Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 1049–1063

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Speech-based recognition of self-reported and observed emotion in a dimensional space

Khiet P. Truong [a,*], David A. van Leeuwen [b], Franciska M.G. de Jong [a]

[a] University of Twente, Human Media Interaction, P.O. Box 217, 7500 AE Enschede, The Netherlands
[b] Radboud University Nijmegen, Centre for Language and Speech Technology, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 3 April 2010; received in revised form 23 April 2012; accepted 24 April 2012
Available online 3 May 2012

**Abstract**

The differences between self-reported and observed emotion have only marginally been investigated in the context of speech-based automatic emotion recognition. We address this issue by comparing self-reported emotion ratings to observed emotion ratings and look at how differences between these two types of ratings affect the development and performance of automatic emotion recognizers developed with these ratings. A dimensional approach to emotion modeling is adopted: the ratings are based on continuous arousal and valence scales. We describe the TNO-Gaming Corpus that contains spontaneous vocal and facial expressions elicited via a multiplayer videogame and that includes emotion annotations obtained via self-report and observation by outside observers. Comparisons show that there are discrepancies between self-reported and observed emotion ratings which are also reflected in the performance of the emotion recognizers developed. Using Support Vector Regression in combination with acoustic and textual features, recognizers of arousal and valence are developed that can predict points in a 2-dimensional arousal-valence space. The results of these recognizers show that the self-reported emotion is much harder to recognize than the observed emotion, and that averaging ratings from multiple observers improves performance.


*Keywords:* Affective computing; Automatic emotion recognition; Emotional speech; Emotion database; Audiovisual database; Emotion perception; Emotion annotation; Emotion elicitation; Videogames; Support Vector Regression

## 1. Introduction

In recent years, there has been a growing amount of research focusing on the automatic recognition of emotion in several communication modalities, e.g., face, body posture, gesture, speech etc. The ability to automatically recognize emotion in speech opens up many research opportunities and innovative applications. For conversational agents, the assessment of the emotional state in the speech of its human interlocutor is one of the key elements in achieving a humanlike conversation – vocal communication is a very natural way for humans to communicate. Further, with the increasing amount of archived speech and audio data available, the need for useful search queries grows. Searching through speech data by the emotion of the speaker is seen as a novel useful feature. Call centers have also shown interest in automatic emotion recognition systems which can be used for automated quality monitoring of incoming calls of customers. As illustrated with these examples, talking is one of the most natural interaction channels for people and as such, many innovative voice-based applications can be targeted. Hence, we focus here on the vocal modality.

We can identify several major challenges in the affect recognition research community. How to obtain reliable

* Corresponding author.
*E-mail addresses:* k.p.truong@utwente.nl (K.P. Truong), d.vanleeuwen@let.ru.nl (D.A. van Leeuwen), f.m.g.dejong@utwente.nl (F.M.G. de Jong).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.04.006

emotion annotations of spontaneous emotional behavior is one of these major challenges. The automatic recognition of *non-prototypical* emotions is another one. This paper addresses these two issues by exploring self-reported emotion ratings, i.e., annotation of emotions by the person who has undergone the emotion him/herself, and by adopting continuous arousal and valence dimension to model non-prototypical emotions. For these purposes, spontaneous audiovisual data was collected through a gaming scenario. Using this data, recognizers were trained with acoustic and lexical features in order to recognize scalar values of arousal and valence.

There is a vast amount of literature available on the modeling of emotional speech (e.g., Williams and Stevens, 1972; Banse and Scherer, 1996) in the speech community. The studies described in this literature usually assume emotion models and descriptions adopted from psychology research. Stemming from Darwin and made popular by researchers such as Ekman and colleagues, the most basic and classical approach to emotion modeling is the use of discrete emotion categories. Ekman (1972) and Ekman and Friesen (1975) applied this approach to the description of facial expressions and proposed six basic emotions ('the big six') that can be assumed universal: happiness, sadness, surprise, fear, anger, and disgust. As an alternative to this theory based on discrete emotions, a dimensional theory of emotion is available which was first described and applied by Wundt (1874/1905) and Schlosberg (1954). In the dimensional approach, emotions are described as points in a multidimensional space. The two main dimensions in this space are the valence dimension (pleasantness ranging from positive to negative) and the arousal dimension (activity ranging from active to passive). Sometimes, a third dimension is used which usually represents the dominance or power dimension. As a third alternative to discrete and dimensional theories of emotion, several researchers (Scherer, 2010) have developed a cognitive approach to emotion. For example, Scherer and colleagues have proposed an appraisal model called the Compoment Process Model. The main assumption here is that an emotion is a reaction (e.g., physiological, feeling) to certain antecedent situations and events that are being evaluated at the cognitive level by the human. In other words, the appraisal (i.e., the evaluation process) of a situation determines how the human is going to react/response to this situation. Componential models emphasize the link between the elicitation of emotion and the response, and as such, these models account for the variability of different emotional responses to the same event that may occur.

One of the attractions of the dimensional approach is that it allows for more flexibility and generality since it provides a way of describing emotions without the use of linguistic descriptors that can be language or culture dependent. Finding category labels to capture every shade of emotion, that frequently occur in everyday daily life, has appeared to be difficult (e.g., Cowie and Cornelius, 2003; Douglas-Cowie et al., 2005). Traditionally, speech-based emotion recognition studies have concentrated on the recognition of discrete emotion categories containing stereotypical emotions. Some of the relevant work include e.g., Batliner et al. (2000), Dellaert et al. (1996), Polzin and Waibel (1998), Petrushin (1999), Devillers et al. (2003), Kwon et al. (2003), Ang et al. (2002), Lee et al. (2002), Liscombe et al. (2003), Nwe et al. (2003), Schuller et al. (2003) and Ververidis and Kotropoulos (2005). Typical emotion categories in these studies are happy, anger, and neutral. Good overviews of these emotion recognition studies can be found in (Cowie et al., 2001; Ververidis and Kotropoulos, 2006). More recently, an increasing number of studies that adopt a dimensional approach to emotion recognition can be observed. Representing (everyday) emotion on a continuous scale could better capture different shades of emotion. Hence, describing emotion by their coordinates in a multidimensional space offers an attractive alternative, especially for computational modeling of emotion. Usually, two dimensions are sufficient to cover the emotions under investigation, where one dimension represents valence and the other dimension represents arousal. Russell (1980) and Schlosberg (1954) have shown that a third dimension, i.e., dominance or power, accounts for only a small proportion of the variance. Hence, the majority of studies have only targeted arousal and valence modeling of emotion. However, one should keep in mind that some information is always lost when mapping to a 2-dimensional emotion space. We give an overview of studies adopting a dimensional approach to emotion recognition in Section 2.

In a slower tempo, progress is also being made in designing procedures for annotation of spontaneous emotion corpora which lead to higher levels of agreement among human labelers and which better reflect the spontaneous nature of the emotion. Emotion annotation is a complex and hard process performed by humans of which the results can have significant impact on the system's performance. Emotion recognition systems need somewhat consistent emotion-labeled data for training and testing. However, it is well-known that the perception of emotion is to a certain extent subjective and person-dependent. In order to deal with this person-dependency and to reach a certain consensus on a specific emotion label, it is common to use several annotators and apply majority voting, i.e., the emotion class with the most 'votes' from the annotators wins (e.g. Batliner et al., 2006). For continuous dimensional annotations, the continuous ratings are usually averaged among the human labelers, see Mower et al. (2009), Truong et al. (2009) and Grimm et al. (2007a). In addition, in order to deal with 'mixed' or 'blended' emotions, which are not uncommon in spontaneous expressive interaction, multi-layered annotation schemes have been proposed (see Devillers et al., 2005). Less attention has been paid in emotion recognition studies to investigate how the annotations from different types of annotators compare to each other. For instance, one could compare annotations from trained emotion labelers to annotations from unexperienced/naïve emotion labelers. Another option is to let the

recorded subject annotate his/her own emotions that were felt during an event and compare these to annotations from outsiders who did not participate in this 'event'. Hypothesizing that people are better decoders of their own emotions and in pursuing the ultimate goal of automatically analyzing a person's felt emotions, it is worthwhile to investigate how ground truth annotations derived from self-report differ from annotations made by other persons, and how these self-reported ratings influence the performances of the recognizers trained.

In this paper, we explore how 'self-annotations' and 'observer-annotations' differ from each other and how the recognizers trained on these annotations differ from each other. We adopt a dimensional description of emotion and represent emotions as points in the 2-dimensional arousal-valence space. First, we review previous studies related to our work. In Section 3, we present the data (TNO-GAMING database) that was collected through a gaming scenario and we describe how the 'self-annotations' were acquired. Section 4 describes how the 'observer-annotations' were added to the corpus and provides comparisons between the obtained 'self' and 'observer-annotations'. In Section 5, we present the recognizers and we report and discuss the results of the classification experiments. Comparative analyses are provided between the performance of the 'self-annotation'-based and 'observer-annotation'-based recognizers. Finally, we summarize and discuss the most important findings in Section 6.

