eNTERFACE'12 Proposal

m2m

“Socially Aware Many-to-Machine Communication”

Principal investigator: Dr. Björn Schuller
Affiliation: Institute for Human-Machine Communication, Technische Universität München
Address: Arcisstr. 21, 80333 München, Germany
Telephone: +49 89 289-28548
Fax: +49 89 289-28535
e-mail: schuller@tum.de

Abstract

The M2M project aims at a first step towards multi-user interaction with emotional virtual agents, targeting the speech input components. It will extend the speech input capabilities of the SEMAINE virtual agent to support hands-free input from multiple users, recognizing speech, personality and affect-related states of individual speakers. The M2M project will develop novel methods to combine speaker diarization with speaker trait and state classification in multi-source environments, and will improve detection of speech utterances directed to the system in a multi-user interaction scenario. The results of the project will be provided as source code and binaries for download. A database of realistic multi-user interaction with a virtual agent will be collected, partially annotated, and published on the workshop web page.

Project's objectives

Social competence, i. e., the ability to permanently analyze and re-assess dialogue partners with respect to their traits (e. g., personality or age) and states (e. g., emotion or sleepiness), and to react accordingly (by adjusting the discourse strategy, or aligning to the dialogue partner) remains one key feature of human communication that is not found in most of today's technical systems. Hence, the SEMAINE project (*S*ustained *Em*otionally colored *M*achine-human *I*nteraction using *N*onverbal *E*xpression) built the world's first fully automatic dialogue system with 'socio-emotional skills' realized through signal processing and machine learning techniques. It is capable of keeping sustained conversations with the user, using very shallow language understanding – basically, reacting to emotional keywords and allowing simple dialogue acts – yet advanced techniques for recognition of affect and non-linguistic vocalizations.

Still, the system is limited to interaction with a single user – however, in many real-world scenarios, human-computer interaction with multiple users, and hence, recognizing traits (e.g., personality) and affect-related states (e.g., interest) of the individuals and of the group as a whole, is desirable. Such scenarios include emotional agents incorporated into robots acting as museum guides, or information kiosks. Yet, the generalization from 1 to *N* system users comes with a variety of 'grand challenges' – the following is to be understood as a non-exhaustive list, reaching from front-end to back-end:

(i) Speech source localization. Among other applications, this is useful for feedback, such as the avatar / robot turning its head to the person speaking.

(ii) Technical robustness to non-stationary background noise (transient noise, background speakers) and reverberation in real-world hands-free application scenarios (such as trade fairs, museums etc.)

(iii) Speaker diarization. This is required for the character to access the interaction history with individual speakers. For instance, it can be used to detect that a person has not been speaking for a longer time; the main challenge is handling overlap between speakers.

(iv) Even in case of perfect speech detection and absence of overlap or background noise, speech may not be addressed to the virtual agent, but to other humans (*side talk*), or simply to the speaker itself (*self directed talk*). This can easily lead to erroneous actions taken by the system.

(v) Multi-talker recognition of affect and speech from cross-talk, i. e., in case that system users are speaking simultaneously.

(vi) Appropriate strategies for dialogue management and adaptation of visual agent behavior, such as 'integrating' users showing a low level of interest while preserving high levels of interest of other users.

Clearly, addressing all these challenges and implementing solutions is beyond the scope of a four week targeted research project. Hence, the M2M project will focus on some aspects of (ii) through (iv) in the above list: Precisely, it will extend the capabilities of the SEMAINE system to cope with a hands-free scenario where multiple users interact with the system in the presence of background talkers, environmental noise and reverberation, yet assuming little to no overlap between the user utterances targeted to the system. In the result, detected keywords, speaker traits and affect-related states will be attributed to different users by means of speaker diarization and visualized appropriately. Utterances not addressed to the system will be rejected. The project's objectives will be verified through a dedicated evaluation work package using objective and subjective measures (cf. page 6). The M2M project will deliver tangible results in the shape of source code and reports (cf. page 8).

