Being bored? Recognising natural interest by extensive audiovisual integration for real-life application

Björn Schuller [a,*,1], Ronald Müller [b], Florian Eyben [a,1], Jürgen Gast [a], Benedikt Hörnler [a], Martin Wöllmer [a,1], Gerhard Rigoll [a], Anja Höthker [c], Hitoshi Konosu [d]

[a] *Institute for Human–Machine Communication, Technische Universität München, D-80333 München, Germany*
[b] *Altran Technologies, Bernhard-Wicki-Str. 3, 80636 München, Germany*
[c] *Toyota Motor Europe, Production Engineering – Advanced Technologies, B-1930 Zaventem, Belgium*
[d] *Toyota Motor Corporation, 1 Toyota-cho, Toyota City, Aichi 471-8571, Japan*

1. Introduction

Information on interest or disinterest of users possesses great potential for general Human–Computer Interaction [1,2] and many commercial applications, such as sales and advertisement systems or virtual guides. Similar to the work introduced in [3], we are likewise interested in curiosity detection, e.g. for topic switching, in infotainment systems, or in customer service systems. Apart from that, also interest detection in meetings [4–6], or (children's) tutoring systems [7] has been addressed so far.

Numerous works exist on the recognition of affective or emotional user-states, which are strongly related to interest. Many use solely acoustic speech parameters [8–10], followed by fewer works which use vision-based features (e.g. [11,12]) or linguistic analysis [5,10]. Only a considerably lower number deals with fusion of these input cues (e.g. [6,13]), even though processing of such complementary information is known to be generally advantageous with respect to robustness and reliability [14–17]. So far, this integration of streams has been fulfilled for acoustic and vision cues, exclusively (e.g. [4,17,13]), without fully automatic integration of textual analysis of spoken content, i.e. by using an Automatic Speech Recognition (ASR) system or considering non-linguistic vocalisations. Linguistic analysis up to date has been performed on ground truth data and not on actual transcripts from an ASR engine, which naturally is much more challenging due to inherent errors in the ASR stage. Further, current models for fusion usually are rather simple, as majority voting or logical operations on a late fusion level [11,18] are implemented.

As shown in many works (e.g. [14,15,11,17,16]), audiovisual processing is known to be superior to each single modality. We therefore propose an attempt to combine features from practically all facial and spoken information available: facial expression analysis based on Active Appearance Models (AAM), eye activity modelling, acoustic and comprehensive linguistic analysis including non-linguistic vocalisations, and additional contextual history information.

In this respect it is received wisdom that a fusion of all accessible information on an early feature level is highly beneficial, as it preserves the largest possible information-basis for the final

* Corresponding author. Tel.: +49 89 289 28548; fax: +49 89 289 28535.
E-mail addresses: schuller@tum.de (B. Schuller), mueller@mmer-systems.eu (R. Müller), Anja.Hoethker@toyota-europe.com (A. Höthker), hitoshi_konosu@mail.toyota.co.jp (H. Konosu).

[1] The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 211486 (SEMAINE).

decision process [19]. The main problem thereby is the asynchronity of the feature streams. Frame-by-frame image analysis operates on 25 frames per second, for example, while speech analysis is term-based and linguistic analyses turn-based [20]. However, so far only fusion of acoustic and linguistic information [21,10], and acoustic and vision-based information [16,17] have each been realised on an early integration level.

Further, practically all results reported are based on databases rather than experience within the use of a real-life demonstrator. These data-sets usually employ a number of idealisations: a fixed head position, no occlusions, constant lighting conditions, no background noises, given pre-segmentation, partly known subject-samples, all modalities showing one clearly assignable affective state at a time, basic, mostly discrete emotions, that are deliberately displayed (as opposed to spontaneous) expressions [22]. If one aims at an automatic system that is capable of fully automatically responding to spontaneous interest, these simplifications clearly need to be overcome and more realistic, large scale databases are required. Moreover, recognition rates such as precision, recall, or accuracy can only report the objective performance of affective computing systems, but not how such a system would be accepted by users and whether it will be useful in a real-world scenario. A system with close to 100% accuracy under laboratory conditions (e.g. by relying on prototypical emotions, as often carried out) will still – in most cases – perform unsatisfactory in real-world scenarios. Thus, actual use-case studies [13,23] must be performed to evaluate the performance and the acceptance of such systems in addition to the objective measures like accuracy.

In contrast to most works in the field of affective computing and interest recognition, we therefore attempt fully automatic audiovisual continuous interest recognition on spontaneous data recorded in a real-world scenario by including information from extensive audiovisual sources via early fusion. In an real-life user-study we evaluate how and if the system provides users a benefit.

The article is structured as follows: in Section 2 the featured approach to multimodal interest recognition is described and discussed in detail. Algorithms implemented for each modality are explained individually and are followed by a description of the multimodal fusion approach and an evaluation of recognition performance using individual modalities as well as various combinations of modalities. The setup and the survey results of the real-life application scenario user-study are discussed in Section 3. The article is concluded by a final discussion in Section 4.

2. Evaluating multimodal interest recognition

The details of the fully automatic approach to human interest detection are presented in this section. After a short description of the recording process and the final database of spontaneous interest data in Section 2.1 we describe the features and algorithms relevant for each modality in Section 2.2. The modalities we considered are as follows: facial expressions in Section 2.2.1, eye activity in Section 2.2.2, acoustics in Section 2.2.3, linguistics in Section 2.2.4, and contextual history integration in Section 2.2.6. The automatic transcription of non-linguistic vocalisations and spoken content for linguistic analysis (Section 2.2.4) is outlined in Section 2.2.5. Multimodal information stream integration on an early feature level is described in Section 2.3, followed by detailed results for various combinations of modalities and full multimodal integration in Section 2.4.

2.1. Spontaneous interest data

In order to overcome today's mostly acted audiovisual databases being only of limited help for real-life emotion recognition [22,24], and due to the lack of a large publicly available audiovisual set dealing with interest, we decided to record a database named *AVIC* (Audiovisual Interest Corpus) in the ongoing. It was firstly introduced in [25]. In the scenario setup, an experimenter and a subject are sitting on opposite sides of a desk. The experimenter plays the role of a product presenter and leads the subject through a commercial presentation. The subject's role is to listen to explanations and topic presentations of the experimenter, ask several questions of her/his interest, and actively interact with the experimenter considering his/her interest in the addressed topics. The subject was explicitly asked not to worry about being polite to the experimenter, e.g. by always showing a certain level of "polite" attention. Visual and voice data is recorded by a camera and two microphones, one headset and one far-field microphone, in this situation. The final *AVIC* recordings are stored with the parameters given by Table 1.

Table 1
AVIC database recording parameters.

Image resolution	720 × 576
Frame rate	25 fps progressive
Colour resolution	24 Bit
Encoder	DV
Audio sampling rate	44,100 Hz
Audio quantisation	16 Bit
Left audio channel	Lapel microphone
Right audio channel	Far-field microphone

Twenty-one subjects participated in the recordings, three of them Asian, the remaining European. The language throughout experiments is English, and all subjects are very experienced English speakers. Three age categories ($<$ 30 years, 30–40 years, $>$ 40 years) were defined during specification phase for balancing. More details on the subjects are summarised in Table 2.

Given by the setting of face-to-face communication, the head poses in the database vary in the approximate ranges of $\pm 20^\circ$ in pitch, $\pm 30^\circ$ in yaw, and $\pm 20^\circ$ in roll rotations.

In the following exclusively the data of the subjects, respectively speaker, is used for analysis. To acquire reliable labels of a subject's "Level of Interest" (LOI) as detailed in the ongoing, the entire video material was segmented in speaker- and sub-speaker-turns and subsequently labelled by four male annotators, independently. The annotators are undergraduate students of psychology in the role of naive assessors: the intention was to annotate observed interest in the common sense. A speaker-turn is defined as continuous speech segment produced solely by one speaker. Back channel interjections (mhm, etc.) are ignored hereby. That is, every time there is a speaker change, a new speaker turn begins. This is in accordance with the common understanding of the term "turn-taking". Speaker-turns thus can contain multiple and especially long sentences. In order to provide Level of Interest analysis on a finer time scale, the speaker turns were further segmented at grammatical phrase boundaries: a turn lasting longer than two seconds is split by punctuation and syntactical and grammatical rules, according to [10], until each segment is shorter than two seconds. These resulting segments are the basis for the experiments in the ongoing and are referred to as sub-speaker-turns.

Table 2
Details on subjects contained in the *AVIC* database. Further details in the text.

Group of subjects	#	Mean age (years)	Rec. time (h)
All	21	29.9	10:22:30
Male	11	29.7	5:14:30
Female	10	30.1	5:08:00
Age <30	11	23.4	5:13:10
Age 30–40	7	32.0	3:37:50
Age >40	3	47.7	1:31:30

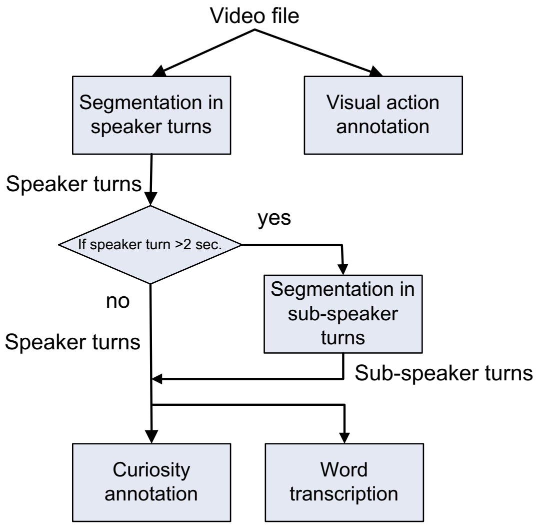

Fig. 1. *AVIC* database annotation workflow.

Table 3
Distribution of non-linguistic vocalisation by type in the *AVIC* database across the five frequently occurring types (only instances longer than 100 ms).

#Breathing	#Consent	#Garbage	#Hesitation	#Laughter	#Sum
452	325	716	1147	261	2901

Table 4
Distribution of sub-speaker turns over LOI-2 through LOI2 and Inter Labeler Agreement (I) with confidence (C). The last coloumn (“?”) shows the number of instances that remain as non-assignable to an LOI at the respective C and I. Likewise, the line total equals 12,839 sub-speaker turns throughout.

#Sub-Speaker Turns	LOI-2	LOI-1	LOI0	LOI1	LOI2	“?”
$I = 50\%, C > 0$	19	383	3602	5386	305	3144
$I = 50\%, C > 0.5$	19	362	3339	5316	305	3498
$I = 50\%, C \geqslant 0.6$	19	261	2832	4603	305	4819
$I = 75\%$	19	185	2226	3741	305	6363
$I = 100\%$	4	19	417	960	25	11,414

Fig. 1 shows the corresponding annotation work flow. The Level of Interest is annotated for every sub-speaker turn. In order to get an impression of a subject’s character and behaviour prior to the actual annotation, the annotators had to watch approximately five minutes of a subject’s video. This helps to find the range of intensity within which the subject expresses her/his curiosity. As the focus of interest based annotation lies on the sub-speaker turn, every such had to be viewed at least once to find out the Level of Interest displayed by the subject.

Five Levels of Interest (LOI) were distinguished *in the first place*:

- LOI-2: *Disinterest* (subject is tired of listening and talking about the topic, is totally passive, and does not follow the discourse).
- LOI-1: *Indifference* (subject is passive, does not give much feedback to the experimenter’s explanations, and asks unmotivated questions, if any).
- LOI0: *Neutrality* (subject follows and participates in the discourse; it can not be recognised, if she/he is interested or indifferent in the topic).
- LOI1: *Interest* (subject wants to discuss the topic, closely follows the explanations, and asks some questions).
- LOI2: *Curiosity* (strong wish of the subject to talk and learn more about the topic).

