A Multimodal Annotation Schema for Non-Verbal Affective Analysis in the Health-Care Domain

Federico M. Sukno Mónica Domínguez Adrià Ruiz
Department of Information and Communication Technologies
Pompeu Fabra University, Spain

Dominik Schiller Florian Lingenfelser
Human-Centered Multimedia
Augsburg University, Germany

Louisa Praagst
Institute of Communications Engineering
Ulm University, Germany

Ekeni Kamateri Stefanos Vrochidis
Information Technologies Institute
Centre for Research & Technology Hellas, Greece

ABSTRACT

The development of conversational agents with human interaction capabilities requires advanced affective state recognition integrating non-verbal cues from the different modalities constituting what in human communication we perceive as an overall affective state. Each of the modalities is often handled by a different subsystem that conveys only a partial interpretation of the whole and, as such, is evaluated only in terms of its partial view. To tackle this shortcoming, we investigate the generation of a unified multimodal annotation schema of non-verbal cues from the perspective of an inter-disciplinary group of experts. We aim at obtaining a common ground-truth with a unique representation using the Valence and Arousal space and a discrete non-linear scale of values. The proposed annotation schema is demonstrated on a corpus in the health-care domain but is scalable to other purposes. Preliminary results on inter-rater variability show a positive correlation of consensus level with high (absolute) values of Valence and Arousal as well as with the number of annotators labeling a given video sequence.

Keywords

valence-arousal, human-machine interaction, multimodal analysis, embodied conversational agents

1. INTRODUCTION

Multimodal systems, especially those involving embodied conversational agents (ECAs), tend to be heterogeneous by definition, as they encompass different areas of expertise (i.e., computer vision, gesture generation, speech technologies, dialogue management), each of which entangles their own idiosyncrasy. This heterogeneity explicitly applies to the standards adopted in each field for data representation of extracted affective states during analysis, in terms of both the generated output and ground-truth data sets used for evaluation of each modality. The latter entails that validation (and training, if applicable) of multimodal systems may require different sets of manual annotations for each modality produced by appropriately trained experts, to serve as ground-truth for each of the subsystems and, additionally, to the final multimodal system as a whole. In order to maximize annotation efforts, and under the premise of considering human non-verbal behavior as a complex whole rather than as an aggregation of isolated events, it seems reasonable to aim at developing a system that extracts and reacts to relevant affective states deploying a unique multimodal representation accounting for changes in affection across discrete time intervals. Such a strategy resembles human behavior when perceiving non-verbal cues and is independent from modality-specific segmentations or representations.

In this paper, we investigate the generation of a unified multimodal annotation schema of non-verbal cues with particular emphasis, but not limited to, the health-care domain. To this end, we have firstly addressed the recording of a naturalistic corpus of spontaneous dialogues in this domain and then created an inter-disciplinary group of annotators with technical expertise in the development and implementation of individual modules to be integrated in an information agent with social competence and human-like interaction capabilities. These modules account for facial expression, gestures, speech prosody, emotion recognition and, more concerned with the functionality of ECAs, ontology representation and dialogue management.

For representing affective states, the well-established Valence and Arousal space is chosen using a discrete scale of non-linear values. Such representation is especially suited to our goals as it is modality-independent and is able to capture changes in the affective state within the dialogue flow. Given the unavoidable subjectivity of affect-related events together with the fact of the highly heterogeneous group of annotators, an objective metric is established as a systematic control measure to assess the consensus level achieved for each dialogue and to determine if corrective measures should be applied in case the consensus score is below a certain threshold. Preliminary results suggest not only a reasonable degree of agreement of the produced annotations, but also a progressive improvement of the annotations as the group evolves through the annotation task.

The rest of the paper is structured as follows: Section 2 provides an overview of previous work related to multimodal affective datasets, their annotation and use in recognition of non-verbal cues; Section 3 is a general overview of the corpus being used in this annotation task; Section 4 includes the description of annotation criteria using Valence and Arousal, with a special focus on Subsection 4.2 on the analysis of joint annotations and Subsection 4.3, where an objective metric is used to assess the optimal number of annotators to achieve a reliable consensus for each video. Finally, conclusions are drawn in Section 5.

2. RELATED WORK

2.1 Multimodal Affective Databases

Affective states are displayed through various channels or *modalities*, including facial expressions [12], vocal prosody [11], gestures and postures [4]. Since the cues involved in non-verbal communication of affect are encoded within multiple modalities, the recognition process should incorporate as much information as possible from each of them [25]. Numerous elaborate methods for fusing multiple modalities in affect recognition have been reported [24, 26, 22, 23, 17] and their results confirm the assumption that multimodal fusion generates more accurate affect recognition systems than unimodal approaches.

