Audiovisual Recognition of Spontaneous Interest within Conversations

Björn Schuller
Institute for Human-Machine Communication
Technische Universität München
D-80333 München, Germany
schuller@tum.de

Ronald Müller
Institute for Human-Machine Communication
Technische Universität München
D-80333 München, Germany
rm@tum.de

Benedikt Hörnler
Institute for Human-Machine Communication
Technische Universität München
D-80333 München, Germany
b@tum.de

Anja Höthker
Toyota Motor Europe
Production Engineering - Advanced Technologies
B-1930 Zaventem, Belgium
Anja.Hoethker@toyota-europe.com

Hitoshi Konosu
Toyota Motor Corporation
1, Toyota-cho
Toyota City, Aichi, 471-8571
Japan
hitoshi_konosu@mail.toyota.co.jp

Gerhard Rigoll
Institute for Human-Machine Communication
Technische Universität München
D-80333 München, Germany
rigoll@tum.de

ABSTRACT

In this work we present an audiovisual approach to the recognition of spontaneous interest in human conversations. For a most robust estimate, information from four sources is combined by a synergistic and individual failure tolerant fusion. Firstly, speech is analyzed with respect to acoustic properties based on a high-dimensional prosodic, articulatory, and voice quality feature space plus the linguistic analysis of spoken content by LVCSR and bag-of-words vector space modeling including non-verbals. Secondly, visual analysis provides patterns of the facial expression by AAMs, and of the movement activity by eye tracking. Experiments base on a database of 10.5h of spontaneous human-to-human conversation containing 20 subjects in gender and age-class balance. Recordings are fulfilled with a room microphone, camera, and headsets for close-talk to consider diverse comfort and noise conditions. Three levels of interest were annotated within a rich transcription. We describe each information stream and a fusion on an early level in detail. Our experiments aim at a person-independent system for real-life usage and show the high potential of such a multimodal approach. Benchmark results based on transcription versus automatic processing are also provided.

Categories and Subject Descriptors

I.5 [**Computing Methodologies**]: Pattern Recognition; J.4 [**Computer Applications**]: Social and Behavioral Sciences

ICMI'07, November 12-15, 2007, Nagoya, Aichi, Japan.

General Terms

Algorithms, Human Factors, Experimentation

Keywords

Affective Computing, Audiovisual, Interest, Emotion

1. INTRODUCTION

Knowledge of a communication partner's interest possesses great potential in many commercial applications. Similar to the work introduced in [9] we are likewise interested in curiosity detection e.g. for topic switching, in infotainment systems or customer service systems. In order to quantify a persons's interest we introduce three levels of interest (LOI) reaching from LOI=0 representing *disinterest*, *indifference*, and *neutrality* over LOI=1 standing for *light interest* to LOI=2 representing *strong interest*. As audiovisual processing is known to be superior to each single modality [8] [4], we propose to combine features derived from acoustic and linguistic analyses, as well as facial expression analysis basing on Active Appearance Models (AAM) and activity modeling. The paper is structured as follows: after a short description of collection of spontaneous interest data in sec. 2 we describe acoustic and linguistic speech processing in sec. 3, Active Appearance Models in sec. 4, Activity Estimation in sec. 5, multimodal information stream integration on an early feature level and experimental fusion results in sec. 6 and a concluding discussion in sec. 7.

