Big data, deep learning: at the edge of x-ray speaker analysis

Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn. 2017. "Big data, deep learning: at the edge of x-ray speaker analysis." In *Speech and Computer: 19th International Conference, SPECOM 2017, Hatfield, UK, September 12-16, 2017*, edited by Alexey Karpov, Rodmonga Potapova, and Iosif Mporas, 20–34. Berlin [u.a.]: Springer. https://doi.org/10.1007/978-3-319-66429-3_2.

Big Data, Deep Learning – At the Edge of X-Ray Speaker Analysis

Björn W. Schuller[1,2,3]

[1] Department of Computing, Imperial College London, London SW7 2AZ, UK
bjoern.schuller@imperial.ac.uk
[2] Chair of Complex and Intelligent Systems,
University of Passau, 94032 Passau, Germany
[3] audEERING GmbH, 82205 Gilching, Germany
http://www.schuller.one

Abstract. With two years, one has roughly heard a thousand hours of speech – with ten years, around ten thousand. Similarly, an automatic speech recogniser's data hunger these days is often fed in these dimensions. In stark contrast, however, only few databases to train a speaker analysis system contain more than ten hours of speech. Yet, these systems are ideally expected to recognise the states and traits of speakers independent of the person, spoken content, language, cultural background, and acoustic disturbances at human parity or even super-human levels. While this is not reached at the time for many tasks such as speaker emotion recognition, deep learning – often described to lead to 'dramatic improvements' – in combination with sufficient learning data satisfying the 'deep data cravings' holds the promise to get us there. Luckily, every second, more than five hours of video are uploaded to the web and several hundreds of hours of audio and video communication in most languages of the world take place. If only a fraction of these data would be shared and labelled reliably, 'x-ray'-alike automatic speaker analysis could be around the corner for next gen human-computer interaction, mobile health applications, and many further benefits to society. In this light, first, a solution towards utmost efficient exploitation of the 'big' (unlabelled) data available is presented. Small-world modelling in combination with unsupervised learning help to rapidly identify potential target data of interest. Then, gamified dynamic cooperative crowdsourcing turn its labelling into an entertaining experience, while reducing the amount of required labels to a minimum by learning alongside the target task also the labellers' behaviour and reliability. Further, increasingly autonomous deep holistic end-to-end learning solutions are presented for the task at hand. Benchmarks are given from the nine research challenges co-organised by the author over the years at the annual Interspeech conference since 2009. The concluding discussion will contain some crystal ball gazing alongside practical hints not missing out on ethical aspects.

1 Introduction

X-radiation – here in the sense of Röntgen radiation is composed of x-rays, which have largely become synonymous of enabling seeing usually hidden aspects via empowering technology. The field of automatic speaker analysis or 'Computational Paralinguistics' dealing with the automatic characterisation of speakers (or authors of written text) such as by assessing states and traits from the voice acoustics and textual cues of an individual is hardly connotated with such 'see-through' abilities in a figurative sense, yet. This comes, as even those tasks which are directly accessible to a human listener can still pose problems to a machine such as when aiming at recognition of human emotion. However, largely unnoticed by the broad public, computers can indeed already provide 'x-ray alike' speaker analysis empowering humans beyond their natural skill-set in terms of listening such as when automatically estimating height or weight of a speaker [3,33] down to a few centimetres or kilograms of error, despite such tasks clearly being challenging [21] also for humans [52].

To be fair, however, humans have an impressive amount of data available to learn on speech and speaker characteristics contained in the signal – simply, as they are constantly exposed to it. Likewise, at the age of just two years, we roughly heard already as much as some thousand hours of speech. At the age of ten years, this has already increased to around ten thousand hours of speech heard [29]. Obviously, these do not come with 'labels' – rather, we learn reinforced and from the situational context on 'recognising', understanding, and analysing the speaker characteristics as conveyed in the speech signal. At the same time, we synthesise speech and learn also from coupling analysis and synthesis efforts.