Despite several datasets have been collected in the past for affect analysis [25], they are usually focused in concrete modalities such as speech [14], facial expressions [15] or body gestures [8]. Moreover, naturalistic dialogues are not present in most of them. In fact, the number of existing multimodal affective databases containing naturalistic human interactions is rather limited. Some exceptions are the SEMAINE [16], RECOLA [18] and Vera am Mittag [7] datasets. However, none of these databases was recorded in the health-care domain. For example, RECOLA includes interactions of people performing collaborative tasks, while Vera am Mittag contains recordings from a German TV Talk-show. Given that the range of emotions and their intensity are very dependent on the context [2], the development of ECAs depends on collecting relevant data envisaging the concrete register and domain where the agent will be deployed. For this reason, in Section 3 we will describe our current efforts in the collection of a new dataset that provides recordings of specific scenarios belonging to the health-care domain.

2.2 Annotating Affective States

Defining an annotation schema is essential for creating multimodal affective datasets. Traditionally, data annotation in this context has followed two main paradigms: categorical and dimensional, each with its own advantages and shortcomings [9]. In the categorical model, affective states are defined using discrete labels (such as *happiness*, *sadness* or *boredom*). In contrast, the dimensional model defines affect in a continuous space where dimensions represent different psychological concepts using a numeric scale. Most relevant dimensions to represent affect using this model are Valence and Arousal. Valence refers to how pleasant or unpleasant is an affective state while Arousal indicates the activation or deactivation level [5].

It is generally accepted that the category-based paradigm is limited for two main reasons [9]. Firstly, affective states involved in every-day life are too complex to be well represented by a limited number of discrete categories; unfortunately, augmenting the number of possible labels complicates the annotation process and lowers inter-annotator agreement [1]. Secondly, while the dimensional model can naturally represent blended emotions because affective states share a common set of continuous dimensions, this blending is not possible in the category-based model, since emotion labels are considered independent and there is not a notion of distance between them.

For these reasons, in this work we have adopted the dimensional paradigm where annotations are provided for the Valence and Arousal space. This approach has also been followed in the SEMAINE and RECOLA datasets [16, 18], but their annotation schema requires annotators to provide continuous frame-to-frame measurements in real-time. Such a strategy is likely to produce lower inter- and even intra-rater agreement as it forces annotators to make instantaneous decisions for labeling. In contrast, our annotation schema is based on labeling short-time segments with a limited number of labels representing a discretization of the continuous Valence and Arousal dimensions. Discrete labels should improve inter-rater agreement [10] and, together with the use of short time segments, facilitates self-revision and correction of the annotated labels.

3. CORPUS OVERVIEW

In this section, specifications for recordings are briefly summarized as they are considered relevant for the acquisition of a naturalistic corpus in a specific domain, in this case, the health-care domain. As corpus recording tasks are currently in progress, a more detailed description of the acquired corpus will be conveniently provided in further publications. Technical specifications on recording equipment used in these tasks are presented in 3.1.

As stated in the previous section, creating a domain-specific corpus is instrumental for the development of ECAs in several aspects, e.g. analysis of communicative cues, modeling, training of algorithms and evaluation of performance. In our specific case, it is of utmost importance to comply with a range of cultural requirements as the final ECA is intended to interact with specific migrant communities (e.g. Polish, Turkish and Arabic) providing information about concrete health-care issues in the host country (i.e. Spain for Arabic Migrants and Germany for Turkish and Polish migrants). Therefore, recording tasks are specifically designed to cover five main requirements:

- Naturalness: spontaneous interaction in dialogue format between two participants: the user (requesting information about concrete topics) and an expert (answering to this information request as the system is expected to perform).
- Participants' profile: recruitment is carried out taking into account gender, age, linguistic proficiency, cultural background and expertise profile according to the intended use case the system is required to cope with. If participants take the system's role, they must have a specific profile in the health-care domain as well.