2. SPONTANEOUS INTEREST DATA

In order to overcome today's mostly acted audiovisual databases , and due to lack of a set dealing with interest, we decided to record a database named AVIC (Audiovisual Interest Corpus) in the ongoing. In the scenario setup, an experimenter and a subject are sitting on both 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 to the addressed topics without respect to politeness. Visual and voice data is recorded by a camera and two microphones, one headset and one far-field mic, in this situation. After the final recording the AVIC database shows the following parameters and statistical figures: image resolution: 720 x 576, frame rate: 25 fps progressive, color resolution: 24 Bit, encoder: DV, audio sampling rate: 44100 Hz, audio quantization: 16 Bit, left audio channel: lapel microphone, right audio channel: far-field microphone. 21 subjects (10 of them female) took part, 3 of them Asian, the others European. The language throughout experiments is English, and all subjects are very experienced English speakers. 3 age categories were defined during specification phase (<30 a, <40 a, >40 a) for balancing. The mean age of male subjects resembles 32.7 a, the mean age of female subjects accordingly 30.1 a. The total recording time for males resembles 5:14:30 h, for females 5:08:00 h. By age categories the recording times are 4:40:40 h for <30 a, 4:10:20 h for <40 a, 1:31:30 h for >40 a. Likewise, a total of 10:22:30 h was recorded. In order to acquire reliable labels of LOI, the entire video material was segmented in speaker and sub-speaker turns and subsequently labeled by 4 male annotators, independently. Figure 1 shows the corresponding annotation workflow. The LOI is annotated for every sub-speaker turn. In order to get an impression of a subject's character and behaviour before the annotation of a person starts, the annotators had to watch approximately 5 minutes of a subject's video. This helps to find out the range of intensity, the subject expresses her/his curiosity. For annotation, every sub-speaker turn has to be viewed at least once to find out the LOI displayed by the subject. 5 LOI were distinguished in the first place: 1 - *Disinterest* (subject is bored listening and talking about the topic, very passive, does not follow the discourse), 2 - *Indifference* (subject is passive, does not give much feedback to the experimentert"s explanations, unmotivated questions if any), 3 - *Neutrality* (subject follows and participates in the discourse, it can not be recognized, if she/he is interested or indifferent in the topic), 4 - *Interest* (subject wants to discuss the topic, closely follows the explanations, asks some questions), 5 - *Curiosity* (strong wish of the subject to talk and learn more about the topic). For automatic processing a fusion of these LOIs 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 3 of 4 annotators: ILA 75%;
  Master LOI := LOI of majority
- Same rating by 2 annotators: ILA 50%
  > If other 2 annotators agree:
  Master LOI := "?" (undefined)
  > If other 2 annotators disagree:
  Master LOI := mean 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 nonverbal interjections have been labeled. These interjections are *breathing*, *confirmation*, *coughing*, *hesitation*, *laughter*, *long pause*, *short pause* and *other human noise*. This additional labeling effort shall demonstrate the potential of such events within higher semantic analysis. Summarized, overall annotation contains subspeaker- and speakerturn segments in msec resolution, spoken content, non-verbals, individual annotator tracks, and Master LOI with confidence in XML-format provided by use of ANVIL [6]. The following table 1 shows the amount of sub-speaker turns (SS-Turns) per master LOI depending on the chosen ILA 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.

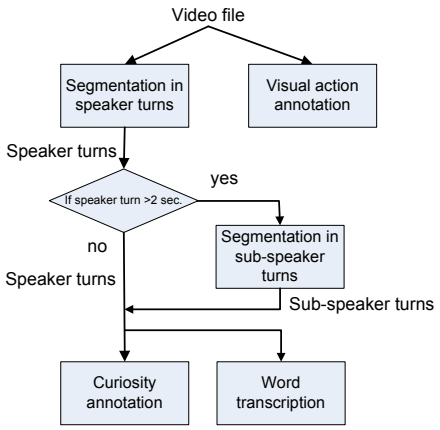

Figure 1: Annotation workflow.

Table 1: Distribution of sub-speaker turns over LOI 1-5 and ILA (I) with confidence (C).

SS-Turns [#]	1	2	3	4	5	"?"
I=50%, $C > 0$	19	383	3602	5386	305	3144
I=50%, $C > 0.5$	19	362	3339	5316	305	3498
I=50%, $C \geq 0.6$	19	261	2832	4603	305	4819
I=75%	19	185	2226	3741	305	6363
I=100%	4	19	417	960	25	11414

As too few items for LOI 1 and 2 have been seen, these were clustered together with LOI 3, and the LOI scale was shifted to LOI 0-2. In order to increase the amount of these low occurence LOIs, further methods of master LOI derivation from the annotator specific LOI will be investigated, if promising for training and evaluation purposes. Overall, the AVIC database is a multimodal data collection of unseen size, quality, realness, and focus, providing un-acted multimodal data for affective computing and especially curiosity detection in human dialogs.

3. SPEECH PROCESSING

3.1 Acoustic Analysis

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 fulfilled. The signal of interest is likewise split into successive 20 ms frames, win-

dowed every 10 ms. In order to obtain a better representation in view of LOI content, feature contours containing information about intonation, intensity, harmonic structure, formants, and spectral development and shape are extracted. In detail these are: pitch based on time-domain calculation by auto-correlation function (ACF), window function normalization and Dynamic Programming (DP) for global cost minimization, energy by frame-based signal-energy computation, formants' 1-5 amplitude, bandwidth, and frequency based on 18 LPC spectrum and DP, Mel-Frequency-Cepstral-Coefficients (MFCC) 1-16, spectral flux, 47 semitone-band interval emphasis and harmonic characteristic based on 1024-point DFT-spectrum, Harmonics-to-Noise Ratio (HNR) based on ACF in the time-domain, window function normalization, shimmer and jitter of periodic parts, and 19 VOC19-coefficients. Secondly, the derivation of speed and acceleration regression coefficients based on these Low-Level-Descriptors (LLD) is fulfilled as further information. By LLD analysis a classification by means of dynamic modeling is already feasible. Yet, basing on our past experience and in accordance with the common practice in the field [10] [4] [8], we decided for a further processing step: In a third stage, statistical functionals f are applied to the LLD in order to project the multivariate time-series F on a static feature vector [7] and thereby become less dependent of the spoken phonetic content:

