Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 343–345

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Introduction to the Special Issue on Broadening the View on Speaker Analysis

In the last five decades, the focus of automatic speech analysis and more recently of automatic singing analysis was on the linguistic and structural content side: words (and note-events) and their semantic interpretation. Yet, when it comes to the human behind speaking and singing, so far, research has been mostly interested in the identity of the person. Only in the last one and a half decades, increasing effort has been invested to computationally analyse a variety of speaker and singer states and traits that do not denote but characterise the person. There are short-term states that are indicated by different speaking and singing styles such as emotions (full blown, prototypical) and emotion-related states or affects (e.g., stress, intimacy, interest, confidence, uncertainty, deception, politeness, frustration, sarcasm, and pain). Medium-term phenomena are between states and traits; they include (partly) self-induced more or less temporary states (e.g., sleepiness, medical, and alcohol intoxication, health state, or moods such as depression) and structural (behavioural, interactional, social) signals (e.g., role in dyads or groups, friendship and identity, positive/negative attitude). Finally, there are long-term traits, such as biological trait primitives (e.g., height, weight, age, gender), group/ethnicity membership (race/culture/social class with a weak borderline towards other linguistic concepts, i.e., speech registers such as dialect, regional accent, or nativeness), or personality traits (such as likability and personality in general) – just to mention a few. All these tasks are so far mostly handled in isolation when it comes to automatic analysis; yet, it seems intuitive that they are highly inter-dependent.

This special issue aims at *Broadening the View on Speaker Analysis*. It focuses on technical issues for highly improved and robust speaker (and singer) state and trait analysis.

The INTERSPEECH 2011 Speaker State Challenge, which has been organised by the guest editors, provided the first forum for a comparison of machine classification results for mid-term speaker states under realistic conditions. In this special issue, we will first summarise the findings of this Challenge: the introducing article *“Medium-Term Speaker States – A Review on Intoxication, Sleepiness and the First Challenge”* by the guest editors in collaboration with Felix Weninger and Florian Eyben aims at providing a broad overview on the state of the art in medium-term speaker state recognition and summarises the Challenge. This article was handled in an independent review process by the editor-in-chief. Then, the winners of the two Sub-Challenges on Intoxication and Sleepiness describe their approaches, followed by a contribution on the winning approach in the previous INTERSPEECH 2010 Paralinguistic Challenge's Affect Sub-Challenge that dealt with the speakers' level of interest. Moreover, there are contributions outside these events dealing with sleep apnoea, vocal fatigue, intelligibility of head and neck cancer patients, emotion, and enthusiasm in singing. Overall, apart from the opening article by the guest editors, out of the many submissions received for this special issue, eight were accepted: three Challenge participants and five general topics in Computational Paralinguistics. Five of the accepted papers underwent two revisions, the other two one and three. In the following, we want to introduce the articles in more detail, starting with the three contributions of the Challenge winners.

(Some of) the winners of the 2011 Intoxication Sub-Challenge – Daniel Bone, Ming Li, Matthew P. Black, and Shrikanth S. Narayanan – describe their approach in *“Intoxicated Speech Detection: A Fusion Framework with Speaker-Normalized Hierarchical Functionals and GMM Supervectors”*. Based on the best performance in the Intoxication Sub-Challenge, the authors build several systems with various representations of prosodic and spectral features and discuss the details of each classifier. The authors address the problem of reducing the associated variability through

0885-2308/$ – see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.csl.2013.11.001

modelling speakers, utterance type, gender, and utterance length. The fusion of the different systems improves the classification results as well as speaker normalisation. Comparable gains to other speaker normalisation techniques are achieved with a held-out set of baseline (sober) data. The combined system improves even upon the previously best results of the authors.

The winners of the 2011 Sleepiness Sub-Challenge – Dong-Yan Huang, Zhengchen Zhang, and Shuzhi Sam Ge – work out the details of their method to deal with unbalanced data sets in *“Speaker State Classification Based on Fusion of Asymmetric Simple Partial Least Squares (SIMPLS) and Support Vector Machines”*. Besides evaluating their approach on the two-class classification problem (sleepy vs. not sleepy) defined in the 2011 Sleepiness Sub-Challenge, they present results for four binary classification tasks for the dimensions activation, expectation, power, and valence on the SEMAINE corpus of emotionally coloured conversations. The authors obtain best results for the sleepiness detection by fusing various systems based on three different feature sets (the official feature sets provided in the INTERSPEECH 2009, 2010, and 2011 Challenge), and two classification methods (ASIMPLS and support vector machines (SVMs) with SMOTE for balancing of instances).

The winners of the 2010 Affect Sub-Challenge – Je Hun Jeon, Rui Xia, and Yang Liu – propose a decision-level fusion approach of acoustic and lexical information for predicting the user’s level of interest in *“Level of Interest Sensing in Spoken Dialog Using Decision-level Fusion of Acoustic and Lexical Evidence”*. In their study, the authors are able to demonstrate that using lexical information improves the detection performance of a system purely based on acoustic features even if the robust lexical features are extracted from an erroneous output of an automatic speech recognition system with a high word error rate.