For automatic processing, a fusion of these Levels of Interest to a “master LOI” was automatically fulfilled. We introduced the following scheme of different cases of Inter Labeler Agreement (ILA) and confidence bounds:

- Same rating by all annotators: ILA 100%;*Master LOI* $:=$*LOI of majority*.
- Same rating by three of four annotators: ILA 75%;*Master LOI* $:=$*LOI of majority*.
- Same rating by two annotators: ILA 50%> If other two annotators agree:*Master LOI* $:=$“?” (undefined)> If other two annotators disagree:*Master LOI* $:=$*median LOI*.In this case an additional confidence measure C is derived from the standard deviation σ of the LOI over all annotators: $C = 1 - 0.5 \cdot \sigma$.

Additionally, the spoken content and non-linguistic vocalisations have been labelled. These vocalisations are *breathing*, *consent*, *coughing*, *hesitation*, *laughter*, *long pause*, *short pause*, and *other human noise* (referred to as *garbage* in the ongoing). This additional labelling effort shall demonstrate the potential of such events within higher semantic analysis. There is a total of 18,581 spoken words, and 23,084 word-like units including non-linguistic vocalisations (19.5%). The latter are distributed as shown in Table 3. Note that only non-linguistic vocalisations with a length greater than 100 ms are considered for automatic detection via our HMM framework and thus only those non-linguistic vocalisations are shown in Table 3. All instances of *coughing* are smaller than 100 ms and thus not considered for automatic recognition.

Summarised, overall annotation contains sub-speaker- and speaker-turn segments in millisecond resolution, spoken content, non-linguistic vocalisations, individual annotator tracks, and Master LOI with confidence in XML-format created with ANVIL [26]. Table 4 shows the amount of sub-speaker turns per master LOI depending on the chosen Inter Labeler Agreement and the bound of confidence C. An LOI of “?” indicates the “undefined class”, i.e. no LOI could be assigned to these samples with the desired confidence. The database comprises 12,839 sub-speaker turns. Overall, a very low kappa-value (κ) of $\kappa = 0.09$ and standard deviation for the Level of Interest of the labelers of $\sigma = 0.54$ is observed for the database at this point.

The Inter Labeler Agreement is therefore pruned of undefined sub-speaker turns (those labelled with “?”) and sub-speaker turns of LOI0 with a confidence $C < 1.0$. Through this reduction of sub-speaker turns, the agreement of the four annotators increases to a substantial kappa-value of $\kappa = 0.62$ with $\sigma = 0.23$. Moreover, the distribution of the instances over the Levels of Interest is more balanced. As too few items for LOI-2 and LOI-1 are present, these were clustered together with LOI0, so that the Level of Interest scale reaches from 0 to 2 in the ongoing for discrete classes[2]. Thereby final values of $\kappa = 0.66$ with $\sigma = 0.20$ are observed. Detailed inter-annotator kappa-values were computed according to [27] and are given in Table 5. The exact LOI-distribution of the single labelers and Inter Labeler Agreement in this reduced set of sections is depicted in Fig. 2. Example video frames for LOI0–LOI2 after clustering and Inter Labeler Agreement based reduction are depicted in Fig. 4.

In order to increase the amount of these low occurrence Levels of Interest (LOI1 and LOI2), further methods of master LOI derivation from the annotator specific LOI need to be investigated, if promising for training and evaluation purposes.

[2] Note that LOI2 has even fewer numbers than LOI-1 in Table 4, first line. Alternatively one could consider clustering LOI2 to LOI1 instead. However, from an application point of view we were rather interested in preserving high interest as separate class. In manifold applications even a two-class decision may however be sufficient while expectantly being more robust at the same time.

Table 5
Kappa-values for the Inter Labeler Agreement.

κ	Labeler 1	Labeler 2	Labeler 3	Labeler 4	ILA
Labeler 1	1.00	0.86	0.62	0.61	0.89
Labeler 2	0.86	1.00	0.72	0.71	0.97
Labeler 3	0.62	0.72	1.00	0.44	0.75
Labeler 4	0.61	0.71	0.44	1.00	0.74
ILA	0.89	0.97	0.75	0.74	1.00

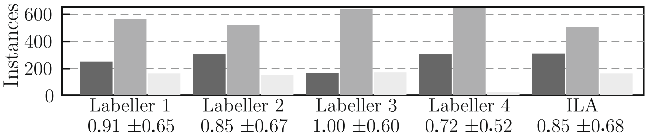

Fig. 2. Distribution of the Level of Interest (LOI) over various labelers; LOI-distribution Inter Labeler Agreement (ILA): 316 510 170, min/mean/max Inter Labeler Agreement confidence: 0.75/0.89/1.0, $\kappa = 0.66$.

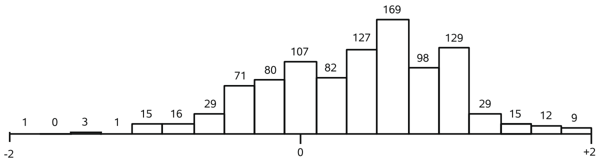

Fig. 3. Mean Level of Interest (LOI) histogram for the full LOI scale [−2,2] as used for regression experiments.

One generally alternative approach is a shift to a continuous scale obtained by averaging the single annotator LOI. The histogram for this mean LOI is depicted in Fig. 3. Note that here the original scale reaching from LOI-2 to LOI2 is naturally preserved. Apart from higher precision, this representation form allows for subtraction of a subject's long-term interest profile [28]. Note that the Level of Interest introduced herein is highly correlated to arousal. However, at the same time there is an obvious strong correlation to valence, e.g. boredom has a negative valence, while strong interest is characterised by positive valence. The annotators however labelled interest in the common sense, thus comprising both aspects.

Overall, the *AVIC* database is a multimodal data collection of unseen size, quality, realness, and focus, providing non-acted multimodal data for affective computing and especially curiosity detection in human dialogues.

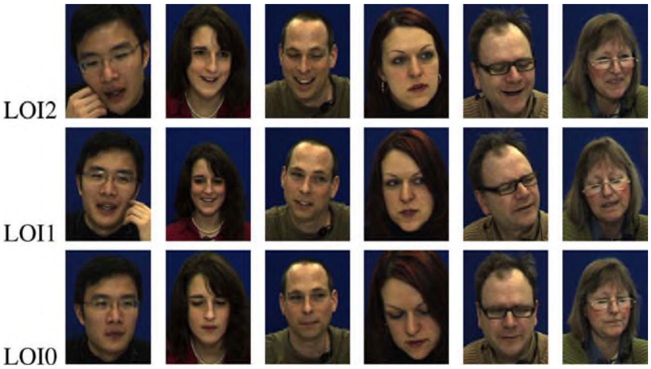

Fig. 4. Example video frames (for better illustration limited to the facial region here) for Level of Interest 0–2 taken from the *AVIC* database. Two subjects in gender balance were chosen from each of the three age groups.

2.2. Modalities and algorithms

In the ongoing we will introduce the four information streams considered for audiovisual fusion and – where appropriate – present individual performance details. Apart from these streams – namely acoustic and linguistic feature information stemming from the audio channel and facial expression and eye movement activity stemming from the video channel – contextual knowledge on the interest level development will be introduced. Finally, we shortly discuss fusion of these knowledge sources.

2.2.1. Facial expression

Apart from being a predominant means in human communication, the human face is also known to portray facial expression related to affective user states [29,30]. There is a considerable number of approaches towards recognition of such states [31], reaching from computationally less demanding low-level features such as global motions [32] over extraction of MPEG7 facial feature points [33] to high-level action units, including such in a 3D space (cf. [34,35,31] for overviews). In particular related to our scenario are works on interest recognition [36] and yawning [37,33]. Herein, we decided for Active Appearance Models (AAM), which are known to be well suited for this task [12]. AAM can be understood as a means of parametrising a face by its shape and texture with respect to a statistical face model. The derived parameters explain, among others, the facial expression of the monitored face. However, AAM serve only for feature provision. The final expression-related analysis is fulfilled in the subsequent classification step (cf. Section 2.4).

Active Appearance Models are statistical models derived from example images of an object class [38], i.e. faces in our case. For a detailed overview on application and variants the reader is referred to [39,40]. In the ongoing we will only give a short introduction: Active Appearance Models assume that the appearance of a face can be described by its two-dimensional shape and its texture within the hull of the shape. Thereby, the shape is defined as the relative position of a set of landmarks, disregarding Euclidean transformations and scaling on the entire shape. The statistical analysis of the shape variations, texture variations and their combination is usually performed by the Principal Component Analysis (PCA). This allows for a compact representation of the obtained variance by a very small set ($\ll 100$) of main components. Now, the appearances of the training objects as well as a great variety of unseen object instances can be synthesised by a linear combination of the main components.

In the application phase of an Active Appearance Model, the coefficients of the linear combination have to be optimised with respect to a maximal similarity between the original object and the artificial object appearance, synthesised by the Active Appearance Models. These optimised coefficients constitute a precise representation of the analysed face and can therefore be considered as features for a statistical classification in facial expressions, head poses, gender, age, etc. Our face analysis system is capable of such pattern recognition tasks due to multiple evaluations of the influence of algorithmic parameters and their optimisation. Exemplary, with Support-Vector Machine based statistical classification, the gender recognition problem is solved with 94.6%, four facial expression classes can be distinguished with 91.8%, and five horizontal head poses ($\pm 30^\circ, \pm 15^\circ, 0^\circ$) can be recognised at 89.8%. Each of the tasks was evaluated on large and public standard databases, namely AR Face Database[3], FG-NET Aging Database[4], and NIFace1[5], with disjunctive data-sets for training and evaluation. Thus, this face

[3] http://cobweb.ecn.purdue.edu/aleix/aleix_face_DB.html

[4] http://www-prima.inrialpes.fr/FGnet/

[5] http://tu-ilmenau.de/fakia/NIFace1.5255.0.html

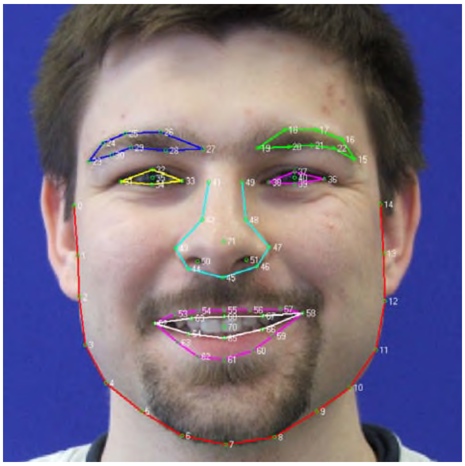

Fig. 5. 2D annotation of a face with 72 landmarks.

analysis system is applied to provide valuable information on a subjects face for the recognition of the interest level.

The statistical analysis via Principal Component Analysis requires a set of *shapes* and corresponding *textures* to build a *shape model*, a *texture model* and finally a *combined model*. First, the training images $\mathbf{p_i} \in \mathscr{P}$ with $0 \leqslant i < p$ have to be manually annotated, producing a set of p corresponding landmark vectors $\mathbf{s_i} \in \mathscr{S}$ with $\mathbf{s_i}$ being the ith landmark vector defined as the concatenation of all landmark coordinates

$$\mathbf{s_i} = (x_0, y_0, x_1, y_1, \ldots, x_{(n/2)-1}, y_{(n/2)-1})^T. \tag{1}$$

Fig. 5 shows an example annotation with $n/2 = 72$ landmarks. These shape vectors are arranged column-wise in the *shape matrix* $\mathbf{S}$.

Additionally, the *mean shape* $\bar{\mathbf{s}}$ is defined as the mean of all shape vectors in $\mathbf{S}$.