## 2. Related work

A number of studies in the field of affective computing have adopted a dimensional approach to emotion recognition. Here, we restrict ourselves mainly to speech-based studies. For an overview that includes visual and physiological cues, the reader is referred to Gunes et al. (2011) and Nicolaou et al. (2011). In previous research, the 2 dimensions arousal and valence were usually discretized (see e.g., Truong and Raaijmakers, 2008) or used to divide the 2-dimensional space into 4 quadrants of Positive-Active, Positive-Passive, Negative-Active and Negative-Passive emotions. Tato et al. (2002) mapped emotion categories such as angry, happy, neutral, sadness and boredom onto three discrete levels of arousal. Yu et al. (2004) classified user engagement in social telephone conversations between friends along arousal and valence scales that were discretized into 5 levels. Kim et al. (2005), Zeng et al. (2005) and Wöllmer et al. (2009) classified emotions in the 4 emotion quadrants of the arousal-valence space. Instead of classifying emotions on discretized scales of arousal and valence, some studies have taken up the challenge to classify emotions on *continuous* scales of arousal and valence. In the context of media content analysis, Hanjalic and Xu (2005) combined video and audio features to model continuous arousal and valence curves for affective video content analysis which allows users to search for funny or thrilling video clips. Grimm and colleagues (Grimm et al., 2007a,b) used fuzzy logic and Support Vector Regression to model continuous dimensions of arousal, valence, and dominance, and applied these methods to a database of dialogues recorded from a German TV Talk show. Giannakopoulos et al. (2009) did something similar and used $k$-Nearest Neighbor rule to model continuous affect in speech from movies. With the aim to build sensitive artificial agents, Wöllmer et al. (2008) and Eyben et al. (2010) addressed the task of continuous affect modeling in human-machine interaction by introducing classification techniques that take into account previous emotion observations. They proposed to use Long Short-Term Memory Recurrent Neural Networks which are able to model long-range dependencies between successive observations. Although progress is being made in terms of performance which is illustrated in the performance scores of the studies described, one still needs to interpret these scores in relation to the way the 'ground truth' annotations are obtained. It is still rather unclear how the performance is affected by the way the 'ground truth' annotations are obtained.

Several methods have been proposed to process continuous affect annotations from multiple coders in order to reach a consensus annotation. The most common method is to average the annotations from multiple coders, either with or without a form of normalization/scaling, as is performed in e.g., Mower et al. (2009), Eyben et al. (2010), Giannakopoulos et al. (2009) and Wöllmer et al. (2008). A weighted average score was proposed by Grimm et al. (2007a) by introducing evaluator-dependent weights, taking into account the subjectivity of each coder. The concept of weighting coders is also applied by Nicolaou et al. (2010) who introduced automatic methods to derive and segment ground truth annotations from multiple continuous annotations. In addressing the question how to obtain a ground truth annotation from multiple continuous annotations, less attention has been paid to whether 'auto-coders', i.e., coders who code their own emotions, are also suitable coders and whether they are different from coders that annotate others' emotions. Under the hypothesis that people are better decoders of their own emotions, the labels derived from 'self-annotation' will better reflect the intended emotions. Auberge et al. (2006) explored this so-called 'auto-annotation' method – the subjects were asked to label what they felt rather than what they expressed – but no conclusive results were reported. The 'self-annotation' method seems to be complicated by the finding that most vocal cues are not likely to be related to speakers' internal states, at least in the case of happiness (Biersack and Kempe, 2005). Busso and Narayanan (2008) compared 'self-assessments' of emotions to assessments made by outside observers and found that there is a mismatch between the expression and perception of emotion. In Truong et al. (2008), similar findings were reported: significant differences were found between 'self-assessments' of emotion and assessments from outside observers. In the current study, we extend the work by Busso and Narayanan (2008) and Truong et al. (2008, 2009) and develop affect

recognizers trained with 'self-annotations' or 'observer-annotations' and investigate how their performances relate to each other. First, we present the audiovisual database collected for this study.

## 3. The TNO-Gaming corpus: a corpus of gamers' vocal and facial expressions

Since there is currently no emotional speech corpus available with continuous emotion annotations made by the subjects themselves and outside observers, we recorded our own corpus. We collected an audiovisual emotion corpus by inviting people to play a videogame.[1]

### 3.1. Audiovisual recordings

Seventeen males and eleven females with an average age of 22.1 years (2.8 standard deviation) participated in the gaming experiment. Participants were recruited in pairs by asking each participant to bring along a friend as team mate since we expect that people are more expressive when they are playing with friends rather than strangers (see Ravaja et al., 2006). A compensation was paid to all participants. Fifteen participants were relatively experienced gamers, while thirteen participants hardly ever or never played videogames.

Speech recordings were made with high quality close-talk microphones that were attached near the mouth to minimize the effect of crosstalk (speech from other speakers) and other background noise. Recordings of facial expressions were made with high quality webcams (Logitech Quickcam Sphere) which allows for multimodal modeling of emotion. The webcams were placed at approximate eye-level on top of the monitor such that a frontal view of the face was captured under an angle that was acceptable for reliable automatic facial recognition. Further, lighting and background conditions were controlled by adjusting the light when needed and by placing evenly colored dark curtains behind the participants to avoid clutter and noise in the background. Noldus' FaceReader (by VicarVision, see Den Uyl and Van Kuilenburg, 2005, an automatic face recognition software application) was used to test the quality of the video recordings under these environmental settings and conditions. Video stills of the hardware setup in the room that was used for the gaming sessions are shown in Fig. 1. The game content itself was also stored by capturing the frames (1 per second) of the video stream during game play.

At the beginning of the gaming experiment, the participants received a general instruction (15 min), a training session to get acquainted with the game (10 min) and instuctions and a training session for the rating task (both 20 min each). During the training sessions, the subjects could try out the game and the annotation tool; the experimenter was also present to address comments and questions. Subsequently, the first session began with a game session (20 min), followed by a questionnaire and a break (25 min), and the annotation tasks which included a ten-minute break (50 min). For the second session, this process was repeated (excluding the training sessions) in the afternoon after a long break of 40 min.

Fig. 1. The hardware setup and the room where the gaming sessions took place.

In summary, three streams of information were recorded: (1) vocal behavior via close-talk microphones, (2) facial behavior via webcams, and (3) context information via screenshots of the videogame-content.

### 3.2. Eliciting emotions

Videogames have previously succesfully been used as an emotion elicitation method, see for example, works by Johnstone et al. (2005), Wang and Marsella (2006) and Yildirim et al. (2005). In our study, the participants played a multiplayer first-person shooter videogame called *Unreal Tournament 2004*, developed by Epic Games. The game-mode 'Capture the flag' was selected: two teams play against each other and the goal is to capture each other's team flag as many times as possible.

Our goal was to evoke a broad range of different emotions. We employed several strategies to evoke these emotions and to stimulate vocal and facial expressive behavior and interaction:

1. Use a multiplayer game where each participant had to bring a friend as team mate. By bringing a friend, we expect to stimulate more interaction as was suggested in Ravaja et al. (2006).
2. Bonuses were granted to the winning team, and the team with 'best collaboration'. We wanted to motivate the subjects to be vocally active – hence the subjects were told that a bonus would be granted to the team with 'best collaboration' which was intentionally not clearly defined and only served as a method to stimulate the subjects to talk to each other.

[1] The gaming sessions and recordings took place at TNO in Soesterberg, The Netherlands.

3. The videogame was manipulated by generating surprising events in the game, for example, sudden deaths, sudden appearances of monsters, and hampering keyboard or mouse controls, were inserted in the game (at an approximate rate of one event per minute).

### 3.3. Rating procedure

After each game session, the participants watched their own videos recorded and judged their own emotions in two different ways: one based on emotion categories and the other one based on emotion dimensions. In addition, the videostream of the game itself was also provided as context information, next to the video recorded, so that all three information streams recorded were available to the participants during rating. We asked the participants to recall what they were feeling during playing. The participants rated the running video and could not pause or rewind the video. If we had allowed this, the rating task would last much longer and the raters would perhaps 'over-analyse' their own emotions. Under the assumption that 'self'-raters know the intentions of their own emotions expressed best, we hence decided not to allow the raters to pause or rewind the video. An alternative annotation method would have been to interrupt the game each time we wanted ratings over the past course of time. However, this would severely interrupt the flow of the game and could influence the interaction and feeling of involvement of the players. Prior to the rating task, the participants had received a training of 20 min duration.

#### 3.3.1. Categories: event/category-based

Participants were asked to select and de-select emotion labels whenever they *felt* the emotion that they experienced at that moment in the game: in other words, they had to click to select an emotion label to mark the beginning of the corresponding emotion and click again on the same label to de-select and to mark the ending of that emotional event. The twelve emotion labels from which the participants could choose are based on the 'Big Six', (universal basic emotions, Ekman, 1972) emotions and are supplemented with typical game-related emotions as described in Lazarro (2004). We expected that these labels, shown in Table 1, would cover most of the emotions that could occur during gaming. The selection of multiple emotion labels at the same time was allowed, which made it possible to have 'mixed' emotions. The participants also had the option to come up with their own emotion label that was not listed in the alternatives, but it appeared that the participants had not used this option.