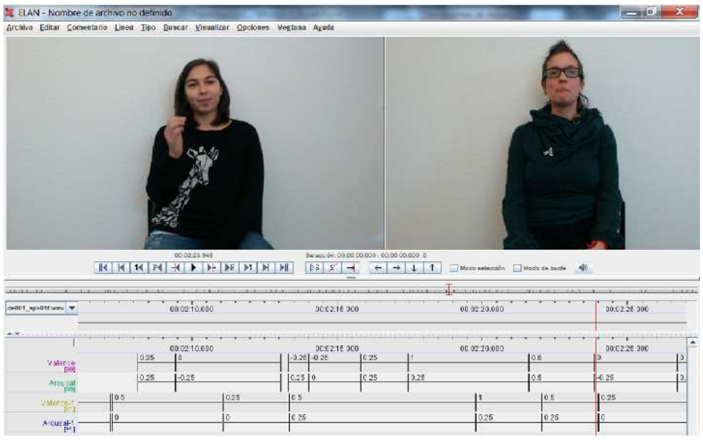
ELAN - Nombre de archivo no definido
Archivo Editar Comentario Línea Tipo Buscar Visualizar Opciones Ventana Ayuda
00:02:23.948
Selección: 00:00:00.000 - 00:00:00.000 0
Modo selección
Modo de bucle
00:02:10.000
00:02:15.000
00:02:20.000
00:02:25.000
Valence
Arousal
Valence-1
Arousal-1
0.25
0
-0.25
0.5
1

Figure 1: Example of the dialogue recordings, together with Valence-Arousal annotations, visualized with ELAN.

Table 1: Corpus Description

Dialogues	Time	Speakers	Culture
66	4h	8	German
36	3.5h	9	Spanish
69	3.5h	6	Polish
9	1.5h	9	Turkish
12	1.5h	8	Arabic
192	14h	40	Total

- Topic preparation and prompting: each dialogue topic is restricted in terms of content (a maximum of five subtopics) and duration (from 2 to 10 minutes). Conveniently selected participants with experience in the topic are given some generic indications (to try keeping a spontaneous behavior) and a list of key ideas as guidance. Further elaboration on the topic and exact wording is strictly avoided when instructing participants before the dialogue is recorded.
- Technical constrains: the recording equipment needs to mimic the setup of the final ECA which is intended to work on a wide range of standard specification equipment such as mobile phones, tablets and PCs. Therefore the choice of sensors used for recording is limited by the subset of commonly available devices, i.e. one webcam and one microphone.

So far, there have been three rounds of recordings, comprising a total of approximately 14 hours of audio-visual material, distributed as indicated in Table 1. Annotation tasks at present have only covered 12 representative dialogues in German, Spanish, Turkish and Arabic to develop the annotation criteria.

3.1 Technical Setup

Audio and video files of each participant are recorded using the open-source Social Signal Interpretation framework (SSI) [20] as the basis for the recording software. Multimodal synchronization of audio and video from the same speaker is automatically handled by the SSI framework. Additionally, the system takes care of starting and stopping the recordings simultaneously for each participant within one session to preserve the chronological order of the dialogues, ensuring simultaneous and synchronized recordings.

The video framing was setup with the premise to guarantee the visibility of the upper part of the body, since face, arms and hands movements are essential for nonverbal analysis. To optimize the recordings for usage with various machine learning techniques audio and video were separately recorded, both with high quality settings: 16 bit at 48 kHz for audio (stored in PCM-WAV format), 720p at 30 FPS for video (stored in H264-MP4 format).

4. MULTIMODAL ANNOTATION OF NON-VERBAL CUES

An initial group of eight human experts is taking part in the annotation of the multimodal corpus described in Section 3. Each expert produces individual Valence-Arousal annotations for a selection of videos using a seven point scale (described below). While all annotators share the interest in non-verbal analysis in the context of multimodal ECAs, there is a considerable variability in their field of expertise, which includes facial expressions, gesture recognition and generation, speech technologies, emotion recognition, ontology representation and dialog management. Thus, the annotation group is quite heterogeneous, which provides an interesting complementarity to their labeling, as each of them analyzes the dialogues from a slightly different perspective.

On the other hand, such heterogeneity can also be an issue in terms of consensus. Recall that, after each expert has annotated a video, our main goal is to fuse those annotations to obtain a final labeling with sufficient confidence to be used as ground-truth. A crucial requirement to achieve this goal is the appropriate definition of the annotation criteria.