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

A systematic generation by calculation of moments, extreme values, and further shape characteristics out of each time series on a phrase basis leads to more than 5k features aiming at broad coverage of prosodic, articulatory and speech quality attributes. In detail the 18 functionals are: extreme values, extreme value positions, range, mean, centroid, standard deviation, quartiles, quartile ranges, 95% roll-off-point, Kurtosis, Skewness, and zero-crossing-rate. The idea thereby is not to extract all these features for the actual LOI detection, but to form a broad basis for self-learning feature-space optimization. In first tests on acoustic features a 10-fold stratified cross validation (SCV) is used. Thereby diverse microphone set-ups are considered. These are close-talk (CT), distant-talk (DT), and mixed channel (MC). Features have been reduced by Information Gain Ratio to 1k. As classifier SVM were chosen with a polynomial Kernel-function. 88.1% mean accuracy is observed for CT, 87.8% for DT, and 87.6% for MC within the discrimination of LOI 0 vs. LOI 2. Likewise rather insignificant influence of microphone positioning in the database can be named.

3.2 Linguistic Analysis

Beyond the analysis of acoustic properties of a speech signal, also the spoken content may carry cues in respect of a speaker's interest, and the combination of both analyses could be shown highly effective in past related works in the field of speech-based emotion recognition. [10] [11]. The precondition of linguistic analysis is to obtain the spoken content out of an audio-file. Within this work once manual annotations have been employed to obtain an impression of performance under idealistic speech recognition conditions, and once a state-of-the-art MFCC and HMM-based tri-phone Large-Vocabulary Continuous Speech Recognition (LVCSR) engine was used. For linguistic analysis a vector-space-representation popular in the field of document retrieval known as Bag-of-Words (BOW) has been chosen [5]. The motivation here fore is the effective fusibility of obtained linguistic features within the acoustic features and later video-based ones on an early level [11]. Likewise, loss of information is postponed to the final decision process allowing for the utmost decision basis. A term w_i within a phrase $\Sigma = \{w_1, \cdots, w_i, \cdots, w_S\}$ with $S = |\Sigma|$ is thereby projected onto a numeric attribute $x_i : w_i \rightarrow x_i$. 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. Throughout feature extraction a value for each term in Θ is calculated: Either 0 in case of no occurrence in the actual phrase, or 1 in case of a binary attribute's type, respectively its term frequency of occurrence (TF) for common BOW representation. A number of further refinement approaches exist as normalization to the phrase 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 to the named alternatives:

$$x_{logTF,i} = \log\left(c + \frac{TF(w_i, \Sigma)}{|\Sigma|}\right) \tag{2}$$

A drawback of this modeling technique is the lack of word order consideration. Still, great flexibility is obtained in comparison to e.g. class-based (back-off) N-Grams. In general, vocabularies will show too high a dimensionality ($> 1k$ terms) and contain many redundancy in view of the aimed at LOI detection. Similar to acoustic feature reduction as described in the next subsection, 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 realized 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. We decided for an Iterated Lovins Stemmer (ILS), here fore, which bases on context-sensitive longest match stemming - a slight enhancement of the very traditional approach to stemming. Table 2 shows the 18 most relevant lexemes after ILS stemming and IGR-FS stopping. The final vocabulary size thereby is 639 lexemes instead of 1,485 terms. Using linguistic features only, maximum mean accuracy within 10-fold SCV, optimal feature type and SVM reached 79.4% accuracy for the full blown LOI analysis based on annotation and 84.2% for discrimination of LOI 0 and LOI 2. Using LVCSR a drop to 69.8% is observed for LOI 0-2. 29.1% of the phrases led to no LVCSR output, as these sub-speaker-turns only consist of non-verbals or are too short. Still, no recognition was used as extra information. One of the main differences of annotation versus LVCSR thereby is the included annotation of non-verbals, interjections or events as described within the database section (sec. 2). While the table above shows the high ranking of four of these events (in italics, as de-

scribed in database section) on the ranks 1, 2, 8, and 9, a reduction of all such only led to an absolute accuracy drop of 1.9% having the same setup as described earlier: 10-fold SCV, SVM and LOI 0-2. Still, this might be of interest considering their automatic recognition within future work. Within linguistic experiments test-runs employing actual LVCSR and annotation-based runs have been fulfilled. Firstly, table 3 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 misrecognized and matched on diverse further terms in the engine vocabulary. Otherwise, this diffusion by word errors also leads to fewer observations of the same terms: Already at a minimum TF of 2 within the database the annotation based level overtakes. Yet, BOW relies on high TF within a data-set. This can partly be repaired by stemming - assuming that phonetic mismatches lead to confusions within a lexeme.