The texture within the annotated shape of each training image is warped to fit the mean shape $\bar{\mathbf{s}}$. For generation of the texture model, we store the obtained set of textures $\mathbf{t_i} \in \mathscr{T}$ as vectors column-wisely in the *texture matrix* $\mathbf{T}$

Further let $\bar{\mathbf{t}}$ be defined as the mean of all textures in $\mathbf{T}$.

The first step of building an Active Appearance Model is the independent application of a Principal Component Analysis to the aligned and normalised shapes in $\mathbf{S}$ and the shape-free textures in $\mathbf{T}$, thus generating a shape and a texture model. Finally these two models are combined to one Active Appearance Model which comprehends the correlated shape and texture variations contained in the training images [41].

The *shape model* is built by applying a Principal Component Analysis to the shape matrix $\mathbf{S}$, i.e. an Eigenvalue Decomposition of the Covariance Matrix over all shapes $\mathbf{s_i}$. The obtained Eigenvectors constitute the *shape basis* $\mathbf{W_s}$, whereas basis vectors are sorted in descending order of the corresponding Eigenvalue $\lambda_{\mathbf{s}i}$. Information reduction is achieved by only selecting the top r_s "most important" basis vectors, discarding those which correspond to principal axes bearing low variance of the data (cf. Fig. 6 for an illustration of the effects of the first two shape model components in our case). Evaluations showed throughout that the remaining basis vectors should explain 98% of the total shape variance. Since the size of the Eigenvalue $\lambda_{\mathbf{s}i}$ indicates the variance explained by the ith Eigenvector, r_s can easily be determined by

$$\frac{\sum_{i=0}^{r_s-1} \lambda_{\mathbf{s}i}}{\sum_{i=0}^{n-1} \lambda_{\mathbf{s}i}} \overset{!}{\geqslant} 0.98. \tag{2}$$

The same method is applied for the texture and combined model. A new shape $\mathbf{s}$ can be synthesised by the linear combination

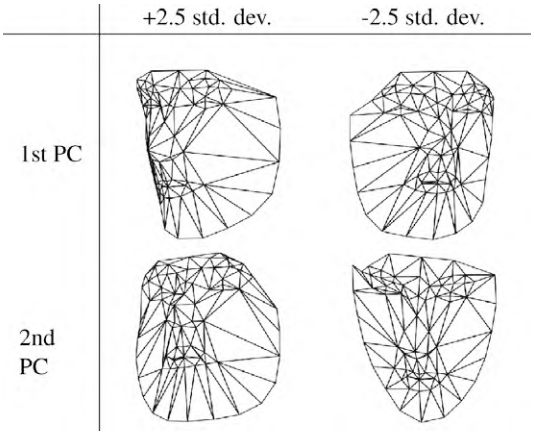

Fig. 6. Effect of the first shape model components.

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{W_s h_s} \tag{3}$$

whereas $\mathbf{h_s}$ contains the *shape coefficients* that control the deformation of the shape model. Note that a zero coefficient vector relates to the mean shape $\bar{\mathbf{s}}$. As $\mathbf{W_s}$ defines an orthonormal basis, the new representation of the known shape $\mathbf{s_i}$ in the new basis can be obtained by

$$\mathbf{h}_{\mathbf{s}i} \approx \mathbf{W_s^T}(\mathbf{s_i} - \bar{\mathbf{s}}). \tag{4}$$

The generation of the *texture model* follows exactly the principle of the shape model generation.

To generate the combined Active Appearance Model, shape and texture correlations are recovered from the so far independent shape and texture models. Let $\mathbf{c_i}$ be the ith vector which contains the concatenated shape and texture coefficient vectors $\mathbf{h_{si}}$ and $\mathbf{h_{ti}}$ for each of the $0 \leqslant i < p$ training samples

$$\mathbf{c_i} = \begin{pmatrix} \mathbf{Eh_{si}} \\ \mathbf{h_{ti}} \end{pmatrix} \tag{5}$$

$\mathbf{E}$ is a diagonal matrix of reasonable weights to equalise the different co-domains of the variance in the shape and the texture model. The vectors $\mathbf{c_i}$ column-wise form the matrix $\mathbf{C}$. Another Principal Component Analysis is applied to the matrix $\mathbf{C}$ producing the *combined basis* $\mathbf{W_c}$, whereas basis vectors are sorted in descending order of their corresponding Eigenvalue $\lambda_{\mathbf{c}i}$, again discarding the "least important" basis vectors. A coefficient vector $\mathbf{c}$ can be synthesised by evaluating

$$\mathbf{c} = \mathbf{W_c h_c} \tag{6}$$

where $\mathbf{h_c}$ contains the *AAM coefficients*. As the matrix $\mathbf{W_c}$ can be split into the shape and texture relevant parts $\mathbf{W_{cs}}$ and $\mathbf{W_{ct}}$

$$\mathbf{W_c} = \begin{bmatrix} \mathbf{W_{cs}} \\ \mathbf{W_{ct}} \end{bmatrix} \tag{7}$$

it is possible to express a new shape $\mathbf{s}$ and texture $\mathbf{t}$ directly as function of $\mathbf{h_c}$ which finally leads to these synthesis rules for a shape and a corresponding texture:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{Q_s h_c}, \quad \mathbf{Q_s} = \mathbf{W_s E^{-1} W_{cs}} \tag{8}$$
$$\mathbf{t} = \bar{\mathbf{t}} + \mathbf{Q_t h_c}, \quad \mathbf{Q_t} = \mathbf{W_t W_{ct}} \tag{9}$$

The Active Appearance Model coefficient optimisation can roughly be understood as a standard multi-variate optimisation problem with the goal to minimise the energy of the difference image $\mathbf{r}(v)$ between this synthesised face and the currently analysed face. This constitutes the error measure with respect to the AAM coefficient vector $\mathbf{v}$ comprising $\mathbf{h_c}$ and the coefficients for translation, rotation, and scale for the shape plus brightness and intensity for the texture.

Due to the high complexity of the face synthesis, a run-time optimised Gauss–Newton gradient descent method by an off-line gradient prediction is applied [38]. Therefore the following steps have to be conducted: definition of an error energy function $E(\mathbf{r}(v))$, estimation of the Jacobian $\mathbf{J} = \frac{\partial \mathbf{r}}{\partial v}$ of the difference function $\mathbf{r}(v)$, as well as calculation of the predictor matrix $\mathbf{R} = (\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T$ used during the coefficient search. The update of the coefficient vector in iteration i follows

$$v(i+1) = v^{(i)} - \alpha \mathbf{R}\mathbf{r}(v^{(i)}) \tag{10}$$

using the step width α. The algorithm terminates when $E(\mathbf{r}(v))$ does not further decrease between iterations. The final value of the error energy serves as confidence measure for the performed Active Appearance Model analysis.

In order to map the Active Appearance Model results to the sub-speaker-turn basis, only the coefficient vector of the video frame with the lowest final $E(\mathbf{r}(v))$ is added to the feature space of the early fusion with the other modalities.

2.2.2. Eye activity

Apart from facial expressions, which are addressed in terms of features by the Active Appearance Model analysis, the level of eye activity is considered, herein, as a criterion for the description of the mental state of a person. Eye activity is widely researched as an important element in the human vision system, e.g. [42]. Psychological studies confirm the existence of a correlation between mental states, such as workload, and eye activity, e.g. in [43]. In the scenario of the *AVIC* database the activity is estimated by a compact description of the body-, and especially the head-movements of the subject over a short video sequence. Note that likewise – in a strict sense – we measure the activity of the head derived from eye position features. Since skin-colour or Viola–Jones [44] based head localisation provides rather rough information about the position and the size of a person's head, we utilise the optimised performance of our eye localisation algorithms from [45]. The localisation of a face and its eyes serves as necessary initialisation of the analysis based on Active Appearance Models as described in the previous section. Hence, we apply an algorithm following the approach presented by Viola and Jones [44] for the face localisation. This localisation algorithm is based on sampling of an image with windows of variable size. From each sample window visual Haar-like wavelet features are extracted. Thereby a Decision Stump as weak classifier operates on single features. These weak classifiers are combined by a Gentle AdaBoost [46], which tries to reject windows without a face at early stages of a cascade. The localisation of the eyes runs on a narrowed area within the face region provided by the face localisation.

We developed several improvements to the standard Viola–Jones algorithm for a more accurate eye localisation. It turned out that the addition of Gabor Wavelet features and the replacement of the Decision Stump as weak classifier by an adaptive interval classifier leads to a localisation at higher efficiency and smaller spatial deviation. Finally, according to our evaluations on the FERET database [47], 98.5% of the eyes in pictures with proximate frontal human faces can be localised with less than five pixel euclidean deviation from the actual eye centre when the face is scaled to a size of 90 × 120 pixels based on the automatic face localisation. The software implementation runs more than five times real-time on images of double VGA resolution. Therefore, the developed improvements provide a reliable basis for our Active Appearance Model analysis as well as for the tracking of the eyes in order to measure their activity.

The derivation of the eye positions, i.e. the speed and direction of the movement of the eyes, and of the eye distance, i.e. change in length and angle of the connecting line between the eyes, are our basic features to describe the person's motion activity. The first measure of the overall motion activity is the mean value. However, homogeneous motion is perceived as less active than heterogeneous motion, although both could lead to the same average value of the derivatives of eye positions and eye distance over a video clip. Therefore, the variance of the motion values should carry important information for activity estimation. On the same account the maxima of each of the motion vector magnitudes are also part of the activity vector. Table 6 lists all examined measures of activity.

Table 6
Features for the estimation of eye activity.

Index	Description
0–2	Eye position δ (maximum, maximum x, maximum y)
4–6	Eye position δ (mean, mean x, mean y)
8–10	Eye position δ (variance, variance x, variance y)
3,7,11	Eye distance δ (maximum, mean, variance)
12	Eye position δ (relative #frames > threshold)

Since head and eye position data is derived from the preceding automatic localisation tasks and is thus not always reliable, a set of conditions must be met for the data to make it into the activity feature vector:

- If the confidence (contained in the AdaBoost meta-data for each region of interest (ROI) type) is less or equal to zero for an eye position, the respective eye data are marked as invalid. This eliminates samples where the eye location could not be determined.
- To avoid wrong tracking results, the change in eye position between two successive frames may not exceed a certain threshold. If the threshold is exceeded, the respective eye position is marked as invalid.

The Head- and Eye-Localisation Module outputs have shown to be noisy quite often. Thus, the eye positions are additionally smoothed over the last three time steps. This of course requires the last three coordinates for the respective eye to be valid. To finally receive a valid derivative of the eye position, two successive smoothed positions of an eye must exist. For the derivative of the eye distance, two successive smoothed values must exist for both eyes.

For the evaluation of the calculated measures of activity, it is mandatory to compare the different image sequences with each other. However, this may not be possible in all cases. For example, different dimensions of the head in the image (originating from different video resolutions) should not influence the resulting measures of activity. Thus, all values are calculated in relation to the dimensions of the head region of interest provided by the head localiser.

The activity vector should give a quantitative statement for the head-motion in closeup views. In the next step the activity vector is used to recognise the Level of Interest (LOI) as it is supposed, that a strong correlation between these two values exists.

2.2.3. Acoustics

There are rather sparse works on recognition of interest from speech in particular. However, as with the vision processing, this can be seen as highly related to the recognition of emotion or affective user states in general. The latter usually relies on prosodic [48], voice quality, and articulatory feature information. Today's systems almost exclusively derive one static feature vector per spoken unit – mostly turns – by application of statistical functionals as linear moments, extremes, ranges, or percentiles to typical acoustic Low-Level Descriptors (LLD) as pitch, energy, duration, or spectral (partly with prior perceptive modelling [49,10]). We adopted these methods for recognition of interest, as successfully shown in [50].

