IKANNOTATE2 – A TOOL SUPPORTING ANNOTATION OF EMOTIONS IN AUDIO-VISUAL DATA

Ingo Siegert, Andreas Wendemuth

Institute for Information and Communications Engineering, Cognitive Systems Group, Otto-von-Guericke University, 39016 Magdeburg, Germany, www.cogsy.de
ingo.siegert@ovgu.de

Abstract: For emotional analyses of interactions, qualitatively high transcription and annotation of given material is important. The textual transcription can be conducted with several available tools, like e.g. Folker or ANVIL. But tools for the annotation of emotions are quite rare. Furthermore, existing tools only allow to select an emotion term from a list of terms. Thus, a relation between the different emotional terms that has been uncovered by psychologists get lost.
In this paper, we present an enhanced version of the tool ikannotate that is able to add an emotional annotation onto already transcribed material. This tool relies on established emotion labelling methods, like the Geneva Emotion Wheel or the Self Assessment Manikins to maintain the relationship. Furthermore, the annotator is guided by a step-wise process to improve the reliability of the emotional annotation. Additionally, the uncertainty in assessing emotions can be covered as well, to evaluate the labels afterwards and exclude samples with to low uncertainty from further analyses. The tool *ikannotate2* can be used under Windows, Linux and macOS. All settings can be changed via corresponding INI-files.

1 Introduction

Automatic emotion recognition is, like several other pattern recognition techniques, data driven. This means that a large number of data samples along with a "class" label is needed to train a model representation. This model can be used afterwards to evaluate unknown data samples of similar type. The method of manually assigning class labels to data samples is often referred as annotation or labeling. It is quite evident, that for robust emotional analyses, high quality labeling of given material is important.

In the case, where acted emotional data should be labeled, the labeling can be solved quite easily. The label is clearly instructed to the actor by the experimenter and thus the label of the emotion utterance is intrinsically given. To guarantee high quality data, the affective quality can be easily assessed afterwards via perception tests. But for data gathered within a naturalistic interaction or if the previous mentioned procedure of perception tests cannot be conducted, an expensive manual labeling is needed. The most promising method of obtaining a valid emotion assignment would be a self-assessment by the observed subject itself [1]. Unfortunately, this is not always feasible, as the experimental setup does not allow an interruption for a self-assessment or the subject is not available for an additional assessment afterwards. This is the case when the data samples were not intended to be used as emotion data. For example the data used in the Emotion Recognition in the Wild Challenge came from cinema movies [2]. In these cases, another subject, called labeler, has to assess the experimental data by assigning an emotional label based on his observation.

Sadly, also this kind of labeling does not necessarily reflect the emotion truly felt by a subject, as felt emotions are not always perceivable by observers [3] and as display rules and cognitive effects influence the assessment [4]. To overcome these issues it is advisable to employ several raters to label the same content, use a majority voting, select a suitable emotion labeling method and optionally record the uncertainty of the labelers.

As the labeling can be conducted together with a textual transcription, tools used for transcription, like e.g. ANVIL [5], Exmaralda [6] or Folker [7], also support a labeling process. But these tools do not support the labeling of emotions using emotional representations. They only allow to select an emotion from a list of terms, which have to be pre-defined. Thus, the relation between the different emotional terms that has been uncovered by psychologists (c.f. [8, 9]) gets lost.

Another aspect, that has to be taken into account, is that the labeling of emotions requires quite a lot of attention of by labeler. Therefore, the computer program used to conduct the labeling should be straightforward. The program should support the labeling process at all stages but should also secure a proper labeling. As the previous mentioned tools are general purpose tools for transcription of interactions, they have to provide much more functionality that is not needed for the labeling process and could in the worst case hinder a proper labeling, as the labeler gets distracted. Therefore, a separate tool specifically designed for the labeling with the support of proper emotion labeling schemes should be used. But, using a separate program to conduct the labeling implies the risk of data loss. To avoid this issue, the labeling tool should use the same database as well-known transcription tools, ideally using the same data-format.

