MORE THAN FIFTY YEARS OF SPEECH AND LANGUAGE PROCESSING - THE RISE OF COMPUTATIONAL PARALINGUISTICS AND ETHICAL DEMANDS

Anton Batliner[1,2], Björn Schuller[3,1]

[1] Technische Universität München, Machine Intelligence & Signal Processing group, Germany, [2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Pattern Recognition Lab, Germany, [3] Imperial College, London, Department of Computing, United Kingdom

Keywords: *Computational Paralinguistics, Automatic Speech Processing, Natural Language Processing, Intrinsic Ethics, Extrinsic Ethics*

Abstract

We sketch the development of Automatic Speech and Language Processing and in particular of Computational Paralinguistics. Recently, both – the larger fields as well as the narrower – reached a level of maturity that allows for machine-based recognition of speech, but also of rich speaker characterisation such as age, height, personality traits, intoxication, typicality in general, or emotion 'in the wild'. Several aspects of this development lead to higher ethical demands. These demands are specified and exemplified with possible applications.

1. Introduction

In this paper, we will address and exemplify the different ethical demands we have to cope with that go along with the chain of processing and with the structure of processing within Computational Paralinguistics. In chapter 2, *What is Computational Paralinguistics?*, we explain and define the field. Chapter 3, *The Development of Automatic Speech and Language Processing and of Computational Paralinguistics,* sketches the development of this field in the last 50 years, chapter 4 its *Chain of Processing,* and chapter 5 the *Ethical Demands* that go along with the *Evolution* of this field. In chapter 6, some *use cases* exemplify different ethical demands*,* whereas in chapter 7, *intrinsic and extrinsic ethics* are set in relation to Computational Paralinguistics and related fields. We conclude with chapter 8 where we look into the *Future of Computational Paralinguistics* and, linked to this, possible higher ethical demands.

2. What is Computational Paralinguistics?

Definitions are notoriously fuzzy when it comes to the border regions separating different fields that might, at first sight, be unambiguous if we only look at the prototypical core area. For us, the problem starts with telling apart speech from language: partly the

same or different? We will use definitions based on the scientific sub-cultures that have evolved in the course of the last fifty years: *Automatic Speech Processing* (ASP) deals with *spoken language*, *Natural Language Processing* (NLP) deals with *written language*. Both address the question *what* has been produced, i.e., what has been spoken or written: phones (underlying: phonemes), words and sequences of words (n-grams, collocations), or the semantics behind these words, e.g., keywords, topic spotting, hot spots, or ontologies. *How* something has been spoken or written – e.g., in which tone of voice, by using which words out of several candidates that denote the same but have different connotations – all this we attribute to the field of *Computational Paralinguistics* (CP). ('Automatic' in ASP and 'Computational' in CP both simply mean that the job is done with the help of or by the computer.) Note that extensionally, all these fields have been defined differently in different sub-cultures as well: sometimes, speech processing is seen as a sub-field of language processing, sometimes, paralinguistics is confined to non-verbal aspects of speech, leaving aside verbal/linguistic aspects. Our motivation to use a rather broad definition of CP is not to annex as much as possible under this heading but simply to mirror 'daily practice' within the computational approaches towards paralinguistics or, in other words, to base our definition on *functional* (what do we want to find out) and not *formal* (which means do we use) aspects (Schuller & Batliner 2014). In the following, we concentrate on analysis, recognition, and classification within CP, leaving aside generation and synthesis. Of course, for the creation of virtual agents, avatars, or robots, all these sub-fields have to be taken into account.