Table 1
The emotion categories used in the category-based rating task.

| | |
|---|---|
| Happiness | Fear |
| Boredom | Anger |
| Amusement | Relief |
| Surprise | Frustration |
| Malicious Delight | Wonderment |
| Excitement | Disgust |

#### 3.3.2. Continuous emotion dimensions: continuity/dimension-based

The participants were also asked to rate their emotions *felt* on two emotion scales namely the arousal scale (active vs passive) and the valence scale (negative vs positive). We believe that these 2 dimensions will capture the majority of emotions occurring in a gaming context (see also Russell, 1980; Schlosberg, 1954). As opposed to the category-based approach where the participants had to mark the beginning and ending of an emotional event, the participants now had to give ratings on emotion scales running from 0 to 100 (with 50 being neutral) each 10 s *separately* (thus not *simultaneously* as is done with some annotation tools such as Feeltrace (Cowie et al., 2000). Each 10 s, an arrow appeared on the screen to signal the participants to give an arousal and valence rating, see Fig. 2.

### 3.4. Processing the ratings

The emotion data collected were not (immediately) ready to use for analysis since we are only interested in emotional *speech* segments. Subsequently, since response times might have played a role here and the category and dimension rating procedures were semi-continuous in time, which resulted in asynchronicity between the speech segments and the ratings, we needed a procedure to link the emotion ratings to speech segments. In short, this procedure can be divided into two parts: (1) obtain speech segments from the data, and (2) link the emotion ratings to these speech segments. The first part was achieved by running an energy-based silence detection algorithm in Praat (Boersma and Weenink, 2009) which resulted in speech segments that were used in subsequent analyses and classification experiments; hence, the silence detection algorithm determined the units of analysis. These speech segments were manually transcribed on word level by the first

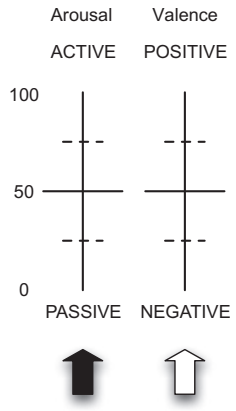


Fig. 2. The emotion scales offered to the participants in the dimension-based rating task.

author. Secondly, the speech segments needed to be linked to emotion ratings. In the category emotion annotation procedure, participants had to mark the beginning and ending of an emotional event. We assumed that the marker of the beginning is more reliable than the ending marker. One of the reasons is that we noticed that some of the emotional events were extremely long; we suspect that participants might have forgotten to de-select the emotion label to mark the ending (in future research this may be solved by making the selected label blink until it is de-selected again). Also, we allow for a delay between the real occurrence of an emotional event and the moment that an emotion label was selected. In the dimensional emotion annotations procedure, people had to give an arousal and valence value when an arrow appeared which happened each 10 s. For both annotation methods, similar 'linking' procedures were applied, taking into account the fact that people are reacting with a certain amount of delay. Fig. 3 shows how we associated speech segments with emotion categories or arousal-valence ratings: check for a maximum number of $N$ segments (we chose $N = 5$) prior to the moment that an emotion label $E$ was selected (or when an arrow appeared) (1) whether a segment $S_i$ ends within a margin of $T$ (we chose $T = 3$ s) before the label was selected (or when an arrow appeared), and (2) whether the segment is labeled as non-silence by the silence detection algorithm.

### 3.5. Distributions of the emotion ratings obtained by 'self'-annotation

The procedure as described above resulted in a set of speech segments that are labeled with an emotion category label and/or an arousal and valence rating. In Fig. 4, we can observe the frequency of emotion category labels as used by the gamers themselves. It seems that Frustration, Excitement, Happiness, Amusement and Surprise are frequently occurring emotions, while Boredom, Fear and Disgust are hardly experienced by the gamers. The black areas represent the number of segments that could be associated with more than one emotion label. Frequent pairs of emotion labels include Amusement & Happiness, Happiness & Excitement, Frustration & Surprise, Happiness & Relief, and Amusement & Excitement. These are not surprising combinations of emotions and they make sense in a gaming context. The question remains how to deal with this blendedness of emotions in a classification context (see also Devillers et al., 2005) which is outside the scope of this paper.

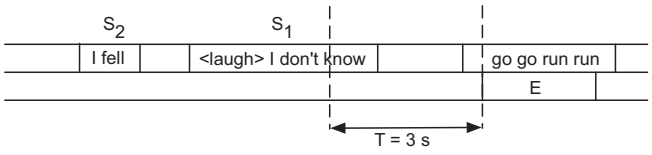


Fig. 3. Procedure for 'linking' speech segments to emotional events or arousal-valence ratings. $S_1$ can be linked to emotion label $E$ (or an arousal-valence rating) because the end time of $S_1$ falls within $T$.

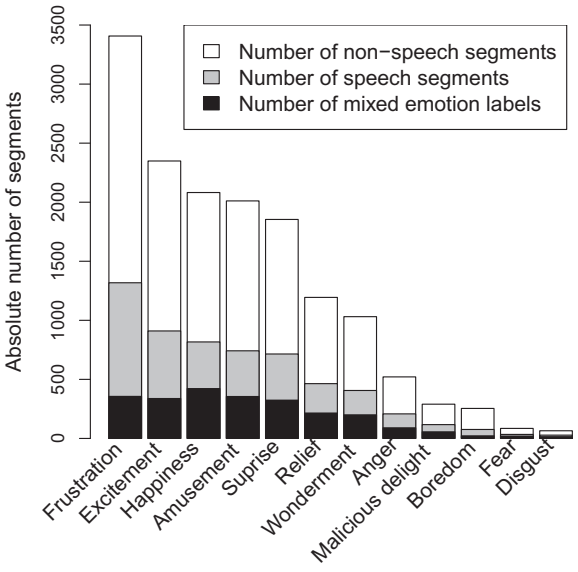


Fig. 4. Numbers of speech (and non-speech) segments that could be associated with a category emotional event or with multiple category emotion events.

The results of the dimension-based rating task are presented in Fig. 5. The figure shows that the majority of emotions felt during playing (while speaking) was situated around Neutral. The Positive-Active quadrant is relatively well-filled with speech segments, followed by the Negative-Active quadrant in the arousal-valence space. There are apparent blank spots in the Positive-Passive and Negative-Passive quadrants. It appears that the participants did not often report feeling very Positive or Negative in a Passive way which resulted in a 'boomerang'-shape-like figure. Whether this shape is the direct result of the gaming context in which the emotions were elicited remains to be

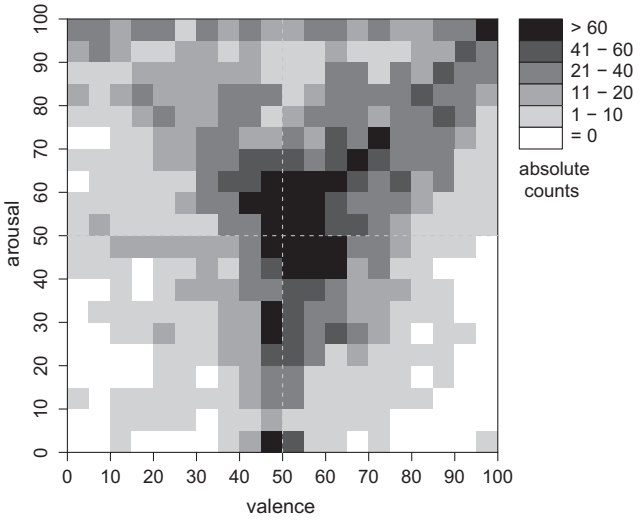


Fig. 5. 2D Histogram plot: the gamers' arousal and valence ratings (i.e., SELF-ratings) that could be associated with speech segments, $N = 7473$.

Table 2
Amount of emotionally labeled speech data according to the gamers' emotion labeling.

| | Duration | | | $N_{segments}$ | $N_{words}$ |
|---|---|---|---|---|---|
| | Total (min) | Mean (s) | Stdev (s) | | |
| Category-based | 78.6 | 1.67 | 1.26 | 2830 | 1322 |
| Dimension-based | 186.2 | 1.50 | 1.12 | 7473 | 1963 |

seen, since this quadratic relationship between arousal and valence has previously been observed in Lang (1995) and Hanjalic and Xu (2005) for similar ratings tasks but in very different contexts, i.e., looking at emotion-evoking pictures and viewing movies or soccer television broadcasts respectively. Finally, our participants mentioned that they sometimes had trouble interpreting the arousal scale: they had some trouble rating something as Passive or Neutral.

In summary, this gaming experiment resulted in a substantial amount of labeled speech data (see Table 2) that can be used for the training and development of automatic speech-based emotion recognizers (approximately 28% and 67% of all recorded audiovisual data for the category- and dimension-based ratings respectively). Due to the spontaneous character of this gaming experiment, we have obtained a corpus that does not always contain extreme emotions, and the corpus is not very well-balanced in the sense that not all areas in the arousal-valence space are uniformly covered with speech segments. One important novelty of the data collected in this gaming experiment, is the fact that all data is rated by the gamers themselves. We will refer to these annotations as SELF-ratings. The participants (i.e., the gamers) who have labeled their own *felt* emotions after playing the videogame are referred to as the SELF-raters. In subsequent experiments, we investigated the relation between these SELF-ratings and ratings given by other observers. We used the data to train and test speech-based affect recognizers. The database collected and described here will be referred to as the TNO-GAMING corpus. In Table 3 some examples of emotional expressions are shown that were captured while the subjects were playing the videogame.

