Developing an Expressive Speech Labeling Tool Incorporating the Temporal Characteristics of Emotion

Stefan Scherer[1], **Ingo Siegert**[2], **Lutz Bigalke**[1], **Sascha Meudt**[1]

[1] Institute of Neural Information Processing, [2] FEIT IESK-Cognitive Systems
Ulm University, Otto-von-Guericke-University Magdeburg
stefan.scherer@uni-ulm.de, ingo.siegert@ovgu.de

Abstract

A lot of research effort has been spent on the development of emotion theories and modeling, however, their suitability and applicability to expressions in human computer interaction has not exhaustively been evaluated. Furthermore, investigations concerning the ability of the annotators to map certain expressions onto the developed emotion models is lacking proof. The proposed annotation tool, which incorporates the standard Geneva Emotional Wheel developed by Klaus Scherer and a novel temporal characteristic description feature, is aiming towards enabling the annotator to label expressions recorded in human computer interaction scenarios on an utterance level. Further, it is respecting key features of realistic and natural emotional expressions, such as their sequentiality, temporal characteristics, their mixed occurrences, and their expressivity or clarity of perception. Additionally, first steps towards evaluating the proposed tool, by analyzing utterance annotations taken from two expressive speech corpora, are undertaken and some future goals including the open source accessibility of the tool are given.

1. Introduction

The collection of expressive corpora comprising emotionally colored data has received a lot of attention in the recent past. Starting from acted collections such as the Berlin Database of Emotional Speech (Burkhardt et al., 2005), current databases are focusing more on realistic emotional expressions (Gnjatovic and Roesner, 2008; Strauss et al., 2008; Douglas-Cowie et al., 2007). However, with an increasing ecological validity of the data the difficulty of finding appropriate labels for the expressions is rising. In the case of acted emotional data the "label" is clearly instructed to the actor and expressive quality can easily be assessed conducting perception tests, whereas in Wizard of Oz like scenarios or other more realistic recordings the expressions are uncontrolled and not obvious at all. Furthermore, the situation is complicated by the fact that non acted expressive data is way more moderate than most of the acted data found (Cowie and Cornelius, 2003). Therefore, the manifold emotional theories that try to cover emotion in general like discrete emotional theories, or the arousal theory, trying to map the expressions into a two dimensional space, are often unsuitable for the task or dataset at hand (Scherer et al., 2003; Craggs and McGee Wood, 2004).

Considering these issues several different methods for labeling realistic expressive data have been developed. In (Douglas-Cowie et al., 2007) for example a mixture of categorical labels in combination with a few other so called trace labels are proposed. By assigning these trace labels, the annotator has to continuously judge the sequence he or she is listening to and adjust the label on a numerical scale. We believe that this method is quite precise and overcomes some of the issues of former approaches, but utilizing it to its full extent (6 different categorical labels, and 8 trace like labels) is quite time consuming for the annotator and renders it somewhat impractical. However, in simpler approaches like using FEELtrace (Cowie et al., 2000) lay annotators are confronted with the task to label expressions in an unfamiliar two dimensional space[1], which may lead to biased answers. Therefore, the goal of the presented approach is to come up with a simple and fast approach, that is still capable of gathering a rich amount of data overcoming known issues.

In this work, we would like to introduce a novel method that is targeted at labeling whole expressive sequences or clips at an utterance level. Being aware of the fact that emotional expressions vary in intensity and meaning very rapidly over time, illustrated by an example found in (Picard, 2000) (see Section 3.), it was necessary to come up with a method that allows the annotators to assign labels to the utterances and a sort of temporal progression. Furthermore, in order to get the best possible results from the annotators the method had to be intuitive and easy to use. Therefore we utilized the Geneva Emotional Wheel (Scherer, 2005) to assign the expressive and emotional labels and for the temporal progress the annotator adjusts bell-shaped beta-distributions with very simple mouse movements.

The remainder of this abstract is organized as follows: In Section 2. the dataset which will be labeled is described in some detail. Furthermore, the developed tool and the underlying theory will be introduced and described in Section 3.. In Section 4. first evaluation experiments with a small number of participants are introduced. And finally, in Section 5. a future outlook expected from this research and a future method to exploit the additional information retrieved from the temporal progress data to classify expressive recordings automatically will be introduced.