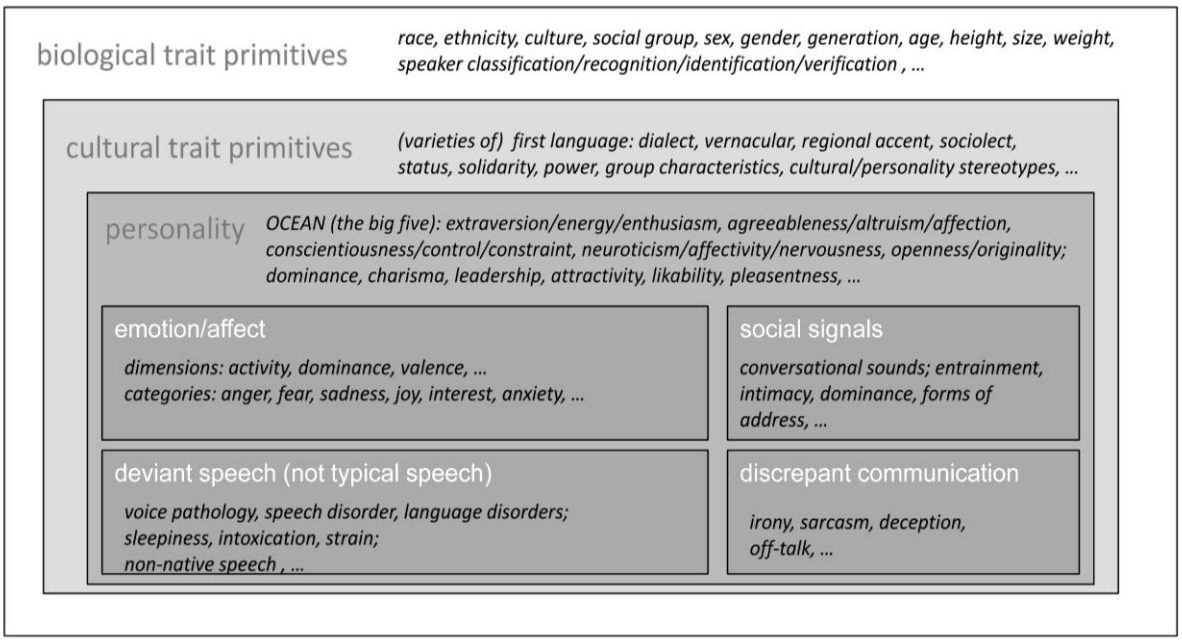

Fig. 1: Layered figure-ground relationship of paralinguistic functions, with examples of traits and states; from (Schuller & Batliner 2014).

Fig. 1 sketches the main functions (fields) of CP in a 'layered figure-ground' relationship, with examples of traits (long-lasting characteristics) and states (short-time). In the lighter background, we see the long-term traits, and in the foreground, the short(er)-term traits and especially states. Everything that relates to characteristics of speakers that are either *biologically* (e.g., sex, age) or *culturally* (e.g., regional accent, status) pre-defined, is called '*trait primitives*'. P*ersonality* evolves and develops based on

these primitives, and figures, in a sort of figure-ground relationship, as the basis of short-term *emotions*, of medium-term *atypical* speech, and of specific communicative behaviour, i.e., either 'normal' interaction (*social signals*) or *discrepant* communication which should not be taken literally. For a full account of all these phenomena and terms, see (Schuller & Batliner 2014).

Thus, in everyday language, we can say that *"You do CP whenever you use acoustic and/or linguistic information and a computer to find out something about the person(s) behind – i.e., the person(s) who produced this information, be this in vocal, non-verbal or in verbal, i.e., in spoken, or in written form."*

3. The Development of Automatic Speech and Language Processing and of Computational Paralinguistics

ASP looks back to some fifty years of research (Furui 2005, Furui 2009), starting with the processing of single digits, produced by single speakers, in the 1960ies; the lexicon grew from some 1000 entries in the 70ies to several 1000 entries in the 80ies; trained dictation in the 90ies was followed by robust processing of millions of words in the first decade of this century; the state of the art approaches real-life recognition and language identification with subsequent automatic translation. NLP evolved on a similar timeline (Sparck Jones 1992, Nadkarni et al. 2011). So far, we addressed the processing of what has been said: the chain of words (i.e., word recognition) and the semantics behind (keywords, topic spotting, hot spots, or ontologies). Within humanities, what has been said is normally dealt with within phonetics and linguistics. Now, we address how something has been said by whom: the term *paralinguistics* dates back to the 50ies (Schuller & Batliner 2014), the field of CP can be traced back to the recognition/verification/identification of speakers, starting in the 70ies; automatic emotion recognition ('affective computing') by using speech emerged in the 90ies and was subsequently complemented by classifying/detecting a plethora of long-term speaker traits (age, height, personality, non-nativeness, dialect, pathology, etc.), of intermediate traits/states (intoxication, sleepiness, etc.), and of short-term states (besides clear emotions: interest, boredom, even heart rate or eye contact by using acoustic information, and alike).