Table 2: Top 18 lexemes after stemming and IGR-FS based stopping. Stems are marked by *.

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

Table 3: Term numbers at diverse minimum TF levels, annotation-based (left) and LVCSR-based (right).

Min. TF	Annotation Terms [#]	LVCSR Terms [#]
1	1,485	1,568
2	645	351
5	277	109
10	149	51
20	98	20
50	48	8

3.3 Feature Space Optimization

So far, we extracted acoustic features, as described in sec. 3.1, and linguistic features as described in sec. 3.2. These are combined in one feature space by simple construction of an acoustic-linguistic super-vector. In order to save extraction effort and likewise reduce high complexity throughout the succeeding classification process, features of high individual information are pre-selected by fast Information Gain Ratio (IGR) calculation and feature selection (FS) of attributes with high IGR (IGR-FS). This method bases on Shannon's information and is suited to find features of high individual relevance. Yet, redundancy of correlated, yet individually effective ones is not filtered thereby.

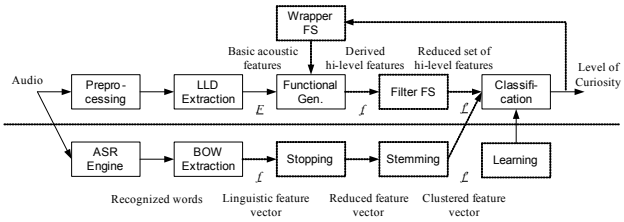

Figure 2: Overview early fusion of acoustic and linguistic analyses.

The reduced set, obtained by elimination of all zero IGR features, is de-correlated and further compressed by application of Sequential Forward Floating Search (SFFS) [12]. This leads to an optimal set as a whole and the overall minimum number of features [11]. SFFS employs a classifier's accuracy, ideally the target one, as optimization criterion. Herein, powerful Support-Vector-Machines (SVM) are used to ensure high quality throughout selection (SVM-SFFS). 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. In figure 2 an overview over the combined acoustic and linguistic processing is depicted. As final classifier SVM with a polynomial Kernel and pair-wise multi-class discrimination is chosen.

4. ACTIVE-APPEARANCE-MODELS

4.1 Introduction

Active Appearance Models (AAM) are statistical models derived from example images of an object class [2], i.e. faces in this case. AAMs 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 synthesized by a linear combination of the main components. In the application phase of an AAM, the coefficients of the linear combination have to be optimized with respect to a maximal similarity between the original object and the artificial object appearance, synthesized by the AAM. These optimized coefficients constitute a precise representation of the analyzed face and contain the relevant information about the face properties such as facial expressions and head pose.

4.2 Data Preparation

The statistical analysis via PCA 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 $\mathbf{p_i} \in \mathcal{P}$ with $0 \leq i < p$ have to be manually annotated, producing a set of p corresponding landmark vectors $\mathbf{s_i} \in \mathcal{S}$ with $\mathbf{s_i}$ being the ith landmark vector defined as the

concatenation of all landmark coordinates

$$\mathbf{s_i} = (x_0, y_0, x_1, y_1, \ldots, x_{(n/2)-1}, y_{(n/2)-1})^T \quad (3)$$

See figure 3 for an example annotation with $n/2 = 72$ landmarks. These shape vectors are arranged in the *shape matrix*

$$\mathbf{S} = [\mathbf{s_0} \mid \mathbf{s_1} \mid \cdots \mid \mathbf{s_{p-1}}] \in \mathcal{R}^{n \times p} \quad (4)$$

Additionally the *mean shape* $\overline{\mathbf{s}}$ is defined as the mean of all shape vectors in $\mathbf{S}$. The shapes are aligned and normalized to each other in order to remove Euclidean transformations and scaling and minimizing the variance to the deformation of the shapes. The texture within the annotated shape of each training image is warped to fit the mean shape $\overline{\mathbf{s}}$. For generation of the texture model, we store the obtained set of textures $\mathbf{t_i} \in \mathcal{T}$ as vectors column-wisely in the *texture matrix*

$$\mathbf{T} = [\mathbf{t_0} \mid \mathbf{t_1} \mid \cdots \mid \mathbf{t_{p-1}}] \in \mathcal{R}^{c \times p} \quad (5)$$

where c is the total number of pixels in each texture multiplied by the number of channels χ. For grayscale textures $\chi = 1$, for interleaved RGB color textures $\chi = 3$. It is suggested to eliminate the texture variance caused by brightness and contrast disparities in the images. Further let $\overline{\mathbf{t}}$ be defined as the mean of all textures in $\mathbf{T}$.

