Evaluation of Vocal-Facial Based Emotion Primitives

Kristian Kroschel, Michael Grimm, Vasilije Krstanovic

Universität Karlsruhe, Institut für Nachrichtentechnik
kroschel@int.uni-karlsruhe.de

Abstract: Emotion detection gains improved importance in man-machine interaction. There are two approaches to represent emotion: first, discrete elementary emotions like fear or joy might be defined, or an *emotion space* is used in which the discrete emotions define a point or a cluster. In this paper the latter method is used with a three-dimensional space spanned by the emotion primitives *valence*, *dominance*, and *activation*.
Synchronously, sequences of speech and facial expressions are extracted from a TV talk-show to generate an audio-video data base of authentic emotions. From these recordings, three data subsets have been generated: video alone, audio alone and combined audio-video. Investigations have shown that the emotions expressed by facial parameters change with a time constant of roughly 1 sec. Thus an additional subset of segmented video data has been generated. These data have been evaluated by 15 persons, i.e., they estimated the intensity of the emotion primitives in the range from -1 to +1 using so-called Self Assessment Manikins (SAM). To test the reliability of these estimates, the *correlation coeffcients* of the estimates of the individual evaluators have been calculated. As a second measure, the *standard deviation* of the estimates for the four subsets and the three emotion primitives has been calculated. These measures have been used to decide whether or not the evaluated data base can be used as the ground truth for automatic emotion classification.

1 Introduction

Human communication is severely controlled by emotions: if one of the communicating individuals gets angry, information transfer might become unreliable, i.e. part of information might get lost so that the terms of communication have to be changed. This is true also for the communication between a human and a humanoid robot. In case that the humanoid robot reacts not fast enough or does not unterstand an explanation of the human, the human might become impatient or angry. The same might happen with a car driver trying to input a new destination into the navigation system if he misspells the name of the destination and the system does not accept the name.

Emotions are expressed by many modalities: by speech, facial expression, physiological data like blood pressure etc. In this paper the first two are inspected in detail because they can be picked up by non-invasive methods. The question is how long a data segment of the observed speech signal should be and whether a single image of the observed facial expression is sufficient for reliable emotion parameter extraction. Even if the semantics of the uttered speech should not be taken into account, this question is of importance. Generally, it can be stated that the observed data segments of the audio and video data stream should be long enough so that the evaluator is able to get an unambiguous estimation of the emotional parameters. On the other hand, the segment should be short enough so that the emotional status of the observed person does not change.

Emotions can be represented in two ways: by a well-defined set of discrete elementa-

ry emotions or by so-called emotion primitives which span an emotion space. In this space, the elementary emotions mark a single point or a well defined cluster of points. In literature, e.g. [1], the six elementary emotions *anger*, *fear*, *surprise*, *disgust*, *joy*, and *sadness* are postulated. On the other hand, the emotion space might be defined by the three emotion primitives *valence*, *dominance*, and *activation* which each take on a value $-1 \leq x^{(i)} \leq +1$ as shown in Fig. 1.

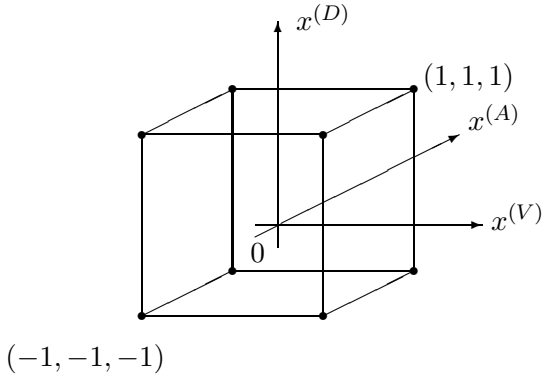


Figure 1: The emotion space with valence, activation and dominance

Due to its flexibility, the emotion space is used throughout this paper. For simple allocation of the weights $x^{(i)}$ - the index i denoting valence, dominance or activation - to the emotion primitives in the range $-1 \leq x^{(i)} \leq +1$, the non-verbal Self Assessment Manikins (SAM) [2] are used so that $x^{(i)}$ is quantized to $x^{(i)} \in [-1,\ -0.5,\ 0,\ +0.5,\ +1]$. For *valence* these values correspond with very negative, negative, neutral, positive and very positive. For *dominance* they correspond with very weak, weak, neutral, dominant and very dominant, and for *activation* with very calm, calm, neutral, excited and very excited.

