Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2014) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge☆

Björn Schuller[a,b,*], Stefan Steidl[c], Anton Batliner[b,c], Elmar Nöth[c,d], Alessandro Vinciarelli[e,f], Felix Burkhardt[g], Rob van Son[h,i], Felix Weninger[b], Florian Eyben[b], Tobias Bocklet[c], Gelareh Mohammadi[f], Benjamin Weiss[j]

[a] *Imperial College London, Department of Computing, England, United Kingdom*
[b] *Technische Universität München, Machine Intelligence & Signal Processing Group, MMK, Germany*
[c] *Friedrich-Alexander Universität Erlangen-Nürnberg, Pattern Recognition Lab, Germany*
[d] *King Abdulaziz University, Jeddah, Saudi Arabia*
[e] *University of Glasgow, School of Computing Science, Scotland, United Kingdom*
[f] *IDIAP Research Institute, Martigny, Switzerland*
[g] *Deutsche Telekom AG Laboratories, Berlin, Germany*
[h] *Netherlands Cancer Institute NKI-AVL, Amsterdam, The Netherlands*
[i] *University of Amsterdam, Phonetic Sciences, Amsterdam, The Netherlands*
[j] *Technische Universität Berlin, Quality & Usability Lab, Germany*

Received 25 July 2013; received in revised form 26 June 2014; accepted 15 August 2014

Abstract

The INTERSPEECH 2012 Speaker Trait Challenge aimed at a unified test-bed for perceived speaker traits – the first challenge of this kind: personality in the five OCEAN personality dimensions, likability of speakers, and intelligibility of pathologic speakers. In the present article, we give a brief overview of the state-of-the-art in these three fields of research and describe the three sub-challenges in terms of the challenge conditions, the baseline results provided by the organisers, and a new openSMILE feature set, which has been used for computing the baselines and which has been provided to the participants. Furthermore, we summarise the approaches and the results presented by the participants to show the various techniques that are currently applied to solve these classification tasks.

Keywords: Computational paralinguistics; Personality; Likability; Pathology; Survey; Challenge

1. Introduction

In 2009–2012, challenges (Schuller et al., 2009, 2010, 2011, 2012, 2013a, 2013) were organised at the INTERSPEECH conferences that featured several different aspects of paralinguistics: topics of interest were not *what* the

☆ This paper has been recommended for acceptance by L. ten Bosch.

* Corresponding author at: Imperial College London, Department of Computing, England, United Kingdom. Tel.: +44 207 59 48357.
E-mail addresses: bjoern.schuller@imperial.ac.uk, schuller@IEEE.org (B. Schuller).

http://dx.doi.org/10.1016/j.csl.2014.08.003

speaker said, i.e., word recognition, or the semantics behind word recognition, e.g., hot spots or ontologies, but *how* it was said; for that, pertinent information can either be found between words (vocal, non-verbal events), it can be modulated onto the word chain (typically supra-segmental phenomena such as prosody or voice quality), or it can be encoded in the (types of) words chosen and in the connotations of these words. Catalogues of (short-term) speaker states such as emotions and of (long-term) speaker traits such as gender or personality are given in Schuller et al. (2013) and Schuller and Batliner (2014). In the 2012 challenge and accordingly in the present article, we want to address speaker traits that were obtained by perceptual annotation and not by some 'objective' measurement such as placing subjects on a scale to find out about their weight, or simply by deciding between male or female.

There are different definitions for the field that deals with 'how' instead of 'what'; traditionally, *paralinguistics* is mostly conceived as dealing with the non-verbal, vocal aspects of communication, sometimes including, sometimes excluding multi-modal behaviour such as facial expression, hand gesture, gait, body posture. Here, we follow the definition given in Schuller and Batliner (2014): paralinguistics is *"[...] the discipline dealing with those phenomena that are modulated onto or embedded into the verbal message, be this in acoustics (vocal, non-verbal phenomena) or in linguistics (connotations of single units or of bunches of units)."* Thus, we exclude multi-modality but include verbal phenomena: although most of the contributions to our challenges so far concentrated on acoustics, i.e. on vocal phenomena modulated onto or embedded into the verbal message, we do not want to exclude linguistic approaches such as the modelling of interjections, hesitations, part-of-speech, or n-grams.

Speech is produced by speakers, and when we aim at paralinguistics, then a specific type of speech (friendly speech, pathological speech) characterises a specific type of speakers – they display friendliness or pathological speech traits. Thus, we could subsume all these phenomena under *Speaker Characterisation* or *Speaker Classification* as was done by Müller (2007, V): *"[...] the term* speaker classification *is defined as assigning a given speech sample to a particular class of speakers. These classes could be Women vs. Men, Children vs. Adults, Natives vs. Foreigners, etc."*. Eventually, it is simply a matter of perspective whether we call the object of our investigation "type of speech" (indicated by specific speech characteristics) or "speaker traits" (indicated by specific speech characteristics extracted from the speech of specific speakers).

Irrespective of the term chosen, it is always about assigning one individual sample (speech or speaker) to $k = 1, \ldots, n$ groups (classes) of speakers; the larger n is, the more likely we may employ regression procedures instead of classification. Of course, it is always possible to map more or less continuous attributions such as rating scales onto a few classes. For challenges like the present one, we as organisers have to know which class a speaker in the test set belongs to. As mentioned above, this 'reference' (or 'ground truth', 'gold standard') can be obtained by (sort of) objective measures (for instance, speaker weight classes by following the 'body mass index') or by using perceptive evaluation. In this challenge on perceived speaker traits, we presented three sub-challenges where all speakers were assigned to (two) different classes, based on perceptive evaluation.

Perceptual judgements as basis for reference classes set specific edge conditions: basically, this mostly results in ranked/ordinal scales; however, often-parametric procedures such as Pearson's correlation are used. Human annotators do not always agree; thus, we do need some measure for agreement, and some method for ending up with one 'unified' label per token. This is normally the mean of the rating scale scores of all annotators. If we aim at classes, we have to partition the scale at appropriate points (mean, median, etc.).

1.1. Why such a challenge: the motivation behind

When some of the authors started organising challenges back in 2009, the main motivation behind was to introduce a certain standard of comparability into the field of Computational Paralinguistics, by introducing concepts like

- a partitioning of the database into train, development, and test data; often, there were only train and test partitions, and researchers defined the partitions of the very same corpus in different ways
- a clearcut stratification of subjects for the partitions, if necessary and feasible, for instance, into male/female, old/young, etc.
- the 'open microphone setting' which means that all data recorded and available should be processed; this pertains especially realistic data that often were preselected, based on labeller agreement, quality of recordings, and alike