4.3 AAM Generation

The first step of building an AAM is the independent application of a Principal Component Analysis (PCA) to the aligned and normalized 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 [3].

4.3.1 Shape Model

The *shape model* is built by applying a (PCA) 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. 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

$$\frac{\sum_{i=0}^{r_s-1} \lambda_{\mathbf{s}i}}{\sum_{i=0}^{n-1} \lambda_{\mathbf{s}i}} \overset{!}{\geq} 0.98 \quad (6)$$

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

$$\mathbf{s} = \overline{\mathbf{s}} + \mathbf{W_s}\mathbf{h_s} \quad (7)$$

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 $\overline{\mathbf{s}}$. 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

$$\mathbf{h_{s}}_i \approx \mathbf{W_s^T}(\mathbf{s_i} - \overline{\mathbf{s}}) \quad (8)$$

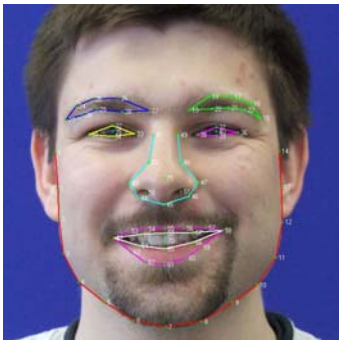

Figure 3: Two-dimensional annotation of a face with 72 landmarks

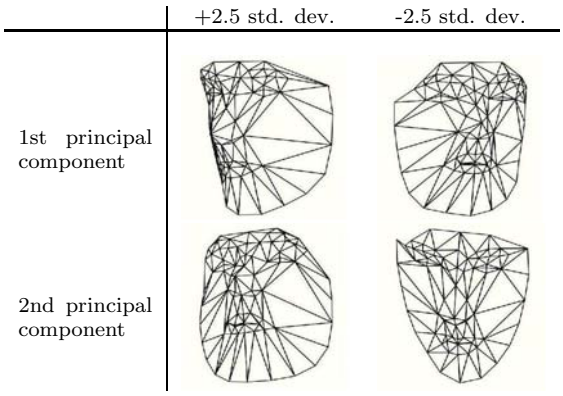

Figure 4: Effect of the first shape model components

4.3.2 Texture Model

The *texture model* is built by applying a PCA to the texture matrix $\mathbf{T}$, resulting in the *texture basis* $\mathbf{W_t}$, whereas basis vectors are sorted in descending order of the corresponding Eigenvalues $\lambda_{\mathbf{t}i}$. Again, the first r_t are used while the "least important" basis vectors of $\mathbf{W_t}$ are discarded. A new texture $\mathbf{t}$ can be synthesized in the shape-free space by

$$\mathbf{t} = \overline{\mathbf{t}} + \mathbf{W_t}\mathbf{h_t} \quad (9)$$

with $\mathbf{h_t}$ containing the *texture coefficients* used to deform the texture model. Note that a zero coefficient vector relates to the mean texture $\overline{\mathbf{t}}$. As $\mathbf{W_t}$ defines an orthonormal basis, the new representation of the known texture $\mathbf{t_i}$ can be obtained by

$$\mathbf{h_{t}}_i \approx \mathbf{W_t^T}(\mathbf{t_i} - \overline{\mathbf{t}}) \quad (10)$$

4.3.3 Combined Model

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 $0 \leq i < p$ training samples

$$\mathbf{c_i} = \begin{pmatrix} \mathbf{E}\mathbf{h_{si}} \\ \mathbf{h_{ti}} \end{pmatrix} \quad (11)$$