## 4. Extending the corpus with perceived affect ratings from external observers

Our goals are to investigate how the ratings given by the gamers themselves differ from ratings given by external observers, and to develop recognizers that continuously can predict arousal and valence ratings. Hence, we focus only on the dimension-based ratings and discard the category-based labels. To these ends, we let a part of the corpus, that part that is rated on dimensions, be (re-)rated by external naïve observers who had not participated in the gaming sessions. Because the number of segments of the whole corpus is relatively large, we decided to make a selection of 2400 segments, out of the original set of 7473 segments (i.e., the speech segments that have arousal and valence ratings), that was offered to a group of naïve observers. The random selection procedure of these 2400 movie clips that were offered to the observers was partly restricted by our criterion to roughly maintain the same proportions of the segments in the arousal-valence space of the original set, and partly driven by the need for a larger number of segments in the lower arousal area to adjust for this strongly imbalanced distribution on the arousal scale. The distribution of the segments selected for re-rating in the arousal-valence space is displayed in Fig. 6. The total length of the whole set of 2400 segments is approximately 76 min. The mean duration and standard deviation of a segment is 1.9 and 1.2 s respectively. The scales of the arousal and valence dimensions are linearly re-scaled from $[0, 100]$ to a range of $[-1, 1]$ which allows for comparison with previous studies (e.g., Grimm et al., 2007b), the linear re-scaling will not affect the analyses or results).

### 4.1. Rating procedure

The set of 2400 emotional speech segments were audiovisually presented to six naïve raters who had not participated in the gaming sessions. The six raters (1 female, 5 male) are on average 25.4 years of age. Similar to the SELF-rating procedure, these raters were asked to rate each audiovisual segment on the arousal and valence scale that runs from 0 to 100, with 50 being Neutral (afterwards we linearly re-scaled to $[-1, 1]$). Although we tried hard to maintain as much as possible the exact same rating procedure that was used for the SELF-raters, practically, this was not possible. The differences with the previous SELF-rating procedure are that (1) the audiovisual segments are already segmented, (2) the raters now can re-play the segment if they like (as a compensation for the fact that these raters do not know the gamers), and (3) no context information was given (one of the reasons for not showing context information was that in our previous perception experiments with the same data, the addition of context information gave mixed results – adding the video game stream did not always increase the agreement among raters, see Truong et al., 2008). We will refer to the individual emotion ratings of the six raters as OTHER.3 ('3' because each segment is rated by 3 different observers, this will be explained below).

Each observer/rater (TH, PI, CO, RA, FR, and AT) rated different parts (A, B, C, and D) of the dataset that overlapped with parts that were rated by other observers. This division ensured that each segment was rated by three different raters. The dataset was divided into four parts, each part consisting of 624 segments. Each observer was assigned to two parts of the database, and thus rated in total $2 \times 624$ segments, see Fig. 7. Of the 624 segments in each part, 24 segments occurred twice and were used to assess the rating consistency of the observer (intra-rater agreement) him/herself. For each observer, it took approximately 4 to 5 h to complete the ratings of all 1248 segments, including breaks. That means that the rating

Table 3
Examples of gamers' word transcriptions and emotion ratings (with an English translation), Val=Valence, Aro=Arousal, NA=Negative Active, NP=Negative Passive, PA=Positive Active, PP=Positive Passive

| | Val | Aro | Transcription |
|---|---|---|---|
| NA | 0 | 99 | 'urgh what are those type of monsters?' |
| | 1 | 100 | 'no yes * *! no! run! no! no!' |
| | 3 | 80 | 'soo irritating' |
| NP | 10 | 31 | 'yes I I try to go there' |
| | 13 | 5 | 'I don't what those * weapons' |
| | 13 | 33 | 'na I don't see anything' |
| PA | 97 | 99 | 'oh that's you sorry [laughter]' |
| | 81 | 98 | 'run run run yes good job' |
| | 97 | 83 | 'score a point score a point [laughter]' |
| PP | 71 | 8 | 'OK now we are going to score a point' |
| | 77 | 18 | 'we are going for the twenty right I have the blue flag' |
| | 74 | 29 | 'I have them I kill them just walk' |

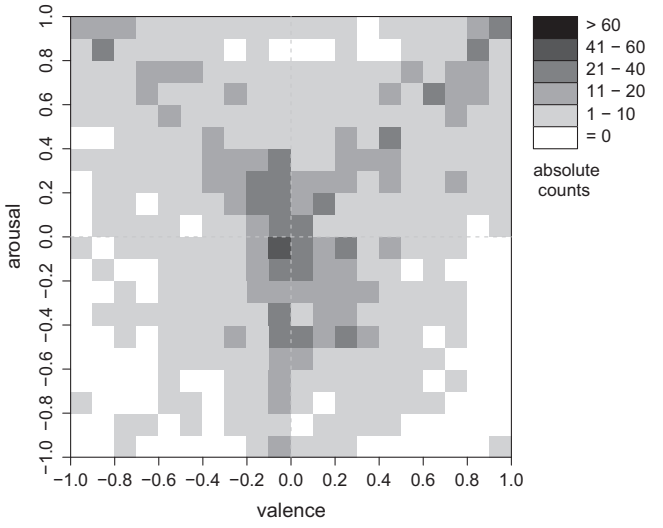


Fig. 6. 2D Histogram plot: the gamers' arousal and valence ratings (i.e., SELF-ratings) that could be associated with speech segments, $N = 2400$, selected as stimuli for the external observers.

procedure was carried out at a rate of approximately 6 times real-time.

### 4.2. Distributions of the affect ratings obtained from the external observers

The OTHER.3 ratings represent the ratings of 3 distinct observers. In order to derive a consensus annotation from these multiple ratings, the ratings of the 3 observers were averaged for each segment (which is a common procedure often applied). We will refer to these ratings as OTHER.AVG ('AVG' stands for 'averaged'). This means that we have 3 types of ratings available that will be used in training and testing our recognizers: SELF, OTHER.3 and OTHER.AVG-ratings.

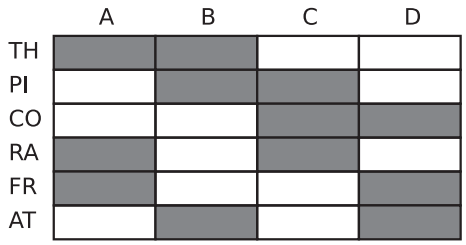


Fig. 7. Division of dataset into several overlapping parts, each observer rated two cells (each cell contains 624 segments) such that each segment is rated by 3 different observers.

By comparing the 2-dimensional histograms based on SELF-ratings and OTHER.AVG-ratings, shown in Figs. 6 and 9 respectively, we can observe that the majority of the segments were, more or less, judged as Neutral by the observers (on average) which differs substantially from the SELF-ratings. The SELF-raters appear to have selected more extreme values for their own felt emotions than the observers have who seemingly did not perceive these emotions as such and who mostly selected values in the vicinity of Neutrality. In addition, the pull towards Neutrality is also partly caused by averaging the ratings, compare Fig. 8 to Fig. 9.

### 4.3. Analysis of 'self' vs 'observer' ratings

How do the SELF-ratings, OTHER.3-ratings, and OTHER.AVG-ratings compare and relate to each other? The amount of agreement between raters was analyzed by assessing Pearson's correlation coefficient and the absolute differences between different ratings (see Eq. (3) and (4)). Since there is no standard measure, we report these several measures to allow for comparison.

First, we assessed the rating consistency of the observers, i.e. the intra-rater consistency. We had included $2 \times 24$ segments that were rated twice by the raters. The intra-rater agreement figures of each individual rater are presented in Table 4. We found that raters are more consistent in rating valence than arousal: $r$ ranges from 0.73 to 0.91, and from 0.46 to 0.64 for valence and arousal, respectively. Given these relatively good intra-rater agreement figures, we considered the raters reliable and hence, all the raters' ratings were considered.