4. The Chain of Processing within Computational Paralinguistics

To start with, there are basic and technical decisions to be taken that mostly are not (yet) highly relevant for ethical considerations: which phenomena do we want to address, which data do we want to use (existing or to be collected/recorded, speech and/or written language, and suchlike), how to record which data, how to pre-process them (orthographic transcription, error detection and correction, definitions of units of analysis (phones, words, stories, persons), annotation and categorisation of phenomena found in the data, technical documentation, and storing. A full account of these steps is given in (Schuller & Batliner 2014).

Ethics comes in when it is not (only) about technicalities but about people – and this means, at many steps, from conceptualisation to publication and releasing the data; some of these steps directly pertain individuals, some other rather societal issues:

1. Is it ok to address the question we are interested in? (*society*)
2. How to guarantee the consent and the privacy of the experimental subjects or of the people we process information from even when they do not know (big data studies using freely available data found on the Internet)? (*individuum*)
3. How to encode the data to guarantee this privacy, even for the future? (*individuum*)
4. How to design possible applications such that they meet ethical demands? (*individuum, society*)
5. How to communicate results such that the public does not have unrealistic expectations? (*society*)

5. The Evolution of Computational Paralinguistics and Ethical Demands

In this chapter, we sketch developments within (ASP/NLP and) CP that quite often lead to higher ethical demands.

what? → how?

Pure ASP or NLP are interested in 'what' has been spoken or written; paralinguistics is interested in 'how' speech or written language have been produced: in which emotion, by whom (i.e., by a non-native, sleepy, intoxicated, nervous, happy, ... person). It is evident that the extension from 'what' onto 'how' something has been produced opens new challenges for ethically acceptable approaches.

basic research → application

Pure research might be considered to do no harm as long as the privacy of (experimental) subjects is guaranteed. Of course, things change if it comes to using results in political debates or decisions that have direct or indirect impact on sub-populations or individuals such as acceptance or rejection of specific therapies. Applications, on the other hand, if they are not only entertainment or harmless edutainment, i.e., in the sense of (Cowie 2012), 'ethically lightweight', can have serious impact on individuals.

uni-modal → multi-modal

In our definition of paralinguistics, it is confined to verbal/vocal (non-verbal) and written phenomena, in a broader one, it encompasses other modalities as well, such as facial expressions, hand/body gestures, and gait. Notwithstanding this definition, ethical considerations become even more important when multi-modality comes into the game, simply because personalisation is easier, thus, anonymisation has to be stricter: audio processing might be technologically more advanced, video processing is, of course, more critical as far as a direct identification of individuals is concerned.

in vitro (lab), individuum → in vivo, crowd (→ big data)

Seen from the point of view that new developments create higher demands to ethics, this development is less unequivocal: In the 'old ages', paralinguistics rather dealt with experimental individuals who, assembled in a small experimental sample, were taken as data basis. Thus, de-personalisation had always been an issue. Big data, i.e., many data obtained from many individuals found on the web, is, at first sight, more anonymous. On the other hand, it is only seemingly harmless – especially in connection with the possibilities to find out other core data which makes personalisation through the backdoor possible as well. In the lab, the individual participant in experiments is known, his/her identity, however, has to be concealed in the following processing, especially when data are passed on to third parties or are made publicly available. In big da-

ta processing, the identity of the individuum is mostly not immediately apparent but can be discovered.

anonymisation → personalisation

Pure research, targeted towards specific phenomena, does not need information on subjects – it is enough when data can be identified unambiguously as belonging to one 'item' or 'subject' in processing. In contrast, follow-ups in tutoring, teaching, or therapy of course need personal information. From a broader point of view, it is of course also advisable to collect as much individual information as possible – data are precious and with their help, it might be possible to address other questions later on. Of course, this conflicts with early anonymisation. 'Big data' can be anonymous or personalised: When we 'only' are interested in whether some specific product or film receives positive or negative reviews, we do not need any personalised information about the people behind these reviews. Of course, the temptation to find out more is high (see the personalised advertisement on the web browser). When we scan for specific people using big data procedures – think of national security agencies – then of course, personalisation is a sine qua non, because the discovery and disclosure of individual information provides possibilities to trace back individuals.