We now continue with the description of the five general contributions, starting with the article *“Analysis of voice features related to obstructive sleep apnoea and their application in diagnosis support”* by Rubén Fernández Pozoa, Doroteo T. Toledanob, José Luis Blanco Murilloa, Ana Montero Benavidesa, Eduardo López Gonzaloa, and Luis Hernández Gómeza. The authors propose a non-intrusive, fast, and convenient screening technique for diagnosing obstructive sleep apnoea. A set of 16 voice features from their own previous studies as well as from studies by other authors is analysed. Finally, eight features are selected and classified with linear discriminant analysis in order to discriminate apnoea patients and healthy subjects. The relatively low classification error rate of 17.1% indicates that this screening technique is helpful before proceeding to the reference diagnostic method, i.e. polysomnography.

Next, in *“Vocal Fatigue Induced by Prolonged Oral Reading: Analysis and Detection”*, Marie-José Caraty, and Claude Montacié analyse the prosody of vocal fatigue using prolonged oral reading corpora. They show that vocal fatigue is not associated with an increase in fundamental frequency and voice intensity. Furthermore, they use the baseline acoustic feature set of the INTERSPEECH 2010 Paralinguistic Challenge in combination with SVMs to automatically classify fatigue and non-fatigue states. The unweighted accuracy on the test set is 68.2%. Last but not least, the authors focus on the estimation of the difference in fatigue level between two speech segments using a combination of multiple phoneme-based comparison functions. The equal error rate is reduced to 19% after filtering phonetic segments and cepstral features.

In *“Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer”*, Catherine Middag, Renee Clapham, Rob van Son, and Jean-Pierre Martens present a new method for the automatic intelligibility assessment of pathological speech (communication deficiencies caused by speech disorders). By combining alignment-free features – that are basically text-independent and applicable to different language varieties or languages – and alignment-based features, they obtain a performance that is comparable to the one of human raters.

The article *“Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications”* by Bogdan Vlasenko, Dmytro Prylipko, Ronald Böck, and Andreas Wendemuth deals with emotion recognition in speech. The authors use phoneme-based classification to discriminate low and high arousal in cross-corpora evaluations using the acted emotions from the Berlin Emotional Speech Database (EmoDB) and the more spontaneous emotions from the Vera-am-Mittag (VAM) corpus. Phonetic-pattern dependent emotion models based on the phonetic transcription of an automatic speech recognition system outperform standard turn-level analysis in this cross-corpora setting.

As a tweak towards the end, Ryunosuke Daido, Masashi Ito, Shozo Makino, and Akinori Ito deal with the singing voice in their article *“Automatic Evaluation of Singing Enthusiasm for Karaoke”*. They use two sets – one set contains several songs, the other one is limited to only one song – of singing along (karaoke) by 24 male and 10 female amateurs. The songs are all in Japanese and by intention in the same key of C major. After partitioning into three sets, 30, 30, and 10 humans labelled ternary enthusiasm in the voice recordings. For acoustic analyses, power, pitch, and vibrato-related features are used. Using linear and logistic regression, the authors reach a correlation coefficient of 0.65 between the estimated value and the gold standard as given by the human raters.

Summing up, this special issue shows a large variety of applications in this emerging field of speaker analysis, ranging from intoxication and vocal fatigue to the level of interest and to emotion, and finally to medical applications such as sleep apnoea and intelligibility ratings of pathological speech. At the same time, the large range of machine learning and feature extraction approaches applied in this field is demonstrated.

Acknowledgements

The guest editors are grateful to the editor-in-chief, Roger Moore, and to the 27 reviewers for their timely and insightful reviews of the submissions: Tobias Bocklet, Felix Burkhardt, Simon Dobrisek, Julien Epps, Florian Eyben, Engin Erzin, Corinne Fredouille, Phil Green, Volker Gross, Christian Heinrich, Denis Jouvet, Wooil Kim, Thomas Kisler, Kornel Laskowski, Anne-Maria Laukkanen, Ching-Ta Lu, Alexandra Makarova, Nicolas Malyska, Erik Marchi, Claude Montacié, Emily Provost, Sebastian Schnieder, Dino Seppi, Milan Sigmund, Louis ten Bosch, Khiet Truong, and Felix Weninger.

Björn Schuller[a,b,*]
[a] *Imperial College London, Department of Computing, United Kingdom*
[b] *Technische Universität München, Machine Intelligence & Signal Processing Group, MMK, Germany*
Stefan Steidl
Friedrich-Alexander University Erlangen-Nuremberg, Pattern Recognition Lab, Germany
Anton Batliner[a,b]
[a] *Technische Universität München, Machine Intelligence & Signal Processing Group, MMK, Germany*
[b] *Friedrich-Alexander University Erlangen-Nuremberg, Pattern Recognition Lab, Germany*
Florian Schiel
Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universität München, Germany
Jarek Krajewski
University of Würzburg, Industrial and Organizational Psychology, Germany

[*] Corresponding author at: Technische Universität München, Institute for Human-Machine Communication, Germany. Tel.: +49 89 289 28548; fax: +49 89 289 28535.
E-mail address: schuller@IEEE.org (B. Schuller)