2 The Audio-Visual Data Base

The goal for the construction of the data base was to include only authentic emotions and to present synchronized audio and video sequences with only one person speaking and no obstruction of the frontal view of the speaker. Due to the required authentic emotions, the German TV show *Vera am Mittag* was used which originally contained 1018 sentences from 36 female and 11 male speakers. The range of emotions was very large, the valence was biased towards negative. From these data only three female and two male speakers with together 78 sentences could be used due to the required quality of the data.

On the basis of these data, four subsets of data were generated:

- unsegmented audio signal
- unsegmented video signal
- segmented video signal
- unsegmented audio-video signal.

As segment length, 1 s has been selected because by subjective evaluation within this time interval the emotion did not seem to change. This assumption will be verified by further analysis of the available data. Thus all together 252 segments were extracted from the 78 available sentences.

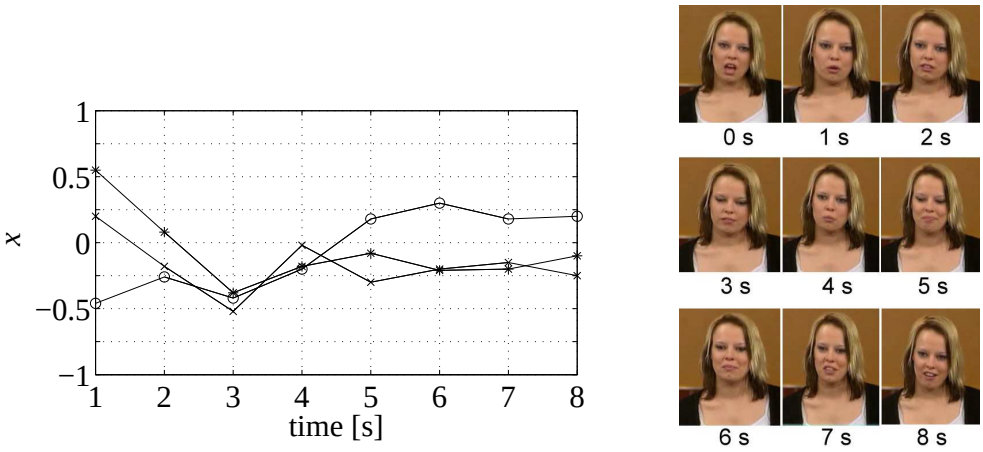
Since no static information was extracted from the video channel and since audio and video information was used synchonously, the question has to be answered which part of the face contributed to the emotion estimation. Due to the movement of lips during

articulation, this part of the face cannot be used for facial emotion extraction. Instead, the region around the eyes and on the forehead is used. This implies a reduction of information which becomes clear from observation of the face in Fig. 2.

3 Data Analysis

The goal of the data analysis was to generate a reliable ground truth as a basis for the design of automatic emotion classifiers. The group of evaluators consisted of 5 female and 10 male young persons of German and non-German background with sufficient knowledge of German.

The analysis of the correlation function of video segments for the three emotion primitives has shown that the correlation time of activation is between 1 and 2 s, of valence is between 2 and 3 s and of dominance is between 4 and 6 s. Thus the minimum segment length was chosen to be 1 s. An example of the evaluation of all evaluators for a typical sequence of segments yields a reasonable result as can be seen from Fig 2.



o: valence, x: activation, *: dominance

Figure 2: Average evaluation x of video segments of a typical sequence

Both, the observation of the correlation function of the evaluation result and the averaged evaluation given in Fig. 2 verify the definition of the segment length by 1 s which has been chosen subjectively to be appropriate.

