The acoustics of eye contact: detecting visual attention from conversational audio cues

Florian Eyben, Felix Weninger, Lucas Paletta, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Eyben, Florian, Felix Weninger, Lucas Paletta, and Björn Schuller. 2013. "The acoustics of eye contact: detecting visual attention from conversational audio cues." In *Proceedings of the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction - GazeIn '13, December 2013, Sydney, Australia*, edited by Roman Bednarik, Hung-Hsuan Huang, Kristiina Jokinen, and Yukiko I. Nakano, 7–12. New York, NY: ACM Press. https://doi.org/10.1145/2535948.2535949.

The Acoustics of Eye Contact – Detecting Visual Attention from Conversational Audio Cues

Florian Eyben
Machine Intelligence and Signal Processing, MMK
Technische Universität München, GERMANY
eyben@tum.de

Felix Weninger
Machine Intelligence and Signal Processing, MMK
Technische Universität München, GERMANY
weninger@tum.de

Lucas Paletta
Joanneum Research
Graz, AUSTRIA
lucas.paletta@joanneum.at

Björn Schuller*
Joanneum Research
Graz, AUSTRIA
schuller@IEEE.org

ABSTRACT

An important aspect in short dialogues is attention as is manifested by eye-contact between subjects. In this study we provide a first analysis whether such visual attention is evident in the acoustic properties of a speaker's voice. We thereby introduce the multi-modal GRAS^2 corpus, which was recorded for analysing attention in human-to-human interactions of short daily-life interactions with strangers in public places in Graz, Austria. Recordings of four test subjects equipped with eye tracking glasses, three audio recording devices, and motion sensors are contained in the corpus. We describe how we robustly identify speech segments from the subjects and other people in an unsupervised manner from multi-channel recordings. We then discuss correlations between the acoustics of the voice in these segments and the point of visual attention of the subjects. A significant relation between the acoustic features and the distance between the point of view and the eye region of the dialogue partner is found. Further, we show that automatic classification of binary decision eye-contact vs. no eye-contact from acoustic features alone is feasible with an Unweighted Average Recall of up to 70%.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Human Factors, Measurement

Keywords

Eye-gaze, attention, visual, acoustic

1. INTRODUCTION

An important aspect in short person to person dialogues is attention as is manifested by eye-contact between subjects. Thus, to replicate human-like behaviour for artificial systems (e. g., humanoid robots or virtual agents) it is believed to be highly important to implement natural patterns of eye contact [4]. Furthermore, consistency between eye contact and acoustic cues emitted by a system, e. g., by means of speech synthesis, should be ensured.

In this study we verify the correlation between visual attention and acoustic cues of a speaker's voice in human to human dialogues. Such information could be used in low-resource, or speech only systems which do not have a camera or eye-tracking device available, e. g., in voice conversations and chats the other partner could be informed about the eye-gaze behaviour of the first partner without actually seeing him/her. Also in forensic analysis these methods could be applied. Further, psychological studies may profit from such knowledge.

The paper is structured as follows: first, we introduce and describe the GRAS^2 database in Section 2. It contains multi-sensor recordings of subjects engaged in a real-life short dialogue (typically one short question and a short answer). Then, we discuss automatic identification of speech segments in the continuous recordings and labelling of them as speech from the test subject (referred to as subject in the ongoing) and speech from dialogue partners (referred to as partner in the ongoing) in Section 4. Next, we present an analysis of the correlations of acoustic features with the eye contact between subject and partner in Section 5 as well as results of experiments where we try to predict whether the subject is looking at the partner's eyes/face or somewhere else just from the acoustics of his/her voice in Section 6. We summarise our findings in Section 7.

*Björn Schuller is also affiliated with the Department of Computing, Imperial College London, U. K.

GazeIn'13, December 13, 2013, Sydney, NSW, Australia

978-1-4503-2563-9/13/12 ...$15.00.
http://dx.doi.org/10.1145/2535948.2535949.

2. THE GRAS^2 DATABASE

The Graz Real-Life Affect in the Street & Supermarket (GRAS^2) corpus is – to the authors' best knowledge – the first database of visual attention recordings with multiple audiovisual, physiological, and movement sensory cues in real-life conversations.

Four subjects took part in the recordings (3 female, 1 male, cf. Figure 1). These were all native Austrian students and they filled a BFI-11 personality questionnaire [7]. The male subject usually wears glasses, the female subjects did not wear glasses.

