Editorial

Introduction to the special issue on Paralinguistics in Naturalistic Speech and Language

Keywords: Computational paralinguistics; Age; Gender; Affect; Paralinguistic Challenge

Paralinguistics deals with all those aspects which are 'besides linguistics'; speech and language do not only convey meaning but much more: We often can tell, when listening to someone, whether they are male or female, native speakers or not, which regional accent they belong to, and whether they are happy or sad, intoxicated or tired, just to mention a few of the manifold aspects encoded in speech, besides pure semantics.

All this information can be employed in every-day and future communication, retrieval, and surveillance systems. However, this requires present research to focus on realistic conditions aiming at 'realistic' occurrences of these phenomena, to focus on real-life requirements such as presence of noise, reverberation or transmission artefacts, and on efficiency requirements. Moreover, optimal embedding of computational estimates in a system context needs to be addressed. To this end, speech and language resources and tailored acoustic and lexical features and machine learning architectures are needed. In addition, synergistic parallel assessment of such phenomena exploiting mutual dependencies may lead to improved performance over isolated consideration. In this vein, this special issue of Computer Speech and Language is devoted to the computational analysis of Paralinguistics in Naturalistic Speech and Language.

The INTERSPEECH 2010 Paralinguistic Challenge, which has been organised by the guest editors and colleagues, 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 and spontaneous speech, recent experience with realistic data, revealing of black holes for future research endeavours, or giving a broad overview.

Apart from the opening article by the guest editors, which was handled in an independent review process by the editor in chief, out of the many submissions received for this special issue eight were accepted – on average, each underwent two revisions – divided into three challenge participants and five general topics in computational paralinguistics.

The introducing article *"Paralinguistics in Speech and Language – State-of-the-Art and the Challenge"* by the guest editors and the other co-organisers of the INTERSPEECH 2010 Paralinguistic Challenge Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan aims at providing a broad overview on the state of the art in the field and summarises the Challenge.

In the following short presentation of the articles in this special issue, we will start with the one by Scherer addressing data issues; the next four articles deal with different types of traits and states, by that giving an exemplary insight into the manifold aspects of the field. The last three articles take up the Challenge topics (Gajšek et al. the one from 2009, the other two dealing with the 2010 Challenge), with a stronger focus on benchmarking issues.

In *"Vocal markers of emotion: Comparing induction and acting elicitation"*, Klaus R. Scherer addresses the ever-lasting question of the type of speech data that is most suitable for investigating emotions encoded in speech. He argues against 'convenience' sampling of rather easily available data from broadcasting and TV, and in favour of tightly controlled inducted or acted emotions obtained in the lab which allow a better access to the speaker's "true" feeling state – by that taking rather a stance against the use of naturalistic data. He claims that enacted data are especially useful "[...] if the purpose of the research is to study the listener attribution of emotion from vocal cues, rather than the diagnosis of symptoms of 'true' underlying emotions." Scherer nicely points out some of the pitfalls of dealing with 'convenience' databases. However, he does not take into account the pivotal problem in application-oriented research that the training data should be as close as possible to the data faced in the application – a scenario sort of simulated by a strict separation of train and test data (from the intended domain) which will only in some exotic constellation be acted data.

In *"Human and computer recognition of regional accents and ethnic groups from British English speech"*, Abualsoud Hanani, Martin J. Russell, and Michael J. Carey deal with long term traits representing language varieties. A language identification (LID) system is applied to 14 regional accents of British English, and to the speech of the two largest ethnic groups in the city of Birmingham (UK). For both tasks, the LID system proved to be very competitive, compared to other established procedures.

In the contribution *"On the Development of an Automatic Voice Pleasantness Classification and Intensity Estimation System"* by Luis Pinto-Coelho, Daniela Braga, Miguel Sales-Dias, and Carmen Garcia-Mateo voice pleasantness is defined in an objective way first, and an approach towards its automatic assessment including its intensity is shown. This is based on a representative set of acoustic, signal periodicity, and phonation speed features. Different classification set-ups comprise artificial neural networks, SVMs, and GMMs. The data stem from Portuguese female voices. The authors report 90.9% accuracy for voice pleasantness classification and 84.3% accuracy for its intensity estimation.

Rajesh Ranganath, Dan Jurafsky, and Dan McFarland describe in their article *"Detecting Friendly, Flirtatious, Awkward, and Assertive Speech in Speed-Dates"* a system for detecting interpersonal stance representing a social role in a dyadic interaction (speed-dates), which is medium-term between long-term traits and short-term states: whether a speaker is flirtatious, friendly, awkward, or assertive. Lexical, prosodic, and dialogue features are used in a Support Vector Machine (SVM) classifier to detect very clear styles (the strongest 10% in each stance) with up to 75% accuracy on seen, and 59% accuracy on unseen speakers. Relevant features are discussed and implications for our understanding of interpersonal stances are addressed.