In terms of sheer amount of data, an automatic speech recogniser's data hunger is these days often fed in similar dimensions. And in fact, also speech recognition engines increasingly learn in weakly supervised ways, exploiting also unlabelled speech data to go from some one or two thousand hours of training material to the order of tens of thousands [53].

This is in stark contrast to the situation in Computational Paralinguistics. There, only few databases allow to train a speaker analysis system based on more than ten hours of speech – ten hours of speech vs several thousand. Yet, expectations are high as to what these systems ideally should be able to recognise: The tasks are often ambiguous such as automatic recognition of emotion or likability or the perceived personality of a speaker – all subjective and therefore ambiguous tasks. At the same time, recognition should be reliable independent of the person, i. e., work also for unknown speakers. Then, such automatic assessment of speaker characteristics should also work independent of the spoken content and by that ideally also independent of the spoken language, i. e., as for the acoustic analysis, there should be no requirement for prompted speech. A potentially even higher depending on the type of information that shall be extracted from spoken language is the desired indifference to varying cultural backgrounds. Then, acoustic disturbances including severe cases such as multiple speakers speaking overlapping should not be in the way of reliable assessment – best at human parity or even super-human levels such as when attempting automatic recognition

of heart-beat down to a few beats of error [19], or recognition of diverse health conditions which at best a trained human could hear from the voice, or even earlier on in terms of age of the affected individual than a human could [32].

Likewise, having only a few hours of learning material at hand, it is not surprising that some automatic recognition tasks have not yet reached or surpassed human parity – an example being the above named emotion recognition from voice acoustics [40,57]. However, the recent advances in processing power, and machine learning methods – most notably deep learning which is often described to lead to 'dramatic improvements' [10] – in combination with sufficient amounts of learning data that can satisfy the 'deep data cravings' [6] that come with deep neural network approaches hold the promise to reach the point of super-human level on most or even all Computational Paralinguistics tasks likely already in the near future.

As to the amount of data available, luckily, every second, more than five hours of video are uploaded to the web. YouTube alone reached 70 million hours of video material by March 2015[1]. This is added by several hundreds of hours of audio and video communication in most languages of the world taking place. If only a fraction of these data would be shared and labelled reliably, 'x-ray'-alike automatic speaker analysis could be around the corner for next gen human-computer interaction, mobile health applications, and many further benefits to society.

In this context, the remainder of this paper is laid out as follows: first, a solution towards utmost efficient exploitation of the 'big' (unlabelled) data available is presented in Sect. 2. Small-world modelling in combination with unsupervised learning help to rapidly identify potential target data of interest. Then, gamified dynamic cooperative crowdsourcing aim at turning its labelling into an entertaining experience, while reducing the amount of required labels to a minimum by learning alongside the target task also the labellers' behaviour and reliability. Subsequently, Sect. 3 introduces increasingly autonomous deep holistic end-to-end learning solutions for the rich speaker analysis. Demonstrating the performance of today's engines, benchmarks are then given in Sect. 4. These stem from the nine research challenges dealing with Computational Paralinguistics held over the years at Interspeech. The concluding discussion will contain some crystal ball gazing alongside practical hints not missing out on ethical aspects.

2 Big Data, Little Labels – Efficiency Matters

While it was outlined above that there is sufficient data for most tasks of interest in Computational Paralinguistics owing to the rich amounts of videos available on social media, it is mostly the labels that lack. Certainly, some tasks of speaker analysis will be hard to find on social media or in conversations of millions of users, such as those dealing with rare diseases or disorders. For others, it may be hard to obtain a 'ground truth' such as accurate height of speakers, accurate heart rate of speakers, etc., from social media and human labelling alone.

[1] https://www.youtube.com/yt/press/de/statistics.html – accessed 1 June 2017.