In the following we present an emotion labeling program that is easy to use, integrates several psychological emotion representations usable for emotion labeling and allows a integration into the data preparation chain, as it can directly work with Folker/Exmaralda XML-transcriptions as well as with utterance-based text files.

2 *ikannotate2* a tool for emotional annotation

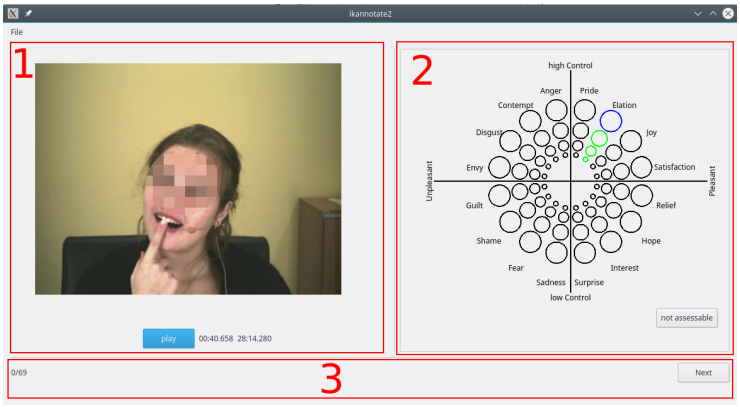


Figure 1 – Screenshot of the *ikannotate2* GUI with the three main parts, the player part (1), the labeling part (2), and the controls part (3)

To overcome the mentioned challenges, we developed the tool *ikannotate2* , specialized on the emotion labeling process, see Figure 1for a screenshot. This version is a successor of the previous developed tool *ikannotate* [10]. In contrast to this older version, *ikannotate2* supports video files, direct processing of Folker and Exmaralda XML files, and a step-wise

labeling of emotional assessments according to the established Geneva Emotion Wheel (GEW). *Ikannotate2* aims to support the labeling process, by making it as smooth as possible using a very easy to understand and intuitive GUI. Thus, the labeler can focus on the labeling process itself. On the other hand, *ikannotate2* has a very strict progress model, guaranteeing a proper labeling. The labeler has to listen to the data in a pre-defined order and is only allowed to choose the label after listening to the actual sample once completely (although the labeler can listen to the sample several times). Only after giving an assessment, the next sample can be processed. A restriction of this approach is that this tool can be only used for the assignment of labels and not to also determine the segments which should be labeled. More specifically, *ikannotate2* is designed to support utterance-based or event-based annotations.

The labeling progress is saved automatically after each assessment and the labeler can interrupt the labeling at any time, without the need to manually save the progress. This procedure does not allow to go back and correct the assignment of a previous sample, but this can be compensated with an anyway necessary training phase. All further settings can be specified using an INI-File.

As this tool is able to use Folker/Exmaralda XML-files for labeling, no data-conversion is needed when *ikannotate2* is used for this type of data. In this case *ikannotate2* uses samples defined as <elements> of a tier. The start and end times of all elements within that tier together with the new emotion label are then stored in a new tier. The attribute `id` of the new tier, displayed as short description within the Folker/Exmaralda GUI can be set within *ikannotate2* . The newly added tier-attribute `ikannotate` indicates which tier should be processed after a restart of *ikannotate2* and also stores the link to the original tier and its elements. If several tiers are processed with *ikannotate2* , a dialog letting the labeler choose the intended tier is displayed. A shortened example of the resulting XML is displayed in Listing 1. The XML file is auto-saved after each labeling step, comprising the result of the actual sample. In addition,

```
<?xml version="1.0" encoding="UTF-8"?>
<basic-transcription>
  <head> ... </head>
  <basic-body>
    <common-timeline>
      <tli id="TLI_0" time="184.17018707797996"/>
      <tli id="TLI_1" time="186.2371779543404"/>
    ...
    </common-timeline>
    <tier id="Utterances" speaker="I" category="v" type="t">
      <event start="TLI_0" end="TLI_1">
        ja also du hast ja grade diesen versuch ge
      </event>
    ...
    </tier>
    ...
    <tier id="Emotion" speaker="I" category="v" type="t" ikannotate="TIE_">
      <event start="TLI_0" end="TLI_1">
        anger
      </event>
    </tier>
  </basic-body>
</basic-transcription>
```