The next analysis of the data is focused on the question whether or not the data represent the emotion primitives in an appropriate way. This question can be answered by observing the distributions of the emotion primitives which are given in Fig. 3.

As can be seen, the availabe data cover more or less the whole range of emotion primitives in all modes. Whereas activation and dominance are symmetrical with respect to the neutral value, this is not true for valence: the negative values are predominant which is caused by the topic discussed in the TV show which coveres family conflicts. But in general, the data can be used as the ground truth for the design of an automatic classifier.

4 Relevance of the Subjective Evaluation

To analyse the evaluation of the $N = 78$ sentences by each of the $K = 15$ evaluators with respect to the ensemble, the correlation coefficient

$$r_k^{(i)} = \frac{\sum_{n=1}^{N}(x_{k,n}^{(i)} - \bar{x}_k^{(i)})(\bar{x}_n^{(i)} - \bar{x}^{(i)})}{\sqrt{\sum_{n=1}^{N}(x_{k,n}^{(i)} - \bar{x}_k^{(i)})^2 \sum_{n=1}^{N}(\bar{x}_n^{(i)} - \bar{x}^{(i)})^2}} \tag{1}$$

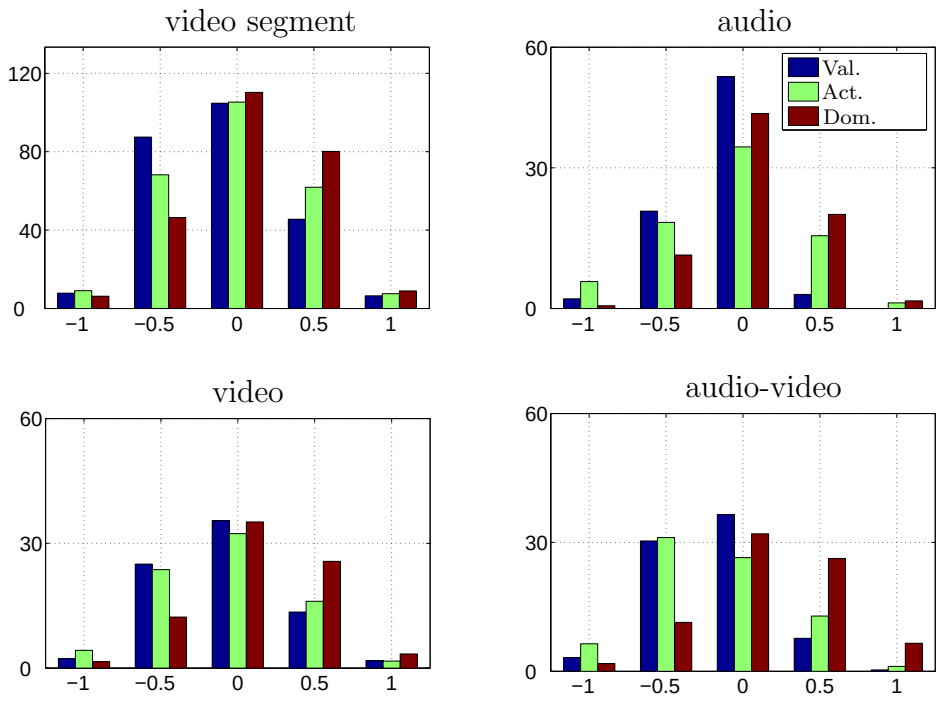


Figure 3: Distribution of the emotion primitives with respect to the modes

is used. The evaluation value of the kth evaluator of the nth sentence is denominated by $x_{k,n}$. The mean $\bar{x}_k^{(i)}$ of the kth evaluator over all $N = 78$ sentences, the mean evaluation $\bar{x}_n^{(i)}$ of all $K = 15$ evaluators for the nth sentence, and the mean $\bar{x}_k^{(i)}$ over all evaluators and sentences is given by

$$\bar{x}_k^{(i)} = \frac{1}{N}\sum_{n=1}^{N} x_{k,n}^{(i)} \qquad \bar{x}_n^{(i)} = \frac{1}{K}\sum_{k=1}^{K} x_{k,n}^{(i)} \qquad \bar{x}^{(i)} = \frac{1}{N}\sum_{n=1}^{N} \bar{x}_n^{(i)}, \tag{2}$$

respectively. As a measure for overall correlation,

$$r^{(i)} = \frac{1}{K}\sum_{k=1}^{K} r_k^{(i)} \tag{3}$$

is used which is given in Tab. 1 together with the standard deviation $\sigma^{(i)}$ for the emotion primitives $i \in \{V, A, D\}$.