However, for practically any task dealing with perceived speaker characteristics and some more, exploiting the data in combination with efficient human labelling mechanisms seems a promising avenue. For the remaining tasks, purely semi-supervised or unsupervised learning approaches may still benefit from sheer endless amounts of speech available [36]. In the ongoing, different ways of reaching utmost efficiency in exploiting big speech data are laid out.

2.1 Network Analysis for Pre-selection of Social Media Data

It seems obvious that labelling social multimedia needs some efficient pre-selection on 'where to start' looking at, e. g., the above named more than 70 million hours of video material available on YouTube alone. At the age of 80, we roughly lived 700 000 hours, i. e., around 1 % of the available video time on YouTube in March 2015. Entering a search term such as 'joy' in a social multimedia platform is unfortunately insufficient to quickly lead to a selection of suited videos (or directly audio streams such as by services as SoundCloud) containing joyful speech, as the retrieved videos may deal with anything related to joy such as movies, songs, etc. that somehow related to joy. This makes it evident that some smart pre-filtering is needed. Such smart pre-filtering could be realised by a 'complex network analysis' to quickly retrieve related videos from social multimedia platforms. Such platforms usually have their own suggestion on the next best related videos to watch, which could be exploited to identify next best options for more data. Unfortunately, the algorithms behind these recommendations are usually unknown, but they are mostly based on the title and description as well as more general (textual) meta-data as well as 'social' data including the viewing statistics including demographic aspects, number of likes/dislikes given by viewers, and related search queries of the users [7]. In particular, the social aspects can be unrelated or even counter-productive if establishing a database for machine learning, as they will likely lead to a biased set of data. Based on existing recommendations, one can aim to reach more suited candidates of videos by providing one's own network analysis to identify relevant videos for database establishment. This can, for example, be based on the assumption of high similarity of videos. An option is then to use interconnections of videos as generated by the social media platform's recommendations such as by small-world models and graph-based analysis finding cliques in the graph. Ideally, some content-based verification check is additionally implemented verifying coarsely that the found videos at least likely contain the desired speech samples. This can contain a speech activity detection engine or even some comparison against an initial or several initial exemplary audio streams.

2.2 Game's On! – Making Crowdsourcing Fun – Seriously

Whether freshly recorded or retrieved from social media, the speech and audio or language data next has to be annotated. Crowdsourcing can be a highly efficient way to label data, but it has also been questioned in terms of ethical aspects [1]. Such concerns touch upon whether the crowd workers are potentially

exploited [11], or "ethical norms of privacy" could be violated – potentially even knowingly by the crowd workers [18]. In addition, unreliable raters can be a severe problem adding noise to the labels [48]. In rather subjective tasks such as observed emotion or perceived personality, it can be particularly difficult to estimate the reliability of raters. Likewise, motivating the crowd worker seems an interesting option for example by gamification of the labour to turn it into fun aiming at lowering the risks of exploitation and unreliable labelling [30]. This may include social elements such as competing against other crowd workers on a leaderboard or in one vs one challenges, a point system and 'badges' or levels such as 'master rater', 'grand master', etc. An exemplary existing platform in the field is given by the iHEARu-PLAY platform [16]. More interestingly, crowd workers could experience how their work empowers Artificial Intelligence by having a gamified crowd-sourcing platform train models exclusively from their labels (or by improving existing systems with their labels) and have these compete against other crowd-workers' engines trained on their respective labels. In automatic speaker analysis, this would mean training engines based on different crowd-workers' labels and having them compete, e. g., on well-defined test-beds such as the challenges introduced in Sect. 4.