typical → atypical

Typicality is a fuzzy concept (Schuller & Batliner 2014): it can mean 'prototypicality' in the sense of 'extreme, very pronounced', thus rather infrequent; it can mean 'very frequent' in the sense of 'typical for a specific sub-population' and thus (mostly) less pronounced. In the context of speech processing, 'typical' often simply means that pertinent data are easily obtainable, 'atypical' are not. Normally, a 'typical' characteristic is not very interesting for paralinguistics; it is rather the deviation, the *atypicality* which is interesting because we always need typical, neutral data (as sort of background model) in order to find out what is deviant, non-typical. This alone is, in a way, a sort of personalisation by building smaller sub-samples consisting of people with atypical characteristics. Ethical considerations already start with the names given to these atypical groups to ensure politically accepted ('correct') terms, cf. 'autism' vs. 'autistic spectrum' or the development from 'negro' to 'black' to 'African American'.

recognition → analysis

CP systems so far are mainly tailored to 'do the recognition' job, i.e., they only target the assignment of a label to a speech or text unit. A very recent and likely future trend, however, is to go beyond and provide additional analysis, such as the confidence level or prototypicality, regulation, feigning, display rules or atypicality. This can go as far as to the feature level by analysing which (acoustic or linguistic) feature is different from the standard case in which way. Such research strategies might contribute to de-anonymisation.

All the developments sketched here are prone to lead to ethically more critical constellations for (1) privacy consideration, and for (2) responsibilities as far as possible implications and consequences of correct or incorrect decisions are concerned. Fig. 2 summaries these aspects.

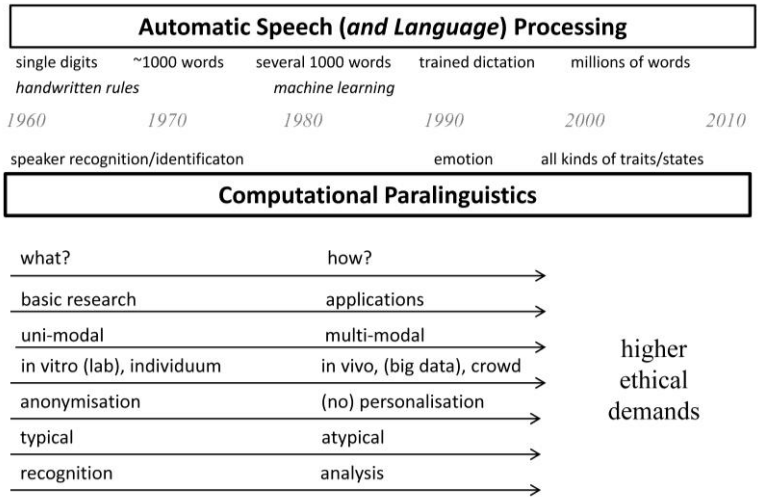

Fig. 2. Evolution of Computational Paralinguistics, leading to higher ethical demands.

6. Different Targets: Some Use Cases

By and large, CP mostly still deals with research and not yet with application. However, this research is often 'application-minded', i.e., motivated by and aimed at possible applications. There might be two main aspects: first, will it be good enough? Second, if it is good enough, will the application pay off and will it be acceptable – especially from the point of view of ethics?

The first example is rather 'harmless': imagine a learner of English as a second language (L2). Normally, in a language course and in special pronunciation exercises, there is a target variety such as British or American English which has to be approximated as good as possible (*single reference*). In a *multiple reference* scenario, we first define the target: should the speaker simply be intelligible, should he/she acquire a 'touch of nativeness', or do they aim at full nativeness? So far, a good teacher might be able to tell apart these different target scenarios, but not yet an automatic program. A wrong target for a learner might produce frustration (near-nativeness as target if the learner is far from achieving this goal) and, as a consequence, lower professional success but this will normally not be a core ethical issue. Normally, a native-like proficiency is aimed at. However, it might be better to aim at a higher diversification of targets, i.e., from *single references* (to the gold standard) towards *multiple references*.

Everything that can be considered being simple, 'tamagotchi-like' entertainment is mostly harmless. This holds for apps like 'lie-detectors' as well (note that the 'serious' lie-detector is profoundly dangerous, cf. (Kreiman & Sidtis 2011): the description of a 'lie detector'[1] states: "Use this Lie Detector on your friends and family and fool them. This is a Prank and Not a real Detector. Lie detector by default says any statement is false."