Background information

Aiming to make interaction with virtual agents more natural, a lot of research effort has been invested to equip dialogue systems with social capabilities that go beyond simple verbal skills. These capabilities include aspects of communication that are emotion-related and non-verbal (Cowie, 2010). So far, most systems are tailored for a one-to-one dialogue situation in which one user has a conversation with one virtual agent. Besides purely speech-based systems, also multimodal frameworks considering for example head movements and facial expressions are becoming popular. The SEMAINE system is one example for a (non-task-oriented) multimodal dialogue system that is sensitive to the user's emotion, non-verbal behavior, and affective cues, trying to recognize the user's state and react to it appropriately via multimodal backchannels (Yngve, 1970) and feedback (Allwood et al., 1992). This also includes natural listener behavior such as head nods, smiles, or short vocalisations such as “uh-huh” or “wow”. Further, the agent has to determine when to 'take the turn' (Sacks et al., 1974) and produce utterances that fit the dialog context. The 'Sensitive Artificial Listener' scenario used in the SEMAINE system (Schröder et al., 2008; Schröder et al., 2012) involves four different virtual characters, each of them representing a different emotional state, i.e., a different quadrant in the valence-arousal space. The virtual agents try to induce 'their' emotion in the user, meaning that they have to recognize and display affect. Emotion recognition in multimodal systems is usually based on low-level features characterizing the user's voice, head movements, and facial expression (Schuller et al., 2011a; Gunes et al., 2011; Valstar et al., 2011). As a first step for speech feature generation, voice activity detection has to be applied in order to extract meaningful acoustic features only in regions where the user is talking. In most speech-based emotion recognition engines, features are generated by applying statistical functionals to contours of acoustic low-level descriptors. Among the low-level descriptors are commonly used features such as loudness, fundamental frequency, probability of voicing, Mel-Frequency Cepstral Coefficients (MFCC), and other features based on the signal spectrum (Schuller et al., 2009a). The functionals include common statistical descriptors such as mean, standard deviation, and other analytical descriptors. Automatic speech recognition (ASR) and keyword spotting systems employed for natural human-machine dialog situations have to be noise robust and tailored for spontaneous and emotional speech containing non-linguistic vocalizations such as laughter, sighing, breathing, etc. (Wöllmer et al., 2009; Wöllmer et al., 2010a). These requirements have motivated a lot of research investigating novel speech recognition approaches that go beyond standard hidden Markov modeling (Wöllmer et al., 2011a; Wöllmer et al., 2011b). In order to enable combined acoustic and linguistic emotion recognition, bag-of-words features can be computed from the ASR output (Eyben et al., 2010). As an alternative to categorical emotion recognition based on classes such as 'happiness', 'anger', 'boredom' etc., emotions can also be modeled in a dimensional way by using a continuous scale for affective dimensions like arousal, valence, expectation, intensity, and power in combination with regression techniques such as Support Vector Regression or neural networks with regression outputs (Eyben et al., 2012). As emotion tends to evolve slowly over time, context-sensitive classification or regression frameworks that model the evolution of emotion usually prevail over static approaches (Wöllmer et al., 2010b; Metallinou et al., 2012).

In addition to typical emotions or affective dimensions, a socially competent virtual agent also profits from recognizing and reacting to user-specific traits such as personality (Mohammadi et al., 2010) and states such as the perceived 'level of interest' or paralinguistic information like age and gender (Burkhardt et al., 2010; Schuller et al., 2009b). In a hands-free interaction scenario, the influence of reverberation and background noise on ASR and affect recognition has to be taken into account (Schuller, 2011b; Wöllmer et al., 2011c). Furthermore, recognition of the traits and affective states of multiple users has rarely, if ever, been investigated, despite the progress in speaker diarization (Reynolds et al., 2009).

Detailed technical description

A. Technical Description

General Strategy of the Work Plan

In the beginning, participants will be familiarized with the SEMAINE system and its technical realization. A technical specification of the components to be developed will be established (WP1). WP2 through WP4 will deal with iterative development of system component, supported by data collection and evaluation in WP5.

To minimize risks associated with insufficient performance of baseline components or system integration, the SEMAINE system will be used as basis for all implementation tasks. Capabilities of the audio component of the SEMAINE system include manifold acoustic feature extraction, tandem LSTM-HMM ASR, voice activity detection including suppression of feedback from agent speech, and classifier frameworks for speaker trait and state recognition.