2.3 Cooperative Learning: The Matrix Needs Us

Aiming to reduce human labelling effort has long since led to the idea of self-learning by machines such as by unsupervised, semi-supervised, or reinforcement learning. This could be shown successful in Computational Paralinguistics tasks starting with the recognition of emotion [64] or the confidence estimation in emotion recognition results [9] exploiting unlabelled data and even earlier on in textual cues' exploitation [14] in sentiment analysis. Purely self-learning hardly seems unsuited, as the risk to run into stagnation of improvement despite adding exponentially more unlabelled data can be high. Furthermore, models could of course also become corrupted by purely semi-supervised learning, if no proper control mechanisms of model performance are in place to monitor the development of the models when adding increasingly more machine-labelled data for model training. Thus, even when aiming at 'never-ending learning' [27], it seems wise to keep the human in the loop by combining semi-supervised learning with active learning – an idea which has been considered early on in general machine learning [66], but only more recently in Computational Paralinguistics [63]. Likewise, rather than to harvest energy from us human beings – a sinister view on future Artificial Intelligence (AI) exploiting mankind taken in Hollywood's "The Matrix" trilogy at the last turn of millennium – AI will indeed profit from human labels.

Active learning, i. e., pre-selecting most informative instances for labelling by humans, has thereby mostly been shown to work well in simulations with 'oracle' labels. This means, experiments were carried out on fully labelled databases blinding part of the labels and revealing them only if the data has been selected for active learning. This may be overly optimistic, as the data likewise has been labelled under comparably controlled conditions, i. e., by the same individuals

on a small dataset in a short time window. However, recently it has been shown that the idea also works well in a crowdsourcing framework for Computational Paralinguistics tasks [17]. In future solutions, learning the labellers, i. e., 'being careful whom to trust when' [48] can play an increasingly important role when it comes to crowdsourcing-based annotation in an active learning manner [58]. This can also help increase efficiency when learning profiles of cross-labeller reliability to get to know optimal patterns of which combination of crowd-workers best to ask to reduce labelling efforts required.

3 Deep Learning, Broad Tasks – Holism Matters

A state-of-the-art (group of) approach(es) to best exploit 'big' data is given by the family of deep learning algorithms that provide a sufficient number of free parameters to be learnt to model complex arbitrary functions for classification or regression of highly non-linear problems [6] in efficient ways. Further, going 'broad' in the sense of widening up of the speaker characteristics targeted – ideally in full parallel – becomes possible with sufficient data. Below, it is argued that this will be beneficial even if interested in only one aspect of the speaker to reduce confusion with effects that other characteristics of the speaker may have on speech acoustics or the choice of words.

3.1 Deep Learning in Computational Paralinguistics

Deep learning has a long tradition in the field of Computational Paralinguistics: the first paper using long-short term memory (LSTM) recurrent neural networks (RNNs) for speech emotion recognition dates back some almost ten years [54], the first to use a deep architecture based on restricted Boltzmann machines – again for speech emotion recognition – appeared some three years later [45]. More recently appeared first works on convolutional neural networks (CNNs) for – speech emotion recognition [26]. However, only last year, the first true end-to-end Computational Paralinguistics system using convolutional layers ahead of LSTM layers [50] appeared. Also there, the task was emotion recognition from speech, making emotion recognition the pioneering task when it comes to deep learning in Computational Paralinguistics. This seems to hold also for one of the latest trends in deep learning – the use of generative adversarial networks [4].

In fact, largely independent of this development in deep learning exploiting acoustic information in Computational Paralinguistics, deep learning is increasingly used in the analysis of textual cues.

LSTM RNNs are for example used in sentiment analysis from textual cues [38,65]. Alternatively, gated recurrent units have been considered to the same task in [47].

CNNs are for example applied for personality analysis [25,35], computation of sentiment [35,46,65], and emotion features [35], or dialect and variety recognition [15].

Adversarial network inspirations can be found on sentiment tasks as well in [22,28].

