THE VERA AM MITTAG GERMAN AUDIO-VISUAL EMOTIONAL SPEECH DATABASE

Michael Grimm, Kristian Kroschel

Institut für Nachrichtentechnik (INT)
Universität Karlsruhe (TH)
Karlsruhe, Germany

Shrikanth Narayanan

Speech Analysis and Interpretation Lab (SAIL)
University of Southern California (USC)
Los Angeles CA, USA

ABSTRACT

The lack of publicly available annotated databases is one of the major barriers to research advances on emotional information processing. In this contribution we present a recently collected database of spontaneous emotional speech in German which is being made available to the research community. The database consists of 12 hours of audio-visual recordings of the German TV talk-show "Vera am Mittag", segmented into broadcasts, dialogue acts and utterances. This corpus contains spontaneous and very emotional speech recorded from unscripted, authentic discussions between the guests of the talk-show. In addition to the audio-visual data and the segmented utterances we provide emotion labels for a great part of the data. The emotion labels are given on a continuous-valued scale for three emotion primitives: valence, activation and dominance, using a large number of human evaluators. Such data is of great interest to all research groups working on spontaneous speech analysis, emotion recognition in both speech and facial expression, natural language understanding, and robust speech recognition.

Index Terms— Data acquisition, Speech analysis, Speech processing, TV, Video signal processing

1. INTRODUCTION

Although in recent years much effort has been invested in collecting and segmenting audio-visual speech data [1, 2, 3, 4], there is still a lack of spontaneous and emotionally rich speech corpora. Such data are essential for enabling studies in the field of speech and emotion analysis. We describe a German audio-visual speech corpus that is being released in conjunction with this paper. It fulfills many desirable criteria suggested by emotion researchers (e.g., [9]) and thus may serve as a valuable common resource for future studies. The unique aspects of this database with respect to other emotional speech corpora is that it is based on spontaneous interactions, it contains both audio and video signals, and it comes with a detailed annotation of emotion labels.

Apart from speech analysis, this data might be useful to challenge the robustness of a variety of speech applications such as automatic speech recognition (ASR), natural language understanding, speaker identification and emotion recognition. Additionally, the video data might be useful for facial expression analysis, lipreading or studies on the synchronization of emotion, speech and facial expression.

Our major interest when collecting this database was the recognition of the emotion conveyed in the dialogues [5, 6]. Therefore, the utterances that are provided in addition to the entire recordings of the talk-shows were segmented only from those speakers who had shown a serious amount of emotion.

For the annotation of the emotion we used a three-dimensional emotion space concept in which an emotion is described in terms of three basic entities (*primitives*): *valence* (negative – positive), *activation* (calm – excited) and *dominance* (weak – strong) [7]. This concept has gained much attention in recent years. Due to its continuous nature it may be used to resolve moderate, authentic emotions, to capture emotion transitions or to model person-specific emotional expression patterns while still being convertible to the well-established emotion categories (e.g., happy, sad, etc.). As an evaluation tool, an icon-based, text-free method using *Self Assessment Manikins* (SAMs) was chosen [6, 8]. This method provides a discretized 5-point scale for each emotion primitive in the range of [-1,+1].

The rest of the paper is organized as follows. Section 2 names other relevant corpora of emotional and spontaneous speech. Section 3 describes the generation of the new database. As a result, in section 4 the new corpus is introduced. Section 5 contains the conclusion.

2. EXISTING DATABASES

In recent years, there has been a considerable amount of work on the collection of emotional speech and facial expression images. Douglas-Cowie *et al.* give a comprehensive overview of 21 different corpora [1]. Important aspects of data collection are highlighted in [2, 9]. An overview, particularly on emotional speech databases, can be found in [3]. Databases with affective facial expressions are summarized in [10]. One of the latest and most extensive data collections can be found in [4]: the HUMAINE database serves as a container for three naturalistic databases and another five databases with induced emotions. However, none of these databases consists of authentic and spontaneous interaction and provides at the same time an emotion annotation within the framework of an emotion space for a large number of utterances.

This work was supported by grants of the Collaborative Research Center (SFB) 588 of the Deutsche Forschungsgemeinschaft (DFG).
Contact: grimm@int.uni-karlsruhe.de

3. DATA COLLECTION

We decided to use a TV talk-show for this data collection because there is a reasonable amount of speech from the same speakers available in each session. Furthermore, the spontaneous discussions between the talk-show guests are often rather affective. Such interpersonal communication leads to a wide variety of emotional states, depending on the topics discussed. These topics were mainly personal issues such as friendship crises, fatherhood questions or romantic affairs. The talk-show guests did not know that the recording was going to be analyzed from a perspective of emotional content. Thus they did not try to particularly mask their emotion. As a conclusion of our previous studies we chose a talk-show in which it was assured that the guests were not being paid to perform as lay actors.

