment where five or more labellers agreed on the same affective state, is used as a valid label. In the case, where only five labellers agreed, than the remaining labellers should not agree on the same other affect.

### 4.1 Influence of Present Channels

Comparing the influence of the available channels in Fig. 1, we notice that the number of majority votes for *concentration* (m) is nearly assesed most for all conditions. We observe an increasing number of votes from the audio-only (1) over the video-only (2) to the audio plus video set (3). The same behavior can be observed for *joy* (g). The opposite effect is noticed for *impatientness* (n). The affective state *surprise* (h) and *relief* (f) gets a rising number of votes comparing

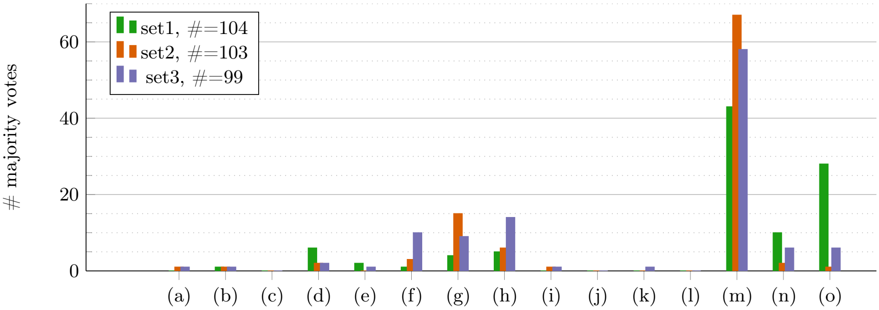

**Fig. 1.** Number of majority votes for different channel informations

the audio only and video only set, but when both channels are present, the number of votes is decreasing. The affective state *anger* (j) is labelled sufficient only, when both channels are present. Remarkable is the varying number of *no emotion* (o) votes, for Set 1 we have the highest number of 28 votes. This effect was expected, as especially in the WA event we used some snippets, where the subject did not talk. For Set 2, where only the video channel was used, only one item was labelled showing no emotion. Finally, Set 3 having both channel informations, we observe again 5 items voted with *no emotion* (o).

### 4.2 Influence of Interaction Course

The next aspect, we want to analyse, is the influence of the knowledge about the interaction course. The resulting numbers of the majority votes are given in Fig. 2. It can be noticed, that the distribution of majority votes does not differ much. In Set 6, utilizing the experimental data in an ordered way, whereas Set 3 uses a random order. Additionally, we count the total number of majority votes, reached for each set. The results are given in Fig. 2. Here it can be noticed, that the number increased from 104 to 131 items, where a majority vote could be drawn. We call this a reduction of variety. We attribute this to the influence of the prior knowledge. The labellers know, which affective state they observed before and therefore take that decision into account.

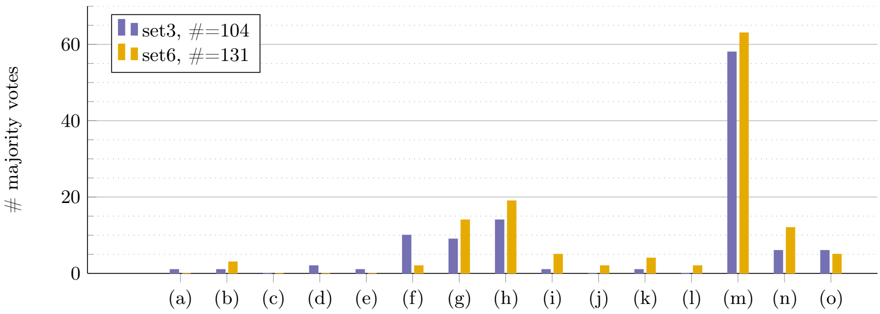

**Fig. 2.** Number of majority votes for ordered/random interaction courses

### 4.3 Influence of Both Present Channels and Interaction Course

When comparing the influence on the interaction course, ordered or random, together with a limited channel, we get the results presented in Fig. 3 and Fig. 4. Here we can notice, that the influence, the ordered presentation has, is stronger than for the contrary presentation of channels. Resulting in a reduction of the chosen labels and a shaping of selected ones. The affective states *surprise* (h), *joy* (g), *concentration* (m), *impatientness* (n) and *no emotion* (o) are labelled most. Whereas especially the number of *no emotion*-votes decreases for the audio onyl set, when using the ordered presentation. Whereas, we can notice an increased