2.1 Recording Devices

All four subjects were equipped with SMI Eye Tracking Glasses able to record both the eyes of the person wearing these and what the person is looking at (static frontal camera, not affected by eye tracker result). They allow for precise measurement of visual attention focus (30 Hz binocular with automatic parallax compensation; pupil/CR by dark pupil tracking, spatial resolution 0.1^o, gaze position accuracy 0.5^o over all distances from 40 cm to infinity with a gaze tracking range of 80^o horizontal and 60^o vertical) in the simultaneously recorded field of vision (recorded in HD 1 280 x 960 pixels at 24 fps compressed with the H.264 codec; viewing field of 60^o horizontal and 46^o vertical). They also feature a monophonic microphone on the left earpiece of the glasses that records in 16 kHz, 16 bit.

The recording of the data from these glasses was carried out on an SMI-ETG laptop (1.3 kg) worn in a backpack. The USB-Cable connection was hidden under hair and clothing as much as possible. Further, subjects were equipped with the Affectiva Q Sensor 2.0, a wearable sensor that measures Electro-Dermal-Activity (EDA) and skin temperature [6] to capture indication on arousal during attention. It was worn on the left hand to resemble a watch in appearance.

To record additional audio data without particularly demanding hardware conditions, an Android smart phone Samsung Galaxy Nexus was used similar as in [8]: The phone was loosely located in a front-pocket of a shirt worn by the subjects. The standard media recording APIs of Android use an AMR codec with poor quality. Therefore, the audio stream was accessed directly and saved uncompressed at 44.1 kHz, 16 bit. The recording component was implemented as a service and thus could run in the background with the phone locked and the screen off. In addition to this, limited motion sensing on the phone is available through an InvenSense MPU-3050 accelerometer unit. This sensor contains a MEMS accelerometer and a gyroscope. Linear and angular accelerations can be captured at a sampling rate of up to 100 Hz. Since the audio recording already puts the processor under considerable load, the actual achievable sampling rate was 10–20 Hz. The MARIA application [5] for public transport guidance was adjusted in a way to log the audio alongside acceleration data from the motion sensor. In addition, a Zoom H2 four-channel recording device recording at 48 kHz, 16 bit was worn on the for high quality audio. Finally, a secondary accelerometer sensor was worn in the backpack: It was contained in the NAVIN Mini Homer GPS tracker that was operated at a sample rate of 0.2 Hz. With this setup, 2-way video, 6 channel audio, EDA, temperature, and twice 3D motion is measured from the subjects. However, in the ongoing only video and audio from the eye tracker glasses and the smart phone will be used.

2.2 Recording Protocol

The subjects were accompanied by a supervisor (27 years, male) who helped with the setup and monitored the progress from a distance. The equipment setup and 3-point calibration was carried out on a parking lot of the Citypark in Graz/Austria. Three-times hand-clapping looking at the hands is used as anchor point for synchronization between those units that are not directly connected. The subjects had to search three stores as a first 'warm-up' task (Le Clou Jewelry, Oxyd fashion store, and the Golden Sun Solarium) to familiarize with the worn equipment that was hidden as much as possible (cf. Figure 1).

The recordings of interest then took place in the Inter-SPAR supermarket, where the subjects engaged in dyadic discourse exclusively with female persons shopping in this supermarket. These are referred to as (dialogue) "partners" in the ongoing as opposed to the knowingly involved and equipped four "subjects". These had no knowledge at first that they were part of the study – 28 persons agreed to provide their recorded audiovisual-footage for scientific purpose (cf. examples in Figure 1, two bottom rows) – data of subjects not agreeing was deleted by the recording subject. The limitation to female subjects was decided upon to reduce gender effects. Further, permission from the site-holders was given to carry out the recordings and use the material.

The subjects followed a study protocol as follows to engage in discussion in German language (Graz-region Styrian dialect) with subjects: They needed to search for *Sauerkraut* and a Swiss chocolate drink (*Ovomaltine*, or US: *Ovaltine*), ask for a SPAR chocolate, a specific Calculator available in the supermarket, a "typical Austrian product", Turkish Ayran, denture adhesive for third teeth, and anti-athlete's foot cream. Thereby, they stuck with one dialogue partner as long as she was willing to help. Subsequently, they immediately asked for written consent explaining the experiment which was also recorded and usually consumed the larger partition of the time. This included a questionnaire on the demographics of the dialogue partners.