Table 7
Low-Level Descriptors used throughout systematic construction of a 1.4k acoustic feature space.

37 Low-Level Descriptors
Formant 1–5: Amplitude, Bandwidth and Position
Pitch (F0), Frame Energy, Envelope
Mel-Frequency Cepstral Coefficients (MFCC) 1–16
Harmonics-to-Noise Ratio (HNR)
Jitter, Shimmer

Table 8
Functionals applied to Low-Level Descriptor contours used for systematic construction of a 1406 dimensional acoustic feature space.

19 Functionals	
Mean, Centroid, Standard Deviance	Quartiles 1,2,3
Skewness, Kurtosis	Quartile 1 – Minimum
Zero-Crossing-Rate	Quartile 2 – Quartile 1
Maximum Value, Minimum Value, Range	Quartile 3 – Quartile 2
Relative Maximum/Minimum Position	Maximum – Quartile 3
Position of 95% Roll-Off-Point	

With respect to the quasi-stationary nature of a speech signal, firstly a pre-processing by windowing the signal with a Hamming-window function is performed. The signal of interest (the audio signal from a complete sub-speaker turn) is split into overlapping frames having a length of 20 ms. The frames are sampled successively at a hop size of 10 ms. Each frame is multiplied by the Hamming-window function. In order to obtain a better representation in view of Level of Interest content, low-level feature contours containing information about intonation, intensity, harmonic structure, formants, and spectral development and shape are extracted. Secondly, delta and acceleration regression coefficients computed from these Low-Level Descriptor (LLD) contours are included as further features.

Using only LLD as features, a classification by means of dynamic modelling (e.g. Hidden Markov Models) is already feasible. Yet, basing on our past experience [8] and in accordance with the common practice in the field [14,15], in a third stage, statistical functionals f are applied to the Low-Level Descriptor contours in order to project the multivariate time-series F on a static feature vector $\mathbb{R}^1$ and thereby become less dependent of the spoken phonetic content Eq. (11):

$$f : F \rightarrow \mathbb{R}^1 \tag{11}$$

A systematic feature generation by calculation of moments, extreme values, and further shape characteristics (see Table 8) of the Low-Level Descriptor contours leads to > 1k features. The length of the analysed time series thereby corresponds to the length of a sub-speaker turn, i.e. Low-Level Descriptor contours are extracted for a complete sub-speaker turn and functionals are applied to these contours, resulting in one feature vector per sub-speaker turn. The idea thereby is not to extract all these features for the actual Level of Interest detection, but to form a broad basis for self-learning feature-space optimisation.

The used basis of 37 typical acoustic Low-Level Descriptors well known to carry information about paralinguistic effects is shown in Table 7. Energy resembles simple log frame energy. Pitch (F0) and Harmonics-to-Noise Ratio (HNR) calculation base on the time-signal Autocorrelation Function (ACF) with window correction. Formants base on 18-point Linear Predictive Coding (LPC) with root-solving and a pre-emphasis factor $\alpha = 0.7$. Mel-Frequency Cepstral Coefficient (MFCC) computation is based on a 26 channel Mel-Scale filter bank. F0 and formant trajectories are globally optimised with respect to the whole sub-speaker turn by the use of Dynamic Programming. The Low-Level Descriptors are smoothed by techniques such as semi-tone-interval filtering or simple moving average low-pass-filtering to overcome noise. Delta coefficients are appended to the 37 descriptors resulting in 2×37 Low-Level Descriptor contours.

In Table 8, a total of 19 statistical functionals chosen is named. The obtained multivariate time series of variable length is thereby projected on a single 1406 dimensional feature vector as used in [51,52]. Here again, we decided for a typical selection of common functionals covering the first four statistical moments, quartiles, extremes, ranges, positions, and zero-crossings. Finally, we use speaker standardisation by mean and standard deviation by a few turns to better cope with speaker independence.

2.2.4. Linguistics

As opposed to "how" something is said as introduced in the last section, "what" was said – i.e. linguistic information – is rather sparsely exploited in the search for affective cues, though it was proved highly beneficial. Approaches to this task vary strongly, reaching from rule-based key-word spotting [53] to more elaborate statistical approaches as (class-based) N-Grams [54] and vector space modelling [50,10]. The spoken content may also in particular carry cues with respect to a subject's interest [50].

The precondition to linguistic analysis is to obtain the spoken content out of an audio-file. Yet, almost all results for emotion or interest recognition based on linguistic analysis reported rely on manual transcription of spoken content rather than on incorporation of an Automatic Speech Recognition (ASR) unit [10]. This comes, as ASR of emotional speech itself is a challenge. In this work, next to linguistic analysis results obtained with ground-truth annotations, we present first results of a fully automatic linguistic analysis based on an ASR engine, which is capable of transcribing non-linguistic vocalisations along with the recognised word chain. This ASR engine is described briefly in Section 2.2.5. The transcription of non-linguistic vocalisations is necessary because they are important linguistic features for interest detection, e.g. sighs and yawns carry information on boredom [55]. This importance is also confirmed by Table 9. Moreover, considering non-linguistic vocalisations is important for correct recognition of spontaneous speech since they are an essential part of natural speech and also carry meaningful information [56–58].

For linguistic analysis a vector-space-representation popular in the field of document retrieval and known as Bag-of-Words (BOW) was chosen [59]. The motivation therefor is the effective fusibility of obtained linguistic features within the combined audiovisual and contextual feature space on an early level. A term w_i within

Table 9
Top 18 lexemes by Information Gain Ratio (IGR) ranking after stemming (transcription based). Stems are marked by *.

Rank	Stem	IGR
1	*coughing*	0.2995
2	*laughter*	0.1942
3	yeah	0.0514
4	oh	0.0474
5	*ver	0.0358
6	if	0.0358
7	*th	0.0337
8	*consent*	0.0325
9	*hesitation*	0.0323
10	a	0.0308
11	that	0.0305
12	car	0.0275
13	*hav	0.0263
14	is	0.0258
15	I	0.0252
16	*s	0.0230
17	and	0.0219
18	it	0.0219

a sub-speaker turn $\Sigma = \{w_1, \cdots, w_i, \cdots, w_S\}$ with $S = |\Sigma|$ is thereby projected onto a numeric attribute $x_i : w_i \rightarrow \mathbb{R}^1$. The precondition is to establish a vocabulary $\Theta = \{w_1, \cdots, w_j, \cdots, w_V\}$, with $V = |\Theta|$, of terms of interest. In a first approach these are all different terms contained in the annotation of the data-set of interest. Each word in the vocabulary serves as a feature. Throughout feature extraction a value for each term in Θ is calculated. This value reflects the frequency of occurrence in the sub-speaker turn. Usual representations for this value are binary and logarithmic frequency measures.

There exist a number of further refinement approaches such as normalisation to the sub-speaker-turn length, the inverse frequency of occurrence in the data-set known as Inverse Document Frequency (IDF), or logarithmic transform (log) to compensate linearity. Thereby an offset-constant $c = 0.5$ is chosen, as many zero-occurrence cases will be observed. Our final per-term feature is calculated as follows and proved superior throughout evaluation (cf. [50]) to the named alternatives:

$$x_{logTF,i} = \log\left(c + \frac{TF(w_i, \Sigma)}{|\Sigma|}\right) \tag{12}$$

A drawback of this modelling technique is the lack of word order consideration. However, use of bags of Back-Off N-Grams by tokenisation did not result in any further gain.

In general, vocabularies will show a too high dimensionality ($> 1k$ terms) and contain much redundancy in view of the sighted LOI detection. Similar to acoustic feature reduction as described in Section 2.3, two standard techniques in linguistic analysis are therefore employed to reduce complexity: stopping and stemming. The first method directly reduces the vocabulary by eliminating terms of low relevance. This is realised based on Shannon's information as described in the ongoing. Stemming on the other hand clusters morphological variants of terms belonging to the same lexeme, i.e. having the same stem. Thereby the hit-rate of such clusters is directly boosted while reducing complexity at the same time. However, danger of over-stemming exists, i.e. clustering of terms that possess different meanings in view of LOI. Therefore, we decided for an Iterated Lovins Stemmer (ILS), which bases on context-sensitive longest match stemming – a slight enhancement of the very traditional approach to stemming.

Table 9 shows the 18 most relevant lexemes after ILS stemming and ranking based on Information Gain Ratio (IGR) by Shannon entropy ratio[6]. The final vocabulary size thereby is 639 lexemes instead of 1485 terms by rejection of all zero Information Gain Ratio terms. Note that the non-linguistic vocalisation *coughing* could not be detected automatically (cf. Section 2.2.5) despite its high relevance for twofold reasons: its occurrences are mostly shorter than 100 ms, which violates our HMM topology, and too few instances are contained for reliable training – IGR does not take overall occurrence into account but measures predictive ability of e.g. *coughing* when it appears. The table also shows the high ranking of four non-linguistic vocalisations (in italics, as described in Section 2.1) on the ranks 1, 2, 8, and 9. Within linguistic experiments test-runs employing actual large vocabulary continuous speech recognition (LVCSR) and annotation-based runs have been conducted. Firstly, Table 10 provides minimum term frequencies within the set and clearly speaks for problems arising when using a real LVCSR engine: more terms of single occurrence are observed than actually contained in the vocabulary when using real LVCSR. This comes, as words are partly mis-recognised and matched on diverse other terms in the engine vocabulary. Otherwise, this diffusion by word errors also leads to fewer observations of the same terms: already at a minimum term frequency of two within the database the annotation based level overtakes. Yet, Bag-Of-Words relies on high term frequency within a data-set. This can partly be repaired by stemming (5.8% absolute gain in accuracy in the setting as shown in Table 12) – assuming that phonetic mismatches lead to confusions within a lexeme. For more information on linguistic processing the reader is referred to [25,50].

Table 10
Term numbers at diverse minimum term frequency levels, annotation-based (left) and LVCSR-based (right).

Min. TF	Annotation #Terms	LVCSR #Terms
1	1485	1568
2	645	351
5	277	109
10	149	51
20	98	20
50	48	8

2.2.5. Non-linguistic vocalisations and speech

At this point we will briefly describe the Large Vocabulary Continuous Speech Recognition (LVCSR) framework used within this work. The main focus will be on the detection of non-linguistic vocalisations, which have been ranked as highly important linguistic features (Section 2.2.4) for interest detection. The detection of non-linguistic vocalisations discussed in [60] has now been integrated into the speech recognition framework. Only the best performing configurations are presented here, since an in depth discussion of speech recognition is beyond the scope of this article.

The LVCSR framework is built using Hidden Markov Models (HMM) [61]. The complete decoding process is organised in a two-step process (Fig. 7): in the first step a discrimination between linguistic and non-linguistic sounds is performed and the latter are spotted and classified. In the second step, the speech in the segments detected as verbal sounds is transcribed.

For automatically dividing the input speech into linguistic and non-linguistic segments in the first step, a modified word-based speech recogniser using HMM is explored. Three word class

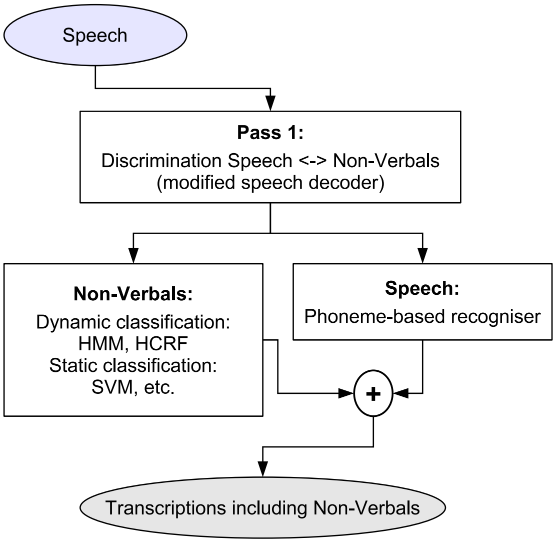

Fig. 7. Framework of proposed two-step method for Large Vocabulary Continuous Speech Recognition and detection of non-linguistic vocalisations.