Listing 1 – Excerpt of an Exmaralda XML-File where *ikannotate2* is used for emotion annotation. The tier Emotion comprises the emotion labels for the tier Utterances.

ikannotate2 allows to use blank text-files for the allocation of separate items. A simple text-file with the suffix `.data` is used, where each line represents a new sample (mediafile representing the sample) that has to be labeled, see Listing 2. The first line indicates the result file. To implement the auto-save functionality, the actual labeled sample is removed from the data-file and added to the results file, see Listing 3. Thus over the labeling process, the data-file is

step-wise decreasing while the result-file is step-wise increasing.

```
resultFilename:TestLabeling
sample1.wav
sample2.wav
...
sampleN.wav
```

Listing 2 – Excerpt of a text-based datafile for *ikannotate2*.

```
sample1.wav;Anger
sample2.wav;Fear
...
sampleN.wav;Disgust
```

Listing 3 – Excerpt of the result file (TestLabeling) for the corresponding datafile.

The tool is developed with the C++ application framework Qt5, thus *ikannotate2* runs natively under Windows, Linux and macOS and is able to play audio and video files in various formats, only limited by the support of the operating system. The Qt Multimedia APIs are built upon the multimedia framework of the underlying platform, which is DirectShow for Microsoft Windows, QuickTime Player for MacOS and GStreamer on Linux. An overview of supported codecs can be found in Table 1. If a codec is missing on Windows, it is sufficient to install a proper codecs packages, like K-Lite [11]. For recent MacOS versions greater than 10.9, only QuickTime Player X is used which cannot play all multimedia formats. Unfortunately to date, a suitable codec package is not known.

Table 1 – Overview of the most important supported multi media formats of *ikannotate2*.

Codec	**Windows 7 DirectShow 9.0**	**Linux Gstreamer 1.10.2**	**MacOS QuickTime X 10.4**
WAV	X	X	X
AIFF	-	X	X
FLAC	-	X	-
AAC	X	X	X
WMA	X	X	-
MP3	X	X	X
OGG Vorbis	-	X	-
Opus	-	X	-
WMV	X	X	-
AVI	X	X	X
H.264	X	X	X
H.265	-	X	-
MPEG1	X	X	X
MPEG2	X	X	X
MPEG4	X	X	X
OGG Theora	X	X	-
Quicktime	-	X	X
VP8	-	X	-
VP9	-	X	-
Shockwave Flash	-	X	X

3 Description of the implemented emotional labeling methods

As *ikannotate2* is specialized for (emotional) labeling, several emotion representations are built-in. A variant of the GEW [9] and the Self Assessment Manikins (SAM) [12] and additionally, self-defined lists of (emotional) category terms can be used. In the following the various emotion representations are described and their implementation in *ikannotate2* are presented.

3.1 Self-defined list of categorical terms

A very common method for assigning emotional labels is the use of (emotional) category terms. Descriptive labels, usually not more than ten, are selected to describe the emotion. These labels can be formed from counterparts like positive vs. negative or comprised from a selection of emotion terms based on an emotion theory. The definition of the labels have to be specified using the text-file "Labels.conf", see Listing 4. Additionally, a free text field can be used to give the labeler the opportunity to indicate a self-chosen category term. This option has to be specified in the INI-file.

```
Title=Basic Emotions % Free choosable title, will be displayed above labels
Label=anger|boredom|disgust|fear|joy|sadness|neutral %Definition of labels separated by |
```

Listing 4 – Example content for the Labels.conf.