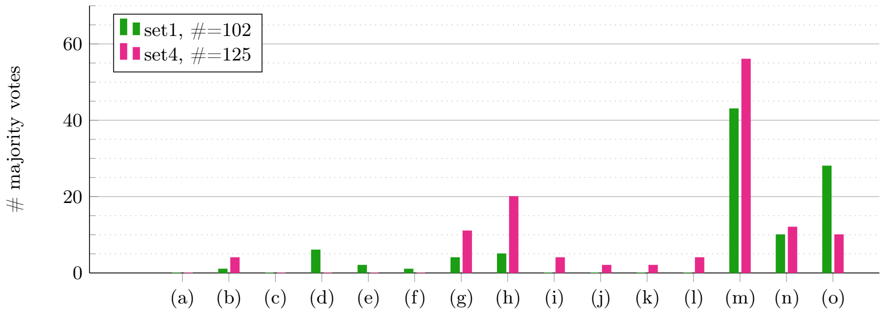

**Fig. 3.** Number of majority votes for ordered/random interaction courses, while using only the audio channel

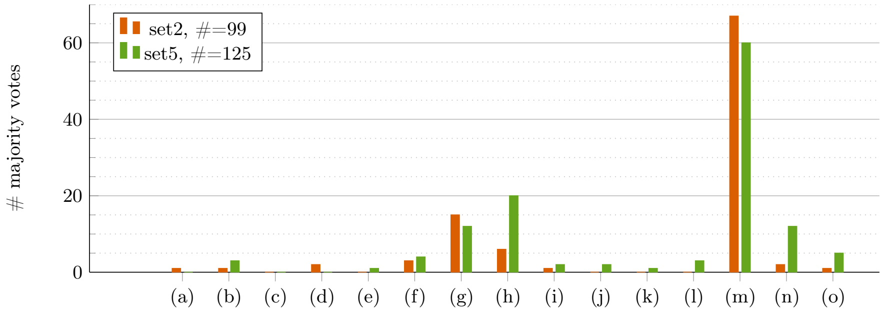

**Fig. 4.** Number of majority votes for ordered/random interaction courses, while using only the video channel

number of majority votes for the affects *concentration* (m) and *joy* (g) comparing the audio only random/ordered sets, we get the opposite result comparing same the video-only sets. But for the affects *surprise* and *impatientness* (n), we notice the same increasing votes. Also the chosen affects are identical for Set 4 and Set 5 despite *relief* (f), which is labelled only for the video only set (Set 5). Furthermore the effect of an increasing total number of given majority votes can be noticed, the number increased from 102 to 125 for the audio only sets (Set 1, Set 4) and from 99 to 125 for the video only sets (Set 2, Set 5). This is similar to the comparison of Set 3 and Set 6.

### 4.4 Notes on the Experimental Events

Comparing the labelled affective states regarding the described experimental events, it can be noticed that the affective labels *surprise* (h), *joy* (g), *confusion* (i), and *concentration* (m) are assessed within all investigated experimental events. From the pure experimental design, the presence of *surprise* (h), *confusion* (m), and *joy* (g) were not expected for BL. The distribution of *joy* and

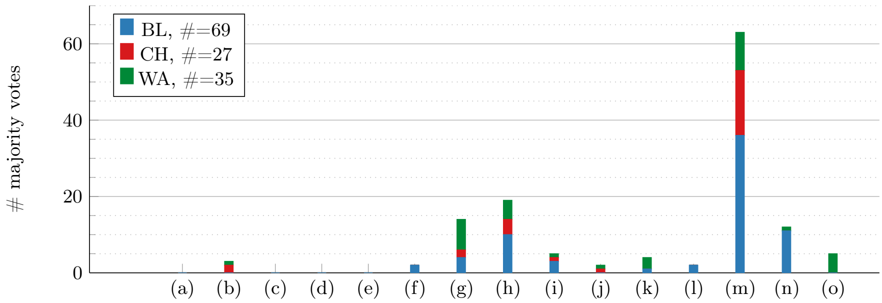

**Fig. 5.** Number of majority votes for the utilized experimental events

*concentration* are in conformity with the experimental design, as CH requires *concentration* whereas WA can create *joy*. The less votes for the affect *surprise* (h) does not match the expected assessment.

Affective states only assigned to BL are *relief* (f), *impatientness* (n) and *stress* (i), with *impatientness* assessed most. This confirms the expected reaction, that should be induced due the mechanical voice and the dominance of the system monologues. The occurence of *contempt* (b) and *anger* (j) during CH is also according the the design, but very rare. The assessment of the WA event snippets with *no emotion* (o) and *shame* (k) indicates, that the intended effect was achieved. The monologue about the travel information should not provoke any affect and the information about the wrong assumed destination provokes shame, as the subject could have know the expected destination.

## 5 Conclusion

We focus on two terms with different conditions, the a) role of available channel information and b) knowledge about the interaction course. These effects where investigated utilizing affective word lists and 10 labellers.

Evaluating the channel information, we can state, that the availability of both channels, audio and video is important. In contrast to [13], we could not get such a confusion between the sets presenting audio only, video only and both channels. This could be due the fact, that our material consist of material where the face is recorded in a frontal view with a very good illumination, so that the labeller could always assess the facial expressions very good.