[6] This ranking was carried out on the transcriptions of the whole *AVIC* database to take all data into account. Feature selections based on the whole corpus are exclusively used in the running system for experiments with new subjects as in the user study presented in Section 3. However, in subject independent performance evaluations in Section 2.4, features have been selected independent of the subject, each, thus resulting in 21 slightly different rankings.

models for speech and four models for the four most important non-linguistic vocalisations plus two silence models are defined. To effectively model words with different durations, the three models have different complexity depending on the word length. The number of letters in the word provides us a rough but sufficient estimate for the acoustical length of the word. Respectively, the three models have $N = 8, N = 13$ and $N = 18$ emitting states for words with less than four, between four and seven and more than seven letters. The model for longer pauses has $N = 3$ emitting states, the model for short pauses has one state. The models for non-linguistic vocalisations have $N = 9$ emitting states with $M = 8$ Gaussian mixture components per state. All HMM have a linear topology. In [60] this configuration among 120 configurations in total has proven to give best results for classification of non-linguistic vocalisations. As acoustic features 12 Perceptual Linear Predictive Cepstral Coefficients (PLPCC) [62] and logarithmic energy extracted from frames of 25 ms length sampled at a rate of 10 ms are used. First and second order regression (δ and $\delta\delta$) coefficients are appended, resulting in a 39 dimensional feature vector. Other approaches, such as those introduced in [63–65], deal only with the detection of one type of non-verbal sound and are more difficult to integrate into a state-of-the-art LVCSR system. Another work by Schultz [66] also deals with non-linguistic vocalisations, however, only to improve ASR performance and not to identify the type. However, the latter was proven relevant for interest recognition based on linguistic features.

For automatic transcription of the segments classified as verbal (speech) in the first step, tied-state decision-tree clustered cross-word context dependent triphone HMM with three emitting states and eight Gaussian mixture components per state are used. The models are trained on the training set of the *AVIC* corpus. As acoustic features, 13 Mel Frequency Cepstral Coefficients (MFCC) [67] are preferred over the PLPCC and log-energy from step 1. The difference in feature sets used can be explained by our findings in [60] where PLPCC slightly outperformed MFCC for the specific task of spotting non-linguistic vocalisations. Further experiments we conducted concerning automatic speech recognition on the *AVIC* corpus have revealed the MFCC feature set being superior to the PLPCC set for the ASR task. The details of these experiments are beyond the scope of the article. The reader shall only be aware of the fact that different feature sets are optimal for spotting and classifying non-linguistic vocalisations on the one hand and for full automatic spoken content transcription on the other hand.

Table 11

Confusion matrix: Hidden–Markov-Model based discrimination between 4 classes of non-linguistic vocalisations (hesitation, consent, laughter, and breathing). Speaker independent cross-validation. Sum over all 3 folds. $N = 9, M = 8$, linear topology Hidden–Markov-Models.

# Classif. as →	Hesitation	Consent	Laughter	Breathing
Hesitation	929	14	13	1
Consent	37	255	3	3
Laughter	1	2	229	12
Breathing	2	1	19	412

For spotting of non-linguistic vocalisations in the first decoding pass as described in the previous section with best parameters a recall rate of 55% and a precision rate of 46% is achieved. It is to note that for this result only non-linguistic vocalisations that are spotted at the correct location within the utterance are scored as correct. If more relaxed method, e.g. a string matching is used the rates are higher, however also of less practical significance if the segmentation is to be used as a basis for further processing.

In order to evaluate which non-linguistic vocalisations are confused most often, the classification of non-linguistic vocalisations is considered as a separate problem. Using the *AVIC* ground-truth transcriptions the non-linguistic segments were isolated. Using the same HMM as in the first step of the decoder, however, now trained on a part of the isolated non-linguistic vocalisation segments, the confusion matrix in Table 11 is obtained.

Many different configurations for speech recognition in the second pass were investigated. For this article only the best configuration was selected, which yields a Word-Error-Rate (WER) of 67.6% on the *AVIC* data. Hereby data from 17 of the 21 speakers was used for training and the data from the remaining four speakers was used for evaluation. Compared to other speech recognition experiments on read data [68] this Word-Error-Rate is high. However, we must consider, that recognising affective, spontaneous speech

Table 12

Subject independent recalls (*rec*) and precisions (*pre*) per Level of Interest (LOI), accuracy (*acc*), mean recall, mean precision, and F_1-measure for Support-Vector Machine (SVM) classification using selected modality combinations with early fusion and late fusion of features from acoustics (A), facial expression (F), eye activity (E), linguistics including non-linguistic vocalisations (L), context (C), and their according combinations; subject-independent Leave-one-Speaker-out (LOSO) evaluation (21-fold); "dummy": a single constant dummy feature for chance reference resulting in picking the majority class (LOI1) at any time.

[%]	LOI0		LOI1		LOI2		MEAN			
	rec	pre	rec	pre	rec	pre	acc	rec	pre	F1
Balanced training and early fusion										
Dummy	0.0	0.0	100	51.2	0.0	0.0	51.2	33.3	17.1	22.6
F	64.2	43.0	38.8	63.5	37.6	29.9	46.6	46.9	45.5	46.2
E	75.7	53.7	31.8	60.2	31.8	19.1	45.6	46.4	44.3	45.4
L	80.7	51.7	50.2	74.0	47.6	51.6	59.4	59.5	59.1	59.3
A	74.1	71.1	66.1	76.1	70.0	53.1	69.2	70.0	66.8	68.4
FE	75.1	53.4	48.8	72.2	44.1	36.1	56.3	56.0	53.9	54.9
LA	79.1	74.0	69.8	79.6	72.9	58.8	73.3	73.9	70.8	72.3
FEL	78.0	57.0	62.2	71.6	46.5	64.8	64.5	62.2	64.4	63.3
FEA	82.7	72.1	68.6	79.5	64.7	56.7	72.4	72.0	69.5	70.7
FELA	81.2	72.4	70.8	79.9	71.8	64.2	74.2	74.6	72.1	73.3
FELAC	81.8	79.5	74.7	82.5	75.9	61.7	**77.1**	**77.5**	**74.6**	**76.0**
Balanced training and late fusion										
FELAC	79.2	73.4	67.7	80.2	73.5	55.3	72.3	73.5	69.6	71.5
Unbalanced training and early fusion										
FELAC	73.5	76.9	82.5	73.7	54.7	75.6	74.9	70.2	75.4	72.2

Table 13
Statistical significance of the accuracy improvement for the different modality combinations (rows) when compared to the accuracy of other modality combinations (columns) at a significance level α of 0.05 (SVM classification, balanced training). “+”: accuracy of the modality combination (row) significantly better than performance of the modality combination in the corresponding column; “o”: no statistical significant difference in performance; “−”: accuracy of the modality combination (row) worse than accuracy of the modality combination in the corresponding column; “$*$”: late fusion; “dummy”: a single dummy feature for chance reference resulting in picking the majority class (LOI1) at any time; “F”: facial expression; “E”: eye activity; “L”: linguistic information; “A”: acoustics; “C”: contextual information.

	Dummy	F	E	L	A	FE	LA	FEL	FEA	FELA	FELAC	FELAC*
F	−	o	o	−	−	−	−	−	−	−	−	−
E	−	−	o	−	−	−	−	−	−	−	−	−
L	+	+	+	o	−	o	−	−	−	−	−	−
A	+	+	+	+	o	+	−	+	−	−	−	−
FE	+	+	+	−	−	o	−	−	−	−	−	−
LA	+	+	+	+	+	+	o	+	o	−	−	o
FEL	+	+	+	+	−	+	−	o	−	−	−	−
FEA	+	+	+	+	o	+	−	+	o	−	−	o
FELA	+	+	+	+	+	+	o	+	o	o	−	o
FELAC	+	+	+	+	+	+	+	+	+	o	o	+
FELAC*	+	+	+	+	o	+	−	+	−	−	−	o

is a great challenge due to its high irregularities [69]. For lower Word-Error-Rates, very large databases are required to model the large variance among all data. *AVIC* originally was not intended as a database for training a speech recogniser for spontaneous speech, and thus is fairly small for a speech database. However, for interest recognition it is not necessary to recognise complete sentences correctly, as long as important keywords and lexemes (such as the ones in Table 9, Section 2.2.4) are correctly recognised.

2.2.6. Context

The *AVIC* database, as the usual application scenario, allows for integration of Level of Interest history, as consecutive sub-speaker turns are observed. Such contextual information has already been proven beneficial in emotion recognition [70]; however, it has not yet been used in a multimodal framework, as few data sets exist that provide sequential turns. Level of Interest context integration can be easily realised in two ways: past feature vectors can be included in the actual feature vector, or simply the last estimates for each Level of Interest or the assumed last Level of Interest can be chosen for feature-vector integration. An alternative could e.g. be a “language model” as used in typical Automatic Speech Recognition (ASR) engines for e.g. Level of Interest bigrams or trigrams. Clearly, there is some danger in context integration, as e.g. by over-modelling sudden changes in Level of Interest may be missed.

To stick with an early integration paradigm and not lay too much feature weight on context integration, we decided – in accordance with preliminary tests – to integrate context in the feature space only by the last estimate. This inherits another danger: the last estimate will regularly be erroneous. However, results indicate that there nonetheless seems to be a benefit in this method on average (cf. Table 12), though not being significant (cf. Table 13).

2.3. Multimodal integration

As motivated in Section 1, we choose an early fusion, whereby feature spaces of all modalities are merged into one space. This space is classified within a single classification process saving all available information for the final decision process. Early fusion further allows for combined feature space optimisation: in order to save extraction effort and reduce high complexity throughout classification, features of high individual relevance are selected and de-correlated for space compression by application of Sequential Forward Floating Search (SFFS) [71]. This leads to an optimal set as a whole and the overall minimum number of features. SFFS employs a classifier's accuracy, ideally the target one, as optimisation criterion. Herein, powerful Support-Vector-Machines (SVM) are used to ensure high quality throughout selection (SVM-SFFS) and ensure no bias with the latter target classifier SVM. SFFS is a Hill-Climbing search, and allows for forward and backward search steps in order to cope with nesting effects. A search function is needed, as exhaustive search becomes NP-hard having such high dimensionality as the extensive audiovisual feature space.

All five feature groups introduced in Sections 2.2.1 (facial expression), 2.2.2 (eye activity), 2.2.3 (acoustic), 2.2.4 (linguistics including non-linguistic vocalisations), and 2.2.6 (context), were intentionally projected onto the sub-speaker turn level. This was realised by multivariate time-series analysis for eye activity and acoustic features, while linguistic features reasonably have to operate on this level at minimum, and AAM features were selected from one best frame match, as described. Likewise, no further synchronisation effort is needed at this point, and fusion is realised by a simple super-vector construction including the context of the last Level of Interest estimate.

As a reference, we also include results of a late fusion though it is well known that highly correlated stream information benefits from early fusion [72,73]. However, to best benefit from correlation and thus exploit synergistic information, we decided for a soft decision fusion: each stream is classified individually with provision of a pseudo class-probability obtained by soft-max normalisation. By this, utmost information is preserved for the final decision process. Instead of simple logic operations or voting, a classifier should learn “which modality to trust when by what weight”. Likewise an additional Support-Vector Machine “on top” of the ones serving as uni-modal classifiers sees only the pseudo-class probabilities as meta-features.