2. Dataset Description

The NIMITEK Corpus used as the underlying database comprises emotionally rich audio and video material (Gnjatovic and Roesner, 2008). To investigate human machine communication a Wizard-of-Oz setup was used, and audiovisual data was gathered. For the experiments a hybrid ap-

[1] Valence-arousal space

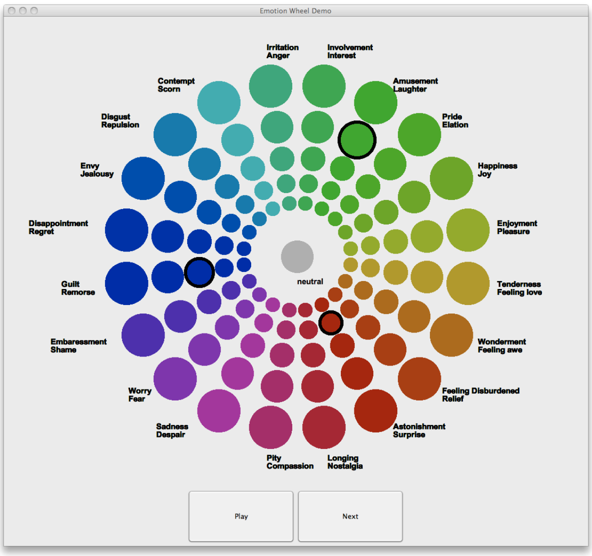

Figure 1: Screenshot of the first step of the labeling process, the Geneva Emotion Wheel.

proach to elicit emotionally colored expressions was used: On the one side a motivating experiment was conducted - the user was told to be participating in an intelligence test, and on the other side different strategies of the wizard were pursued to stress and annoy the user further.

For example, for a short period in which the user gets to know the system the wizard recognizes the users input correctly, and the system performs the right actions and provides useful comments and answers. However, this strategy changes in the second part of the experiment. In which, the wizard starts to simulate malfunctions of the system, and to provokes emotional reactions of the user by behaving in inappropriate ways.

The language of the corpus is German. Ten native German speakers, 3 male and 7 female with an average age of 21.7 (ranged from 18 to 27) participated in the experiments. None of them has an educational background with spoken dialogue systems and they were not aware of the wizard at any time. Furthermore, the corpus consists of 10 sessions with an approximate duration of 90 minutes.

Since the task of the experiment, namely simulating an intelligence test with special questions to solve, is very specific the vocabulary is limited to a rather small number. However, the comments of the user are recorded as well, and since the wizard stimulates the user to express emotions verbally too, the corpus is emotionally rich and valuable for the task at hand.

3. Labeling Tool

In order to be able to label the utterances taken from the NIMITEK corpus appropriately we developed a labeling tool that is founded on the theoretical findings for the Geneva Emotional Wheel (Scherer, 2005). The advantages of this approach include the natural way of labeling expressions using discrete categorical words, such as “relief”, “worry”, or “pleasure”, and the possible mapping of discrete labels into a two dimensional space (valence - dominance space) (Scherer, 2005) and. While labeling the annotators do not know about the mapping from their decision onto the common two dimensional space since most of them are not familiar with the underlying theory. Furthermore, as it can be seen in Figure 1, it is possible to assign differently sized circles to the utterances. The annotators are beforehand instructed to use the larger circles if they feel sure about what they perceived and smaller circles if they are uncertain or the expression is not perceived as clearly. Using this method it is possible to assign labels with different values for the certainty or clarity of the expressed emotion. Even mixed emotional states can be assigned using this method, by simply selecting more than one circle in the GEW, which was already a problem identified by (Plutchik, 1966):

> It is then possible to show systematically that mixtures of two or more primary emotions (dyads and triads) produce the many hundreds or even thousands of mixed emotions we encounter in daily life.

After selecting one of the circles a second frame is shown to the annotator and a beta-distribution function has to be adapted to his or her likings using the mouse and mouse-wheel. As it can be seen in Figure 2 a temporal progress with either flat or steep characteristics can be adjusted by the user by simply dragging the curve to the intended position and adjusting the gradient using the mouse wheel. The so called beta-distributions are preferred over the standard normal distribution since they are defined on a closed interval and the integral of the curve is constant over this interval.

This temporal annotation was integrated, considering the example found in (Picard, 2000), where a tennis player suddenly feels a strong pain in the back of his legs and turns around in rage about the carless passerby and immediately turns his anger into compassion for the handicapped person in the wheelchair who was not able to stop the chair in time. We find it very important to be able to label multiple emotions with different temporal characteristics on an utterance based level, since emotion is not a constant signal but a highly dynamical one, as nicely written in (Ortony et al., 1988):

> In fact, some of our research has focused on naturally occurring sequences of emotions, and has led us to the conclusion that emotional experiences probably occur in complex sequences much more often than they do in isolation.