Of intermediate ethical importance is, for instance, the so-called call-centre-scenario: Estimating the length of conversation as a measure of felicity/success, or estimating the interest of the caller with the help of audio features, is ethically not critical. However, trying to detect callers' anger or sleepiness, or trying to assess agents' friend-

[1] https://play.google.com/store/apps/details?id=com.km.prank.liedetector, retrieved on March 23th, 2014.

liness might be OK, if anonymous, but highly critical, if this information is harnessed for decisions on employment.

This use case leads to the area of *affective computing* which arguably is the most important field within CP where we have to consider ethics, both when collecting data and recording subjects in experiments, and when generating (embodied) agents or robots that interact with humans not only by exchanging information but by exchanging emotions as well. Ethical issues within affective computing are discussed in (Reynolds & Picard 2004, Cowie 2011, Cowie 2012, Döring, Goldie & McGuinness 2011, Goldie, Döring & Cowie 2011, Sneddon, Goldie, & Petta 2011).

The following example is definitely not harmless: imagine a serious game intended for teaching autistic children to understand and to produce emotions. In the ASC project (Schuller et al. 2014), the children have to look at and listen to actors producing different emotions, and then, they have to produce these emotions themselves. Here, what is the criterion for right or wrong? A real 'hit', i.e., a correspondence with the prototypical acted emotion? Or the emotion can be perceived as such but is not very pronounced/typical? Or it is rather atypical and ambiguous but not outright wrong, i.e. indicating another, opposite emotion (happy instead of angry)? And most important, in the case of erroneous recognition or classification, when we teach awkward or wrong expressions of emotions, risks are high that the outcome of such a therapeutic game is not only irrelevant but outright harmful. This is an example for applications that are aimed at individual patients (here: children) with possible ethically critical impact for these individuals (Morrow & Richards 1996, Ragin & Amoroso 2011).

'Big data' approaches such as the screening of face book entries and their linguistic analysis belonging to a specific population – in (Lewis et al. 2008), Harvard students – or activities of national security agencies are definitely critical, as far as ethics in general and privacy of the individuals observed are concerned.

All these considerations pertain at the same time good research and appropriate ethics.

7. Intrinsic and Extrinsic Ethics and Computational Paralinguistics

In short, *intrinsic ethics* aims at producing sound scientific results; *extrinsic ethics* aims at the societal requirements that scientific results have to meet. (We do not tell apart intrinsic from procedural ethics as is done in (Schienke et al. 2011) because both pertain only slightly different aspects of conducting scientific research.)

Fig. 3 left, displays the relationship between CP, other most relevant scientific fields, and application domains. CP and multimodal approaches denote methodologies, health/biology/psychology/sociology both scientific fields and possible application domains, and infotainment/edutainment/entertainment denote application domains. This illustrates a close relationship and by that, the fact that ethical demands are very similar or identical across all these fields and domains. All this can be seen as 'science and its possible applications'. Fig. 3, right, shows the relationship between science and by that, intrinsic ethical demands, on the one hand, and extrinsic ethics with the two closely related but different individual and societal aspects. A few examples for individual and societal demands have been given above.