2.4. Results and discussion

Now, we present a number of experimental results for diverse multimodal setups. For subject independent testing we use Leave-One-Subject-Out (LOSO) evaluation, and provide mean results over the 21 subjects. All classification and regression results base on the sub-speaker turn level. Note, that in this section all feature selections are carried out in each cycle, to ensure full subject independence at any time. However, for the running system used in the user study in Section 3, the system is trained on the whole corpus in order to exploit the maximum amount of available training material.

We first consider Support-Vector Machines (SVM) to cover the traditional approach to affect and emotion recognition, where discrete classes (here: {LOI0, LOI1, LOI2}) are used [15,10]. As

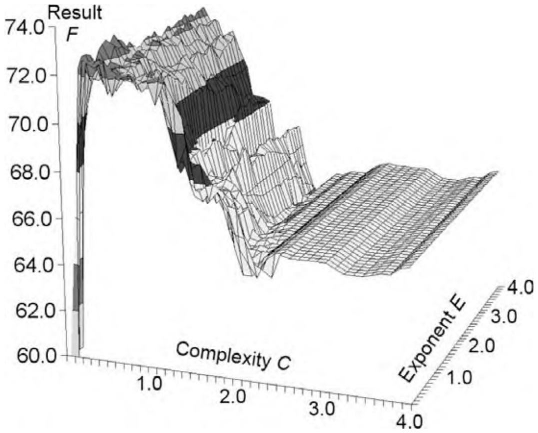

Fig. 8. 3D visualisation of polynomial Support-Vector Machine parametrisation effect on the F_1-value. Integration of the full audio-visual feature space in balanced training, subject-independent Leave-one-Subject-out evaluation.

Table 14
Subject independent cross-correlation for Support-Vector Regression (SVR) of selected modality combinations in early fusion out of facial expression (F), eye activity (E), acoustics (A), linguistics including non-linguistic vocalisations (L), context (C), and their according combinations; Mean Linear Error (MLE) and Correlation Coefficient (CC); subject-independent Leave-one-Speaker-out (LOSO) evaluation (21-fold); performance of a classifier using a single constant "dummy" feature for chance reference: $MLE = 0.77; CC = 0.16$.

	F	E	L	A	FE	LA	FEL	FEA	FELA	FELAC
MLE	0.69	0.72	0.48	0.40	0.69	0.31	0.64	0.55	0.54	0.55
CC	0.29	0.36	0.51	0.71	0.44	0.72	0.56	0.68	0.70	0.69

performance measure we use accuracy[7] (acc), recall[8] (rec), precision[9] (pre), and F_1 measure[10], as introduced in [10].

In Fig. 8 the influence of the SVM parameters complexity (C) and the polynomial exponent (E) on the F_1-measure can be seen. SVM are thereby trained by Sequential Minimal Optimisation (SMO) and a pairwise discrimination for multi-class problems is employed [71]. The optimal parameter-set is chosen for each individual modality combination in the final comparison.

Secondly, we consider a continuum instead of discrete classes, as often proposed but seldom used in related emotion recognition (e.g. valence and arousal, [14], or dominance [28]): the Level of Interest (LOI) is within the interval $LOI \in [-2;2]$, and the master LOI is obtained by mean over the four labelers as described in Section 2.1. This demands for a regression approach for detection of the current LOI, as by the chosen Support-Vector Regression (SVR). Still, parametrisation is accordingly. The evaluation is carried out by cross-correlation (CC) of the estimated LOI and the mean ground truth master LOI, whereby $CC \in [-1;1]$, and a high CC value of $CC = 1$ is optimal. As further performance measure we calculated the Mean Linear Error (MLE) as used in [74] which is the mean deviation of the estimated LOI with respect to the ground truth in the interval [−2;2].

As not all modalities are present at any time (often no speech, especially in the case of boredom, i.e. LOI0), only a considerably lower number of instances as depicted in Fig. 2 can be used for multimodal evaluation. However, note that a real-life system profits from multi-modality in spite of the partial lack of modalities. Also note, that the number of instances among classes for training are highly unbalanced. Therefore, we also consider uniformly distributed training sets obtained by deterministic up-sampling. This is, however, only possible in a straight-forward manner in the case of classification (by SVM): instead of discrete classes only a mean Level of Interest is available in the case of regression (SVR). Balancing is therefore not carried out for SVR.

As can be seen in the summarised result Table 12, the usage of balanced training sets leads to a significantly more satisfying result with respect to balance among recall rates. Stemming could improve the accuracy by 4.4% when using linguistic information only. The combination of all groups of features prevails. However, the combination of facial expression, eye activity and acoustics alone does not fall far behind. Yet, all possible combinations do not satisfyingly solve the problem of LOI0 and LOI2 being discriminated more easily than each one from LOI1. However, in many applications a discrimination of boredom vs. interest may be sufficient.

Table 13 indicates the statistical significance of the performance improvements obtained through multimodal classification. Statistical significance of the performance gain of every modality combination (row) is evaluated with respect to every other modality, modality combination, as well as with respect to chance (column), according to [75]. As significance level we chose the common value of 0.05. Note that the significance test is based on the assumption that the error rates of the compared classifiers are independent, which can only be fulfilled if the classifiers are evaluated on different data-sets. Since we used the same data-set for all test runs, the premise to reject the null hypothesis [75] was comparably strict.

In Table 14 the results for Support-Vector Regression are given. For the regression experiments, the full Level of Interest spectrum (−2 to 2) was used. The higher resolution of the regression approach (providing "inbetween" LOI values such as 1.5) has the downside of yielding a slightly lower accuracy: if we discretise the regression output into the discrete classes {LOI0, LOI1, LOI2} and compare it with the discrete master LOI, an F_1 measure of 69.1% is obtained for the optimal case of fusion of all information instead of 76.0% for the directly discrete classification. Note that a single dummy feature for the regression approach leads to a correlation coefficient of 0.16. Interestingly, the combination of acoustic and linguistic processing could not be improved by the additional visual and contextual knowledge provision in this modelling approach. However, their consideration is still justified (only) by the cases of speech absence for the regression analysis.

3. Evaluating real-life usage

Apart from figures "in the lab" as provided in typical works on interest and human affect recognition, it seems to be important for real-life usage to test systems with the user in the loop. We therefore present results of a study in which the system described so far (Section 2) is tested in a real usage study.

The primary aim of this survey is to measure whether a recognition system can indeed already improve a virtual agent-based presentation according to a person's level of interest in a topic. For comparison, the presentation is carried out without any adaptation as lower benchmark. The upper benchmark is obtained by a human conductor – the "Wizard-of-Oz" (WoO) [76,77] – that estimates a subject's interest.

An agent-based system thereby inherits a shift of paradigm: so far we had investigated the performance of interest detection based on natural human-to-human conversation. In the ongoing

[7] number of correctly classified cases divided by total number of cases or weighted average

[8] recall is the "class-wise" computed accuracy (note, that the mean recall rate differs from the accuracy as it is not weighted by total instance numbers; thus, it provides a better performance measure in the case of unbalanced class distributions)

[9] shows in how many cases the classifier is right when claiming a certain class

[10] trade-off between rec and pre by harmonic mean of these two: $F_1 = 2 \cdot rec \cdot pre/(rec + pre)$

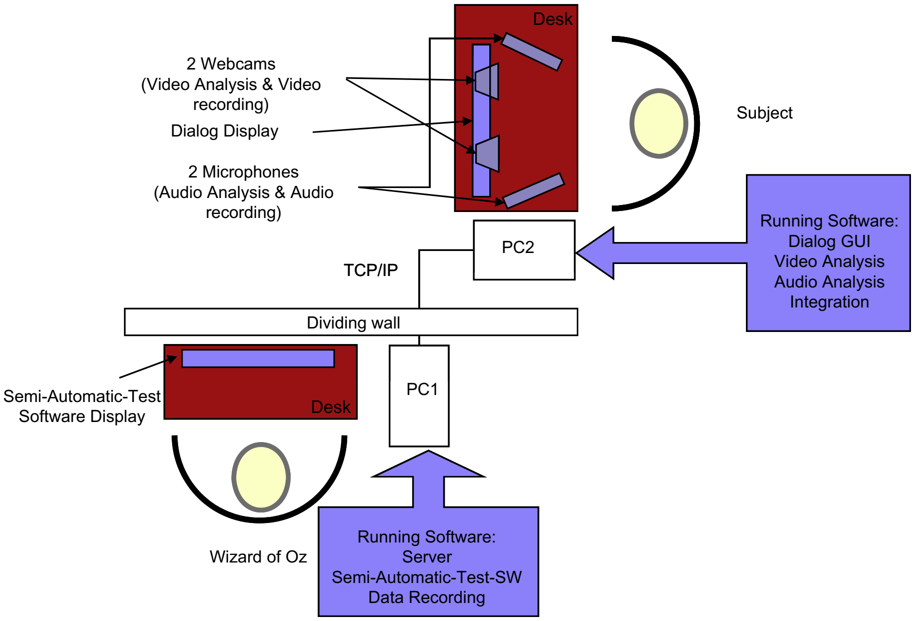

Fig. 9. Schematic view of survey setup.

the trained system will be used in Human–Computer Interaction to reveal benefits in this area.

3.1. Experimental setup

Nine topics are used in a virtual product and company tour (Toyota Museum, Safety, Intelligent Transport System, Toyota Production System, Environment, Motor sports, Toyota History, Toyota Partner Robot, and Toyota Prius). The user is thereby guided by a male animated embodied conversational agent (ECA) [78] that is visualised by a photo-realistic face with lip movements synchronised and speech is synthesised by concatenative synthesis.

User feedback is requested in average 8.7 times per topic (maximum 14 times, minimum 5 times) including yes/no and knowledge-based questions. The system reacted depending on a "true" or "false" type of answer with two different dialogue alternatives, accordingly. In the case of knowledge-based questions it reacted with a predefined standard phrase independent of the actual answer.

The experimental setup (as depicted in Fig. 9) consists of two standard office PCs. One for the Wizard-of-Oz and the other for the subject. The vision, speech and integration software were running together with the dialogue-display on the subject PC. The dialogue has been presented to the user on a wide-screen (16:10) display, showing an animated male character (30 year old looking) on the left and a still image or video on the left adequate to the current topic in 50:50 split. The Wizard-of-Oz PC was hosting the communication server and a semi-automatic test-environment software (cf. [79]), which was used to control the dialogue for the user, based on his answers and current Level of Interest. It was also used for audio and video recording of the subject. Two pairs of microphone and web-cam were used. One pair was needed for audio and video based Level of Interest and speech recognition and the other one to record videos of the subject. The web-cams have a resolution of 320 × 200 and the videos were encoded as Windows Media Video (WMV).

The interest recognition system, as described so far, consists of four modules, namely AAM-based facial expression analysis, eye activity, acoustic analysis, and speech recognition with integrated non-linguistic vocalisations and subsequent linguistic analysis. The four modules run in real-time on one desktop PC by making additional use of the graphics co-processor (GPU) for calculations. For the study we train on the full *AVIC* database. This is in no contradiction to subject independence as the test subjects are fully disjunctive. Further, the system is used in its optimal configuration, that is early integration with balanced training. Note that it performs fully automatic interest recognition without any manual help. Each module forwards its features by TCP/IP socket communication to an integration server. The stream segmentation is achieved by audio as first priority followed by video segmentation if no audio is available. Audio segmentation is realised by two-fold dynamic energy thresholding with subsequent speech/non-speech verification by the acoustic feature module using the described 1.4k feature space. Video segmentation is obtained by Bayesian Information Criterion (BIC) as described in more detail in [32].

3.2. Subjects and experiments

The survey was conducted in two experiments. For every experiment 20 persons with a desired gender and age balance had to listen to the dialogue and afterwards fill out a questionnaire and a multiple choice test.