While the research goals pursued in M2M are ambitious, the work package structure has been carefully designed so as to minimize dependencies between component development work packages (2-5) to ensure a successful project outcome even in the case of failure of individual research tasks.

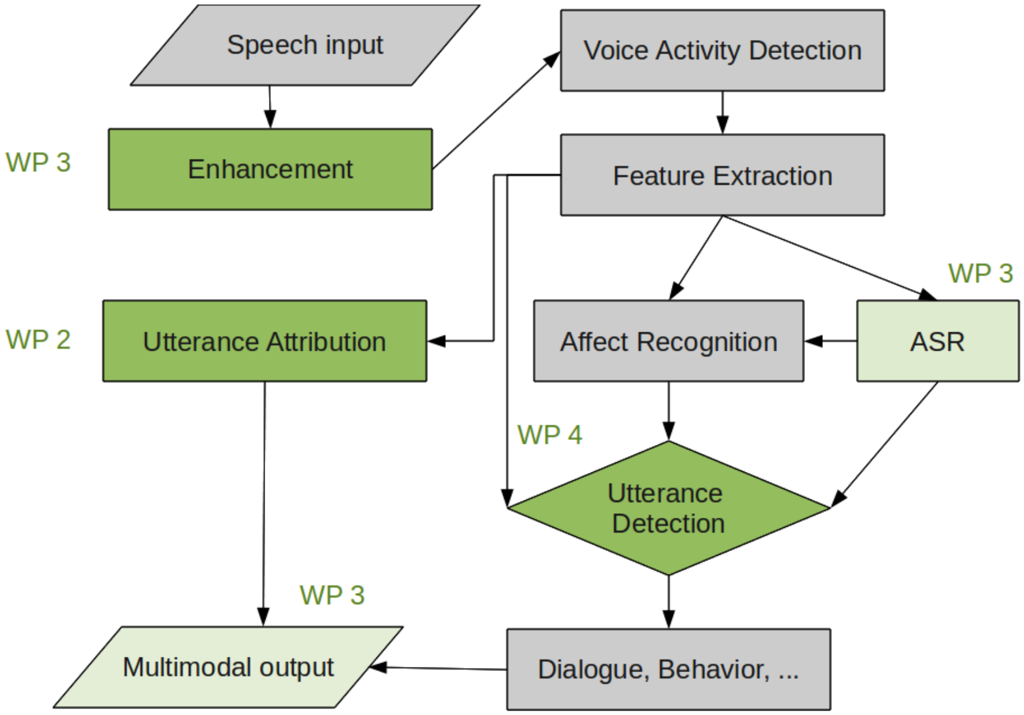

Figure 1: Simplified flowchart of the SEMAINE system with speech input. Gray: Existing components; green: new components added by M2M; gray/green: components improved by M2M.

List of Work Packages (WP)

WP no.	Name	Estimated person-hours
1	System Design and Integration	220
2	Multi-User Affect Recognition	300
3	Environment Adaptation	240
4	High-Level Utterance Detection	220
5	Data Collection and Evaluation	300
Sum		1280 = 32 person-weeks

WP 1: System Design and Integration

In this work package, requirements for the software components will be established by use cases and detailed technical specification, and will be iteratively refined. An architecture based on the ActiveMQ message passing system in SEMAINE will be defined to enable communication between the existing SEMAINE components and the new components developed in WPs 2 and 4, specifically, to integrate speaker attribution into the XML messages sent to the dialogue manager, and to reject utterances not directed to the system.

WP 2: Multi-User Affect Recognition

This work package deals with attributing detected user utterances to different users. Standard approaches to on-line speaker diarization can be combined with additional features available from the SEMAINE feature extraction back-end openSMILE, such as prosodic or voice quality features, that can enable more robust modeling in the presence of noise and reverberation, than conventional cepstral features. As an optional task in this work package, the usage of speaker diarization results for on-line adaptation of affect and speech recognition models can be investigated.

WP 3: Environment Adaptation

The goal of this work package is to ensure technical robustness of the affect and speech recognition systems in hands-free multi-user interaction scenarios, especially against non-stationary noise types (such as cross-talk, transient noise), and reverberation. Standard techniques for on-line speech enhancement, noise suppression and echo cancellation will be combined with model-based approaches, such as context-sensitive modeling in LSTM-RNN or hybrid architectures, which are already featured in the SEMAINE system; these can be extended, for example, to a data-based approach to recognize the foreground speaker in the presence of cross-talk.