The choice of items to ask for and the sequential asking for continued help as well as the surprising revealing of them being recorded in an experiment are intended to elicit a range of affect including besides neutral also joyful, uncertain, surprised, confused, and negative emotional behaviour in diverse real-life blend.

This was further benefited by the condition that the subjects addressed their dialogue partners with the second personal pronoun "Du" (you) as usually used with friends and familiar persons as opposed to the formal and polite German "Sie" (also translates as you in English–but usually equivalent to addressing a person with the last name only). In the questionnaire, five dialogue partners would usually prefer the casual form "Du", four the formal "Sie", two would not have cared, and 17 made no statement.

The age range of the dialogue partners that agreed is as follows (in years): 18–25 (3x), 26–35 (2x), 36–45 (4x), 46–55 (6x), 56–65 (4x), and no mention (9x). The four subjects which carried out the recordings are referred to as subjects A, B, C, and D in the ongoing, where A is the male subject and the other three are female subjects.

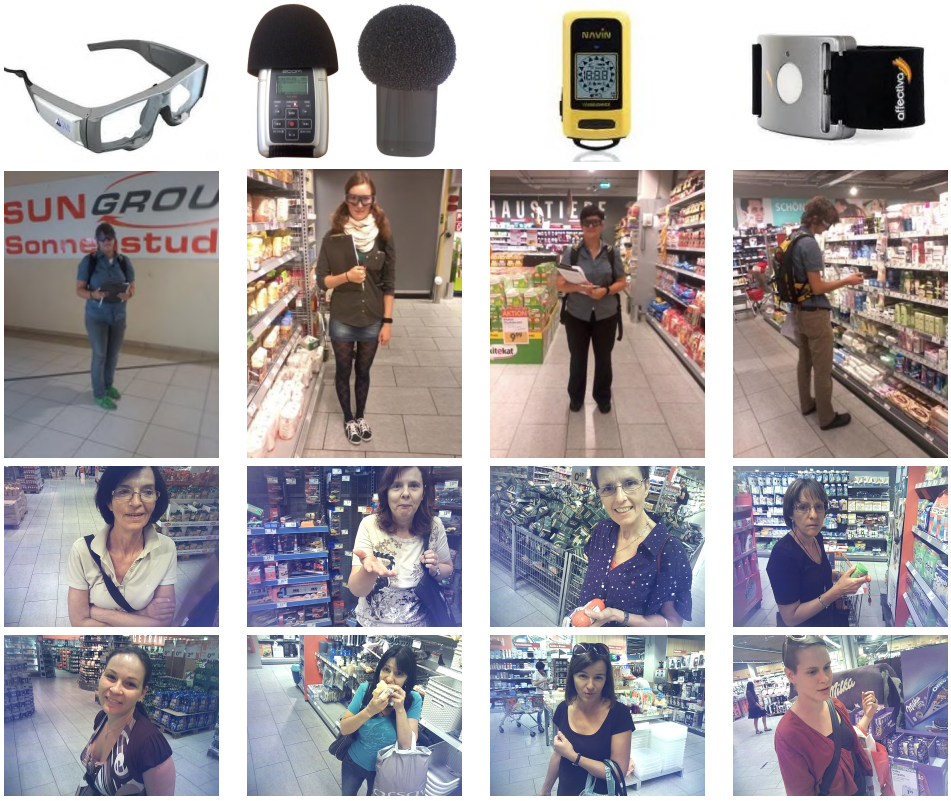

Figure 1: Top-most row: Recording equipment worn by the subjects (from left to right: eye tracking glasses, audio recorder and smart phone, GPS tracker (used for extra accelerometer in backpack), EDA sensor (Affectiva) – details in the text). Second row: The four participating subjects as equipped on site. Bottom two rows: Examples of recorded dialogue partners as seen by the four subjects through their worn eye tracking glasses.