One disadvantage of such data collection from a TV show is that it is obviously not possible to control the affective states that would occur in the cause of the dialogues. However, it is always a trade-off between controllability of (affective) content and naturalness of the interaction. Furthermore, the spectrum of emotional states in one dialogue is often restricted due to the discussed topics.

3.1. Selection of the talk-show

For this database we recorded 12 broadcasts of the talk-show “Vera am Mittag” (in English, “Vera at noon”). The acronym *VAM Corpus* is derived from this title. The recorded shows were broadcasted on the German TV channel Sat.1 between December 2004 and February 2005.

Each broadcast consists of several dialogues between two to five persons each. The discussions were moderated by the anchorwoman, Vera. In addition to the spontaneous and unscripted nature of the discussions, this well-structured set-up with few persons having rather long dialogues of up to 20 minutes led us to use this talk-show.

The speakers were between 16 and 69 years old. 70% were 35 or younger at the time of recording.

3.2. Segmentation

Ten of the broadcasts (of approx. 1 hour each) were first segmented into discussions containing the whole dialogue between a limited number of 2-5 talk-show guests. The remaining two broadcasts were not segmented since they contained rather serious interviews which were not relevant with respect to the affective content. In total, 45 such discussions were extracted as videos containing the audio and video signals.

In a second step the dialogues were segmented into utterances. The audio signal was stored separately for each sentence. For the visual signal a series of still images was extracted in the Portable Network Graphics (png) format. An identification was given to each speaker and to each utterance. Finally, the sentences were rearranged to form speakerwise groups, which led to a hierarchical data structure.

A complete listing including the transcription and additional comments can be found in the documentation that comes along with the data. In the sequel this large amount of data was reduced to have most relevant data only with respect to speech emotion analysis.

Usability	Gender	Num. Speakers	Num. Sentences	Num. Sentences per Speaker
++	m	4	91	22.8
++	f	15	408	27.2
++	m+f	19	499	26.3
+	m	7	109	15.6
+	f	21	410	19.5
+	m+f	28	519	18.5
○	m	6	79	13.2
○	f	4	78	19.5
○	m+f	10	157	15.7
–	m	27	129	10.8
–	f	20	117	9.0
–	m+f	47	246	9.8
++, +, ○, –	m	44	408	14.1
++, +, ○, –	f	60	1013	19.1
++, +, ○, –	m+f	104	1421	17.3

Table 1. Selection of usable speakers for emotion analysis in the VAM database: Summary of *very good* (++), *good* (+), *usable* (○) and *not usable* (–) speakers.

3.3. Speaker selection

During segmentation the speakers were roughly classified into four classes: *very good*, *good*, *usable* and *not usable*. This classification reflects the usability for audio-visual emotion analysis and was based on how many sentences a speaker produced, how emotional these utterances were, and which spectrum of emotions was covered by this person. *Very good* speakers were characterized by a high activity with many emotional sentences of different emotion qualities. *Good* speakers typically showed high activity, but only a small set of emotions, such as only angry. *Usable* speakers did not talk a lot in general during the recording while still showing some emotion, whereas *not usable* speakers neither talked much nor showed many emotions.

Table 1 shows the selection of the speakers on the basis of these criteria. It reveals that it is very difficult to find *very good* speakers in spontaneous interactions; this is not surprising since the frequency of occurrence of any specific emotion type in a dialog is fairly small compared to the neutral state in spontaneous unscripted interactions. 47 speakers comprise the set of *very good* and *good* speakers taken together, which is a reasonable amount of authentic emotional data in the state of the art.

3.4. Recording quality

The video files are MPEG-coded image sequences of 352x288 pixels with a frame rate of 25 fps. A constant code rate of 1.15 Mbit/s was used. The audio stream was included at 44.1 kHz, stereo. For the extracted wav files, the signal was downsampled to 16 kHz (16 bit). The png images have a resolution of 352x288 pixels (24 bit).

4. THE VAM CORPUS

In this section the individual parts of the *VAM Database* are introduced.

4.1. The VAM-Video Database

As explained in the previous section, the audio-visual signal of each utterance was segmented from the dialogues. These 1421 videos of the 104 speakers are called *VAM-Video* database in the following. An evaluation of the emotion was not carried out for this database. The whole dialogues and the whole broadcasts are also included in this corpus. The two other database modules are derived from this database.

4.2. The VAM-Audio Database

The second database, *VAM-Audio*, contains only the acoustic signal of the recorded utterances. These utterances were mainly complete sentences, but sometimes also just exclamations, affect bursts or grammatically incomplete sentences which were due to the spontaneous nature of the interactions. Many utterances had to be discarded because of background music, applause from the audience or other interruptions.