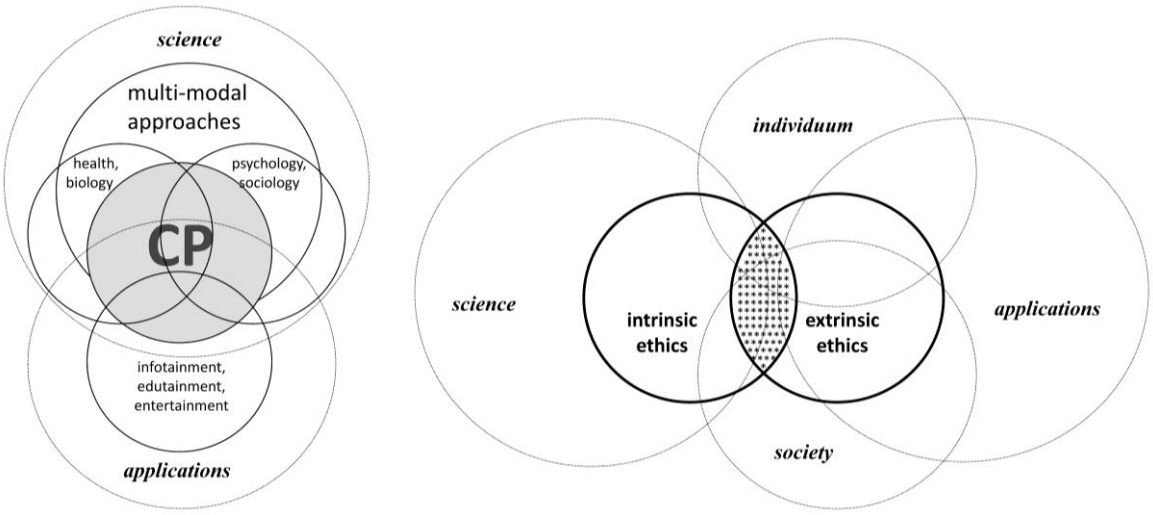

Fig. 3, left: *Domain diagram of Computational Paralinguistics (CP) and (selected) related fields;* right*: intrinsic and extrinsic ethics between science and applications, and between individual and societal demands; intersection of intrinsic and extrinsic ethics (starred).*

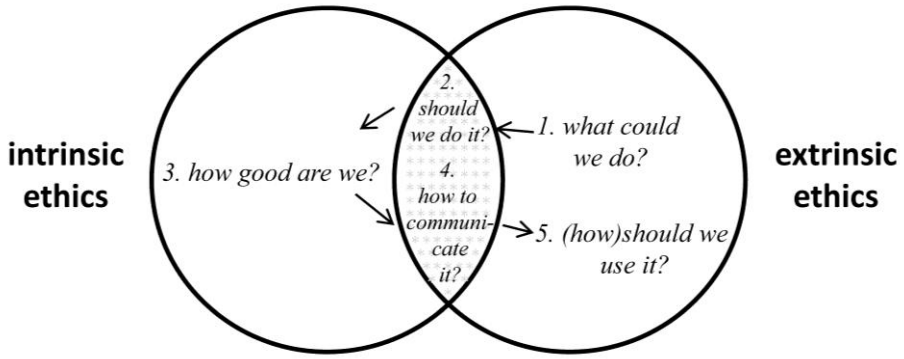

Fig. 4: *Questions to be asked within and across (intersection) intrinsic and extrinsic ethics – the 'ethical loop'.*

Fig. 4 takes up again the chain of processing, this time concentrating on the interplay of (science-) intrinsic with extrinsic ethics. We always start with the question *What could we do* (within extrinsic ethics, above) meaning 'What is promising to be addressed and then harnessed in real life', and this should be accompanied by an ethical assessment whether *we should do it,* i.e., whether our scientific aims are in line with societal ethical demands – another way to express Goethe's Dr. Faust's dilemma whether he should create a homunculus or not. Intrinsic to science are the models and the experiments that lead to an answer to the question *how good are we*. A peculiarity of the phenomena addressed by CP is that normally, they cannot be obtained in any objective but only in an inter-subjective way; thus, our gold standard is mostly based on human subjective labels and assessments (fuzzy and/or moving targets). Computer performance has to be compared with this human performance. Based on this performance and on ethical reasoning, we can find an answer to the questions whether at all *we should use it* and if yes, *how we should use it*. In between, at the intersection of intrinsic and extrinsic ethics, is the question *how to communicate* how good we are; this we now want to concentrate on. This question foremost targets how to communicate measures of performance in scientific publications and in public. The basic requirements should simply be

that we have to convey what's possible and what's not, and this has to be done in a terminology which can be understood by the public. Within all the sub-fields of CP, we are far from a standard that uses *common language* measures (McGraw & Wong 1992) which anyway only are common for highly debated topics such as change of climate, mammography or Prostate-specific antigen (PSA) values and risk of cancer where we find statements such as: out of *1000* subjects, *k* are conspicuous, and out of these *k* subjects, (only) *n* subjects do have cancer, or: within the next 30 years, temperature will rise by more than *X* degrees Celsius. More common is the statement *"We can detect X"* – per se not valid, because it sort of implies the generic statement *"We always can detect (any) X."* In fact, this should be translated into: *"For specific samples out of the whole population, we so far found out that we could detect X with an accuracy of Y%."* Here, 'accuracy' stands for different measures used in the community. In CP, we mostly deal with atypical phenomena in a figure-ground relationship to typical phenomena which are much more frequent (sparse data problem). For such a constellation, we strongly argue in favour of *Unweighted Average Recall* (UAR) which reports the mean of the diagonal in a confusion matrix in percent. This measure does not depend on the most frequent class as *Weighted Average Recall* (WAR) that is more common in ASP; chance level for UAR is always 50% for two classes, 33.3% for three classes, and so on (Schuller & Batliner 2014, Rosenberg 2012). UAR can be taken as *effect size* to be reported, with a range between 50% and 100%. Additionally, some confidence measure should be reported in order to provide more information beyond strict certainty (not yes/no decision). Note that statements about *significance* are often inherently vague – when not accompanied by information which test has been used, which alpha-level has been assumed beforehand, etc., and if they are accompanied with all this necessary information, they are still vague because they cannot be taken as statements about how good we are.