3. AUDIO TRACK ALIGNMENT

The audio tracks from the eye tracker and the smart phone (and also the Zoom H2 recorder – however, this is not used here) needed to be aligned for three reasons: a) the start time of the recordings differs, as recording on the devices was started sequentially by hand, b) the sampling clocks of the devices were not synchronized and thus drifted significantly over the course of a one hour recording, and c) the recording app on the mobile phone occasionally dropped audio frames at random locations – presumably due to high system load – of more than 1 second. To be able to process as much audio data as possible and ideally have all audio tracks aligned as perfectly as possible at every time instant, we used an automatic alignment algorithm. This algorithm is capable of aligning the audio tracks completely unsupervised. We consider two tracks, where one is referred to as the master track, and the other as slave track. The goal is to align the slave track to the master track. The master track is not modified in any way. The algorithm consists of three steps:

1. Finding the initial displacement of the tracks at the beginning of the recording
2. Finding frame drops and sudden misalignments within the recording
3. Estimating sample-wise displacements (compensating drift of sampling clocks).

Once the displacement values for each sample are known, spline interpolation is applied on a sample level to align the slave track to the master track. The initial displacement is estimated via a window based cross-correlation search, which finds the position of the first 8 seconds of master audio in the slave signal. The slave audio before this position is truncated before the other two steps are executed. In step (2) a large sliding window (8 s) is used for cross-correlations.

The windows of the master signal are sampled at a constant rate of 8 s, while the corresponding windows in the slave signal are dynamically shifted by the current displacement, which is initially 0 for the first window, and for the second window equal to the displacement found by cross-correlation of the first window of master and slave, etc. Due to the large window, discontinuities caused by frame drops of up to 2 seconds in both directions can be robustly detected, which is sufficient for the GRAS^2 corpus. As an example, the result of the displacement analysis between the eye tracker's audio (master) and the Phone's audio (slave) is shown in Figure 2 for subject A. In step (3) the locations of the frame drops causing jumps in the track displacement function (Figure 2) are estimated with a better temporal resolution by a smaller search window (0.25 s). Next, the accuracy of the small drift occurring by the sample clock de-synchronization is refined with the same 0.25 s search windows in regions where no jump occurs. The lag of the cross-correlation thereby is constrained by the upper and

Subject	A (m)	B (f)	C (f)	D (f)
Recording duration [min]	85	61	81	67
# Subject speech segments	611	480	566	329
Subject speech segments duration [min]	20	14	20	13
% segments with face present	91.2	92.5	95.1	95.1
Duration face detected [min]	9.5	7.1	11.1	7.9
% segments with eye – eye view (V_c)	10.3	5.6.	3.0	13.7
% segments with eye – face view (V_{b+c})	47.5	27.3	24.9	37.1
% segments with eye – near face (V_{a+b+c})	64.0	56.9	48.6	58.4
Per turn mean length of case V_c [s]	.39	.11	.11	.16
Per turn mean length of case V_b [s]	.57	.21	.20	.48
Per turn mean length of case V_a [s]	.30	.23	.24	.34
Mean speech segment length [s]	2.0	1.8	2.1	2.3
Max speech segment length [s]	15.6	12.5	14.5	19.0
Std. dev. of speech segment length [s]	1.8	1.7	1.9	2.4
Energy difference (M-S) bias	.006	-.012	.007	.022

Table 1: Data statistics for the four subjects A-D (1 male (m), 3 female (f)). The last row gives energy difference threshold between two recording microphones which was used for the automatic segmentation of subject utterances (see text on automatic segmentation for details).

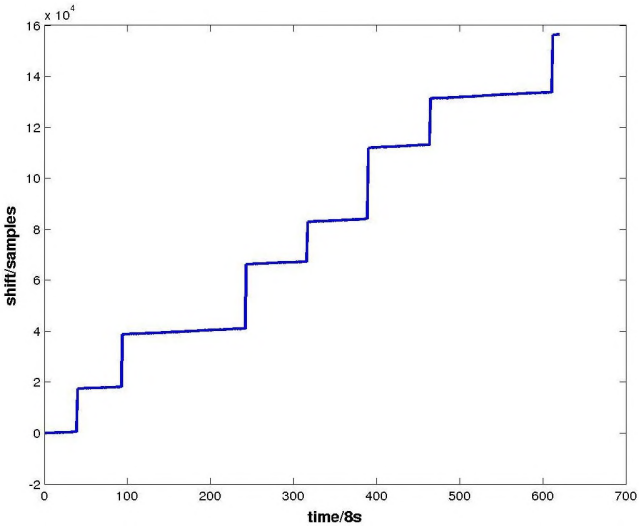

Figure 2: Displacement between eye tracker and Phone audio signals. The amount of samples by which the Phone audio signal needs to be shifted to match the eye tracker audio signal is shown on the y-axis. The x-axis shows the time in units of 8 s windows.

lower bounds estimated for the previous and the following frame in step (2).