Mode	$r^{(V)}$	$\sigma^{(V)}$	$r^{(A)}$	$\sigma^{(A)}$	$r^{(D)}$	$\sigma^{(D)}$
audio	0.38	0.202	0.54	0.153	0.44	0.193
video	0.69	0.1	0.5	0.111	0.44	0.169
video segment	0.62	0.116	0.53	0.096	0.47	0.124
audio-video	0.55	0.159	0.5	0.147	0.51	0.165

Table 1: Correlation coefficient $r^{(i)}$ and its standard deviation $\sigma^{(i)}$

As can be seen, the correlation coefficient for valence is the weakest in the audio mode, followed by the audio-video and being best and close together for the video and the video

segmented mode. This correlates with the standard deviation $\sigma^{(i)}$ of the evaluators: it is largest for the audio mode and smallest for the video mode. For the other emotion primitives the correlation is similar in all modes and the standard deviation is lower. But the correlation is positive, and thus the trend of the evaluation points into the same direction. This shows again that the data can be used as ground truth for the design of an automatic classifier.

5 Distribution of the Subjective Evaluation

Because of the limitation of the evaluation values within $-1 \leq x_{k,n} \leq +1$ and the quantization of the individual values due to the application of the SAMs, some statistical properties can be given a priori with which the output values of the subjective evaluation can be compared [3]. For this purpose, the unbiased standard deviation $\sigma_n^{(i)}$ of each sentence will be calculated:

$$\sigma_n^{(i)} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(x_{k,n}^{(i)} - \bar{x}_n^{(i)})^2} \tag{4}$$

with the mean value $\bar{x}_n^{(i)}$ given by (2). Since the individual evaluation values are quantized with $x_{k,n}^{(i)} \in \{-1,\ -0.5,\ 0,\ +0.5,\ +1\}$, the true evaluation value has a maximum distance of $x_{k,n}^{(i)} - \bar{x}_n^{(i)} = 0.25$ with respect to the given discrete values. From (4) follows for $K = 15$ evaluators:

$$\sigma_{minmax}^{(i)} = \sqrt{\frac{K}{K-1} \cdot 0.25^2}\Bigg|_{K=15} = 0.259. \tag{5}$$

On the other hand, if the the distance ist $x_{k,n}^{(i)} - \bar{x}_n^{(i)} = 1$, i.e. a neutral emotional state is evaluated to be one of the extremes, the standard deviation is in this case

$$\sigma_{extreme}^{(i)} = \sqrt{\frac{K}{K-1} \cdot 1^2}\Bigg|_{K=15} = 1.035. \tag{6}$$

In Fig. 4 the standard deviation $\sigma_n^{(i)}$ given by (4) as a function of the sentence index n for the audio data is plotted. Furthermore, the limits according to (5) and (6) are shown. The mean of the standard deviation is $\sigma^{(V)} = 0.26$, $\sigma^{(A)} = 0.36$, and $\sigma^{(D)} = 0.3$, respectively. There are a few sentences with the standard deviation equal to zero, and no standard deviation exceeds the value 0.5. In the right lower corner the distribution of sentences is shown which display a standard deviation below a given threshold $\sigma_n^{(i)} < \theta$. By this it can be seen that for a given threshold θ most sentences are found for valence followed by activation and dominance.

The representation of the video mode is given in Fig. 5. Here some values are larger than 0.5 which means that some sentences have been evaluated with a deviation of more than one of the quantized values with a spacing of 0.5. Compared to the audio mode, valence shows a similar behaviour whereas dominance and activation are closer together. The results for segmented video are not shown here because there is not much difference compared to unsegemented video. Of course, the reliability in the statistical sense is higher due to the higher number of 252 segments compared to the 78 sentences. But obviously the lower number of 78 sentences is sufficient to extract a reliable statistic.