4.1 Annotation Schema

A set of guidelines were defined after discussion among all annotators. One of the essential elements of these guidelines concerns our aim to produce truly multimodal annotations, which implies labeling Valence and Arousal from an integrative view of the subject as a whole, rather than considering specific cues that are typically attached to a particular modality. For example, our interpretation of a smile (which could be unique from the point of view of facial expressions)

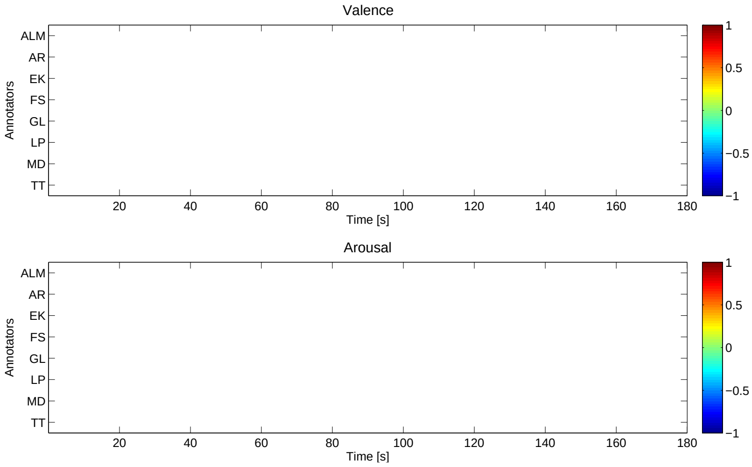

Figure 2: Example of Valence-Arousal labels: the x-axis represents the timeline; the y-axis contains rows for each annotators' labeled values, encoded as indicated by the color bars on the right.

can and should be modulated by the tone of voice or body gestures.

In the same line as the above, given our focus on non-verbal communication, annotators are instructed not to pay attention to the linguistic content (i.e. to ignore what the speakers are actually saying). Admittedly, this is a controversial aspect because, even though instructions are clear and well-agreed among all annotators, we cannot ignore a potential bias in perception when annotators understand the language spoken in the videos. As a side note, this corpus might allow for some quantitative analysis of such bias given that not all annotators were acquainted with the languages used by the speakers.

Once these preliminary basics are set, there are also some technical aspects to consider with regard to the actual annotation schema used to define and produce annotations:

- Axes: Valence and Arousal are labeled in different axes. Each axis is represented as a timeline with its own set of labels.
- Labels: 7 points are established in each axis, in a discrete but non-linear setting, as follows: 0, ±0.25, ±0.5, ±1. The choice of discrete labels is considered more appropriate to achieve higher inter-rater agreement when compared to continuous labels. The non-linearity of the scale is introduced to try capturing subtle deviations from the neutral state, which seem to occur more frequently than larger Valence-Arousal values, especially in a naturalistic setting as the one targeted by this corpus, probably due to the rather formal interaction within the health-care domain.
- Segmentation: a time division based on segments is established. These segments are freely defined by each annotator, asynchronously in terms of their duration but synchronously between axes. That is, the duration of each segment is defined as the time interval in which Valence and Arousal levels did not change, as determined by the perception of the annotator. A change in affective state (even if it involves only one axis) implies creating a new time segment and labeling both axes according to the new affective state. Finally, it was also agreed that, while time was considered a continuous variable, no segment should have a duration below 0.5 seconds.
- Default state: based on the expectation that for a large proportion of the recordings speakers would be in neutral state (both Valence and Arousal equal zero), such a state was defined as the default one, with no need to explicitly annotate it.

After evaluating multiple software tools regarding their fitness for our annotation requirements we found ELAN [21] to be the best suitable solution. ELAN is a flexible tool for assigning labels to a freely selectable time interval in audio- and video-files. Contrary to other popular tools, such as Gtrace [3] or CARMA [6], which are based on a continuous labeling approach, the discrete annotation schema of ELAN allows the annotator to adjust the time interval and the value of each label until it coincides exactly with his/her observation. This feature is particulary useful when reviewing annotations based on discrete labels (i.e. to establish common ground between all annotators). It is worth mentioning that ANVIL [13] would also be a reasonable choice. However, at least for the software versions we evaluated, ELAN is preferred given its better performance handling audio-visual formats.

4.2 Analysis of Joint Annotations

Fig. 2 shows an example of the Valence-Arousal labels produced by the different annotators for a 3-minute video. As expected, it is easy to identify discrepancies due to the subjective nature of perceived affective states in human behavior; there are at least two different labels for any selected time interval, with a few exceptions largely dominated by neutral states (recall that this is the default one and hence is not explicitly annotated).