4. AUTOMATED SEGMENTATION

As can be seen in Table 1, the time the subjects talk is much less than the total recording time. Therefore annotations are needed for speech and non-speech segments as well as whether speech comes from the subject (wearing the eye tracking glasses) or the partner or some other person close by. To be able to annotate large amounts of data in a short amount of time, we used an automated annotation method: To robustly detect speech segments in a high level of background noise (supermarket) we used our highly accurate Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) multi-condition Voice Activity Detector (VAD), pre-trained as described in [1]. The background noise contains babble from other people in the store, announcements, background music, children playing, and shopping carts moving around. For increased robustness, we apply the VAD to both, the phone and eye tracker audio track.

To detect whether the subject or someone else is speaking, we rely on the relative energy differences in voice segments between the phone and the eye tracker audio tracks. For this, both audio tracks were normalized to 0 dB peak amplitude before frame-wise (25 ms, sampled every 10 ms) root-quadratic energy was computed. In the cases where the subject is speaking, the energy level in a 500 ms sliding window is generally larger for the eye tracker recordings than for the phone recordings. A small adjustment of a bias of the level differences was needed independently for each subject. These biases were found by empirically looking at the number of detected segments and the balance between subject and other segments for energy level biases in the range from -0.05 to +0.05. The threshold yielding the maximum number of segments and at the same time yielding a higher number of subject segments than partner segments was used (cf. Table 1, last row). As the segmentation is automatic and unsupervised, there are errors in the detected segments. However, a manual inspection of a subset of the detected segments confirmed that the automatic segmentation has a high accuracy and the segments can be used in further experiments.

5. EYE CONTACT AND ACOUSTICS

From the eye tracking glasses we can extract the position where the subject is looking at in the coordinate system of the eye tracker frontal camera. From the video of the frontal camera we detect the presence of a face (frontal view) with the openCV face detector based on Local Binary Pattern (LBP) features and try to estimate the eye region within the face with a Haar-wavelet based eye detector also available in openCV. If no eye region was detected in the face (e. g.,

if people wear glasses), we estimate the eye region from the face region as:

$$Xe = xf + 0.25wf \quad (1)$$

$$Yh = yf + 0.25hf \quad (2)$$

$$We = 0.5wf \quad (3)$$

$$He = 0.16hf, \quad (4)$$

where the subscript e indicates the eye region bounding box and the subscript f the face region bounding box. X, y, w, and h are the coordinates of the upper left corner, the width, and the height of the bounding box, respectively.

By combining the eye tracker coordinates with the detected face and eye region, we can define three classes for where the subject is looking with respect to the partner: Direct eye contact – i. e., looking into the eye region (V_c), looking into the face region (V_b), or looking next to the face region in a corridor with 0.5 width/height to the left, top, right, bottom of the face region (V_a). Additionally, we compute the Euclidean distance between the centre of the detected eye region and the point the subject is looking at. This is referred to as eye-eye distance in the following. If no face is detected in the image, a maximum value is filled in for this distance.

To produce an eye-contact ground truth per speech segment, we apply the following rule in this particular order: If for at least 2 frames there is direct eye contact (V_c), we assign the V_c label to the whole segment. Otherwise, if for at least 2 frames there is case V_b, we assign label V_b, and otherwise the same for V_a. If neither case is present in the segment we assign the label V_n for no eye contact. Detailed statistics on the amount of eye contact in the segments where the subject is talking are found in Table 1. There are notable differences between the subjects in terms of eye contact behaviour. Subject A apparently has the most eye contact with his partners, durations of cases V_a and V_b are almost 1 second on average for a two second average segment duration, while for subjects B and C it is only .3 seconds and for D .7 seconds.

6. EXPERIMENTS AND RESULTS

In order to explore correlates between the acoustic and vocal properties of speech with the location of where the subject is looking at in a conversation with another person, we present an analysis of the correlations between acoustic features and the eye-eye distance. The audio recorded via the eye tracker microphone is used for this purpose. As acoustic feature set, we use a large standard set of acoustic features, as used for the baseline results in the Interspeech 2013 ComParE Challenge [9].