Finally, the combined audio-video mode is investigated with the results given in Fig. 6. Again, valence is extracted with higher reliability because the evaluation yields a lower standard deviation than dominance and activation.

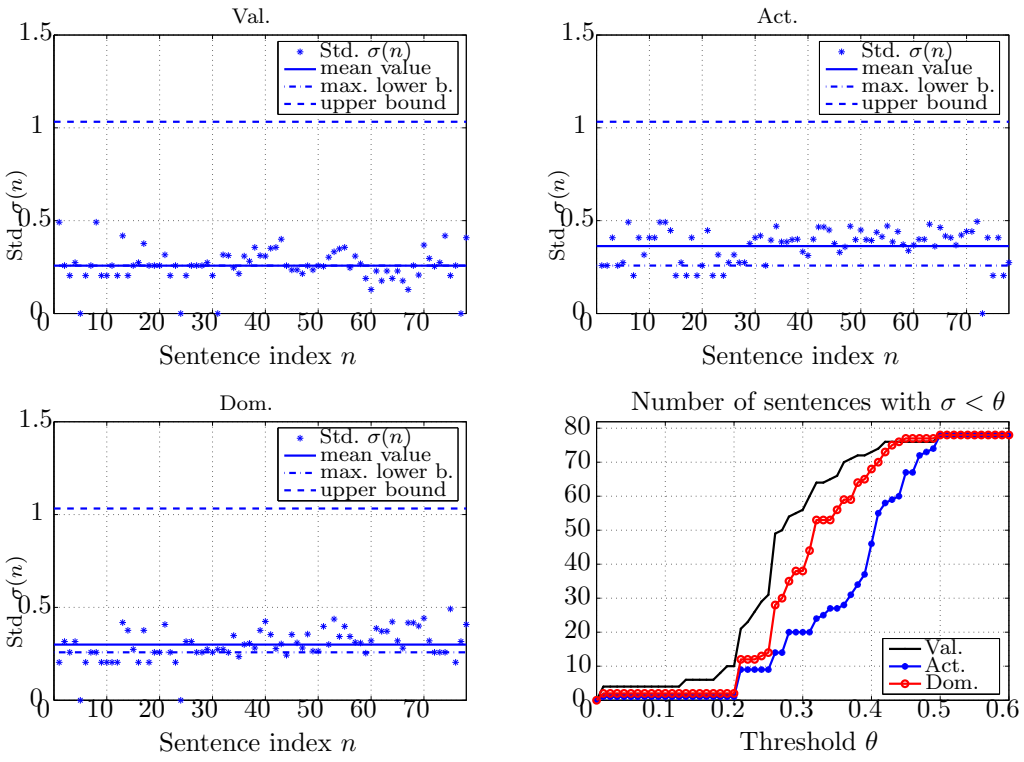


Figure 4: Standard deviation for the audio mode

For comparison, Tab. 2 summarizes the average standard deviation for the emotion primitives. As can be seen, audio seems to be the most reliable mode to extract emotion

mode	$\sigma^{(V)}$	$\sigma^{(A)}$	$\sigma^{(D)}$
audio	0.26	0.36	0.30
video	0.29	0.39	0.37
video segment	0.33	0.37	0.37
audio-video	0.29	0.40	0.39

Table 2: Mean standard deviation of the three emotion primitives

primitives. A reason for this result might be that not only the acoustic parameters of the speech signal are the basis for the evaluation result but also the semantics expressed by the evaluated sentences. Only a blind test with evaluators unable to understand German might support this assumption. On the other hand, the correlation coefficients from Tab. 1 tells another story: here the correlation for audio was lowest. Since the differences are not very high, also the reduced statistical basis might be the reason for this contradiction. Despite this discrepancy, all parameters, i.e. the correlation coefficient and the standard deviation, yield the same result: it is possible to extract emotion primitives from authentic speech by untrained evaluators. This is the basis to generate a data set of audio-video data as ground truth for the design of an automatic classifier for emotion primitives.