On the other hand, it is also clear that for several time segments annotators shared at least a common tendency in their annotations. For example, slightly after 80 seconds, all raters agreed on assigning high values for Valence and moderately positive Arousal; two raters disagreed on the exact label for each axis (interestingly, they are not the same two for Valence and Arousal), but still chose the immediately

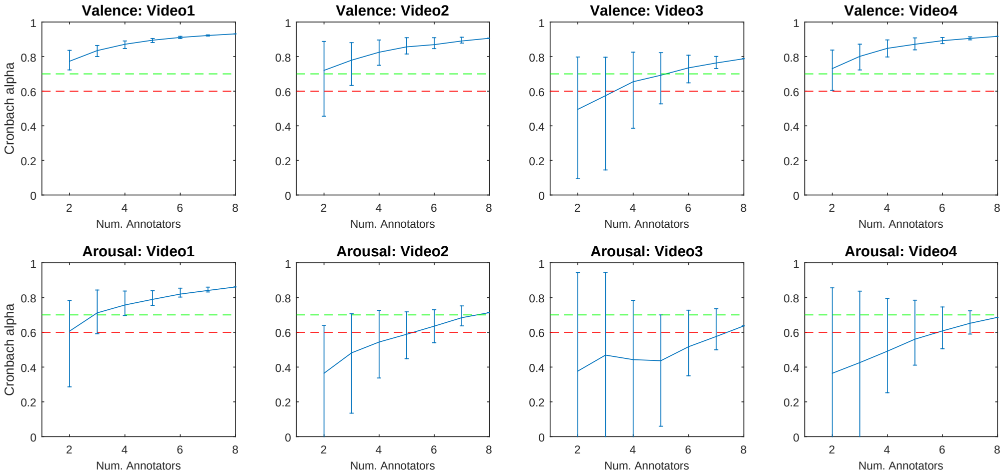

Figure 3: Inter-annotator agreement for 4 selected videos. Agreement is evaluated in terms of Cronbach's alpha by increasing the number of annotators. Values below the red line indicates not acceptable agreement whereas values above the green line indicate acceptable ones.

neighboring label. This high agreement is evident at several points in the video, with a remarkable alignment in the limits for the selected time segments (recall that time segments are freely chosen by each annotator on a continuous basis).

While Fig. 2 is just a particular example, it serves to illustrate the tendency that we observed throughout the corpus recordings processed so far:

- Certain time segments (*events*) obtained highly consensuated annotations. These events were typically of short duration and had rather well defined time limits.
- There were no remarkable deviations from the neutral state in the majority of videos and most raters assigned either zero or the lowest non-zero labels available. The latter was one of the greatest factors of disagreement, including not only a difference in the Valence and Arousal labels but also in the definition of the time segments (for example, the region between 20 and 60 seconds in Fig. 2).
- Because of the above, dialogues in which participants showed more intense affective states resulted in annotations with higher consensus, while *flat* or more subtle affective behaviors tended to produce lower agreement. Interestingly enough, the latter were subjectively perceived as more difficult to label by the annotators.
- In general, the agreement of the annotations was higher for Valence than for Arousal.

4.3 Objective Metric to Assess Consensus

While the analysis presented in the previous section helps to illustrate the generated annotations in qualitative terms, we have also conducted a number of preliminary tests to quantitatively evaluate inter-annotator agreement.

We used Cronbach's alpha coefficient [19] to compute the consensus level of each video. This measure has been used in other affective databases [16, 18] and provides a single value in the range between 0 and 1 for a set of K annotations. Values above 0.7 are considered acceptable whereas values below 0.6 can be considered as practically unacceptable.

Cronbach's alpha has been computed for all possible combinations of K annotators, varying K from 2 to 8. Fig. 3 shows the average, minimum and maximum alpha coefficients obtained for each set of possible annotator combinations given a fixed K. Note that for $K = 8$, only one combination is possible, therefore the three values are equivalent. The curves shown in Fig. 3 suggest that:

- The level of agreement strongly depends on the specific video that is annotated. However, there seems to be a clear tendency of improved agreement as more experts produce annotations for the same video.
- Looking at minimum values, between 4 and 6 annotators per video should be used to minimize the risk of low agreement. On the other hand, as this varies from video to video, a reasonable strategy could be assigning additional annotators to videos with insufficient agreement scores.
- Arousal was consistently found harder to annotate than Valence. This has been previously reported in the literature and it has been attributed to the fact that, in general, the concept of Valence is more intuitive and easier to understand/perceive than Arousal.