One disadvantage of category terms is the missing relationship between the labels. This makes it difficult for the labeler to give an evaluative assessment. The labels and their meaning also have to be introduced to the labeler as the subjective interpretation can differ from labeler to labeler. Furthermore, the application of category terms to other languages requires a complex task of translation and validation. Another flaw is that the selection of emotional terms is mostly limited to a specific domain or selection of emotions. This can cause that some emotional observations are missing a proper label and thus get lost or will be merged into inappropriate emotional terms. This may later complicate the emotion recognition since the emotional characteristics of these merged phenomena may also differ.

3.2 Geneva Emotion Wheel (GEW)

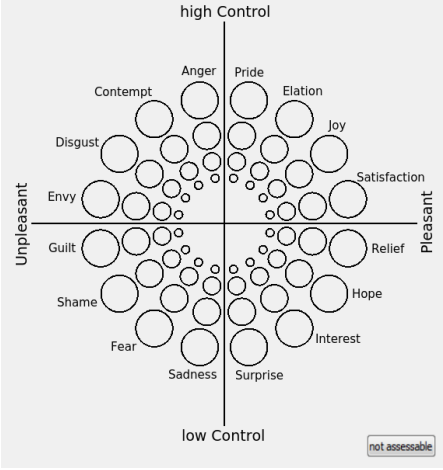
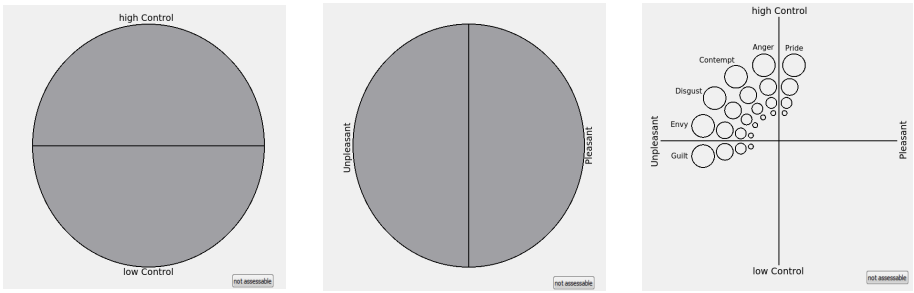


Figure 2 – The *ikannotate2* GUI part of the built-in GEW labeling method.

A prominent solution to overcome the mentioned problems of the category terms is the GEW introduced by Scherer [13]. It is a theoretically derived and empirically tested instrument to measure emotional reactions to objects, events, and situations. The GEW consists of 16 emotion categories, called "emotion families", each with five degrees of intensity, arranged in a wheel shape in the control and pleasantness space. Additionally, the option "not assessable" is added to provide the labeler with the opportunity to assign unspecific situations. This

arrangement supports the labeler in assessing a single emotion family with a specific intensity by guiding him with the axes and quadrants [14]), see Figure 2.

Unfortunately, the labeling effort using GEW is quite high since the labeler has to mark the emotion via a multi-step approach: 1) decide on the control axis to get the semicircle, 2) choose the value for pleasantness to get the quadrant, and 3) decide between the remaining four emotion families. This gets even more complex, when the intensity of an emotion should be assessed, too. Therefore, *ikannotate2* can support the labeling by offering a three-step approach. First, the labeler decides between high and low control. Then, the labeler decides between pleasant and unpleasant. Afterwards the resulting quadrant is displayed and the annotator selects one of the emotion families with the corresponding intensity, see Figure 3. To cover borderline case, the emotion families adjacent to the resulting quadrant are displayed as well. By this method the complexity of the annotation is reduced. The labeler first has to make two decoupled binary decisions, only in the last step a selection of emotion families reduced by the decisions of two previous steps are presented. Thus, the labeler only needs to resolve 6 emotion families rather than 16. In *ikannotate2* it is possible to use both versions of GEW, the three-step version and the one-step full version. This behavior can be managed by the INI-File.