Now, what is especially 'intrinsic' to CP as far as ethics is concerned? Not much if we compare it to other modalities such as vision (facial expression, body posture and gestures, and gait): types of data collection, annotation, and computational procedures, and impact on individuals and society are very similar. Thus, especially when CP is embedded in multi-modal approaches, ethical demands are the same across modalities. Specific for scientific fields is, however, how they communicate their results; this leads back to our discussion of most appropriate effect size measures in the last paragraph: within artificial intelligence, we might more often encounter statements like "we are able to do X", without any measure of performance. Within clinical, psychological, and sociological studies, we might often encounter complex inferential statistics and statements about significance, not yet too often accompanied by appropriate effect size measures. Within ASP and partly CP, different other performance measures are used; as mentioned above, we argue in favour of UAR, if possible.

8. The Future of Computational Paralinguistics and Higher Ethical Demands

In a bird's eye view, automatic CP procedures can be conceived as modelling correct/adequate human behaviour when taking over the following sequence of tasks: they should *survey* a population (a *crowd*), *screen* a specific *sub-population* in order to find out something about specific *individuals*, *recognise/assess* – based on (mostly) human *annotations* – characteristics of these *individuals*, and then *communicate* with these individuals and (help to) *decide* about suitable strategies.

At all these steps, we will see new developments in the future, with higher technical and ethical demands:

big data: This strategy surveys the crowd and selects individuals; of course, the new possibilities will create severe privacy issues for CP as well; **autonomous learning systems** tend towards taking a life of their own, thus, we have to introduce some 'check of balances' and regular review whether they meet the criteria of 'traditional' science, cf. the discussion of Google flu trends failures in (Lazer et al. 2014).

distributed learning: New methods for distributing algorithms between local devices (embedded systems) and servers, or simply distribution of data (open or semi-open access) creates higher demands on depersonalisation and anonymisation. For instance, it might be advisable not to share raw audio data together with personal information but only to share feature vectors in such a case.

crowd sourcing: this is a rather new way of 'distributed' transcription and annotation via the internet; this makes higher numbers of annotators possible but poses new questions as for monetary compensation, contractual considerations of workers (Adda & Mariani 2013); at the same time, annotation quality will differ much from the more traditional type of annotations with a few number of experts who look at the data in one pass. This is crucial in so far as the performance of CP systems is typically measured against the same type of labelling. (cf. the demand for industrial benchmarks in the following).

industrial benchmarks: Benchmarks according to standards are needed to testify expectable performance of CP applications.

corrective feedback: the better and more elaborated applications are, the more dangerous a wrong decision can be, e.g., an erroneous corrective feedback.

For all these steps, the vision is to substitute humans with computers. The most human-like application is 'embodied conversational agents'; the one which is more or less pure paralinguistics, i.e., using only language and speech – are 'communicative but non-embodied agents'. We can start this story with Eliza (a primitive but successful NLP program, pretending to be a psychotherapist, written by Joseph Weizenbaum in the 1960ies), and let it end so far with 'her', a seemingly intelligent female operation system calling herself 'Samantha' in a film by Spike Jonze from 2013, without body, just with speech, but with full human capacities to speak, listen, understand, feel, and respond. At the time being, this is pure fiction. More modest applications can be expected to work in the near future, such as assessments of atypical speech (for example, non-native or pathological speech) with a close-to-human performance. Problems are still not enough trainings data, and prejudices on part of professionals. To be on the safe side, final decisions should not be taken by the computers but they should help in finding the right ones. Hopefully, the scientific community (in general and/or within CP) will introduce more common and generally applicable ethical standards in the future.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements No. 338164 (ERC Starting Grant iHEARu) and No. 289021 (STREP ASC-Inclusion), and from the German Research Council (DFG) under grant agreement KR 3698/4-1. The responsibility lies with the authors.