5. CONCLUSIONS AND FUTURE WORK

In this paper we report ongoing work on the creation of a multimodal annotated database in the context of ECAs in the health-care domain. We define an annotation schema with the aim to produce a set of annotations with sufficient consensus to be trusted as reliable ground-truth. To this end, we propose a set of guidelines that can be conceptually summarized around the following key points: (i) holistic assessment of affective states, considering a unique multimodal annotation for each time segment rather than individual modality-dependant labels; (ii) focus exclusively on nonverbal communication; (iii) use of discrete labels within the Valence-Arousal space, with a scale especially designed to emphasize subtle variations from the *neutral* affective state.

Annotations produced so far showed a mixture of time segments with strong and weak agreement, that seem to correlate well with high and low (absolute) values of Valence and

Arousal. Overall, computation of Cronbach's alpha coefficient showed acceptable agreement levels, as long as videos are labeled by a sufficient number of annotators. Indeed, a positive correlation was observed between the level of agreement and the number of annotators. Taking into account the heterogenous background of the annotators, this is an interesting point and it supports the suitability of the annotation schema that has been adopted.

The work described here is currently being extended to incoming corpus recordings. Future work aims to complete a reasonably large annotated database, which is expected to include various languages and cultural backgrounds.

Acknowledgments

This work is partly supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and is part of the Kristina project funded by the European UnionŠs Horizon 2020 research and innovation programme under grant agreement No 645012.

6. ADDITIONAL AUTHORS

The following authors were actively taking part in the annotation experiments described in this paper: Alex de Mulder (Almende B.V., The Netherlands), Gerard Llorach (Department of Information and Communication Technologies, Pompeu Fabra University, Spain) and Thodoris Tsompanidis (Information Technologies Institute, Centre for Research & Technology Hellas, Greece).

7. REFERENCES

[1] R. Cowie and R. R. Cornelius. Describing the emotional states that are expressed in speech. *Speech communication*, 40(1):5–32, 2003.

[2] R. Cowie, et al. The essential role of human databases for learning in and validation of affectively competent agents. *A Blueprint for Affective Computing: a Sourcebook and Manual*, pages 151–165, 2010.

[3] R. Cowie, G. McKeown, and E. Douglas-Cowie. Tracing emotion: an overview. *International Journal of Synthetic Emotions*, 3(1):1–17, 2012.

[4] N. Dael, M. Mortillaro, and K. R. Scherer. The body action and posture coding system: Development and reliability. *J. of Nonverbal Behavior*, 36:97–121, 2012.

[5] A. J. Gerber, et al. An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces. *Neuropsychologia*, 46(8):2129–2139, 2008.

[6] J. M. Girard. CARMA: Software for continuous affect rating and media annotation. *Journal of Open Research Software*, 2(1):e5, 2014.

[7] M. Grimm, K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In *IEEE Int. Conf. Multimedia and Expo*, pages 865–868, 2008.

[8] H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Int. Conf. Pattern Recognition*, vol 1, pages 1148–1153, 2006.

[9] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.

[10] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 13:120–136, 2013.

[11] P. N. Juslin and K. R. Scherer. Vocal expression of affect. *The New Handbook of Methods in Nonverbal Behavior Research*, 2005.

[12] D. Keltner and P. Ekman. Facial expression of emotion. *Handbook of Emotions*, pages 236–249, 2000.

[13] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Eurospeech*, pages 1367–1370, 2001.

[14] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.

[15] P. Lucey, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 94–101, 2010.

[16] G. McKeown, et al. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[17] M. A. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *Int. Conf. Pattern Recognition*, pages 3695–3699, 2010.

[18] F. Ringeval, et al. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *IEEE Int. Conf. Automatic Face and Gesture Recognition*, pages 1–8, 2013.

[19] J. R. A. Santos. Cronbach alpha: A tool for assessing the reliability of scales. *Journal of extension*, 37(2):1–5, 1999.

[20] J. Wagner, et al. The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. In *ACM Int. Conf. Multimedia*, 2013.

[21] P. Wittenburg, et al. ELAN: a professional framework for multimodality research. In *Int. Conf. on Lenguage Resources and Evaluation*, pages 1556–1559, 2006.

[22] M. Wöllmer, et al. A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. *Neurocomputing*, 73:366–380, 2009.

[23] M. Wöllmer, et al. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing*, 4:867–881, 2010.

[24] Z. Zeng, et al. Audio-visual based emotion recognition - a new approach. In *IEEE Conf. Computer Vision and Pattern Recognition*, vol 2, pages 1020–1025, 2004.

[25] Z. Zeng, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:39–58, 2009.

[26] Z. Zeng, et al. Audio-visual affective expression recognition through multistream fused hmm. *IEEE Transactions on Multimedia*, 10:570–577, 2008.