$\mathbf{E}$ is a diagonal matrix of reasonable weights to equalize unit differences between the shape and the texture model. As Cootes and Taylor [3] suggest, a simple approach is to set $\mathbf{E} = q\mathbf{I}$ where q is the ratio of the total intensity variation of the textures to the total shape variation. The vectors $\mathbf{c_i}$ form the matrix of concatenated coefficient vectors $\mathbf{C} = [\mathbf{c_0} \mid \ldots \mid \mathbf{c_{p-1}}]$ which can be written as

$$\mathbf{C} = \begin{bmatrix} \mathbf{E}\mathbf{W_s}^T[\mathbf{S} - \overline{\mathbf{s}}\mathbf{1}^T] \\ \mathbf{W_t}^T[\mathbf{T} - \overline{\mathbf{t}}\mathbf{1}^T] \end{bmatrix} \tag{12}$$

where $\mathbf{1}^T$ is the vector containing all ones and $\mathbf{1} \in \mathcal{R}^n$ or $\mathbf{1} \in \mathcal{R}^c$ respectively. Since the shape coefficients $\mathbf{h_{si}}$ and texture coefficients $\mathbf{h_{ti}}$ are already mean-free, so is $\mathbf{C}$. Another PCA is applied to the matrix $\mathbf{C}$ producing the *combined basis* $\mathbf{W_c}$, whereas basis vectors are sorted in descending order of the corresponding Eigenvalue $\lambda_{\mathbf{c}i}$, again discarding the "least important" basis vectors. A coefficient vector $\mathbf{c}$ can be synthesized by evaluating

$$\mathbf{c} = \mathbf{W_c}\mathbf{h_c} \tag{13}$$

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 $\mathbf{W_{ct}}$

$$\mathbf{W_c} = \begin{bmatrix} \mathbf{W_{cs}} \\ \mathbf{W_{ct}} \end{bmatrix} \tag{14}$$

it is possible to express a new shape $\mathbf{s}$ and texture $\mathbf{t}$ directly as function of $\mathbf{h_c}$ by combining eq. 12 with eq. 7 and 9 which finally leads to these synthesis rules for a shape and a corresponding texture:

$$\mathbf{s} = \overline{\mathbf{s}} + \mathbf{Q_s}\mathbf{h_c} \quad , \quad \mathbf{Q_s} = \mathbf{W_s}\mathbf{E}^{-1}\mathbf{W_{cs}} \tag{15}$$

$$\mathbf{t} = \overline{\mathbf{t}} + \mathbf{Q_t}\mathbf{h_c} \quad , \quad \mathbf{Q_t} = \mathbf{W_t}\mathbf{W_{ct}} \tag{16}$$

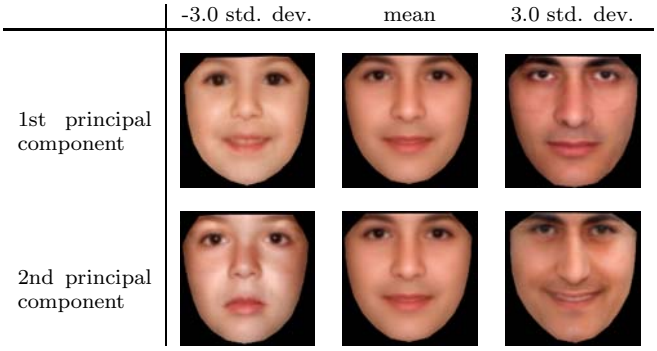

Figure 5: Effect of the first two combined model components of an AAM

4.4 AAM coefficient optimization

The AAM coefficient optimization can roughly be understood as a standard multi-variate optimization problem with the goal to minimize the energy of the difference image $\mathbf{r}(\mathbf{v})$ between this synthesized face and the currently analyzed 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 runtime optimized Gauss-Newton gradient descent method by an offline gradient prediction is applied [2]. Therefore the following steps have to be conducted: definition of an error energy function $E(\mathbf{r}(\mathbf{v}))$, estimation of the Jacobian $\mathbf{J} = \frac{\partial \mathbf{r}}{\partial \mathbf{v}}$ of the difference function $\mathbf{r}(\mathbf{v})$, 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

$$\mathbf{v}(i+1) = \mathbf{v}^{(i)} - \alpha\mathbf{R}\mathbf{r}(\mathbf{v}^{(i)}) \tag{17}$$

using the step width α. The algorithm terminates when $E(\mathbf{r}(\mathbf{v}))$ does not further decrease between iterations. The final value of the error energy serves as confidence measure for the performed AAM analysis. In order to map the AAM results on sub-speaker turn basis, only the coefficient vector of the video frame with the lowest final $E(\mathbf{r}(\mathbf{v}))$ is added to the feature space of the early fusion with the other modalities.

5. ACTIVITY ESTIMATION

Apart from facial expressions, which are addressed by the Active Appearance Model analysis, the level of Activity is considered, herein, as a criterion for the description of the emotional state of a person. In the scenario of the AVIC database the activity shall be estimated by a compact description of the body-, and especially head-movements of the subject over a short video sequence. Since skin-color or Viola-Jones based head localization provides rather rough information about the position and the size of the person's head, we utilize the optimized performance of our eye localization algorithms. 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, homogenous motion is perceived as less active than heterogenous 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 4 lists all examined measures of activity.