From these 40 participants in total, 29 were < 30 years, six were between 30 and 40 years and the remaining five were > 40 years. The gender was nearly equally balanced with 18 females and 22 males. Two subjects were Asian, the rest were European. Their profession ranged from students, secretary, dress designers to people with graduate profession. Again, all subjects were very experienced English speakers, and all dialogue was carried out in English. See Table 15 for more details.

Table 15
Statistics of the 40 subjects that participated in the user-study.

Group of subjects	# Subjects	Mean age (years)
All	40	30.6
Male	22	31.0
Female	18	30.1
Group1	12m/8f	33.7
Group2	10m/10f	27.5

The first dialogue was presented to the subject without any adaptation. All nine prepared dialogue topics were presented to the subject. Besides the first topic (Toyota Museum), all were presented in random order. The first 20 participants (Group1) had a mean age of 33.7. The subjects were picked in according age-class balance per group as mentioned in Section 2.1.

In the second experiment the topic "Toyota Museum" was presented in full length as introduction to the second group of subjects (Group2) that had a mean age of 27.5 years. The next four topics were changed automatically by the system (as described in Section 3.3). In the second half, the remaining four topics have been changed manually by the wizard, based on the camera image of the subject and the answers of the subject. Furthermore, all topics had a minimum presentation length of two minutes.

3.3. Automatic topic change

The automatic change of the randomly chosen four topics is based on the computed Level of Interest from the integration module, which combines the visually and acoustically computed Level of Interest. The topics have been presented for a minimum time span of two minutes. This fixed period provides time for the subject to settle down in a new Level of Interest according to the currently presented topic. After this time span the current Level of Interest was compared to an empirically determined threshold to decide whether the topic should be changed or not. This current Level of Interest ($\mathrm{LOI}(t)$) was calculated over the last i sub-speaker turns since the end of the last system uttered dialogue event by

$$\mathrm{LOI}(t) = \frac{\sum_{i=0}^{N} length(i) \cdot \mathrm{LOI}(i)}{\sum_{i=0}^{N} length(i)} \tag{13}$$

where $\mathrm{LOI}(i)$ is the Level of Interest computed by the integration module and $\mathrm{length}(i)$ its corresponding segment length. Thus, a mean Level of Interest is considered having a typical speaker-turn as unit of analysis. This leads to more stable estimates functioning as filtering. Finally the topic was changed either when this mean Level of Interest dropped under the threshold or when all available material for the current topic has been presented to the subject.

3.4. Results and discussion

Table 16 shows selected results from the evaluation of this survey.

As can be seen from the results, the wizard-based detection of interest clearly improved the dialogue quality with respect to how much information was subjectively understood. It may thereby be speculated that more is understood if one is interested in a topic. More users further felt that the system was taking their interest into account in the wizard-based experiment, though they were never told whether it actually does, and that they could provide sufficient feedback which was used in the way expected. For questions like "How interesting was the information that the system gave you" or "Did you get enough possibility to give feedback to the system" the answers did not differ significantly among the three variants of topic switching.

Table 16
Selected significant results of a questionnaire in a real-life study on interest recognition. Virtual tour through a museum. Forced choice reaching from 1 (best) to 5 (worst) and percentage of users, depending on the question type. The tour was carried out with topic switching after a fixed number of 29 dialogue elements (each element consisting of one or two sentences) per topic (FND) whereby the average topic length resembles 3:57 min, if the user was bored by subject-independent fully automatic interest recognition (AIR), and by a human test-conductor, the "Wizard of Oz" (WoO). 40 subjects are contained in gender and age-class-balance. Note that one person occasionally abstained from judging the system concerning some of the questions.

Questions	FT	AIR	WoO
Did you think the system was taking into account your interest?	35%	63%	84%
How much of the info the system gave you did you understand?	1.90	1.53	1.74

However, as far as taking into account the users' interest is concerned, the fully automatic system, lying in between no topic switching and wizard-based topic switching, falls only slightly behind the wizard. Likewise, real practicability seems to be proven.

It remains to be stated that by this experiment a further audiovisual data-set of 20:58:03 h of user-interaction of 40 subjects (mean time per subject: 0:34:00 h, minimum time 00:12:41 h, maximum time 00:58:59 h) was recorded which can be used in future studies after it was transcribed and annotated.

4. Concluding remarks

In this work a fully automatic system for extensive audiovisual stream integration based recognition of human interest, which is able to operate in real-time, was presented. The approach proposed, integrates numerous streams to enable most reliable automatic determination of interest-related user states, such as being bored or being curious. For training and testing the *AVIC* database was recorded. The database contains data of a real-world scenario where an experimenter leads a subject through a commercial presentation. The subject interacts with the experimenter and thereby naturally and spontaneously expresses different levels of interest.

Two different visual feature streams are used: facial expression is covered through Active Appearance Models and application of Principal Component Analysis for statistical analysis of shape and texture variations. Furthermore, eye activity is determined considering various characteristics of eye movement.

From the audio signal acoustic features are extracted by systematically applying statistical functionals to acoustic Low-Level Descriptors in combination with a self-learning feature-space optimisation. Further, linguistic features are included in the multimodal recognition framework. Thereby Bag-of-Words are used as efficient vector space representation while the strategy of stopping and stemming reduces the complexity of linguistic analysis. As extra-linguistic information, such as non-linguistic vocalisations can also reveal a subject's Level of Interest, a two-step Hidden Markov Model for the detection and discrimination of non-linguistic vocalisations in combination with speech recognition is included in fully automatic processing.

Additionally, contextual interest information is integrated in the feature space by using the last estimate of the Level of Interest as feature. This slightly improved recognition performance on average, but was not observed as significant.

An early fusion strategy is used before the multimodal feature space is optimised in a combined manner. Support-Vector Machines discriminate three discrete levels of interest in fully subject-independent experiments. The results presented, show that by early fusion of all information sources the maximum accuracy is obtained: a remarkable subject-independent F_1-measure of 72.2% is achieved for unbalanced training. Performance could be further boosted through balanced training resulting in an F_1-measure of 76.0%. In comparison a late semantic fusion based on meta-classification led to only 71.1% F_1 measure. As alternative modelling approach a continuous Level of Interest scale was used for Support-Vector Regression. Here, the best cross-correlation value reached 0.72. In a significance analysis, the integration of additional streams was shown to be significant for the classification task. However, more experience will be needed in the case of regression where no gain could be reached yet.

To determine whether such a recognition system can already improve a virtual agent-based presentation [80] according to the Level of Interest the user shows, the system was additionally tested in a real-life use-case study. There, the automatic recogniser performed significantly better than a system simply ignoring the users' interest. However, there is still a small gap between the performance of the automatic interest recogniser and a human operator acting as a "wizard" that decides upon the interest of users.

Overall, spontaneous interest could be automatically detected, independent of the subject, in human conversation by the proposed extensive audiovisual and contextual information carried out for the first time on such broad basis and in fully automatic, yet real-time-capable processing. While there is clearly room left for further improvement, it seems that present technology has matured to a degree that allows us to take affective computing technology into real-life Human–Computer Interaction systems.

Future works will have to deal with improved discrimination of the subtle difference of the border class between strong interest and boredom. Also, in this respect more instances of strongly expressed boredom should be recorded in future efforts to broaden the scope of use-cases: in the face-to-face communication captured herein, these did not occur sufficiently often – potentially due to subject's minimum politeness. Also the detection of a group's interest seems important in many application scenarios.

For many applications detection of boredom or high interest moments may be sufficient. As these were observed to lie in-between comparably long sequences of normal interest, a detection approach may be an interesting alternative to the classification shown. However, the further introduced continuous modelling by regression allows for a threshold definition, already.

Moreover, in this work no heavy noise, occlusions or failure of whole components have been investigated. In case of noted heavy disturbance of single modalities, however, a shift to others can be performed – being one additional strength of a multimodal approach. Automatically noticing such events and performance with automatic modality selection will be one future research issue.

Furthermore, it would be interesting to examine the benefits of integrating knowledge about the context not only as a feature, but in the architecture of the recogniser. Thereby classifiers which model long-range dependencies such as Conditional Random Fields for discrete classification or Long-Short Term Memory Recurrent Neural Networks for continuous prediction could be used [81].

A refinement of the presented system will also have to allow for stronger pitch, yaw and roll rotations of the head pose, as the current system is trained and focused on face-to-face conversation. Thus, in present status it is in particular useful in similar settings such as a user at a traditional computer or e.g. in an automotive setting.

Finally, the measures of interest herein were observational: an automated system was trained with observations. As interest in a broader sense refers to a cognitive-motivational state, additional focus may be laid on subjective interest rather than its appearance.

References

[1] A. Pentland, A. Madan, Perception of social interest, in: Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI), Beijing, 2005.

[2] E. Shriberg, Spontaneous speech: How people really talk and why engineers should care, in: Proc. European Conf. on Speech Communication and Technology (Eurospeech), Lisbon, Portugal, 2005.

[3] P. Qvarfordt, D. Beymer, S.X. Zhai, Realtourist – a study of augmenting human–human and human–computer dialogue with eye-gaze overlay, in: INTERACT 2005, vol. LNCS 3585, 2005, pp. 767–780.

[4] R. Stiefelhagen, J. Yang, A. Waibel, Modeling focus of attention for meeting indexing based on multiple cues, IEEE Transactions on Neural Networks 13 (4) (2002) 928–938.

[5] L. Kennedy, D. Ellis, Pitch-based emphasis detection for characterization of meeting recordings, in: Proc. ASRU, Virgin Islands, 2003.

[6] D. Gatica-Perez, I. McCowan, D. Zhang, S. Bengio, Detecting group interest-level in meetings, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, 2005.

[7] S. Mota, R. Picard, Automated posture analysis for detecting learners interest level, in: Proc. Workshop on CVPR for HCI, Madison, 2003.

[8] B. Schuller, G. Rigoll, M. Lang, Hidden markov model-based speech emotion recognition, in: Proc. ICASSP 2003, vol. II, 2003, pp. 1–4.

[9] A. Batliner, S. Steidl, C. Hacker, E. Nöth, H. Niemann, Private emotions vs. social interaction – towards new dimensions in research on emotion, in: Proceedings of a Workshop on Adapting the Interaction Style to Affective Factors, 10th International Conference on user Modelling, Edinburgh, 2005.

[10] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, V. Aharonson, Combining Efforts for Improving Automatic Classification of Emotional User States, in: Proceedings of IS-LTC 2006, Ljubliana, Slovenia, 2006, pp. 240–245.

[11] A. Kapoor, R.W. Picard, Y. Ivanov, Probabilistic combination of multiple modalities to detect interest, in: Proc. of the 17th Int. Conf. on Pattern Recognition, ICPR 2004, vol. 3, 2004, pp. 969–972.

[12] A. Ashraf, S. Lucey, T. Chen, K. Prkachin, P. Solomon, Z. Ambadar, J. Cohn, The painful face: pain expression recognition using active appearance models, in: Proc. 9th Int. Conf. on Multimodal Interfaces (ICMI), Special Session on Multimodal Analysis of Human Spontaneous Behaviour, ACM SIGCHI, Nagoya, Japan, 2007, pp. 9–14.

[13] L. Maat, M. Pantic, Gaze-x: Adaptive affective multimodal interface for single-user office scenarios, Artificial Intelligence for Human Computing, vol. 4451/2007, Springer LNCS, Berlin, Heidelberg, 2007, pp. 251–271.

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

[15] M. Pantic, L. Rothkrantz, Toward an affect-sensitive multimodal human–computer interaction, in: Proceedings of the IEEE 91 (September 2003) 1370–1390.

[16] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, B. Radig, Audiovisual behaviour modeling by combined feature spaces, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), vol. II, Honolulu, HY, 2007, pp. 733–736.