Furthermore, it is more and more argued that emotions are occurring in mixed states (Scherer et al., 2009), which is also possible to label with this novel tool. Additionally, for unclear or very moderate expressions it is possible to assign a neutral label which does not require a temporal flow labeling. This neutral label should however, only be assigned if none of the others is possible to be assigned.

A few technical details are given in the following paragraph: The tool is entirely platform independently developed using Java, which is ideal to distribute the tool among a many annotators. Furthermore, the categories (the spokes of the wheel) as seen in Figure 1 are predefined using a

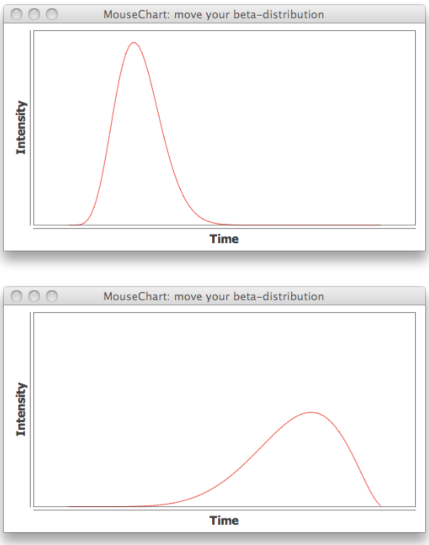
MouseChart: move your beta-distribution
Intensity
Time
MouseChart: move your beta-distribution
Intensity
Time

Figure 2: Examples gradients of the beta-distribution representing the temporal intensity progress of the emotion in a whole utterance. The charts are produced using the open source software JFreeChart.

property file and no further compilation is required in order to apply changes. Not only the number of categories may be varied, but also their labels, and the number of differently sized circles. Finally, the output that is produced using this tool contains the following items:

- ID and location of the audio or video file.
- Selected emotion blobs with ID, label, and selected circle size.
- Values of the beta-distribution, typically α and β, and estimated mean and variance.

4. Preliminary experiments

Driven by the question which emotion theory and related methods are most suitable and relevant for human-machine-communication, we developed a tool for comparing the three most used methods, ”Basis” Emotions (BE), Geneva Emotion Wheel (GEW) and Self Assessment Manikins (SAM) (Scherer, 2005; Darwin, 1978; Lang, 1980).

The developed tool provides three methods for emotion labeling in a similar manner. So that we could minimize the variance of influence due to different tools. We used two sessions, each about 40 minutes, from the NIMITEK corpus (Gnjatović and Rösner, 2008) with different speakers. So that at the end we have eight annotated sessions for each emotion method. Additionally we asked the annotators to fill a questionnaire.

As with this small amount of data there is only a quantitative prediction possible, we limit our investigation to some single points in the comparison. Here we focus on this main aspects:

- Which method reflects real emotions best,
- is it possible to label emotions with different intensities or mixture emotions and
- does it provide a robust method, that is manageable by annotators?

This investigation does not aim to offer the best labeling for emotions, but it tries to enhance the emotion categories for HCI. How can we answer the questions above? First we will compare Basic Emotions (e.g. the so called big six: anger, fear, happiness, sadness, disgust, surprise, etc. (Keltner et al., 2003)) and the GEW. Here we included a option called “other”. If the available categories cover the expressed emotions well enough this option should be chosen less often and of course vice versa. Furthermore, if the options and vocabulary is increased the fraction of a ”neutral” label should also decrease. Both of these assumptions could be verified by comparing the behavior of the annotators as shown in Table 1. Additionally, the comments regarding the questionnaires showed, that the annotators could handle the GEW much better, as finer labels are possible. The basic emotions were more often identified with high arousal emotions, such as anger or happiness.

Table 1: Proportion of category “other” in Basis Emotions and GEW

	Basis Emotions	GEW
other	16.2%	1.6%
neutral	46.0%	20.3%

Since SAM doesn’t provide categorical emotion labels we couldn’t compare it in the same way. However it is left to say, that SAM doesn’t allow mixed emotions and it is not intuitive for the annotators according to the ratings in the questionnaire.

We plan further tests with more subjects, to give a more reliable conclusion.