Table 4: Features for the estimation of activity.

Index	Description
0-2	eye position delta (maximum, max. x, max. y)
4-6	eye position delta (mean, mean x, mean y)
8-10	eye position delta (variance, var. x, var. y)
3,7,11	eye distance delta (max., mean, var.)
12	eye position delta (rel. # frames > threshold)

Since head and eye position data is derived from the preceding automatic localization tasks and 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 metadata for each region of interest (ROI) Type) is less or equal to 0 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-Localization 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 ROI provided by the head localizer.
The activity vector should give a quantitative statement for the head-motion in closeup views. In the next step the activity vector was used to recognize the level of interest (LOI) of the analyzed person. It is supposable, that there exists a strong correlation between this two values. Since finally a multi-modal fusion of AAM analysis, activity estimation, and speech analysis is planned, we assume that the single activity features contribute to a improved LOI recognition performance. Best classification results were obtained with the following configuration: SVM with polynomial kernel with exponent 5 [1]. A feature selection (feature indices 8,12 and 9,10 left out) caused a small but significant increase of performance.

6. MULTIMODAL FUSION

Two main types of multi-modal fusion exist: early- and late-fusion. Both types combine different modalities of data. In the case of late-fusion, a classification is performed for each modality separately. The results of each modality are fused to a final class-prediction accuracy. During early fusion, the feature spaces of all modalities are merged into one feature space. This space is classified within a single classification process. In the evaluation of LOI recognition on the AVIC database, we focus on early fusion, as it saves all available information for the final decision process, allows for combined feature space optimization, and due to the highly unbalanced datasets. For the late-fusion, this latter problem occurs twice: first, during the training of the classifiers for each information stream, and second, during the fusion of the class-prediction accuracy of the modality classifiers to the final LOI.
All four feature groups introduced in sections sec. 3, sec. 4, and sec. 5, namely acoustic, linguistic, AAM, and activity features, were intentionally projected onto the sub-speaker-turn level. This was realized by multivariate time-series analysis for acoustic and activity 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 synchronization effort is needed at this point, and fusion is realized by a simple super-vector construction.
Next, we present a number of experimental results for diverse multimodal setups. Linguistic analysis bases on LVCSR output in the ongoing, and is always handled together with acoustic features as *Audio*. For testing, the AVIC database was split into 3 stratified and subject disjunct test and training sets: as 21 subjects are contained, 7 subjects were used per test set, and accordingly 14 per training set. No subject belongs to more than one set during one run. Likewise, a 3-fold subject independent SCV (SI-SCV) is performed. Table 5 shows the number of sub-speaker turns for each class

[#]	**0**	**1**	**2**	**Overall**
before fusion				
Test set 1	771	951	41	1703
Test set 2	1086	1665	83	2834
Test set 3	860	1348	124	2332
Training set 1	1946	3013	207	5166
Training set 2	1571	2299	165	4035
Training set 3	1797	2616	124	4537
after fusion				
Test set 1	130	404	27	561
Test set 2	116	663	44	823
Test set 3	67	686	99	852
Training set 1	183	1349	143	1675
Training set 2	197	1090	126	1413
Training set 3	246	1067	71	1384

Table 5: Number of sub-speaker turns for each LOI in the different data sets, database AVIC.

before and after fusion. As not all modalities are present at a time (here: often no speech, especially in the case of boredom, that is LOI 0), only the considerably lower number of instances after fusion can be used for multimodal evaluation in the ongoing. However, note that a real-life system profits from multimodality also with respect to such 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 random down-sampling. In tables 6 and 7 different feature spaces are evaluated. Table 6 shows the recall-rates of each class for training performed on unbalanced data. Note that in every training set, LOI 1 is pre-dominant. Table 7 shows the recall-rates of each class for training performed on uniformly distributed data by random down-sampling. SVM with a polynomial kernel in constant parameterization are thereby used throughout. As can be seen, use of

Accuracy [%]	**0**	**1**	**2**	**RR**	**CL**	**F**
unbalanced						
Full space	47.3	87.2	36.8	77.7	57.1	65.8
Audio+Acti.	45.4	93.0	32.9	**81.8**	**57.1**	**67.3**
Audio+AAM	46.5	88.1	39.5	78.6	58.0	66.7
AAM+Acti.	22.7	92.0	00.9	75.4	38.5	51.2

Table 6: Results early fusion. Recall values for the reduced set LOI 0-2, overall mean - weighted by instance number (RR) and non-weighted (CL), and the harmonic mean F=2 RR CL/(RR+CL). Unbalanced training. SVM, FS, database AVIC, 3-fold SI-SCV.