In a first iteration, only those sentences were selected that were by speakers who had been roughly classified as *very good* with respect to the emotions conveyed. This part contains 499 utterances of 19 different speakers (4m/15f), and may be denoted as *VAM-Audio I*. On average, we recorded 26.3 sentences per speaker.

Here is one example of a dialogue fraction, in addition with a first emotional impression that had been noted during segmentation:

> Melanie: “Sie hat gesagt, sie überlässt ihn mir, sie würde mir sogar helfen.” [1] – *serious*
> Kristin: “Findest du?” [2] – *somewhat amused*
> Melanie: “Kristin, das hast du aufm Schiff gesagt!” [3] – *somewhat angry*
> Kristin: “Melanie, ich weiß genau, was ich auf dem Schiff gesagt habe!” [4] – *angry*

This database was evaluated by 17 human listeners. They assessed the emotional content in terms of the emotion primitives *valence*, *activation*, and *dominance* in each sentence using the SAMs [6, 8].

Since the evaluation results suggested a rather unbalanced distribution of the emotions, we decided to further include the sentences by those speakers who had been classified as *good* speakers. This second part contains 519 additional sentences by 28 speakers (7m/21f) with 18.5 sentences per speaker on average. It will be denoted *VAM-Audio II* in the following.

Since not all evaluators were still available at the time we extended the database by this second module, only 6 evaluators assessed the emotional content in *VAM-Audio II*.

The joint database *VAM-Audio* thus includes both modules. It contains 1018 emotional utterances by 47 speakers (11m/36f) with an average of 21.7 sentences per speaker. The average sentence duration was 3.0 s.

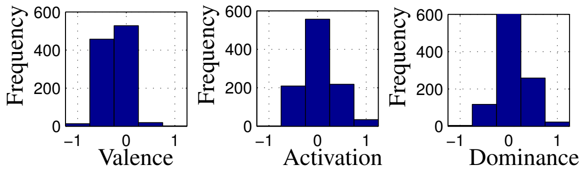

Fig. 1. Histogram of the emotions in the database *VAM-Audio*, containing both modules *VAM-Audio I* and *II* .

Figure 1 shows the histograms of the emotion primitives in the joint database *VAM-Audio*. It can be seen that a wide range of emotions is covered by the database. However, there is still a major part of the emotions in the neutral or negative area. Such a distribution was mainly due to the topics discussed in the talk-show. Often, the show was finished just when the situation was turning to be positive, or some romantic music was superimposed to highlight a happy outcome of the discussion.

4.3. The VAM-Faces Database

Besides the speech signal, the facial images of the speakers were also extracted from the original database *VAM-Video*. Strictly keeping every single frame as a still image was not reasonable since often the camera was directed to the audience or to the anchorwoman. For a database of facial images, we used only those sentences in which the camera was directed towards the speaker for at least a great portion of the sentence. Furthermore, those sentences were neglected in which the speaker was visible only from a profile view.

Thus, to have a relevant set of sentences we decided to keep 20 speakers for which the facial image sequence was extracted. This collection contains a set of 1867 images (93.6 images per speaker on average). This database is called *VAM-Faces* in the following.

For each speaker, the data was sub-structured into sentences. For each sentence we provide the audio-visual signal as a synchronization reference, and in addition the facial images as a png file sequence. Figure 2 shows some examples taken from the database *VAM-Faces*.

The emotion was also annotated for this corpus [11]. Since all major works on facial expression recognition are based on Ekman’s list of six basic emotion categories, *happiness, anger, sadness, disgust, fear, surprise* [12], together with *neutral*, we asked the evaluators for a rating on this category list in addition to the rating along the emotion primitives. To allow for the annotation of complex emotions, two emotion categories could be selected as an emotion mixture with one being the *major emotion* and the other (optionally) being the *minor emotion*. Since the emotional content obviously did not change every frame (frame delay 0.04 s), only every third frame of a sentence was evaluated, and the result was assigned to its neighbors as well. The number of evaluators varied from 8 to 34 (average 13.9), since not all evaluators were available to assess all images.

[1] She said, she would leave him to me, and she would even help me.

[2] Do you really think so?

[3] Kristin, this is what you told me on the boat!

[4] Melanie, I do know very well what I told you on the boat!

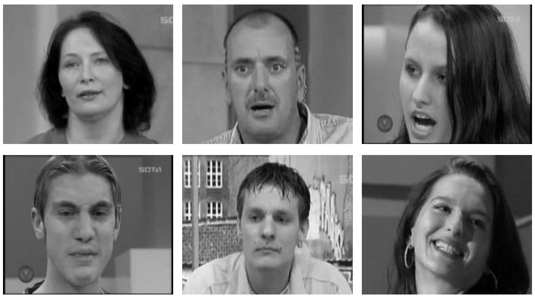

Fig. 2. Sample images contained in the database VAM-Faces.