5. Outlook

As this is ongoing work, the current status of the tool is that it is capable of assigning the aforementioned labels to any number of audio or video files specified in a properties file or stored in a specific folder. Furthermore, we are considering releasing the software as open source for other research projects looking for a flexible and theoretically funded labeling tool capable of dealing with non prototypical emotion expressions as well as with the temporal characteristics of emotion. We have developed this tool since we believe that a lot of information is being lost if simply discrete labels are assigned to realistic expressions considering the existence of emotion dynamics and mixed states. Additionally, we humans are not always sure about the expressed emotion and uncertainty is a big issue if labeling crisply, however using the proposed method we might extract some additional information from multiple labels for one expression and the temporal progress of the label. Therefore, we also want to compare the labeling of the six basis emotions against our method, using a questionnaire to see if deciding between different intensities, having more variabilities

and using additional temporal progression gives advantages to emotional classification with a much larger subject pool. Furthermore, a test verifying the human capabilities to label expressions using the provided temporal progress labeling tool using acted emotional data will be performed and evaluated in the near future. In this test the annotators will be listening to several pseudo word groups taken from the WaSeP dataset (Wendt and Scheich, 2002). In each group one or two emotional and objective words with no emotional coloring will be played back and the annotator will be asked to identify the emotional peaks. The results will then be verified against the ground truth.
Additionally in the future, we will be using the extracted labels using fuzzy machine learning techniques such as the fuzzy-input fuzzy-output support vector machines proposed in (Thiel et al., 2007), where fuzzy labels were already essential for improving the results of automatic classification tasks.

Acknowledgment

The presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

6. References

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of german emotional speech. In *Proceedings of Interspeech 2005*.

Roddy Cowie and Randolph R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Commun.*, 40(1-2):5–32.

Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle Mcmahon, Martin Sawey, and Marc Schröder. 2000. 'feeltrace': An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 19–24, Belfast. Textflow.

R. Craggs and M. McGee Wood. 2004. A two dimensional annotation scheme for emotion in dialogue. In *AAAI Spring Symposium: Exploring Attitude and Affect in Text 2004*.

C. Darwin. 1978. *The expression of emotion in man and animals*. HarperCollins, London, 3rd edition.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. 2007. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII'07)*, pages 488–500, Berlin, Heidelberg. Springer-Verlag.

M. Gnjatovic and D. Roesner. 2008. On the role of the nimitek corpus in developing an emotion adaptive spoken dialogue system. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

M. Gnjatović and D. Rösner. 2008. On the role of the nimitek corpus in developing an emotion adaptive spoken dialogue system. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

D. Keltner, P. Ekman, G. C. Gonzaga, and J. Beer, 2003. *Handbook of Affective Sciences - Facial expression of emotion*, chapter 22, pages 415–432. Affective Science. Oxford University Press.

P.J. Lang, 1980. *Behavioral Treatment and Bio-behavioral Assessment: Computer Applications*, pages 119–137. Ablex Publishing, Norwood (NJ), USA.

A. Ortony, G.L. Clore, and A. Collins. 1988. *The cognitive structure of emotion*. Cambridge University Press, Cambridge, UK.

R. W. Picard. 2000. *Affective Computing*. MIT Press Cambridge.

R. Plutchik. 1966. Psychophysiology of individual differences with special reference to emotions. *New York Academy Sciences Annals*, 134:776–781, February.

K. R. Scherer, T. Johnstone, and G. Klasmeyer, 2003. *Handbook of Affective Sciences - Vocal expression of emotion*, chapter 23, pages 433–456. Affective Science. Oxford University Press.

S. Scherer, E. Trentin, F. Schwenker, and G. Palm. 2009. Approaching emotion in human computer interaction. In *International Workshop on Spoken Dialogue Systems*, pages 156–168. IEEE.

K. R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):693–727.

P.-M. Strauss, H. Hoffmann, W. Minker, H. Neumann, G. Palm, S. Scherer, H. Traue, and U. Weidenbacher. 2008. The pit corpus of german multi-party dialogues. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

C. Thiel, S. Scherer, and F. Schwenker. 2007. Fuzzy-input fuzzy-output one-against-all support vector machines. In *11th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems KES 2007*, volume 3, pages 156–165. Springer. note: Won the Best Paper award of the conference!; key: F2SVMS.

B. Wendt and H. Scheich. 2002. The "magdeburger prosodie korpus" - a spoken language corpus for fmri-studies. In *Speech Prosody 2002*. SProSIG.