When we look at the inter-rater agreement among the human raters, we can see that this amounts to correlations of 0.64 and 0.32 for valence and arousal respectively, see Table 5. The correlations (and $e_{avg}$) were calculated between each possible pair of raters and subsequently averaged. Although the correlation coefficients found here are slightly lower than the ones reported in Busso and Narayanan (2008), who reported correlation coefficients of 0.79 and 0.59 for valence and arousal respectively, we see that Busso and Narayanan (2008) also report higher agreement for valence than for arousal. A possible explanation for their correlation coefficients being higher than ours could be that Busso and Narayanan (2008) used acted emotion data. Grimm et al. (2007a) report different results for the German talk show data: they found higher average correlation coefficients for arousal, 0.72, than for valence, 0.48. It should be noted that Grimm et al. (2007a) presented audio

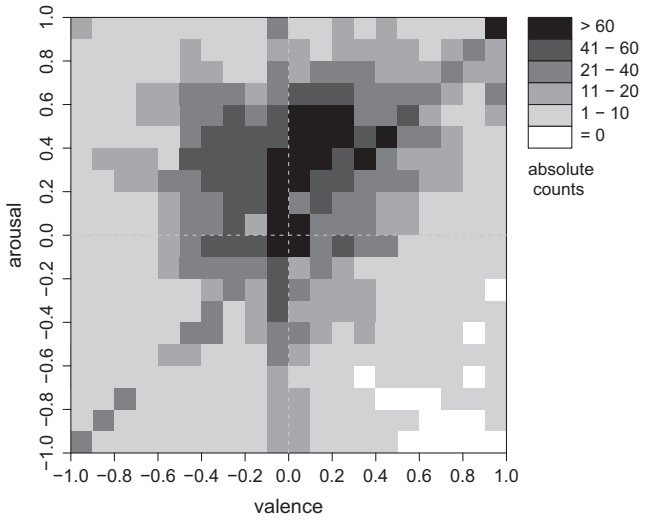


Fig. 8. 2D Histogram: the distribution of the 2400 selected speech segments in the arousal-valence space, rated by 6 different observers (i.e., the OTHER.3-ratings, $N_{\text{ratings}} = 3 \times 2400 = 7200$).

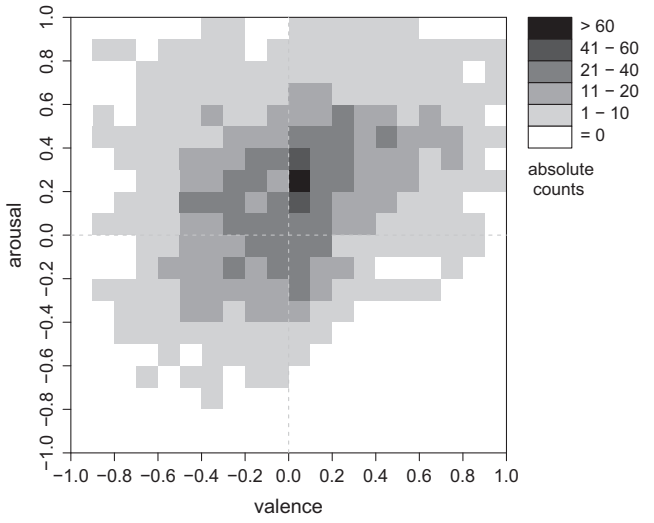


Fig. 9. 2D Histogram: the distribution of the 2400 selected speech segments in the arousal-valence space, based on the averaged ratings of the 6 observers (i.e., the OTHER.AVG-ratings, $N_{\text{ratings}} = 2400$).

only to the human listeners, whereas in Busso and Narayanan (2008) and our case, audiovisual data was presented.

Table 4
Intra-rater agreement, based on 48 doubly-rated segments.

| Rater | $e_{\text{avg}}$ | | Pearson's $r$ | |
|---|---|---|---|---|
| | Valence | Arousal | Valence | Arousal |
| TH | 0.12 | 0.19 | 0.91 | 0.55 |
| PI | 0.17 | 0.34 | 0.84 | 0.46 |
| CO | 0.08 | 0.16 | 0.87 | 0.64 |
| RA | 0.17 | 0.33 | 0.84 | 0.56 |
| FR | 0.27 | 0.43 | 0.73 | 0.54 |
| AT | 0.16 | 0.21 | 0.91 | 0.64 |
| Mean | 0.16 | 0.28 | 0.85 | 0.57 |

How do the gamers' own ratings compare to the observers' ratings? This question was approached in two ways. Firstly, we looked at what the effects are on the averaged level of agreement when different types of ratings are added to the group of OTHER.3-ratings, see Table 5. One would expect that when a rater is added to a group of raters who in general is disagreeing with this group of raters, the level of agreement among these raters will decrease (and vice versa). When the SELF-rater is added, $e_{\text{avg}}$ increases slightly for arousal and more substantially for valence. The numbers suggest that the SELF-rater is not agreeing with the OTHER-raters: the addition of the SELF-rater decreases agreement among the raters. As expected, adding OTHER.AVG or a fictional rater who perfectly agrees with one of the OTHER.3-raters increases agreement. Adding a fictional rater who completely disagrees with one of the OTHER.3-raters decreases agreement. A 'perfect' disagree-er disagrees as much as possible with one of the three raters and chooses $-1$ if a rater's rating is $>0$, and chooses $+1$ if a rater's rating is $<0$. These fictional raters were added to illustrate how the agreement numbers fluctuate under the influence of added agreeing or disagreeing raters.

Secondly, we calculated agreement directly between the SELF-ratings, and the OTHER.3 and OTHER.AVG-ratings, see Table 6. Based on these results and Table 5, we can conclude that there is relatively low agreement between the SELF-ratings, and the OTHER.3 and OTHER.AVG-ratings. Furthermore, there is higher agreement for valence than for arousal.

We have established that there is a discrepancy between the different types of ratings: the observers who have rated *perceived* affect show relatively low agreement with the gamers who have rated their own *felt* affect. What do these observations mean for the development of speech-based affect recognizers that will use these ratings for training and testing? And what does this mean for the concept of 'ground truth'? These aspects are discussed in the following sections.

## 5. Automatic recognition experiment: recognizing felt and perceived affect

In order to find out how automatic recognizers deal with these seemingly subjective affect ratings, we trained and tested 3 types of recognizers in parallel to recognize affect in speech: one is based on the SELF-ratings, one is based on the individual OTHER.3-ratings and the final one is based on the OTHER.AVG-ratings. Using regression techniques, the task of the recognizers is to estimate scalar values of arousal and valence. We used acoustic and textual features. Note that our main interest and goal were not to optimize and tweak classification algorithms to achieve the highest performance possible (for that reason, thorough comparisons between other regression techniques and features were not included in this study), but rather to see how performances change under influence of felt and observed annotations.

Table 5
Addition of different type of ratings to the OTHER.3 ratings.

| | OTHER.3 | | + SELF | | + OTHER.AVG | | + agree-er | | + disagree-er | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $e_{avg}$ | $r$ | $e_{avg}$ | $r$ | $e_{avg}$ | $r$ | $e_{avg}$ | $r$ | $e_{avg}$ | $r$ |
| Aro | 0.40 | 0.32 | 0.43 | 0.26 | 0.32 | 0.50 | 0.34 | 0.39 | 0.82 | −0.04 |
| Val | 0.28 | 0.64 | 0.33 | 0.47 | 0.22 | 0.73 | 0.23 | 0.66 | 0.76 | 0.00 |

Table 6
Inter-rater agreement (averaged) between SELF and OTHER.3 and between SELF and OTHER.AVG.

| | OTHER.3 | | | | OTHER.AVG | | | |
|---|---|---|---|---|---|---|---|---|
| | Aro | | Val | | Aro | | Val | |
| | $e_{avg}$ | $r$ | $e_{avg}$ | $r$ | $e_{avg}$ | $r$ | $e_{avg}$ | $r$ |
| SELF | 0.46 | 0.24 | 0.38 | 0.35 | 0.32 | 0.33 | 0.23 | 0.41 |

Table 7
Material used in automatic affect recognition experiments.

| | |
|---|---|
| Number of speech segments | 2400 |
| Number of unique speakers | 11 females/17 males |
| Size of vocabulary | 1141 |
| Total length | appr. 76 min |
| Mean length segment | 1.9 s ($\sigma = 1.2$ s) |

### 5.1. Material

As reference annotations, the SELF (Fig. 6), OTHER.3 (Fig. 8) and OTHER.AVG-ratings (Fig. 9) as described in Section 4 were used. The differences between the SELF-ratings and OTHER.AVG-ratings can be seen when comparing Fig. 6 to Fig. 9; we can observe that the gamers have rated their own emotions in a more extreme way than the observers have done. The variances for the arousal SELF- and OTHER.AVG-ratings are 0.25 and 0.10, respectively, for the valence SELF- and OTHER.AVG-ratings these are 0.20 and 0.12, respectively. The total length of the material comprises approximately 76 min with a mean length of 1.9 s for a segment (see Table 7).

### 5.2. Features and method

#### 5.2.1. Support Vector Regression

Since our goal is to predict real-valued output rather than discrete classes, we used a learning algorithm based on regression. Support Vector Regression (SVR) was employed to train regression models that can predict arousal and valence scalar values on a continuous scale. Similar to SVMs, SVR is a kernel-based method and allows the use of the kernel trick to transform the original feature space to a higher-dimensional feature space through a (non-linear) kernel function. For a more in-depth description of Support Vector Machine and Support Vector Regression techniques, readers are referred to Smola and Scholkopf (2004) and Vapnik (2002). We used $\epsilon$-SVR available in *libsvm* (Chang and Lin, 2001) to train our models. In SVR, a margin $\epsilon$ is introduced and SVR tries to construct a discriminative hyperplane that has at most $\epsilon$ deviation from the original training samples. In our emotion prediction experiments, the RBF kernel function was used. The parameters $c$ (cost), $\epsilon$ (the $\epsilon$ of the loss function), and $\gamma$ were tuned on a development set (see Table 9) via a simple grid search procedure that evaluates all possible combinations of $c$ (with exponentially growing values between $2^{-4}$ and $2^{4}$), $\epsilon$ (with exponentially growing values between $10^{-3}$ and $10^{0}$), and $\gamma$ (with exponentially growing values between $2^{-10}$ and $2^{2}$).