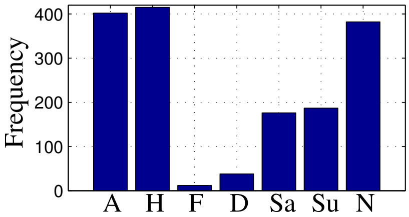

Fig. 3. Distribution of the emotions in *VAM-Faces*, classified as *anger (A), happiness (H), fear (F), disgust (D), sadness (Sa), surprise (Su)* and *neutral (N)*.

Figure 3 shows the histogram of the *major emotion* categories contained in *VAM-Faces*. It can be seen that our selection of the speakers led to many facial images expressing *anger, happiness* and *neutral* (approx. 400 images for each category). *Sadness* and *surprise* were shown in almost 200 images each, while *fear* and *disgust* were hardly shown by the speakers.

In addition, the evaluation of emotion in the facial expressions was carried out using the emotion primitives *valence*, *activation*, and *dominance*. The histogram is shown in figure 4. It can be seen that the extreme values of the primitives are not shown as often as the neutral values. In particular for *dominance* most images were assessed as being neutral, which might be due to the fact that *dominance* in general is rather conveyed in speech than in the facial expression.

5. CONCLUSION

In this paper, we presented the collection, segmentation and emotional labeling of many samples of spontaneous speech extracted from unscripted, natural discussions in a TV talkshow. Three modules of the *VAM Corpus* were introduced: *VAM-Video*, *VAM-Audio* and *VAM-Faces*. The emotion histograms showed a large spectrum of different emotions in this database. In contrast to other available databases, the emotion annotation is directly available along with the raw data, which may be of significant value for facilitating different types of studies on audio-visual speech in spontaneous interactions.

At the time of the conference, this data will be available to the research community through the online portal of the HUMAINE Association: `http://emotion-research.net`.

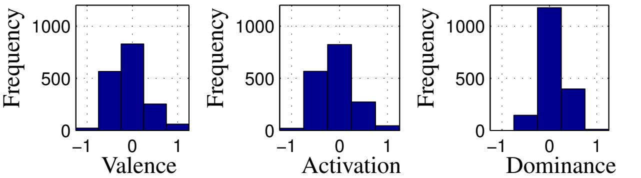

Fig. 4. Histogram of the emotions in the database *VAM-Faces*, evaluated using the emotion primitives.

6. REFERENCES

[1] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, “Emotional speech: Towards a new generation of databases,” *Speech Communication*, vol. 40, pp. 33–60, 2003.

[2] N. Campbell, “The recording of emotional speech: Jst/crest database research,” in *Proc. of LREC’02*, 2002, vol. 6, pp. 2029–2032.

[3] D. Ververidis and C. Kotropoulos, “A state of the art review on emotional speech databases,” in *Proceedings of the 1st Richmedia Conference*, Lausanne, Switzerland, 2003, pp. 109–119.

[4] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRoie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, “The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data,” in *Proc. ACII, LNCS 4738*, A. Paiva, R. Prada, and R. Picard, Eds. 2007, pp. 488–500, Springer Berlin Heidelberg, Germany.

[5] M. Grimm, K. Kroschel, and S. Narayanan, “Support vector regression for automatic recognition of spontaneous emotions in speech,” in *Proc. ICASSP*, 2007, vol. 4, pp. IV–1085 – IV–1088.

[6] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[7] R. Kehrein, “The prosody of authentic emotions,” in *Proc. Speech Prosody Conf.*, 2002, pp. 423–426.

[8] M. Grimm and K. Kroschel, “Evaluation of natural emotions using self assessment manikins,” in *Proc. IEEE ASRU*, 2005, pp. 381–385.

[9] E. Douglas-Cowie, R. Cowie, and M. Schröder, “A new emotion database - considerations, sources and scope,” in *Proc. ISCA ITRW on Speech and Emotion*, Newcastle, UK, 2000, pp. 39–44.

[10] L. Yin, X. Wei, Y. Sun, J. Wang, and M.J. Rosato, “A 3D facial expression database for facial behavior research,” in *Proceedings of the 7th Int. Conference on Automatic Face and Gesture Recognition (FGR)*, 2006, pp. 1–6.

[11] M. Grimm, D.G. Dastidar, and K. Kroschel, “Recognizing emotions in spontaneous facial expressions,” in *Proc. Int. Conf. on Intelligent Systems And Computing (ISYC)*, Ayia Napa/Cyprus, 2006.

[12] P. Ekman, “Universals and cultural differences in facial expressions of emotion,” in *In J. Cole (Ed.), Nebraska Symposium on Motivation*, Lincoln, NE, 1972, vol. 19, pp. 207–283, University of Nebraska Press.