In *"Quantification of Speech Dysfluency as a Marker of Medication-Induced Cognitive Impairment: An Application of Computerized Speech Analysis in Neuropharmacology"*, Serguei V. S. Pakhomov, Susan E. Marino, and Angela K. Birnbaum use automatic speech recognition procedures to characterise spontaneous speech disfluency induced by topiramate, an anti-epileptic medication with language related side-effects. It turns out that spontaneous speech characteristics such as filled pauses, false starts, and repetitions, are sensitive to the effects of this medication, and that they are associated with the topiramate concentration in the blood. These results demonstrate the potential use of speech technology for neuropsychological testing and neuropharmacology.

A commonly used approach in various speech analysis tasks is the adaptation of a Universal Background Gaussian Mixture Model (UBM-GMM) and the subsequent classification of the adapted mean vectors of the mixtures, the so-called supervectors, with SVMs. In their article *"Speaker State Recognition Using an HMM-Based Feature Extraction Method"*, Rok Gajšek, France Mihelič, and Simon Dobrišek extend this approach by estimating Hidden Markov Model (HMM)-UBMs. They evaluate two different techniques: (1) transforming the monophone-based segmented HMM-UBM into a GMM-UBM and proceeding with the standard adaptation scheme and (2) adapting the HMM-UBM directly. The authors applied both approaches to the 2-class emotion classification problem defined in the INTERSPEECH 2009 Emotion Challenge (negative vs. idle) and to an alcohol/non-alcohol classification problem using the VINDAT database of Slovenian speech at different levels of alcohol intoxication. Both approaches outperformed the standard approach.

The following two contributions are in the context of the INTERSPEECH 2010 Paralinguistic Challenge; the first one is an extended version of the authors' Challenge paper.

In *"Automatic Speaker Age and Gender Recognition Using Acoustic and Prosodic Level Information Fusion"*, Ming Li, Kyu J. Han, and Shrikanth Narayanan present a system for speaker age and gender identification. They combine seven different sub-systems by score level fusion. In addition to the three standard systems (Mel-Frequency Cepstral Coefficients modelled with GMMs, GMM mean supervectors classified with SVMs, and 450-dimensional utterance level features classified with SVMs), the authors propose an SVM system based on UBM weight posterior probability supervectors using the Bhattacharyya probability product kernel, a sparse representation system based on UBM weight posterior probability supervectors, an SVM system based on GMM MLLR matrix supervectors, and an SVM system based on the polynomial expansion coefficients of the syllable level prosodic feature contours in voiced speech segments. Experimental results are reported for the aGender corpus used in the INTERSPEECH 2010 Age and Gender Sub-Challenge.

The article *"Automatic Detection of Speaker State: Lexical, Prosodic, and Phonetic Approaches to Level-of-Interest and Intoxication Classification"* by William Yang Wang, Fadi Biadsy, Andrew Rosenberg, and Julia Hirschberg focuses on the level-of-interest classification and reports results for the TUM AVIC corpus used in the INTERSPEECH 2010 Affect Sub-Challenge. Furthermore, the authors present a system to detect intoxication as addressed in the INTERSPEECH 2011 Intoxication Sub-Challenge. In the level-of-interest classification task, the authors propose a novel discriminative Term Frequency Inverse Document Frequency (TFIDF) linguistic feature and a novel prosodic event detection approach using AuToBI in combination with acoustic features. For the detection of intoxication, they evaluate the performance of prosodic event-based, phone duration based, phonotactic, and phonetic-spectral based approaches.

Summing up, the articles contained in this special issue demonstrate nicely and exemplary the arguably most important aspects of dealing with paralinguistics in speech and language: Which type of data to use (induced, acted, or 'naturalistic'); which type of phenomena to address (long term traits such as voice pleasantness or regional variants, interpersonal stances such as flirtatious behaviour, short term states such as emotions, deviant speech influenced by medication); and finally, which types of procedures to employ for improving performance, and this done within a strictly controlled setting such as provided by the Challenge.

Acknowledgements

The guest editors are grateful to the editor in chief, Roger Moore, and to the 25 reviewers who undertook timely and insightful reviews of the submissions: Tobias Bocklet, Felix Burkhardt, Christophe d'Alessandro, Julien Epps, Florian Eyben, Raul Fernandez, Bruce R. Gerratt, Christer Gobl, Christian Hacker, Mark Huckvale, Christian Kaernbach, Marcel Kockmann, Jody Kreiman, Kornel Laskowski, Robin Lickley, David Loring, Florian E. Metze, Emily Mower Provost, Elmar Nöth, Andrew Rosenberg, Marc Schröder, Dino Seppi, Nigel Ward, Felix Weninger, and Martin Wöllmer.

Björn Schuller[a,b,*]
[a] *CNRS-LIMSI, Spoken Language Processing Group, Orsay, France*
[b] *Technische Universität München, Munich, Germany*
Stefan Steidl[c,d]
[c] *International Computer Science Institute (ICSI), Berkeley, CA, USA*
[d] *Friedrich-Alexander University Erlangen-Nuremberg, Pattern Recognition Lab, Germany*
Anton Batliner
Friedrich-Alexander University Erlangen-Nuremberg, Pattern Recognition Lab, Germany

[*] Corresponding author at: CNRS-LIMSI, Spoken Language Processing Group, Orsay, France.
Tel.: +33 1 69 85 80 08;
fax: +33 1 69 85 80 88.
E-mail address: schuller@IEEE.org (B. Schuller)
Available online 21 June 2012