WP 4: High-Level Utterance Detection

The goal of this WP is to distinguish between given utterances directed to the agent, side talk and self directed talk. As opposed to voice activity detection, this WP relates to higher level features including prosodic and lexical information. The current voice activity detection in the SEMAINE system is based on LSTM modeling of MFCC features and energy, and multi-condition training to provide robustness. Based on this 'low-level' VAD, this WP will add a second stage for detecting user turns directed to the agent. Based on data collected in WP5, features will be investigated that allow discrimination between the above named classes of utterances. These include prosody, voice quality and ASR confidence measures, based on the assumption that speaking style changes when people are speaking to an agent, as opposed to talking to other humans or themselves. Besides, usage of context information and model-based approaches, e. g., by LSTM-RNN modeling, will be investigated to allow better rejection of background speakers than is possible by simple energy thresholds or similar methods. The result of this work package will be a component that is integrated into the SEMAINE system between the ASR component and the dialogue manager; it will contain a trained classification model that decides whether the detected and decoded utterance is to be further processed or rejected. Existing technology in the SEMAINE system, including the acoustic-linguistic feature extraction and classification frameworks (e. g., SVM, GMM or LSTM-RNN) can be used 'out of the box' for developing this component. Extraction of additional features such as ASR confidences will be added as needed.

WP 5: Data Collection and Evaluation

In WP 5, suitable data for training and evaluating classifier models in the M2M system will be collected. These will span interaction of multiple users with the emotional virtual agent, including real background noise and reverberation occurring in the workshop environment. Parts of the data will be annotated semi-automatically by manual correction of ASR and affect recognition. Personality will be assessed by standard self-assessment questionnaires. To augment such real-life data, data will be synthesized in order to minimize risks related to delayed data collection and to foster early system integration and evaluation. Furthermore, in this WP, iterative evaluation of the system components will be carried out by suitable objective measures such as accuracy, source-distortion-ratio etc. A final evaluation of the system will include subjective measures such as impression of ASR and affect recognition performance, overall user satisfaction etc. on a Likert scale.

B. Resources Needed

Hardware and Software

- 5 installations of SEMAINE system (one per work package):
  - quad-core CPU; >= 4 GB RAM
  - standard room microphone
- Operating system: Windows Vista or higher
- Development environment: Microsoft Visual Studio 2005 or higher
- SVN repository or similar

Human Resources (Participants)

The effort is estimated at 32 person-weeks (8 participants). The envisioned research work requires complementary skills covering a wide range of research areas. Thus, each team member should have excellent background in as many as possible of the following areas:

- monaural speech enhancement and/or de-reverberation;
- speaker diarization;
- affective computing, emotion and personality models and annotation;
- machine learning in signal processing;
- object oriented software development in C++.

Work plan and implementation schedule (max. 1 page)

At least one member of the research staff will be present at all times. For each work package, one participant will be appointed as work package leader and will prepare a short (15 minute) presentation at the end of each week, on advances in the project work.