References

Adda, G. & Mariani, J.J. (2013). Economic, Legal, and Ethical Analysis of Crowdsourcing for Speech Processing. In: Eskénazi, M., Levow, G-A., Meng, H., Parent, G., & Suendermann, D. (eds). *Crowdsourcing for speech processing: applications to data collection, transcription, and assessment.* 303-334. Wiley.

Cowie, R. (2011). Editorial: 'Ethics and good practice' – Computers and forbidden places: Where machines may and may not go. In P. Petta, C. Pelachaud, & Cowie, R. (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, 707–712. Springer, Berlin.

Cowie, R. (2012). The Good Our Field Can Hope to Do, the Harm It Should Avoid. *Affective Computing, IEEE Transactions on*, vol.3, 410-423.

Furui (2005). 50 Years of Progress in Speech and Speaker Recognition Research. *ECTI Transactions on Computer and Information Technology*, vol. 1, 64-74.

Furui, S. (2009). 40 Years of Progress in Automatic Speaker Recognition**.** *Advances in Biometrics***,** Lecture Notes in Computer Science Volume 5558, 1050-1059.

Döring, S., Goldie, P., & McGuinness, S. (2011). Principalism: A method for the ethics of emotion-oriented machines. In P. Petta, C. Pelachaud, & Cowie, R. (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, 713–724. Springer, Berlin.

Goldie, P., Döring, S., & Cowie, R. (2011). The ethical distinctiveness of emotion-oriented technology: Four longterm issues. In P. Petta, C. Pelachaud, & Cowie, R. (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, 725–734. Springer, Berlin.

Kreiman, J. & Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Wiley-Blackwell, Malden, MA.

Lazer , D. , Kennedy, R., King, G. and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science, vol.* 14, 343, 1203-1205.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A. & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks,* vol 30, 330–342.

McGraw, K.O. & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, vol. 111, pp.361-365.

Morrow, V. & Richards, M. (1996). The Ethics of Social Research with Children: An Overview. *Children & Society*, 10, 90-105.

Nadkarni, P.M., Lucila Ohno-Machado, L. & Wendy W Chapman, W.W. (2011). Natural language processing: an introduction*. J. Am. Med. Inform. Assoc*., vol.18, 544-551.

Reynolds, C. & Picard, R. (2004). Affective sensors, privacy, and ethical contracts. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '04). ACM, New York, NY, USA, 1103-1106.

Ragin, C. C. & Amoroso, L. M. (2011). The ethics of social research.In: Ragin, C. C. & Amoroso, L. M. (eds.), *Constructing Social Research: The Unity and Diversity of Methods*. 59–89. Sage, Publications, Thousand Oaks, CA.

Rosenberg, A. (2012). Classifying Skewed Data: Importance Weighting to Optimize Average Recall. *Proceedings of Interspeech,* Portland, 2242-2245.

Schienke, E.W., Baum, S.D., Tuana, N., Davis, K.J. & Keller, K., (2009). Intrinsic Ethics Regarding Integrated Assessment Models for Climate Management. *Science and Engineering Ethics*, vol. 17, 503-523.

Schuller, B. & Batliner, A. (2014). *Computational paralinguistics: emotion, affect, and personality in speech and language processing.* Wiley.

Schuller, B., Marchi, E., Baron-Cohen, S., O'Reilly, H., Pigat, D., Robinson, P., Davies, I., Golan, O., Fridenson, S., Tal, S., Newman, S., Meir, N., Shillo, R., Camurri, A., Piana, S., Staglianò, A., Bölte, S., Lundqvist, D., Berggren, S., Baranger, A., & Sullings, N. (2014). The state of play of ASC-Inclusion: Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions. *Proc. of the 2nd International Workshop on Digital Games for Empowerment and Inclusion (IDGEI 2014) held in conjunction with the 19th International Conference on Intelligent User Interfaces (IUI 2014),* ACM, Haifa, Israel.

Sneddon, I., Goldie, P., & Petta, P. (2011). Ethics in emotion-oriented systems: The challenges for an ethics committee. In P. Petta, C. Pelachaud, & Cowie, R. (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, 753–768. Springer, Berlin.

Sparck Jones, K. (1992). Natural language processing: an overview, *International encyclopedia of linguistics* (ed W. Bright), New York: Oxford University Press, Vol. 3, 53-59.