6 Correlation of the Modes

The question to be answered by the evaluators was to decide whether or not the impression of emotion primitives expressed in the four modes was contradictive or congruent.

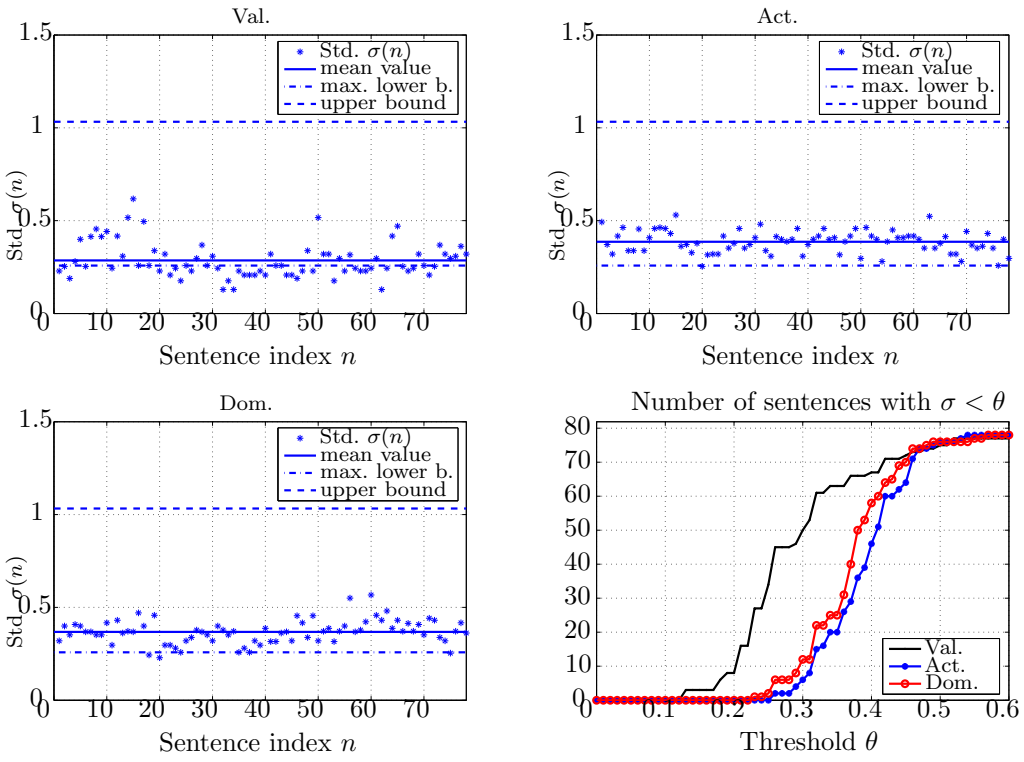


Figure 5: Standard deviation for the video mode

If there was no contradiction, the grade 1 was chosen, in case of total contradiction the grade 5. For detailed analysis, the correlation coefficient for all judgements was calculated according to (1). The result is given in Tab. 3. The mean of all judgements was

	audio V A D	video V A D	video segment V A D	audio-video V A D
audio	1.00 1.00 1.00	0.12 0.53 0.36	0.12 0.44 0.22	0.34 0.58 0.49
video	0.12 0.53 0.36	1.00 1.00 1.00	0.79 0.66 0.80	0.81 0.40 0.72
video segment	0.12 0.44 0.22	0.79 0.66 0.80	1.00 1.00 1.00	0.65 0.34 0.65
audio-video	0.34 0.58 0.49	0.81 0.40 0.72	0.65 0.34 0.65	1.00 1.00 1.00

Table 3: Correlation coefficients between the modes

$\mu = 3.1$, i.e. neither a total contradiction nor a total coincidence. The standard deviation of all judgements was $\sigma = 0.96$, i.e. a significant trend was found. As is expected, video correlates most with video segment and audio-video, and least with audio and vice versa. This again is an argument for the reliability of the data base.