#### 5.2.2. Acoustic features

The acoustic feature extraction was performed with Praat (Boersma and Weenink, 2009). Prior to feature extraction, a voiced-unvoiced detection algorithm (available in Praat) was applied to find the voiced units. To avoid the use of an automatic speech recognizer (ASR), that can provide word alignments, the features were extracted over each *voiced unit* of a segment. We made a selection of features based on previous studies (e.g., Batliner et al., 2006; Banse and Scherer, 1996), and grouped these into features related to *pitch* information, *energy/intensity* information, and information about the *distribution of energy in the spectrum*. The spectral features MFCCs (Mel Frequency Cepstrum Coefficients) as commonly used in ASR were also included. And finally, global information calculated over the whole segment (instead of per voiced unit) about the speech rate and the intensity and pitch contour was included. An overview of the features used is given in Table 8.

Pitch and energy/intensity information are known to be useful in emotion recognition and are thus very commonly used. MFCCs are powerful speech features and are commonly used in automatic speech recognition and speaker and language recognition technologies. The distribution of energy in the spectrum can give information about the vocal effort: in general, when speakers increase their vocal effort, the energy in the higher frequency regions of the long-term spectrum increase which results in a less steep spectral slope. The Hammarberg index is a measure that measures differences of the energy in different frequency regions of the long-term spectrum: it is defined here as the maximum energy measured in the frequency region 0–2000 Hz minus the maximum energy measured between 2000 and 4000 Hz. The features 'speech rate1' and 'speech rate2' are calculated per segment and are defined as the number of voiced units divided by the segment duration without and with unvoiced regions respectively. The mean positive and negative slopes of pitch and intensity are

Table 8
Acoustic features used for emotion prediction with SVR.

| Level | Features | | $N_{feat}$ |
|---|---|---|---|
| Voiced unit | Pitch (PITCH) | Mean, standard deviation, range (max-min), mean absolute pitch slope | 4 |
| Voiced unit | Intensity (INTENS) | Root-Mean-Square (RMS), mean, range (max-min), standard deviation | 4 |
| Voiced unit | Distribution energy in spectrum (ESPECTR) | Slope Long-Term Averaged Spectrum (LTAS), Hammarberg index, standard deviation, center of gravity (cog), skewness | 5 |
| Voiced unit | MFCC (MFCC) | 12 MFCC coefficients, 12 deltas (first order derivatives) | 24 |
| Whole segment | other | speech rate1, speech rate2, mean positive slope pitch, mean negative slope pitch, mean positive slope intensity, mean negative slope intensity | 6 |

calculated by summing and averaging all the positive and negative changes in pitch and intensity measured framewise over the voiced parts.

The majority of our acoustic features were measured per voiced unit. The features extracted on voiced-unit-level were aggregated to segment-level by taking the **mean**, **minimum**, and **maximum** of the features over the voiced units. Hence, we obtain per segment a feature vector with $(3 \times (4+4+5+24))+6 = 117$ dimensions. These features were normalized by transforming the features to $z$-scores: $z = (x-\mu)/\sigma$, with $\mu$ and $\sigma$ calculated over a development set.

### 5.2.3. Lexical features

As SVMs (and SVRs) do not naturally take raw text (words) as input, we used lexical features that are based on a continuous representation of the textual input (similar to Truong and Raaijmakers, 2008). The textual input in our case is a manual word-level transcription made by the author herself (but could eventually be made by an ASR system). A fairly standard method to build features from textual input, and that has successfully been applied to text and document classification/retrieval (see e.g., Salton and Buckley, 1988; Joachims, 1998) was employed, namely a *tf-idf* weighting scheme (**t**erm **f**requency-**i**nverse **d**ocument **f**requency). The *term frequency* $tf_{w,s}$ is defined as the number of times a given word $w$ appears in a segment $s$ (i.e., an utterance) and reflects its importance to that specific segment. The *document frequency* $df_w$ is defined as the number of segments containing word $w$. The *tf-idf* weight for each word $w$ is then computed by:

$$\mathrm{tf}-\mathrm{idf}_{w,s} = \mathrm{tf}_{w,s} \times \mathrm{idf}_w = \mathrm{tf}_{w,s} \times \log\left(\frac{N}{\mathrm{df}_w}\right) \quad (1)$$

where $N$ is the total number of segments in the training set. The weights tend to filter out common words. Words that appear frequently in one utterance ($= tf$), but rarely in the whole collection of utterances ($= idf$) are more likely to be relevant to that utterance and thus have a high *tf-idf* weight. In addition, to adjust for differences in utterance length, the feature vectors were normalized to unit length by L2-normalization.

$$x_n = \frac{x_n}{\sqrt{\sum_{i=0}^{N} x_i^2}} \quad (2)$$

where $x_n$ is a value in a vector with $N$ dimensions. To give an idea of the size of $N$, the number of unique words in the whole corpus is 1963. Note that these features are normalized over the entire corpus.

Table 9
Experimental setup of the material for $N$-fold cross-validation experiments.

| Gender | Total segments | $N_{fold}$ | Splits (approximately) in training–development–testing sets |
|---|---|---|---|
| Female | 1048 | 11 | 80%–10%–10% |
| Male | 1352 | 17 | 87%–8%–5% |

## 5.3. Experimental setup

The automatic affect recognition experiments were carried out speaker-independently, but separately for female and male speakers. We performed $N$-fold cross-validation, where in each fold, we left out one specific speaker for testing. In each fold, the data set was divided into three sets: a training, development and test set (see Table 9), where the training and test sets are disjoint. The test set consists of speech segments from a specific speaker that is excluded from the training and development set. The development set is comprised of randomly chosen segments, drawn from the remaining segments after the test speaker has been filtered out.

The development set is used for parameter tuning and $z$-scoring. The features were normalized by $z$-scoring ($z = (x-\mu)/\sigma$) where the $\mu$ and $\sigma$ were calculated on the development set. In parameter tuning, the parameter set that achieved the lowest error rate, averaged over $N$ folds, was selected to use in the final testing. The error rate is a simple measure based on the absolute difference between the reference and the predicted value, see Eq. (3) and (4).

Three prediction experiments using different types of annotations were performed. With these 3 experiments, we compared the added value of annotation of felt emotion versus annotation of perceived emotion, and we assessed the effect of averaging annotations. The SELF-ratings refer to the annotations that were made by the gamers themselves which are most likely to reflect 'felt' emotions. The OTHER.-AVG-ratings refer to the averaged ratings of 3 different observers. The OTHER.3-ratings are the individual ratings

of the observers. Not each segment was rated by the same 3 observers as each of the 6 observers rated different parts of the data, see Fig. 7. Given this situation, the training and testing procedure for OTHER.3 were not as straightforward as for SELF and OTHER.AVG. Each observer complements another observer such that all segments are rated once: TH pairs with CO, PI pairs with FR, and RA pairs with AT (see Fig. 7). For OTHER.3, the recognizers were trained on each of these pairs' ratings and tested using ratings drawn randomly from the other 2 observers not used during training (to make the tests more conservative, in contrast with the OTHER.AVG experiments where the raters are drawn from the same pool in training and testing). The performance reported are the evaluation metrics' averages taken over the 3 pairings of observers.

We report several evaluation metrics so that these can be compared to other results in the literature. Firstly, we used a relatively simple evaluation metric (similar to Grimm et al., 2007b) that measures the absolute difference between the predicted output and the reference input:

$$e_i = \left| x_i^{\text{pred}} - x_i^{\text{ref}} \right| \tag{3}$$

$$e_{\text{avg}} = \frac{1}{N} \sum_{i}^{N} e_i \tag{4}$$

and we report the error $e_{\text{avg}}$ that is averaged over a total of $N$ segments. The lower $e_{\text{avg}}$, the better the performance. Secondly, Pearson's $r$ correlation coefficient was also reported.

### 5.4. Results

Here, we present the results of our automatic affect recognition experiments which were performed separately for female and male data, and separately for arousal and valence dimensions. The affect recognizers were developed with either acoustic information or lexical information. The main evaluation metrics are $e_{\text{avg}}$ and $r$. We present the results for the acoustics-based and text-based arousal and valence recognizers in Table 10.