[17] M. Wimmer, B. Schuller, D. Arsic, B. Radig, G. Rigoll, Low-level fusion of audio and video features for multi-modal emotion recognition, in: Proc. 3rd International Conference on Computer Vision Theory and Applications (VISAPP 2008), Funchal, Madeira, Portugal, 2008.

[18] N. Sebe, I. Cohen, T.S. Huang, Multimodal emotion recognition, in: Handbook of Pattern Recognition and Computer Vision, World Scientific, 2005.

[19] M. Paleari, C.L. Lisetti, Toward multimodal fusion of affective cues, in: Proc. 1st ACM Int. Workshop on Human-centered Multimedia, Santa Barbara, CA, 2006, pp. 99–108.

[20] N. Bianchi-Berthouze, C. Lisetti, Modeling multimodal expression of user's affective subjective experience, User Modeling and User-Adapted Interaction 12 (2002) 49–84.

[21] B. Schuller, R. Müller, M. Lang, G. Rigoll, Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles, in: Proc. Interspeech 2005, ISCA, Lisbon, Portugal, 2005.

[22] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1) (2009) 39–58.

[23] L. Axelrod, K. Hone, E-motional advantage: performance and satisfaction gains with affective computing, in: Proc. Conference on Human Factors in Computing Systems, ACM, Portland, OR, 2005, pp. 1192–1195.

[24] B. Schuller, D. Seppi, A. Batliner, A. Maier, S. Steidl, Towards more reality in the recognition of emotional speech, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), vol. IV, Honolulu, HY, 2007, pp. 941–944.

[25] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, G. Rigoll, Audiovisual recognition of spontaneous interest within conversations, in: Proc. 9th Int. Conf. on Multimodal Interfaces (ICMI), Special Session on Multimodal Analysis of Human Spontaneous Behaviour, ACM SIGCHI, Nagoya, Japan, 2007, pp. 30–37.

[26] M. Kipp, Anvil – a generic annotation tool for multimodal dialogue, in: Proc. ISCA EUROSPEECH 2001, 2001, pp. 1367–1370.

[27] J. Carletta, Assessing agreement on classification tasks: the kappa statistic, Computational Linguistics 22 (2) (1996) 249–254.

[28] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, T. Moosmayr, On the necessity and feasibility of detecting a drivers emotional state while driving, in: Proc. of 2nd Int. Conf. on Affective Computing and Intelligent Interaction (ACII 2007), ACM, Springer, Lisbon, Portugal, 2002, pp. 126–138.

[29] P. Ekman, Facial expressions, in: T. Dalgleish, M. Power (Eds.), Handbook of Cognition and Emotion, John Wiley & Sons Ltd, New York, NY, 1999. Ch. 16.

[30] J. Cohn, Foundations of human computing: facial expression and emotion, in: T. Huang, A. Nijolt, M. Pantic, A. Pentland (Eds.), State of the Art Survey, Lecture notes in artificial intelligence, Springer, Berlin Heidelberg, 2007, pp. 1–16.

[31] M. Pantic, M. Bartlett, Machine analysis of facial expressions, in: K. Kurihara (Ed.), Face Recognition, Advanced Robotics System, Vienna, Austria, 2007, pp. 327–366.

[32] F. Wallhoff, B. Schuller, M. Hawellek, G. Rigoll, Efficient recognition of authentic dynamic facial expressions on the feedtum database, in: Proc. of the IEEE Int. Conference on Multimedia and Expo (ICME 2006), Toronto, Ontario, 2006, pp. 493–497.

[33] D. Arsic, J. Schenk, B. Schuller, F. Wallhoff, G. Rigoll, Submotion for hidden markov model based dynamic facial action recognition, Tech. Rep., Technische Universität München, 2006.

[34] A. Samal, P. Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey, Pattern Recognition 25 (1) (1992) 65–77.

[35] Y. Tian, T. Kanade, J. Cohn, Facial expression analysis, in: S. Li, A. Jain (Eds.), Handbook of Face Recognition, Springer, New York, NY, 2005, pp. 247–276.

[36] R. El Kaliouby, P. Robinson, Real-time inference of complex mental states from facial expressions and head gestures, in: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR 2004), vol. 3, Washington, DC, 2004, pp. 154–157.

[37] H. Gu, Q. Ji, An automated face reader for fatigue detection, in: Proc. IEEE Int. Conf. Face and Gesture Recognition (FGR 2004), Seoul, Korea, 2004, pp. 111–116.

[38] T. Cootes, G. Edwards, C. Taylor, A comparative evaluation of active appearance model algorithms, in: P. Lewis, M. Nixon (Eds.), Proc. of British Machine Vision Conference, vol. 2, Southampton, UK, 1998, pp. 680–689.

[39] A.U. Batur, M.H. Hayes, Adaptive active appearance models, IEEE Transactions on Image Processing 14 (11) (2005) 1707–1721.

[40] R. Müller, A system for automatic face analysis based on statistical shape and texture models, Ph.D. thesis, Technische Universität München (TUM), Munich, Germany, 2008.

[41] T.F. Cootes, C.J. Taylor, Statistical models of appearance for computer vision, Tech. rep., University of Manchester, UK, March 2004.

[42] M. Poletti, M. Rucci, Fixational eye movements and retinal activity during a single visual fixation, Journal of Vision 8 (6) (2008) 280.

[43] D.L. Neumann, O.V. Lipp, Spontaneous and reflexive eye activity measures of mental workload, Australian Journal of Psychology 54 (3) (2002) 174–179.

[44] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: CVPR, 2001.

[45] F. Wallhoff, M. Ablasmeier, G. Rigoll, Multimodal face detection, head orientation and eye gaze tracking, in: Proceedings IEEE International Conference on Multisensor Fusion and Integration, 2006.

[46] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting.

[47] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The feret evaluation methodology for face-recognition algorithms, PAMI 22 (10) (2000) 1090–1104.

[48] R. Kompe, Prosody in Speech Understanding Systems, Springer Verlag, 1997.

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

[50] B. Schuller, N. Köhler, R. Müller, G. Rigoll, Recognition of interest in human conversational speech, in: Proc. INTERSPEECH 2006, ISCA, Pittsburgh, PA, 2006, pp. 793–796.

[51] A. Batliner, B. Schuller, S. Schaeffler, S. Steidl, Mothers, adults, children, pets – towards the acoustics of intimacy, in: Proc. ICASSP 2008, Las Vegas, Nevada, 2008, pp. 4497–4500.

[52] B. Vlasenko, B. Schuller, A. Wendemuth, G. Rigoll, On the influence of phonetic content variation for acoustic emotion recognition, in: Proc. 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems (PIT 2008), Springer, LNCS, Kloster Irsee, Germany, 2008.

[53] D. Litman, K.M. Forbes Riley, S. Silliman, Towards emotion prediction in spoken tutoring dialogues, in: Proc. Human Language Technology Conference North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2003), 2003.

[54] C.M. Lee, R. Pieraccini, Combining acoustic and language information for emotion recognition, in: Proc. of the International Conference on Speech and Language Processing (ICSLP 2002), Denver, CO, 2002.

[55] J. Russell, J. Bachorowski, J. Fernandez-Dols, Facial and vocal expressions of emotion, Annual Review of Psychology 54 (2003) 329–349.

[56] N. Campbell, On the use of nonverbal speech sounds in human communication, in: COST 2102 Workshop, 2007, pp. 117–128.

[57] M.W. Decaire, The detection of deception via non-verbal deception cues, Law Library 1999–2001, 2000.

[58] R. Lickley, R. Shillcock, E. Bard, Processing disfluent speech: How and when are disfluencies found? in: Proc. of the European Conference on Speech Technology, vol. 3, 1991, pp. 1499–1502.

[59] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, Tech. Rep., LS-8 Report 23, Dortmund, Germany, 1997.

[60] B. Schuller, F. Eyben, G. Rigoll, Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech, Proc. 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems (PIT 2008), vol. LNCS 5078, Springer, LNCS, Kloster Irsee, Germany, 2008, pp. 99–110.

[61] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (v3.4), Cambridge University Press, Cambridge, UK, 2006.

[62] H. Hermansky, Perceptual linear predictive (plp) analysis of speech, Journal of the Acoustical Society of America 87 (4) (1990) 1738–1752.

[63] M. Knox, M. Mirghafori, Automatic laughter detection using neural networks, in: Proc. of the INTERSPEECH-2007, 2007.

[64] N. Campbell, H. Kashioka, R. Ohara, No laughing matter, in: Proc. of INTERSPEECH-2005, 2005, pp. 465–468.

[65] M. Goto, K. Itou, S. Hayamizu, A real-time filled pause detection system for spontaneous speech recognition, in: Eurospeech'99, 1999, pp. 227–230.

[66] T. Schultz, I. Rogina, Acoustic and language modeling of human and nonhuman noises for human-to-human spontaneous speech recognition, in: Proc. ICASSP-1995, vol. 1, Detroit, Michigan, 1995, pp. 293–296.

[67] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[68] T. Hain, P.C. Woodland, G. Everman, D. Povey, The cu-htk march 2000 hub5e transcription system, in: Proc. Speech Transcription Workshop, 2000.

[69] B. Schuller, J. Stadermann, G. Rigoll, Affect-robust speech recognition by dynamic emotional adaptation., in: Proc. ISCA Speech Prosody 2006, ISCA, Dresden, Germany, 2006.

[70] J. Liscombe, G. Riccardi, D. Hakkani-Tür, Using context to improve emotion detection in spoken dialog systems, in: In INTERSPEECH-2005, 2005, pp. 1845–1848.

[71] I.H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, second ed., Morgan Kaufman, San Francisco, 2005.

[72] H. Gunes, M. Piccardi, Affect recognition from face and body: early fusion vs. late fusion, in: IEEE International Conference on Systems, Man and Cybernetics, 2005, vol. 4, 2005, pp. 3437–3443.

[73] W. Lizhong, S. Oviatt, P.R. Cohen, Multimodal integration – a statistical view, in: IEEE Transactions on Multimedia, vol. 1, 1999, pp. 334–341.

[74] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, T. Moosmayr, On the necessity and feasibility of detecting a driver's emotional state while driving, in: Proc. 2nd Int. Conf. on Affective Computing and Intelligent Interaction ACII 2007, Lisbon, Portugal, Vol. LNCS 4738, Springer Berlin, Heidelberg, 2007, pp. 126–138.

[75] L. Gillick, S.J. Cox, Some statistical issues in the comparison of speech recognition algorithms, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1989), vol. 1, 1989, pp. 23–26.

[76] J. Nielsen, Usability Engineering, Academic Press, Inc., 1993.

[77] R. Nieschulz, B. Schuller, M. Geiger, R. Neuss, Aspects of efficient usability engineerings, it+ti Journal, Usability Engineering 44 (1) (2002) 23–30.

[78] I. Poggi, C. Pelachaud, B. de Carolis, To display or not to display? Towards the architecture of a reflexive agent, in: Proc. of the 2nd Workshop on Attitude, Personality and Emotions in User-adapted Interaction, User Modeling 2001, Sonthofen (Germany), 2001, p. 7.

[79] B. Schuller, M. Lang, G. Rigoll, Towards automation of usability studies, in: Proc. of IEEE International Conference on Systems, Man and Cybernetics (SMC 2002), vol. 4, Yasmine Hammamet, Tunesia, 2002.

[80] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, B. Schuller, Towards responsive sensitive artificial listeners, in: Proc. 4th Intern. Workshop on Human–Computer Conversation, Bellagio, Italy, 2008.

[81] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, Abandoning emotion classes – towards continuous emotion recognition with modelling of long-range dependencies, in: Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia, 2008, pp. 597–600.