7 Conclusion

The following conclusions can be drawn from the investigations documented in this paper: it is possible to construct an annotated data base which can be used as ground truth for the design of an automatic classifier for emotion primitives. Furthermore, video sequences might be cut into segments of 1 s duration because the value of the emotion primitives might change for observations larger than 1 s. All modes - audio, video and audio-video - yield results for the extraction of emotion primitives which are on almost the same level

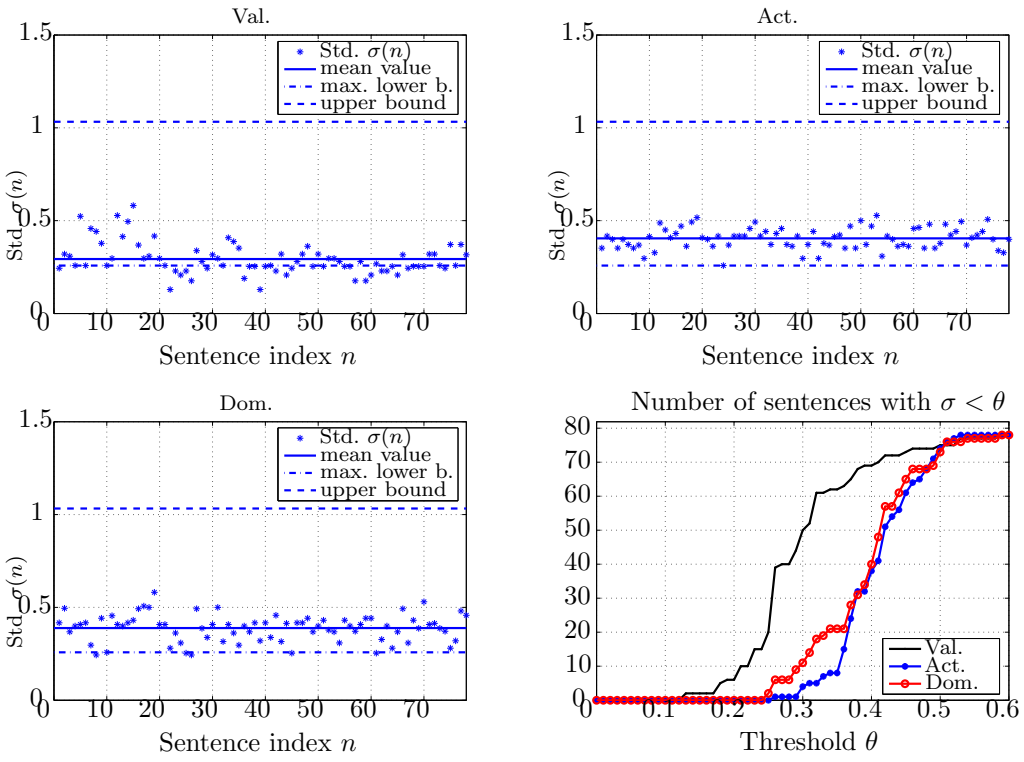


Figure 6: Standard deviation for the audio-video mode

of reliability. The combination of modes - audio and video - does not improve the level of reliability.

Acknowledgement

The work documented in this paper has been partly supported by the DFG (German Science Foundation) within the SFB 588 *Humanoid Robots* for which the authors want to thank.

References

[1] Ekman, P., Friesen, W.: Contstants Across Cultures in the Face and Emotion. Journal of Personality and Social Psychology, vol. 17, no. 2, 1971, pp. 124-129.

[2] Lang, P.: Behavioral Treatment and Bio-behavioral Assessment. In Sidowski, J.B. et al. (eds.) Technology in Mental Health Care Delivery Systems. Ablex Publishing, Norwood, NJ, 1980, pp. 119-137.

[3] Grimm, M., Kroschel, K.: Evaluation of Natural Emotions Using Self Assessment Manikins. Proceedings IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), San Juan, Puerto Rico, 2005, pp. 381-385