The features were extracted with our open-source feature extraction and affect recognition toolkit openSMILE [2]. The ComParE feature set contains 6 373 features, which are functionals of acoustic low-level descriptors (LLDs). The LLDs include prosodic features (signal energy, perceptual loudness, fundamental frequency), voice quality features (jitter and shimmer of the fundamental frequency, voicing probability, and the harmonics-to-noise ratio), spectral features (spectrum statistics such as variance and entropy and energies in relevant frequency bands), and cepstral features (Mel-Frequency cepstral coefficients – MFCC). From these LLDs, the first order delta coefficients are computed and both LLDs and delta coefficients are smoothed with a 3 tap moving average filter over time. Then, functionals are applied to the LLDs and their delta coefficients over a complete speech segment resulting in one final 6 373 dimensional feature vector for the particular segment. The functionals include statistical measures such as moments (means, variances, etc.), statistics of peaks (mean amplitude of peaks, mean distance between peaks, etc.), distribution statistics such as percentiles (especially quartiles and inter-quartile ranges), regression coefficients obtained by approximating the LLD over time as linear or quadratic function and the errors between the approximation and the actual LLD, temporal characteristics such as positions of maxima and the percentage of values above a certain threshold, and modulation characteristics expressed as linear predictor (autoregressive) coefficients of a predictor of five frames length.

In this study, rather than simply computing the Pearson correlation coefficients (CC) across all subjects and taking those with the highest absolute correlation, we use a selection criterion that rewards consistent correlation across subjects and penalizes inconsistencies such as a feature being correlated for one person yet inversely correlated for another. This leads to the following criterion for feature f:

$$CC'_f = \frac{\sum_{s=1}^{S}\sum_{t=s+1}^{S}\left(|CC_f^{(s)} + CC_f^{(t)}| - |CC_f^{(s)} - CC_f^{(t)}|\right)}{S\,(S-1)} \quad (5)$$

where S is the number of subjects. This criterion ranges from -1 to +1, with -1 indicating strong inconsistency, zero indicating low correlation or medium inconsistency, and one indicating perfect and consistent correlation across all subjects. One of the best correlated acoustic features from the ComParE set (CC' = 0.21; max. CC = 0.37 for subject B) is the gain of the linear prediction on the voicing probability. In speech analysis this gain resembles the energy of the 'predictable' (i. e., correlated and generated by the human vocal tract) signal parts. As we are applying linear predictive coding to the contour of the voicing probability, the gain has a different meaning: it resembles the energy of predictable modulation of the voicing probability and is therefore related to speech rhythm caused by the sequence of voiced and unvoiced phonemes. The more regular the rhythm, the higher the gain is. The best negatively correlated feature (CC' = -0.22) is the range of the peak amplitudes relative to the arithmetic mean for the 6^{th} critical band (approximately 500 – 620 Hz) of an auditory filter bank after applying a RASTA-style band-pass filter to emphasize speech-rate modulations in the range from 4 – 8 Hz. This frequency range corresponds to a frequency relevant for the first formant of vowels. Thus, if the range of peaks in this frequency band is high, there is a high variation of articulatory strength of individual vowels, which corresponds to a sloppy style of articulation, or might resemble general level variations due to quickly changing acoustic conditions (e. g., a person moving relative to the microphones). Altogether the results indicate that modulation descriptors (functionals) are the most relevant. This might indicate that if we have eye contact with a person, we articulate clearer and with a different rhythm than if we do not have eye contact.

Let us now turn to the feasibility of fully automatic attention recognition based on selected acoustic features. In preliminary experiments, we found regression on the actual eye-eye distance too challenging, and four-way classification of V_a, V_b, V_c and V_n to suffer from data sparsity in the V_b and

Subject	A	B	C	D	Mean
UAR [%]	69.6	67.0	64.8	68.2	67.4 ± 2.0
AUC	.765	.707	.679	.775	.732 ± .046

Table 2: Results of automatic classification of $V_{a,b,c}$ (Looking at eyes, head, or near head) vs. V_n (looking somewhere else) on the GRAS^2 corpus, using leave-one-subject-out cross-validation and SVM classifiers. Evaluation in terms of unweighted average recall (UAR) and area under the receiver operating curve (AUC). Mean and standard deviation across four subjects (A–D). Chance level for AUC and UAR is .5 and 50 %, respectively.

