Editorial

Introduction to the special issue on sensing emotion and affect – Facing realism in speech processing

Human-machine and human-robot dialogues in the next generation will be dominated by natural speech which is fully spontaneous and thus driven by emotion. Systems will not only be expected to cope with affect throughout actual speech recognition, but at the same time to detect emotional and related patterns such as non-linguistic vocalisation, e.g., laughter, and further social signals for appropriate reaction. In most cases, this analysis clearly must be made independently of the speaker and for all speech that 'comes in' rather than only for pre-selected and pre-segmented prototypical cases. In addition – as in any speech processing task – noise, coding, and blind speaker separation artefacts, together with transmission errors need to be dealt with. To provide appropriate back-channelling and socially competent reaction fitting the speaker's emotional state in time, on-line and incremental processing will be among further concerns. Once affective speech processing is applied in real-life, novel issues such as standards, confidences, distributed analysis, speaker adaptation, and emotional profiling are coming up next to appropriate interaction and system design. In this respect, the INTERSPEECH 2009 Emotion Challenge, which has been organised by the guest editors, provided the first forum for comparison of results, obtained for exactly the same realistic conditions. In this special issue, on the one hand, we will summarise the findings from this challenge, and on the other hand, provide space for novel original contributions that further the analysis of natural, spontaneous, and thus emotional speech by late-breaking technological advancement, recent experience with realistic data, revealing of black holes for future research endeavours, or giving a broad overview.

The last special issue on emotion processing in Speech Communication appeared in 2003, i.e., before automatic emotion processing really turned from 'yet another exotic speech topic' into mainstream. However, what still can be observed nowadays is on the one hand, a more and more sophisticated employment of statistical procedures; on the other hand, these procedures do not keep up with the requirements of processing application-oriented, realistic speech data, and too often, the data used are still un-realistic, i.e., prompted and acted data, and thus not representative for real-life: In comparison to related speech processing tasks such as Automatic Speech and Speaker Recognition, practically no standardised corpora and test-conditions exist to compare performances under exactly the same conditions. Instead a multiplicity of evaluation strategies employed – such as cross-validation or percentage splits without proper instance definition – prevents exact reproducibility. Further, in order to face more realistic use cases, the community is in desperate need of more spontaneous and less prototypical data. To address these problems and at the same time, to set an exemplar, we organised the INTERSPEECH 2009 Emotion Challenge to help bridging the gap between excellent research on human emotion recognition from speech and low compatibility of results: For the FAU Aibo Emotion Corpus of emotionally coloured spontaneous children's speech, benchmark results of the two most popular approaches were provided to the participants. Three sub-challenges were addressed in two different degrees of difficulty by using non-prototypical four or two emotion classes versus a garbage model: the acoustic Feature Sub-Challenge, the Classifier Sub-Challenge, and the Open Performance Sub-Challenge.

Apart from the opening article by the guest editors, which was handled in an independent review process by the editor in chief, 25 submissions were received for this special issue, out of which 10 were accepted (i.e., 40% acceptance rate).

The introducing article "*Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge*" by the guest editors and Dino Seppi aims at providing a broad overview on the state of the art in the field and summarises the Emotion Challenge held at INTERSPEECH 2009.

Next come five articles dealing with different aspects of emotion in speech: In "*Recognizing Affect from Speech Prosody Using Hierarchical Graphical Models*" Raul Fernandez and Rosalind Picard focus on the prosodic

component based on local and global phenomena. They are using first acted data, then transfer the approach onto realistic data, one from a natural social setting, and one from a natural business setting. Acoustic parameters reflect intonation, loudness, rhythm, and voice quality. The model is hierarchical, determined by prosodic constituents but based on a blind, bottom-up segmentation. They obtain a 70% detection rate with 30% false alarms when detecting high arousal negative valence speech in call centres. Next, Simon F. Worgan and Roger K. Moore consider the role of paralinguistic features in their article "*Towards the detection of social dominance in dialogue*". They do so especially for pitch and Long-Term Average Spectrum, in establishing and manipulating social dominance which is claimed to be established as a feature of rapport not of the individual: Other adjust their speech to match the (current) dominant individual. For illustration, the authors use the American English Map Task (AEMT) database. Katherine Forbes-Riley and Diane Litman contribute on "*Benefits and Challenges of Real-Time Uncertainty Detection and Adaptation in a Spoken Dialogue Computer Tutor*". They evaluate the performance of a spoken dialogue system in the tutoring domain; the system detects uncertainty in each student turn. It uses acoustic-prosodic, lexical, and dialogue features, and yields higher performance than two non-adaptive control systems if adapting to uncertainty and correctness detected in the students turns. In "*Achieving Rapport with Turn-by-Turn, User-Responsive Emotional Coloring*" Jaime C. Acosta and Nigel G. Ward use a corpus of conversations with students about graduate school to analyse the emotional states of the interlocutors; they found that the emotional colouring of the speaker's utterance could be largely predicted from the emotion shown by her interlocutor in the immediately previous utterance. The authors developed a spoken dialogue system that reacts adequately to the interlocutor's emotion. This system was evaluated favourably in a user study. Then, Ammar Mahdhaoui, and Mohamed Chetouani describe the development of an infant-directed speech discrimination system for parent-infant interaction analysis in their article "*Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis*". Using real-life family home movies, they investigate different feature sets in both supervised and semi-supervised settings. The proposed dynamic weighted co-training approach combines various features and classifiers and was shown to be effective for dealing with the real-life data employed.