From Table 10, we can make several observations. First of all, we observe that performances are highest when OTHER.AVG-ratings are used, followed by OTHER.3 and SELF, respectively (note: the lower $e_{\text{avg}}$ the better, the higher $r$ the better). This suggests that emotions as perceived by observers can be better modeled than emotions as felt and reported by the gamers themselves: predicting individual observers' ratings is easier than predicting the gamers' own ratings. In addition, averaging ratings from multiple raters results in better recognition performances than using individual-specific ratings such as the SELF and the OTHER.3-ratings (which is in line with Mower et al., 2009). Secondly, the arousal dimension is better modeled by acoustic features, while the valence dimension is better modeled by textual features, see Table 11 re-confirming Grimm et al. (2007a) and Truong and Raaijmakers (2008) (a feature analysis of the acoustic and lexical features is out of scope for the current paper, however, Truong and Raaijmakers, 2008, provide a small feature analysis of the lexical features used although performed with a different learning algorithm). Finally, we note that in general, performance is relatively low, but that the majority of recognizers perform better than the baseline. One aspect that may have contributed to these relatively low performance scores for the perceived affect recognition is that the observers rated the stimuli on the basis of audiovisual information whereas our recognizers are based on audio information only. The

Table 10
Results (averaged over male and female performances) of the *acoustics-based* and *text-based* arousal and valence recognizers: the last column under 'Baseline' represent results from a baseline recognizer that always predicts Neutrality.

| | Reference | $\text{Test}_{\text{SVR}}$ | | $\text{Test}_{\text{SVR}}$ | | Baseline |
|---|---|---|---|---|---|---|
| | | $e_{\text{avg}}$ | $r$ | $e_{\text{avg}}$ | $r$ | $e_{\text{avg}}$ |
| | | Acoustic | | Textual | | |
| Aro | SELF | 0.41 | 0.25 | 0.44 | 0.01 | 0.45 |
| | OTHER.3 | 0.32 | 0.31 | 0.34 | 0.04 | 0.39 |
| | OTHER.AVG | 0.21 | 0.55 | 0.24 | 0.29 | 0.31 |
| Val | SELF | 0.36 | 0.18 | 0.36 | 0.14 | 0.36 |
| | OTHER.3 | 0.30 | 0.32 | 0.28 | 0.48 | 0.31 |
| | OTHER.AVG | 0.26 | 0.41 | 0.21 | 0.62 | 0.28 |

Table 11
Summary of several comparable speech-based studies working with dimensional ratings.

| Study | Data | | Human–interrater agreement | | Human–machine agreement | |
|---|---|---|---|---|---|---|
| | | | Val | Aro | Val | Aro |
| Grimm et al. (2007a) | German talk show | $r$ | 0.48 | 0.72 | 0.34 | 0.73 |
| | | $e_{\text{avg}}$ | | | 0.34 | 0.19 |
| Busso and Narayanan (2008) | Dyadic interactions | $r$ | 0.79 | 0.59 | | |
| Current study (OTHER.AVG) | Video game | $r$ | 0.64 | 0.32 | 0.41 | 0.55 |
| | | $e_{\text{avg}}$ | 0.28 | 0.40 | 0.26 | 0.21 |

Table 12
$e_{avg}$ for cross-rating experiments: train on one type of ratings and test on another type (error rates averaged over acoustics and text-based recognizers).

| Training | Testing | | | | | | Baseline | |
|---|---|---|---|---|---|---|---|---|
| | SELF | | OTHER.3 | | OTHER.AVG | | | |
| | Aro | Val | Aro | Val | Aro | Val | Aro | Val |
| SELF | 0.41 | 0.36 | 0.35 | 0.31 | 0.25 | 0.26 | 0.45 | 0.36 |
| OTHER.3 | 0.42 | 0.36 | 0.33 | 0.29 | 0.24 | 0.24 | 0.39 | 0.31 |
| OTHER.AVG | 0.39 | 0.35 | 0.32 | 0.29 | 0.21 | 0.21 | 0.31 | 0.28 |

expectation is that adding a facial expression classifier will additionally increase performance (which is future work).

When we inspect the errors that the recognizers produce, we notice that the largest errors are made in the extremeties of the arousal-valence space. On the one hand, this makes sense since the chance of large errors is highest in the extremeties. On the other hand, when we assume that the annotations correctly reflect the emotions expressed, then we would expect that the errors in the extremeties would be smaller since extreme emotions are expected to be easier to detect.

### 5.5. Cross-rating emotion recognition experiments

We performed cross-rating recognition experiments, training on one type of ratings and testing on another type of ratings, in order to see whether there are ratings that are more 'robust' than others; can we, for example, use OTHER.AVG ratings to predict SELF ratings? According to the error rates shown in Table 12, it appears that OTHER.AVG-ratings are easiest to predict and that they are most robust, i.e., they can also be used to recognize SELF. Conversely, SELF-ratings are most difficult to model.

## 6. Conclusions and discussions

### 6.1. Summary

We have presented a spontaneous audiovisual emotion database that was collected in a videogame environment and that has some unique properties: (1) part of the corpus is rated by both the gamers themselves and observers on continuous arousal-valence scales, and (2) the elicitation method used in this corpus exploits the advantages of multiplayer videogames. By putting friends together in one room and letting them play a (manipulated) multiplayer videogame, a natural environment is created in which spontaneous, affective vocal and facial interaction can easily take place. With this corpus, we explored several research questions. Under the assumption that people are the best decoders of their own emotions, we compared self-reported and observed emotion ratings to each other. We found confirmations that there are discrepancies between SELF-ratings and OTHER.AVG-ratings. The SELF-raters appeared to rate their own emotions much more extremely than the OTHER.AVG-raters. This observation suggested that the emotions felt by the gamers were not always perceivable with the observers. Furthermore, human observers have difficulty agreeing with each other on the perception of spontaneous affect. In general, the human-human agreement scores among the 3 observers were relatively low, especially for arousal – $e_{avg}$ of 0.40 ($r = 0.32$) and 0.28 ($r = 0.64$) for arousal and valence respectively. The human raters showed more agreement among each other along the valence dimension than the arousal dimension.

The differences between 'self' and 'observer' ratings influenced the development and performance of automatic affect recognizers. The results, obtained with acoustics-based and text-based regression models, showed that the observed emotion ratings were much better predicted than the self-reported emotion ratings. Here, we should remark that in the SELF condition, by design, the ratings were made by different raters and hence, the recognizers are performing 'rater-independent' recognition, whereas in the OTHER.AVG condition, the raters are drawn from the same pool in training and testing; this means that the task in the SELF condition is presumably more difficult than the task in the OTHER.AVG condition. We also have to keep in mind that despite our efforts, there are some differences between the way the SELF- and OTHER.AVG-ratings are obtained (see Section 4) which may have affected performance. Recognition experiments were also performed with the individual OTHER.3-ratings which resulted in intermediate recognition performances, illustrating that integrating different views from multiple raters by averaging increases recognition performance, and also results in more 'robust' ratings as illustrated by the cross-data recognition experiments we performed.

In conclusion, the differences between human self-reported and observed emotion ratings also lead to large differences in performances of the emotion recognizers: the self-reported ratings were much harder to recognize than the observed ratings. The results raise the question whether future machine recognizers should be able to recognize 'felt' or 'perceived' affect, and how machine recognizers should learn to recognize 'felt' affect.

### 6.2. Discussion and future research

We suggest that the validation of human affect ratings and how these subjective ratings influence the behavior of automatic affect recognizers should require more attention. The way the data is rated is of much importance, especially in the case of annotation of spontaneous affect where a high level of subjectivity is intrinsic to the data. We have seen that it matters who you ask to annotate: whether you ask people to report their own emotions or ask other people to rate observed emotions makes a big difference. Other annotation aspects may affect the emotion ratings as well. For example, during the rating processes, we have also experienced that some participants expressed difficulties with the interpretation of the arousal scale – they found it

difficult to make a distinction between neutrality and passiveness. Whether this was due to the specific gaming context or the specific annotation tool that was used and how much this has affected the ratings remains to be seen. In addition, the way we measured the 'felt' emotion of the participants was constrained by practical issues and can hence be improved: one could for example add physiological measures, such as heart rate, to obtain a more reliable assessment of 'felt' emotion. As a consequence, there were (unavoidable) small differences between the rating procedures of the self-raters and the observers which should be addressed in future work. Furthermore, it would be interesting to see whether more training of the raters in emotion annotation would improve agreement and consistency of the annotations. However, this would mean that the recognizer will learn to recognize the views of trained emotion experts rather than the 'layman's' view from naive observers which is not always what one wants. As mentioned earlier, the large differences between self-reported emotion and observed emotion, and the relatively low recognition performance for the self-reported emotion, makes one think about whether an emotion recognizer should aim to detect felt or perceived emotion. For speech, few investigations have been performed on the relation between vocal characteristics and the *felt* emotion, i.e., the internal emotional state (e.g., Biersack and Kempe (2005)). Humans have more control over the speech that comes from the vocal tract than they have over physiological effects such as heart rate which raises the expectation that the internal emotional state is difficult to measure in speech. However, this requires more investigation. Furthermore, the search for more reliable acoustic correlates and modeling of emotion continues. Moreover, it is still unknown how acoustic and other multimodal cues of emotion interact with each other and how this interaction can be modeled computationally for recognition purposes (the TNO-GAMING corpus allows for an audiovisual analysis). Finally, we have approached the recognition of spontaneous affect with a dimensional approach and adopted regression techniques to recognize arousal and valence values separately. This is a relatively new approach that has not been fully explored and matured yet, and there are many aspects up for discussion: e.g., how realistic is it to develop models to recognize spontaneous affect in terms of a pair of arousal-valence coordinates, or what other techniques than regression techniques can we use to model ordered scales of affect, or what other ways can be used to obtain reliable 'ground truth' dimension-based labels? As illustrated, many of these issues need more investigation and addressing these issues will gradually take us one step closer to understanding how we can develop more adequate spontaneous (speech-based) affect recognizers.

## Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work was supported by MultimediaN and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet).

## References

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002), pp. 2037–2040.

Auberge, V., Audibert, N., Rilliard, A., 2006. Auto-annotation: an alternative method to label expressive corpora. In: Proc. Fifth Internat. Conf. on Language Resources and Evaluation (LREC 2006).

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70, 614–636.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2000. Desperately seeking emotions: actors, wizards, and human beings. In: Cowie, R., Douglas-Cowie, E., Schröder, M. (Eds.), In: Proc. ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, pp. 195–200.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2006. Combining efforts for improving automatic classification of emotional user states. In: Language Technologies (IS-LTC), pp. 240–245.

Biersack, S., Kempe, V., 2005. Tracing vocal expression of emotion along the speech chain: do listeners perceive what speakers feel? In: Proc. ISCA Workshop on Plasticity in Speech Perception (PSP2005), pp. 211–214.

Boersma, P., Weenink, D., 2009. Praat: doing phonetics by computer (Version 5.1.07). [Computer Program]. <http://www.praat.org/> Retrieved 16.06.09.

Busso, C., Narayanan, S.S., 2008. The expression and perception of emotions: comparing assessments of self versus others. In: Proc. Interspeech 2008, pp. 257–260.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. Speech Commun. 40, 5–32.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schroeder, M., 2000. FEELTRACE: an instrument for recording perceived emotion in real time. In: Proc. ISCA ITRW on Speech and Emotion, pp. 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. 18 (1), 32–80.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP 1996), pp. 1970–1973.

Den Uyl, M.J., Van Kuilenburg, H., 2005. The facereader: online facial expression recognition. In: Proc. Measuring Behavior, pp. 589–590.

Devillers, L., Lamel, L., Vasilescu, I., 2003. Emotion detection in task-oriented spoken dialogues. In: Proc. IEEE Internat. Conf. on Multimedia and Expo (ICME'03), pp. 549–552.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Davvidou, S., Abrillian, S., Cox, C., 2005. Multimodal databases of everyday emotion: facing up to complexity. In: Proc. Interspeech 2005, pp. 813–816.

Ekman, P., 1972. Universals and cultural differences in facial expressions of emotion. In: Cole, J. (Ed.), Nebraska Symposium on Motivation, pp. 207–283.

Ekman, P., Friesen, W., 1975. Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions. Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R., 2010. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. J. Multimodal User Interf. 3, 7–19.

Giannakopoulos, T., Pikrakis, A., Theodoridis, S., 2009. A dimensional approach to emotion recognition of speech from movies. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP'09), pp. 65–68.

Grimm, M., Kroschel, K., Mower, E., Narayanan, S., 2007a. Primitives-based evaluation and estimation of emotions in speech. Speech Commun. 49, 787–800.

Grimm, M., Kroschel, K., Narayanan, S., 2007b. Support Vector Regression for automatic recognition of spontaneous emotions in speech. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007), pp. 1085–1088.

Gunes, H., Schuller, B., Pantic, M., Cowie, R., 2011. Emotion representation, analysis and synthesis in continuous space: a survey. In: Proc. IEEE Internat. Conf. on Automatic Face & Gesture Recognition and Workshops (FG2011), pp. 827–834.

Hanjalic, A., Xu, L.-Q., 2005. Affective video content representation and modeling. IEEE Trans. Multimedia 7, 143–154.

Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Proc. 10th European Conf. on Machine Learning (ECML-98), pp. 137–142.

Johnstone, T., van Reekum, C.M., Hird, K., Kirsner, K., Scherer, K.R., 2005. Affective speech elicited with a computer game. Emotion 5 (4), 513–518.

Kim, J., André, E., Rehm, M., Vogt, T., Wagner, J., 2005. Integrating information from speech and physiological signals to achieve emotional sensitivity. In: Proc. Interspeech, pp. 809–812.

Kwon, O.-W., Chan, K., Hao, J., Lee, T.-W., 2003. Emotion recognition by speech signals. In: Proc. Eurospeech, pp. 125–128.

Lang, P., 1995. The emotion probe. Amer. Psychol. 50, 371–385.

Lazarro, N., 2004. Why We Play Games: Four Keys to More Emotion Without Story.

Lee, C.M., Narayanan, S., Pieraccini, R., 2002. Classifying emotions in human-machine spoken dialogs. In: Proc. IEEE Internat. Conf. on Multimedia and Expo (ICME '02), pp. 737–740.

Liscombe, J., Venditti, J., Hirschberg, J., 2003. Classifying subject ratings of emotional speech using acoustic features. In: Proc. Eurospeech, pp. 725–728.

Mower, E., Mataric, M.J., Narayanan, S.S., 2009. Evaluating evaluators: a case study in understanding the benefits and pitfalls of multi-evaluator modeling. In: Proc. Interspeech 2009, pp. 1583–1586.

Nicolaou, M.A., Gunes, H., Pantic, M., 2010. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In: Proc. Internat. Workship on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, pp. 43–48.

Nicolaou, M.A., Gunes, H., Pantic, M., 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. Affect. Comput. 2, 92–105.

Nwe, T.L., Foo, S.W., De Silva, L.C., 2003. Speech Commun. 41, 603–623.

Petrushin, V.A., 1999. Emotion in speech: recognition and application to call centers. In: Proc. 1999 Conf. on Artificial Neural Networks in Engineering (ANNIE'99).

Polzin, T., Waibel, A., 1998. Detecting emotions in speech. In: Proc. Cooperative Multimodal Communication (CMC'98).

Ravaja, N., Saari, T., Turpeinen, M., Laarni, J., Salminen, M., Kivikangas, M., 2006. Spatial presence and emotions during video game playing: does it matter with whom you play? Presence: Teleoper. Virtual Environ. 15, 381–392.

Russell, J.A., 1980. A circumplex model of affect. J. Pers. Soc. Psychol. 39, 1161–1178.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manage. 24 (5), 513–523.

Scherer, K.R., 2010. The component process model: architecture for a comprehensive computational model of emergent emotion. In: Scherer, K.R., Bänziger, T., Roesch, E. (Eds.), Blueprint for Affective Computing: A Sourcebook, pp. 47–70.

Schlosberg, H., 1954. Psychol. Rev. 61, 81–88.

Schuller, B., Rigoll, G., Lang, M., 2003. Hidden Markov model-based speech emotion recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003), pp. 1–4.

Smola, A.J., Scholkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.

Tato, R., Santos, R., Kompe, R., Pardo, J.M., 2002. Emotional space improves emotion recognition. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP 2002), pp. 2029–2032.

Truong, K.P., Neerincx, M.A., Van Leeuwen, D.A., 2008. Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data. In: Proc. Interspeech 2008, pp. 381–321.

Truong, K.P., Raaijmakers, S., 2008. Automatic recognition of spontaneous emotions in speech using acoustic and lexical features. In: Proc. Fifth Joint Workshop on Machine Learning and Multimodal Interaction (MLMI 2008), pp. 161–172.

Truong, K.P., Van Leeuwen, D.A., Neerincx, M.A., De Jong, F.M.G., 2009. Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion. In: Proc. Interspeech, pp. 2027–2030.

Vapnik, V.N., 2002. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Ververidis, D., Kotropoulos, C., 2005. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In: Proc. IEEE Internat. Conf. on Multimedia and Expo (ICME 2005), pp. 1500 –1503.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features, and methods. Speech Commun. 48 (9), 1162–1181.

Wang, N., Marsella, S., 2006. Introducing EVG: an emotion evoking game. In: Proc. Internat. Conf. on Interactive Virtual Agents (IVA 2006), pp. 282 – 291.

Williams, C.E., Stevens, K.N., 1972. Emotions and speech: some acoustical correlates. J. Acoust. Soc. Amer. 52 (4B), 1238–1250.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R., 2008. Abandoning emotion classes – towards continuous emotion recognition with modeling of long-range dependencies. In: Proc. Interspeech, pp. 597–600.

Wöllmer, M., Eyben, F., Schuller, B., Douglas-Cowie, E., Cowie, R., 2009. Data-driven clustering in emotional space for affect recognition using discriminatively trained lstm networks. In: Proc. Interspeech, pp. 1595–1598.

Wundt, W., 1874/1905. Grundzüge der physiologischen Psychologie, [Fundamentals of pshysiological psychology]. Engelmann, Leipzig.

Yildirim, S., Lee, C.M., Lee, S., Potamianos, A., Narayanan, S.S., 2005. Detecting politeness and frustration state of a child in a conversational computer game. In: Proc. Interspeech 2005, pp. 2209–2212.

Yu, C., Aoki, P., Woodruff, A., 2004. Detecting user engagement in everyday conversations. In: Proc. Interspeech, pp. 1329–1332.

Zeng, Z., Zhang, Z., Pianfetti, B., Tu, J., Huang, T.S., 2005. Audio-visual affect recognition in activation-evaluation space. In: Proc. IEEE Internat. Conf. on Multimedia and Expo (ICME'05), pp. 828–831.