balanced training sets leads to a significantly more satisfying result with respect to balance among recall rates. Looking at table 7 to find the best possible combination of the

available modalities, the combination of audio and activity features after individual pre-selection of features prevails. This super-vector has a size of 109. When AAM feature information is fused with the audio features and the activity features, the recall value of all classes decreases. This comes, as complexity for the classifier is raised without provision of significantly novel and valuable information. The best single modality is clearly audio. Yet, all possible combinations of modalities do not satisfyingly solve the problem of LOI 0 and LOI 2 being discriminated more easily than each one from LOI 1. However, in many applications a discrimination of boredom vs. interest may be sufficient. The confusion matrix of the best result of early fusion of audio and activity is shown in table 8, to further illustrate this problem.

Accuracy [%]	**0**	**1**	**2**	**RR**	**CL**	**F**
uniformly distributed training						
Full space	75.3	55.5	59.3	58.6	63.4	60.9
Audio+Acti.	79.2	60.2	73.5	**63.9**	**71.0**	**67.3**
Audio+AAM	71.5	56.8	60.6	59.1	63.0	60.1
AAM+Acti.	66.4	37.3	41.9	41.7	48.5	44.8
Audio	70.0	59.4	75.3	62.1	68.2	65.0
Activity	58.0	35.7	38.9	39.1	44.2	41.5
AAM	71.1	20.8	65.4	31.2	52.4	39.1

Table 7: Results early fusion. Recall values for the reduced set LOI 0-2, overall mean - weighted by instance number (RR) and non-weighted (CL), and the harmonic mean F=2 RR CL/(RR+CL). Training uniformly distributed. SVM, FS, database AVIC, 3-fold SI-SCV.

[%] classified as >	**0**	**1**	**2**
early fusion			
0	**79.2**	16.5	4.3
1	21.7	**60.2**	18.1
2	3.3	23.2	**73.5**

Table 8: Confusion matrix of the best early fusion (audio and activity) using the reduced LOI set 0-2. Training uniformely distributed. SVM, FS, database AVIC, 3-fold SI-SCV.

7. DISCUSSION AND CONCLUSIONS

The summary of the results of sec. 6 leads to this five key points: early fusion seems to be a promising approach with respect to high accuracies and combined feature space optimization. The training data must be uniformly distributed and random sampling seems a reasonable solution. Best single modality is audio by a combination of acoustic and linguistic feature information, which can be processed in real-time on a state-of-the art desktop. Combination of the modality audio and the information stream activity achieves better results than the single modalities, and is still real-time capable. Yet, Active-Appearance-Models could not help to further increase the accuracy. Overall, spontaneous interest could be detected subject independently in human conversation by the proposed audiovisual processing. Also, an early feature level fusion of acoustic, but as special novum also linguistic features with vision-based features and fully automatic processing could be demonstrated. Yet, future works will have to deal with improved discrimination of the subtle difference of the border class between strong interest and boredom. Further interesting topics will be investigation of individual segmentation for each modality in an asynchronous manner and shifting from a classification to a regression approach.

8. REFERENCES

[1] R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. Technical Report 02-46, IDIAP, 2002.

[2] T. Cootes, G. Edwards, and C. Taylor. A comparative evaluation of active appearance model algorithms. In P. Lewis and M. Nixon, editors, *Proceedings of British Machine Vison Conference*, volume 2, pages 680–689, Sept 1998.

[3] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, UK, Mar. 2004.

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

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

[6] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proc. ISCA EUROSPEECH 2001*, pages 1367–1370, 2001.

[7] I. Mierswa. Automatic feature extraction from large time series. In *Proceedings of the 28th Annual Conference of the GfKl 2004*, pages 600–607. Springer, 2004.

[8] M. Pantic and L. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proccedings of the IEEE*, 91:1370–1390, September 2003.

[9] P. Qvarfordt, D. Beymer, and S. X. Zhai. Realtourist - a study of augmenting human-human and human-computer dialogue with eye-gaze overlay. In *INTERACT 2005*, volume LNCS 3585, pages 767–780, 2005.

[10] B. Schuller, M. Ablaßmeier, R. Müller, S. Reifinger, T. Poitschke, and G. Rigoll. Speech communication and multimodal interfaces. In K.-F. Kraiss, editor, *Advanced Man Machine Interaction*, pages 141–190. Springer, Berlin, Heidelberg, 2006.

[11] B. Schuller, R. Mueller, M. Lang, and G. Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proc. Interspeech 2005*, Lisbon, Portugal, 2005. ISCA.

[12] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition, 2005.