Following, five contributions on vocal emotion recognition are contained. Four of these are from participants of the INTERSPEECH 2009 Emotion Challenge, presenting classification results on the data set used in the Emotion Challenge – the FAU Aibo Emotion Corpus, which has been released for scientific research after this event. For a fair comparison of the results presented in this special issue and the results of the challenge, it has to be pointed out that after the challenge the whole data set including the emotion labels of the test set has been made available. The ground truth of the test set was not available during the challenge. In addition to the results on the FAU Aibo Emotion Corpus, most authors present results on other databases, too, in order to demonstrate the usefulness of their proposed classification systems.

The following four contributions are from participants of the challenge: Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan focus on new classification techniques in "*Emotion Recognition Using a Hierarchical Binary Decision Tree Approach*". They use the official feature set of the Emotion Challenge provided by the organisers and feature selection and apply a hierarchical classification system of binary classifiers, where the easiest classification task is solved first. Results are presented for the 5-class problem of the Emotion Challenge and additionally on the USC IEMOCAP database, a database of both scripted and spontaneous dialogues between a female and a male actor. For these experiments, a subset containing only angry, happy, sad, and neutral utterances was selected. Marcel Kockmann, Lukas Burget, and Jan Černocký next discuss "*Application of Speaker- and Language Identification State-of-the-Art Techniques for Emotion Recognition*". They use various Gaussian Mixture Modelling techniques to classify Mel frequency cepstral coefficients (MFCC) and prosodic features. Results are presented for the 5-class problem of the Emotion Challenge and, in addition, for the well-known Berlin Database of Emotional Speech, which is a database of German emotional speech produced by actors. Then, Elif Bozkurt, Engin Erzin, Cigdem Eroglu Erdem, and Arif T. Erdem present "*Formant Position based Weighted Spectral Features for Emotion Recognition*". The article proposes a variant of the MFCC features, where critical spectral bands around formant locations are emphasised during the MFCC calculation. These features are fused with standard spectral and prosodic features on the decision level and results are presented on the FAU Aibo Emotion Corpus for both the 5-class (Anger, Emphatic, Neutral, Positive, Rest) and the 2-class (negative vs. idle) problem. "*Anger Recognition in Speech Using Acoustic and Linguistic Cues*" by Tim Polzehl, Alexander Schmitt, Florian Metze, and Michael Wagner is one of the few studies that use both acoustic and linguistic features. The authors focus on the detection of anger in interactive voice response (IVR) systems. Hence, they present results for the 2-class problem negative vs. idle of the Emotion Challenge. Furthermore, they use two IVR databases of real application data, one with recordings from a German voice portal, and one with data from an US-American portal.

The fifth paper is not related to the Emotion Challenge, but as the previous one, it focuses on emotion recognition for spoken dialogue systems and uses both acoustic and linguistic features: In "*Enhancement of Emotion Detection in Spoken Dialogue Systems by Combining Several Information Sources*" Ramón López-Cózar Delgado, Jan Silovsky, and Martin Kroul train different classification systems for prosodic features, MFCC features, lexical features, and

dialogue act information. After evaluating these different systems independently of each other, the authors apply different late fusion techniques. Results are presented on their own database, which contains telephone calls between students of the university and the dialogue system of the authors, for a 3-class (angry, tired, neutral) and a 2-class (negative vs. non-negative) problem.

Finally, we want to give an overview of databases and approaches used in the studies included in this special issue. A necessary precondition for inclusion was the use of realistic data; apart from the FAU Aibo database, we find interactions between users and real or simulated automatic dialogue systems (call centres or specifically tailored systems, e.g., tutoring dialogues), map task dialogues, and interactions within families. As could be expected, the languages dealt with are mostly English, but German, Italian, and Spanish as well. Numbers of subjects and annotators vary widely, from 8 speakers to 51 or more, and from 2 judges/annotators to 29. Mostly, a dimensional approach has been chosen (e.g., activation, valence, power), or categories/dimensions which are not prototypical such as certainty, dominance, or motherese. Besides some classical inferential statistics, we see the usual candidates such as Gaussian Mixture Models, Support Vector Machines, Multi-Layer Perceptrons, and alike, often from the Weka toolbox. In Emotion Challenge articles, quite often MFCC were used as features; in the other articles, prosodic features, sometimes together with other acoustic or lexical/dialogue features, prevail. For the specifically tailored dialogue systems, the use of emotion modelling has been evaluated within the systems.

Summing up, it is encouraging that it has been possible to collect so many high quality research articles that broaden the view on automatic emotion modelling, away from acted data and only prototypical categories/dimensions, towards a more realistic view, tailored for specific applications.

Acknowledgements

The guest editors are grateful to the editor in chief, Marc Swerts, and to the 39 reviewers who undertook timely and insightful reviews of the submissions: Noam Amir, Elisabeth André, Plinio Barbosa, Mirjam Broersma, Felix Burkhardt, Carlos Busso, Christophe d'Alessandro, Laurence Devillers, Ellen Douglas-Cowie, Pierre Dumouchel, Kjell Elenius, Julien Epps, Florian Eyben, Raul Fernandez, Katherine Forbes-Riley, Christian Hacker, Helen Hastie, Marcel Kockmann, Kornel Laskowski, Florian Metze, Wolfgang Minker, Elmar Nöth, Nicole Novielli, Olivier Pietquin, Santiago Planet, Tim Polzehl, Norbert Reithinger, Marc Schröder, Dino Seppi, Julia Sidorova, David Suendermann, Louis ten Bosch, Khiet Truong, Thurid Vogt, Nigel G. Ward, Karl Weilhammer, Felix Weninger, Martin Wöllmer, and Simon F. Worgan.

Björn Schuller
Institute for Human-Machine Communication, Technische Universität München, Germany
Tel.: +49 89 289 28548; fax: +49 89 289 28535
E-mail address: schuller@tum.de

Anton Batliner
Stefan Steidl
Pattern Recognition Lab, University of Erlangen-Nuremberg, Germany