Evaluating the influence of the interaction course, we can state, that this is important, too. But the differences between the majority votes of the ordered and unordered set, where both channels are present, is quite small. Nevertheless, presenting the interaction in an ordered way can support the labellers in situations, where one channel is partly missing. This is supported, by the comparison of Set 1 and Set 4 or Set 2 and Set 5, respectively. Here the resulting majority votes of the ordered sets is similar to Set 6, presenting both channels in an ordered way.

These results show that the affective states of observed subjects within a naturalistic human computer interaction are assessable, but consist mostly of states with low expressiveness and indicating affects, pointing on a potential problematic dialogue are very rare.

A very interesting additional investigation, we want to follow-up with, is the investigation of the detailed changes of affective state for the different experimental conditions. This could help to get a deeper understanding which affective states are often confused and which influence specific contextual informations have.

## References

1. Batliner, A., Seppi, D., Steidl, S., Schuller, B.: Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach. In: Advances in Human-Computer Interaction 2010 (2010)
2. Böck, R., Siegert, I., Vlasenko, B., Wendemuth, A., Haase, M., Lange, J.: A processing tool for emotionally coloured speech. In: Proc. of the 2011 IEEE International Conference on Multimedia & Expo., Barcelona, Spain (July 11-15, 2011)
3. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. In: Proc. of Interspeech (2005)
4. Callejas, Z., López-Cózar, R.: Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Com. 50, 416–433 (2008)
5. Cauldwell, R.T.: Where did the anger go? the role of context in interpreting emotion in speech. In: Proc. of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, pp. 127–131 (September 2000)
6. Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech Commun. 40(1-2), 5–32 (2003)
7. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: towards a new generation of databases. Speech Com. Special Issue Speech and Emotion 40, 33–60 (2003)
8. Douglas-Cowie, E., et al.: The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 488–500. Springer, Heidelberg (2007)
9. Douglas-Cowie, E., Devillers, L., Martin, J.C., Cowie, R., Savvidou, S., Abrilian, S., Cox, C.: Multimodal databases of everyday emotion: facing up to complexity. In: European Conference on Speech Com. and Technology, pp. 813–816 (2005)
10. Frommer, J., Michaelis, B., Rösner, D., Wendemuth, A., Friesen, R., Haase, M., Kunze, M., Andrich, R., Lange, J., Panning, A., Siegert, I.: Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus. In: Proc. of the Eight International Conference on Language Resources and Evaluation (LREC 2012), ELRA, Istanbul, Turkey (May 2012)
11. Gnjatović, M., Rösner, D.: On the role of the NIMITEK corpus in developing an emotion adaptive spoken dialogue system. In: Proc. of the Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco (2008)
12. Krell, G., Glodek, M., Panning, A., Siegert, I., Michaelis, B., Wendemuth, A., Schwenker, F.: Fusion of Fragmentary Classifier Decisions for Affective State Recognition. In: Schwenker, F., Scherer, S., Morency, L.-P. (eds.) MPRSS 2012. LNCS, vol. 7742, pp. 116–130. Springer, Heidelberg (2013)

13. Lefter, I., Rothkrantz, L.J.M., Burghouts, G.J.: Aggression detection in speech using sensor and semantic information. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 665–672. Springer, Heidelberg (2012)
14. Panning, A., Siegert, I., Al-Hamadi, A., Wendemuth, A., Rösner, D., Frommer, J., Krell, G., Michaelis, B.: Multimodal affect recognition in spontaneous hci environment. In: IEEE International Conference on Signal Processing, Communications and Computings (ICSPCC), pp. 430–435 (2012)
15. Rösner, D., Friesen, R., Otto, M., Lange, J., Haase, M., Frommer, J.: Intentionality in interacting with companion systems – an empirical approach. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part III, HCII 2011. LNCS, vol. 6763, pp. 593–602. Springer, Heidelberg (2011)
16. Siegert, I., Böck, R., Philippou-Hübner, D., Vlasenko, B., Wendemuth, A.: Appropriate Emotional Labeling of Non-acted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins. In: Proc. of the IEEE International Conference on Multimedia and Expo., ICME 2011, Barcelona, Spain (2011)
17. Siegert, I., Böck, R., Wendemuth, A.: The influence of context knowledge for multimodal annotation on natural material. In: Proc. of the First Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3), Santa Cruz, USA (September 2012)
18. Vaughan, B., Kosidis, S., Cullen, C., Wang, Y.: Task-based mood induction procedures for the elicitation of natural emotional responses. In: The 4th International Conference on Cybernetics and Information Technologies, Systems and Applications, Orlando, Florida (2007)
19. Vidrascu, L., Devillers, L.: Real-life emotion representation and detection in call centers data. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 739–746. Springer, Heidelberg (2005)
20. Wendemuth, A., Biundo, S.: A Companion Technology for Cognitive Technical Systems. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 89–103. Springer, Heidelberg (2012)
21. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence 31, 39–58 (2009)