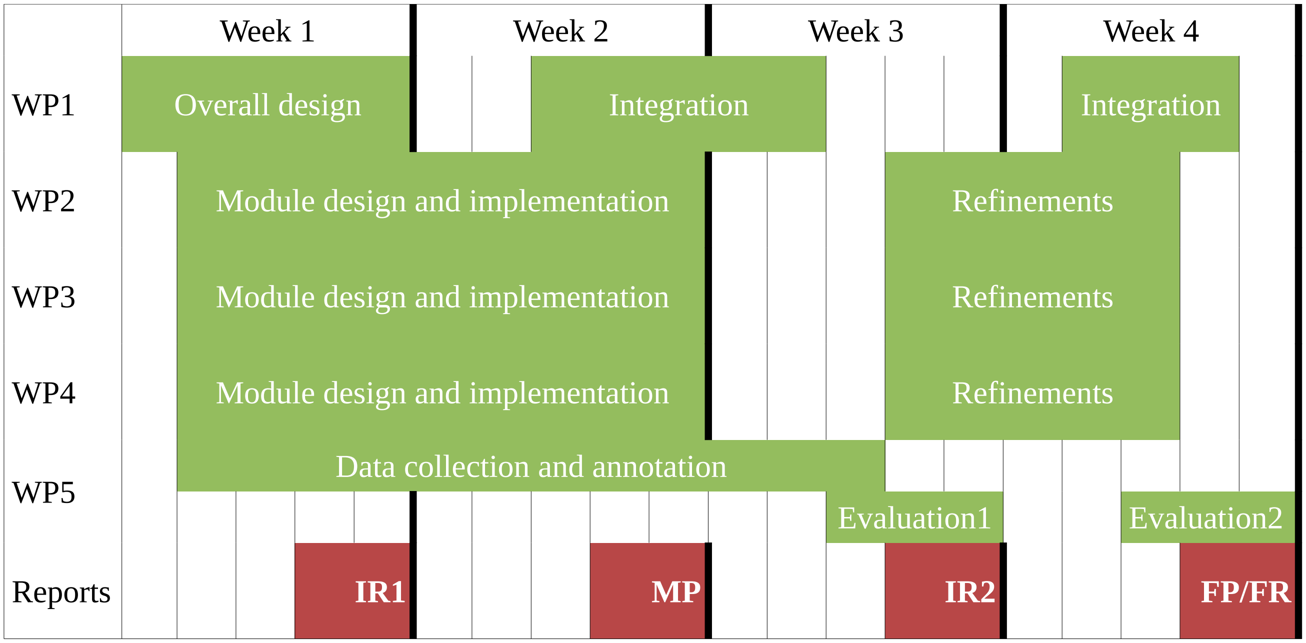

Figure 2: Work schedule (Gantt chart)

IR1: Internal report 1 MP: Mid-term Presentation
IR2: Internal report 1 FP/FR: Final Presentation / Final Report

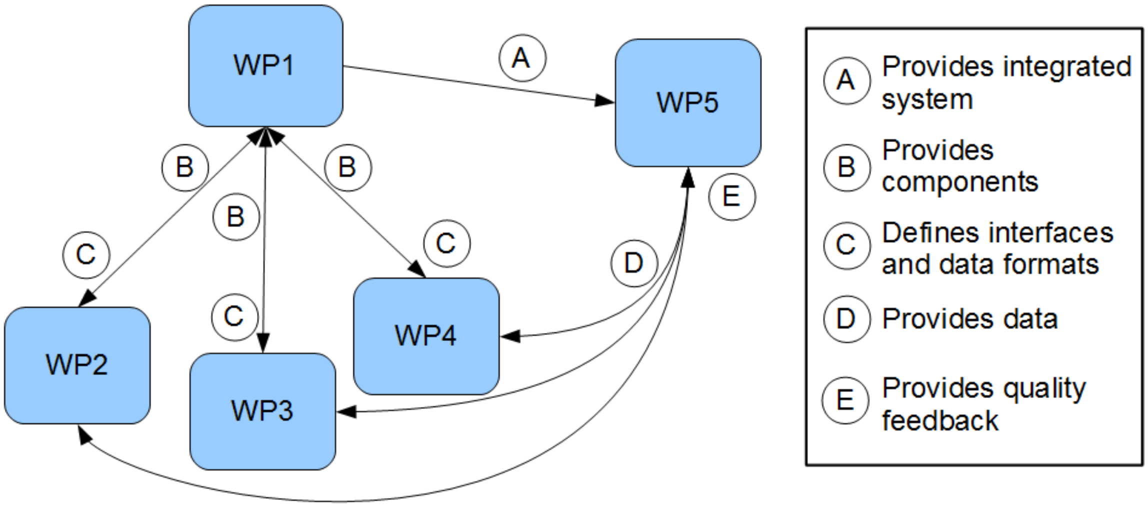
WP1
WP2
WP3
WP4
WP5
A
B
B
B
C
C
C
D
E
A Provides integrated system
B Provides components
C Defines interfaces and data formats
D Provides data
E Provides quality feedback

Figure 3: Interrelations between work-packages (Pert chart)

Benefits of the research (max. 1 page)

Impact

The project outcome will result in a major advance in the state-of-the-art of today's emotional virtual agents, paving the way for deployment for speech-based agents in real-life contexts such as museum guides or information kiosks. Scientifically, we are looking forward to exploiting synergies between the speech signal processing (speaker diarization, speech enhancement) and affective computing domains. We expect a significant number of publications in high-profile journals and conferences resulting from the project work in these areas.