(a) Step 1 **(b)** Step 2 **(c)** Step 3

Figure 3 – Stepwise labeling using the built-in GEW of *ikannotate2*.

3.3 Self Assessment Manikins (SAM)

Another emotion assessment designed to circumvent the challenges caused by a verbal description, is the picture-oriented instrument invented by Lang. The so-called SAM can be used to assess the pleasure, arousal, and dominance dimension directly [15]. It depicts the three dimensions by 3×5 figures, see Figure 4. These figures depict the main characteristic for each dimension in changing intensity, for instance changing from a happy smiling manikin to a weeping, unhappy one to represent pleasure. Furthermore, this method is also usable with labelers that are not "linguistically sophisticated", like children. One disadvantage of this method is the missing ability to evaluate distinct or blended emotions.

4 Including uncertainty labeling

Another aspect where *ikannotate2* supports the emotion annotation is the indication of the labeler's uncertainty. Despite the variety of labeling methods, it is hard to assign the correct emotion. Thus, the labeling of emotions is accompanied with uncertainty. To avoid confusions in later analyses, and to take into account the labeler's uncertainty, *ikannotate2* provides the possibility to indicate the degree of uncertainty for the current labeled sample. The uncertainty

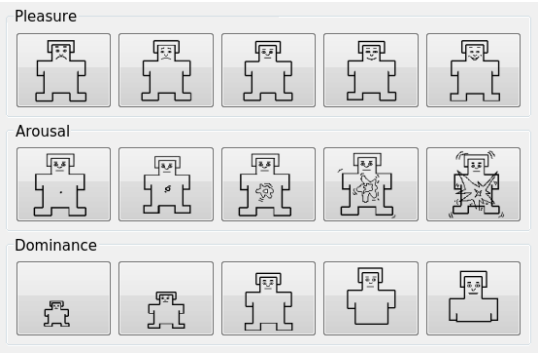


Figure 4 – The *ikannotate2* GUI part of the built-in SAM labeling method.

values provide the researcher with additional information about the labeling and can be used to combine labels with high uncertainty into a separate "unidentifiable" class. Furthermore, it gives an indication of the quality of the assignment. The scale division can be specified within the INI-File. The uncertainty is implemented as a scroll bar. The uncertainty value is stored as an additional number together with the emotion assessment: `sample.wav:anger;1`.

5 Conclusion

In this paper we introduced the tool *ikannotate2*, having a simple GUI for emotional labeling while also offering many possibilities for the researcher to control the annotation. The step-wise process of the labeling is implemented in such a way that the labeler can fully concentrate on the labeling process.

The designer or developer of the annotation task can define the labeling method and the order of samples. Furthermore, he can specify the emotion terms for the Emotional Word Lists. When using the GEW the designer can decide if the labeling should be conducted in one single step or via a step-wise process. Additionally, the uncertainty of the labeler in assigning a label can be recorded and used to develop classifiers, which handle and incorporate this information. Furthermore, the categorical terms can also be used to annotate not only emotions but also dialog functions, sentence types, etc.

The tool is able to use text files as well as Folker/Exmaralda XML-files as input for the labeling of whole pre-defined segments. It has been successfully used in master's theses and research projects, e.g. [16, 17]. The current version of *ikannotate2* is available for academic research on demand from ingo.siegert@ovgu.de.

Acknowledgments

The work presented in this paper was done within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation (DFG).

References

[1] GRIMM, M. and K. KROSCHEL: *Evaluation of natural emotions using self assessment manikins*. In *Proc. of the IEEE ASRU*, pp. 381–385. Cancún, Mexico, 2005.