V_c classes. Hence, we unified the V_a, V_b, and V_c classes and considered their discrimination from V_n as a binary classification task. For choosing the most relevant features for the attention recognition task at hand, we perform a straightforward ranking based selection, taking into account the CC' criterion with the minimum eye-eye distance as in the feature relevance analysis described above, but applying feature selection in a cross-validation scheme (leave-one-person out) to reduce the danger of over-fitting to the four test subjects. In particular, for testing on each of the four subjects in the database, we use the remaining three subjects as training data. In this way, we select the 200 most relevant features by CC' on the training data, and train a support vector machine (SVM) classifier using the Sequential Minimal Optimization (SMO) algorithm implemented in the Weka toolkit [3]. SVMs are particularly suited to learn from large feature sets with probably inter-correlated features. After classification, the unweighted average recall (UAR) of the classes is computed, as well as the area under the receiver operating curve (AUC). We obtain the results shown in Table 2. Both, UAR and AUC are significantly above chance level (.5) according to a one-tailed z-test ($p < .001$), indicating that the selected features generalize across recordings from different subjects. The low inter-subject deviations of the UAR and UAC further indicate the robustness of the obtained classification results.

7. CONCLUSIONS

We have introduced the GRAS^2 corpus, a multi-modal and multi-sensory corpus of real-life interactions of people seeking for help and directions from strangers in a public shopping centre. The corpus has been recorded for the purpose of analysing the role of visual attention and dialogue behaviour in such interactions. Using information from multiple audio tracks we were able to automatically label when the subject carrying the recording equipment or his or her dialogue partner is talking. The analysis of correlations between acoustic features of the voice of the subject and the visual attention (eye contact with dialogue partner) has revealed a low, but meaningful correlation between the acoustics and the distance of the point at which the subject is looking and the eye region of the dialogue partner. Yet, the correlations are strong enough, such that an automatic classification of whether a subject is looking at or close by the head of the dialogue partner or somewhere else based only on automatically extracted acoustic speech parameters is feasible with up to 70 % unweighted average recall rate (the chance level would be 50 %).

In future work we aim at significantly increasing the size of the corpus by conducting new recordings with the same setup. We will further manually correct the automatic segmentation and conduct experiments on the short interactions to look at the style of the interactions and analyse the reactions and emotions of the dialogue partners.

8. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements No. 289021 (ASC-Inclusion) and No. 288587 (MASELTOV) as well as from the Austrian FFG via contracts No. 832045 (FACTS) and No. 836270 (EVES), and by the Provincial Government of Styria (NeoAttrakt). We kindly thank INTERSPAR Graz and CITYPARK GmbH for the permission to capture the data.

9. REFERENCES

[1] F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life Voice Activity Detection with LSTM recurrent neural networks and an application to Hollywood movies. In *Proc. ICASSP, Vancouver, Canada*. IEEE, 2013.

[2] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. 9th ACM Multimedia, Florence, Italy*, pages 1459–1462. ACM, 2010.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.

[4] D. Miyauchi, A. Sakurai, A. Nakamura, and Y. Kuno. Active eye contact for human-robot communication. In *Proc. CHI 2004*, pages 1099–1102. ACM, 2004.

[5] L. Paletta, R. Sefelin, J. Ortner, J. Manninger, R. Wallner, M. Hammani-Birnstingl, V. Radoczky, P. Luley, P. Scheitz, O. Rath, M. Tscheligi, B. Moser, K. Amlacher, and A. Almer. MARIA – Mobile Assistance for Barrier-Free Mobility in Public Transportation. In *Proc. CORP, Vienna, Austria*, pages 1151–1155, 2010.

[6] M. Z. Poh, N. C. Swenson, and R. W. Picard. A wearable sensor for unobtrusive, longterm assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, 57(5):1243–1252, May 2010.

[7] B. Rammstedt and O. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41:203–212, 2007.

[8] B. Schuller, F. Pokorny, S. Ladstätter, M. Fellner, F. Graf, and L. Paletta. Acoustic geo-sensing: Recognising cyclists' route, route direction, and route progress from cell-phone audio. In *Proc. ICASSP, Vancouver, Canada*, pages 453–457, 2013.

[9] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. Interspeech 2013, Lyon, France*, pages 148–152. ISCA, 2013.