Public deliverables

The expected outcomes of the projects, which will be available at the end of the fourth week for publication on the workshop web page, are:

- A final technical report, which will detail the work realized in the project and the obtained results
- A demonstrator in the shape of a ready-to-install binary package for Windows Vista or higher
- A module for multi-user affect recognition (source code)
- A module for environment adaptation (source code)
- A module for high-level utterance detection (source code)
- An annotated database for multi-user machine interaction

While we envision Windows Vista or higher as target platform for system deployment, platform independence of the new source code developed in the project will be enforced wherever possible. Since the M2M source code will integrate GPL licensed SEMAINE source code, it will necessarily be licensed under the GPL license as well.

The annotated database will be licensed under the Creative Commons BY-NC-ND license or similar. The written consent of all individuals recorded in the database will be sought, and individuals will be enabled to revoke their consent at any time, resulting in removal of the corresponding data from the public repositories.

Profile of the team

A. Leader

Björn W. Schuller received his diploma in 1999 and his doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, both in electrical engineering and information technology from TUM (Munich University of Technology), one of Germany's repeatedly highest ranked and among its first three Excellence Universities.

He is tenured as Senior Lecturer in Pattern Recognition and Speech Processing heading the Intelligent Audio Analysis Group at TUM's Institute for Human-Machine Communication since 2006. From 2009 to 2010 he lived in Paris/France and was with the CNRS-LIMSI Spoken Language Processing Group in Orsay/France dealing with affective and social signals in speech. In 2010 he was also a visiting scientist in the Imperial College London's Department of Computing in London/UK working on audiovisual behaviour recognition. In 2011 he was guest lecturer at the Università Politecnica delle Marche (UNIVPM) in Ancona/Italy and visiting researcher of NICTA in Sydney/Australia. Best known are his works advancing Human-Computer-Interaction, Semantic Audio and Audiovisual Processing, Affective Computing, and Music Information Retrieval.

Dr. Schuller is a member of the ACM, HUMAINE Association, IEEE and ISCA and (co-)authored two books and more than 250 publications in peer reviewed books (20), journals (31), and conference proceedings in the field of signal processing, and machine learning leading to more than 2,300 citations - his current H-index equals 25. He serves as member and secretary of the steering committee, associate editor, and guest editor of the IEEE Transactions on Affective Computing, associate and repeated guest editor for the Computer Speech and Language, associate editor for the IEEE Transactions on Systems, Man and Cybernetics: Part B Cybernetics and the IEEE Transactions on Neural Networks and Learning Systems, and guest editor for the IEEE Intelligent Systems Magazine, Speech Communication, Image and Vision Computing, Cognitive Computation, and the EURASIP Journal on Advances in Signal Processing, reviewer for more than 40 leading journals and 30 conferences in the field, and as workshop and challenge organizer including the first of their kind INTERSPEECH 2009 Emotion, 2010 Paralinguistic, 2011 Speaker State, and 2012 Speaker Trait Challenges and the 2011 and 2012 Audio/Visual Emotion Challenge and Workshop and programme committee member of more than 30 international workshops and conferences. Steering and involvement in current and past research projects includes the European Community funded ASC-Inclusion STREP project as coordinator and the awarded SEMAINE project, and projects funded by the German Research Foundation (DFG) and companies such as BMW, Continental, Daimler, HUAWEI, Siemens, Toyota, and VDO. Advisory board activities comprise his membership as invited expert in the W3C Emotion Incubator and Emotion Markup Language Incubator Groups, and his repeated election into the Executive Committee of the HUMAINE Association where he chairs the Special Interest Group Speech.