[2] DHALL, A., R. GOECKE, G. T., and N. SEBE: *Emotion recognition in the wild. Journal on Multimodal User Interfaces*, 10, pp. 95–97, 2016.

[3] TRUONG, K. P., A. DAVID LEEUWEN, VAN, and F. M. G. JONG, DE: *Speech-based recognition of self-reported and observed emotion in a dimensional space*. *Speech Commun*, 54, pp. 1049–1063, 2012.

[4] FRAGOPANAGOS, N. F. and J. G. TAYLOR: *Emotion recognition in human-computer interaction*. *Neural Networks*, 18, pp. 389–405, 2005.

[5] KIPP, M.: *Anvil - a generic annotation tool for multimodal dialogue*. In *Proc. of the INTERSPEECH-2001*, pp. 1367–1370. Aalborg, Denmark, 2001.

[6] SCHMIDT, T. and K. WÖRNER: *EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research*. *Pragmatics*, 19, pp. 565–582, 2009.

[7] SCHMIDT, T. and W. SCHÜTTE: *FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction*. In *Proc. of the 7th LREC*, pp. 2091–2096. Valletta, Malta, 2010.

[8] PLUTCHIK, R.: *Emotion, a psychoevolutionary synthesis*. Harper & Row, New York, USA, 1980.

[9] SCHERER, K. R.: *What are emotions? and how can they be measured? Soc Sci Inform*, 44, pp. 695–729, 2005.

[10] BÖCK, R., I. SIEGERT, M. HAASE, J. LANGE, and A. WENDEMUTH: *ikannotate - a tool for labelling, transcription, and annotation of emotionally coloured speech*. In S. D'MELLO, A. GRAESSER, B. SCHULLER, and J.-C. MARTIN (eds.), *Affective Computing and Intelligent Interaction.*, vol. 6974 of *Lecture Notes in Computer Science*, pp. 25–34. Springer Berlin, Heidelberg, 2011. doi:10.1007/978-3-642-24600-5.

[11] CODEC GUIDE: *K-Lite Codec Pack - For Windows 10 / 8.1 / 7 / Vista / XP*. `https://www.codecguide.com/`, 2016. [Online; accessed 27th-December-2016].

[12] BRADLEY, M. M. and P. J. LANG: *Measuring emotion: The self-assessment manikin and the semantic differential*. *J Behav Ther Exp Psy*, 25, pp. 49–59, 1994.

[13] SCHERER, K. R.: *Unconscious Processes in Emotion: The Bulk of the Iceberg*, pp. 312–334. Guilford Press, New York, USA, 2005.

[14] SACHARIN, V., K. SCHLEGEL, , and K. R. SCHERER: *Geneva Emotion Wheel rating study*. Tech. Rep., Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences, 2012.

[15] LANG, P. J.: *Behavioral treatment and bio-behavioral assessment: Computer applications*, pp. 119–137. Ablex Publishing, New York, USA, 1980.

[16] SIEGERT, I., D. PHILIPPOU-HÜBNER, M. TORNOW, R. HEINEMANN, A. WENDEMUTH, K. OHNEMUS, S. FISCHER, and G. SCHREIBER: *Ein Datenset zur Untersuchung emotionaler Sprache in Kundenbindungsdialogen*. In W. GÜNTHER (ed.), *Elektronische Sprachsignalverarbeitung 2015. Tagungsband der 26. Konferenz*, vol. 78 of *Studientexte zur Sprachkommunikation*, pp. 180–187. TUDpress, Eichstätt, Germany, 2015.

[17] SIEGERT, I., A. F. LOTZ, M. MARUSCHKE, O. JOKISCH, and A. WENDEMUTH: *Emotion intelligibility within codec-compressed and reduced bandwith speech*. In *ITG-Fb. 267: Speech Communication : 12. ITG-Fachtagung Sprachkommunikation 5. – 7. Oktober 2016 in Paderborn*, pp. 215–219. VDE Verlag, 2016.