B. Staff Proposed by the Leader

Cyril Joder received the engineering degree from the École Polytechnique and Telecom ParisTech, and the M.Sc. degree in acoustics, signal processing and computer science applied to music from the university Pierre et Marie Curie, Paris, France, in 2007. He received his PhD degree in Signal Processing from Telecom ParisTech, Paris, in 2011. Since October 2011, he is working as a post-doctoral researcher in the Institute for Human-Machine communication of the Technical University Munich, Germany. His research interests include audio signal processing, machine learning, statistical models and source separation. He is the author of several

international peer-reviewed conference and high-impact journal articles in the field.

Florian Eyben obtained his diploma in Information Technology from TUM. He is currently pursuing his PhD degree in the Intelligent Audio Analysis Group within the Institute for Human-Machine Communication at TUM, working on audio feature extraction for affect recognition from speech. He is one of the main contributors to the implementation of the SEMAINE dialogue system, and is the main author of the award-winning openSMILE feature extractor which has become a standard in computational paralinguistics research through the series of INTERSPEECH and other Challenges (2009-2012). His research interests include large scale hierarchical audio feature extraction and evaluation, automatic emotion recognition from the speech signal, recognition of non-linguistic vocalizations, automatic continuous large vocabulary speech recognition, statistical and context-dependent language models, and Music Information Retrieval. Teaching activities of his comprise Pattern Recognition and Speech and Language processing. He has over 60 publications in peer-reviewed books, journals and conference proceedings covering many of his areas of research, leading to over 500 citations and an H-index of 12.

Felix Weninger received his diploma in computer science (Dipl.-Inf. degree) from Technische Universität München in 2009. He is currently pursuing his PhD degree as a researcher in the Intelligent Audio Analysis Group at TUM's Institute for Human-Machine Communication, focusing his research on multi-source speech and audio recognition, including signal separation and robust back-ends for automatic speech recognition and paralinguistic information retrieval. He is the main author of the open-source BliSSART toolkit for monaural source separation and speech enhancement. Mr. Weninger is a member of the IEEE and serves as a reviewer for the IEEE Transactions on Audio, Speech and Language Processing, IEEE Transactions on Affective Computing, Speech Communication, and Computer Speech and Language as well as the International Society for Music Information Retrieval (ISMIR) conference.

C. Team as a Whole

The proposed staff of researchers provides a unique combination of expertise in affective computing and emotional virtual agents with background in robust speech recognition and signal enhancement.

Name	**Expertise**	**Role in the project**
Björn Schuller	Affective Computing	Principal investigator
Florian Eyben	Audio Feature Extraction; Real-time Audio Processing; Emotional Virtual Agents	Supervision of implementation and integration
Cyril Joder	Speech Enhancement; Source Separation	Front-end design and implementation
Felix Weninger	Source Separation; Automatic Speech Recognition; Robustness in Affective Computing	Back-end design and implementation

References

J. Allwood, J. Nivre, E. Ahlsén: On the Semantics and Pragmatics of Linguistic Feedback, in *Journal of Semantics*, vol. 9, no. 1, pages 1-26, 1992.

F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, “A database of age and gender annotated telephone speech,” in Proc. LREC (Language Resources Evaluation Conference), 2010.

R. Cowie: Describing the Forms of Emotional Colouring that Pervade Everyday Life, in *The Oxford Handbook of Philosophy of Emotion*, Oxford University Press, pages 63-94, 2010.

F. Eyben, M. Wöllmer, B. Schuller: A Multi-Task Approach to Continuous Five-Dimensional Affect Sensing in Natural Speech, to appear in *ACM Transactions on Interactive Intelligent Systems, Special Issue on Affective Interaction in Natural Environments*, ACM, 2012.

F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, R. Cowie: On-line Emotion Recognition in a 3-D Activation-Valence-Time Continuum using Acoustic and Linguistic Cues, in *Journal on Multimodal User Interfaces (JMUI), Special Issue on Real-time Affect Analysis and Interpretation: Closing the Loop in Virtual Agents*, Springer, vol. 3, no. 1-2, pages 7-19, 2010.

H. Gunes, B. Schuller, M. Pantic, R. Cowie: Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey, in *Proc. International Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE (EmoSPACE 2011), held in conjunction with the 9th International IEEE Conference on Face and Gesture Recognition 2011 (FG 2011)*, IEEE, Santa Barbara, CA, pages 827-834, 2011.

A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, S. Narayanan: Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification, to appear in *IEEE Transactions on Affective Computing*, IEEE, 2012.

G. Mohammadi, M. Mortillaro, and A. Vinciarelli, “The voice of personality: Mapping nonverbal vocal behavior into trait attributions,” in Proc. International Workshop on Social Signal Processing, Florence, Italy, pp. 17–20., 2010.

D. Reynolds, P. Kenny, F. Castaldo: A Study of New Approaches to Speaker Diarization. In: Proc. INTERSPEECH, Brighton, UK, pp. 1047-1050, 2009.

H. Sacks, E. A. Schegloff, G. Jefferson: A simplest systematics for the organization of turn-taking for conversation, in *Language*, vol. 50, no. 4, pages 696-735, 1974.

M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, B. Schuller: Towards responsive Sensitive Artificial Listeners, in *Proc. Fourth International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, M. Wöllmer: Building Autonomous Sensitive Artificial Listeners, to appear in *IEEE Transactions on Affective Computing (TAC)*, IEEE, 2012.

B. Schuller, M. Wöllmer, F. Eyben, G. Rigoll: Spectral or Voice Quality? Feature Type Relevance for the Discrimination of Emotion Pairs, in *The Role of Prosody in Affective Speech, Linguistic Insights, Studies in Language and Communication*, vol. 97, Slyvie Hancil (ed.), Peter Lang Publishing Group, pages 285-307, 2009a.

B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu: Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application, in *Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, Elsevier, vol. 27, no. 12, pages 1760-1774, 2009b.

B. Schuller, A. Batliner, S. Steidl, D. Seppi: Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge, in *Speech Communication (SPECOM), Special Issue: Sensing Emotion and Affect – Facing Realism in Speech Processing*, vol. 53, no. 9/10, pages 1062-1087, 2011a.

B. Schuller: *"Affective Speaker State Analysis in the Presence of Reverberation"*, International Journal of Speech Technology, Springer, Vol. 14, Issue 2, pp. 77-87, 2011b.

M. Valstar, B. Jiang, M. Mehu, M. Pantic, K. Scherer: The First Facial Expression Recognition and Analysis Challenge, in *Proc. of IEEE Int'l. Conf. Face and Gesture Recognition*, Santa Barbara, CA, 2011.

M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, G. Rigoll: Robust Discriminative Keyword Spotting for Emotionally Colored Spontaneous Speech Using Bidirectional LSTM Networks, in *Proc. of ICASSP 2009*, IEEE, pages

3949-3952, Taipei, Taiwan, 2009.

M. Wöllmer, F. Eyben, A. Graves, B. Schuller, G. Rigoll: Bidirectional LSTM Networks for Context-Sensitive Keyword Detection in a Cognitive Virtual Agent Framework, in *Cognitive Computation, Special Issue on "Non-Linear and Non-Conventional Speech Processing"*, Springer, New York, vol. 2, no. 3, pages 180-190, 2010a.

M. Wöllmer, B. Schuller, F. Eyben, G. Rigoll: Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening, in *IEEE Journal of Selected Topics in Signal Processing (JSTSP), Special Issue on Speech Processing for Natural Interaction with Intelligent Environments*, IEEE, vol. 4, no. 5, pages 867-881, 2010b.

M. Wöllmer, F. Eyben, B. Schuller, G. Rigoll: A Multi-Stream ASR Framework for BLSTM Modeling of Conversational Speech, in *Proc. of ICASSP*, IEEE, Prague, Czech Republic, pages 4860-4863, 2011a.

M. Wöllmer, B. Schuller, G. Rigoll: A Novel Bottleneck-BLSTM Front-End for Feature-Level Context Modeling in Conversational Speech Recognition, in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2011)*, IEEE, pages 36-41, Waikoloa, Big Island, Hawaii, 2011b.

M. Wöllmer, F. Weninger, S. Steidl, A. Batliner, B. Schuller: *"Speech-based Non-prototypical Affect Recognition for Child-Robot Interaction in Reverberated Environments"*, Proc. INTERSPEECH 2011, ISCA, Florence, Italy, pp. 3113-3116, 2011c.

V. H. Yngve: On getting a word in edgewise, in *Chicago Linguistic Society. Papers from the 6th regional meeting*, vol. 6, pages 567-577, 1970.