3.2 Learning End-to-End

The learning of feature representations from the data seems attractive in a field that has been coined by huge efforts put into the design of acoustic features over the years. Indeed, as outlined above, last year first efforts in doing so were successfully reported [50]. In the work, the authors train an emotion recogniser to learn directly from the raw audio signal waveform. Furthermore, via correlation analysis, they show that the network seems to learn features that relate to the 'traditional' ones extracted by experts such as functionals of the fundamental frequency or energy contours. In [39], this is broadened up to three more paralinguistic tasks providing a benchmark of a challenge event by end-to-end learning among other ways of establishing a benchmark. While the approach is not always superior to traditional methods in these works, it shows that indeed, meaningful feature representations can be learnt from the data. One can assume that given the above named small size of corpora is the major bottleneck when it comes to reaching much more competitive results.

3.3 Borrowing Pre-trained Models from Computer Vision

This bottleneck of little data for pre-training is yet overcome in computer vision, where large pre-trained netowrks such as AlexNet [20] or VGG19 [43] exist. In [2], these are for the first time exploited for Computational Paralinguistics showing the power of the approach on the Interspeech 2017 Computational Paralinguistic Challenge's [39] snoring sub-challenge: image classification CNN descriptors are extracted from audio spectrograms called "deep spectrum features" in the paper. They are extracted by forwarding the audio spectrograms through the very deep task-independent pre-trained CNNs named previously to build up feature vectors. In this first paper, the authors evaluate the use of different spectrogram colour maps and different CNN topologies. They beat the conventionally established baseline in the challenge by a large margin, which the authors can further increase by suited feature selection by competitive swarm optimisation in [13], rendering this approach highly promising and likely supporting the claim that it is mostly about the amounts of data needed to fully exploit deep learning in Computational Paralinguistics.

3.4 Going Broad – Holistic Speaker Analysis

As the characteristics of a speaker are usually 'all present' or 'all on' more or less at the same time, it appears crucial to address them in parallel rather than one by one in isolation ignorant to potential other ones. This seems relevant even if one is only interested in one speaker characteristic, e. g., emotion of the speaker, to avoid confusion by interfering other speaker states or traits such as being tired, intoxicated by alcohol, being under a certain cognitive load, or simply with one's personality type. There are only a few approaches, yet, considering this mutual dependency of speaker characteristics, mostly based on multi-task

learning with neural networks. Examples in acoustic speech information exploitation include simultaneous assessment of age, gender, height, and race recognition [41], age, height, weight, and smoking habits recognition at the same time [34], emotion, likability, and personality assessment in one pass [62], commonly targeting deception and sincerity [60] or drowsiness and alcohol intoxication [61] in the recognition, as well as assessment of several emotion dimensions or representations in parallel [12,55,56,59], and aiming at speaker verification [5] co-learning other aspects.

Similar approaches can be found in text-based information exploitation [22].

4 Where Are We on Automatic Speaker Analysis?

The above sections laid out options for future improvement of Computational Paralinguistics, mainly by collection of more data and training deeper and 'broader' models to best exploit these data. But what are current performances? To provide an impression of what today's speaker analysis systems can reach in a nutshell, Table 1 shows the baseline results of the Interspeech challenges centred on Computational Paralinguistics. Since 2013, these are running under the unified name of Interspeech Computational Paralinguistics Challenge or INTERSPEECH COMPARE for short. Previous events were the 2009 Emotion Challenge, the 2010 Paralinguistic Challenge, and the 2011 and 2012 Speaker State and Speaker Trait Challenges[2].

In these challenges, weight is put on realism in the sense of assessing the speaker from a short snippet of audio only (usually around one to a few seconds), independent of the speaker, in mostly real-world conditions such as telephone or broadcast speech. Different measures were used over the different tasks in the 'sub-challenges' per year respecting the different type of representation or task such as classification, regression, or detection. Explanations on these are given in the caption.

The baselines have been established under somewhat similar conditions over the years based on the openSMILE toolkit[3] for large-scale acoustic feature space brute forcing with standardised feature sets (which, however, grew over the years from 384 features (2009) over 1 582 (2010), 3 996 (2011), 6 125 (2012), to 6 373 (since 2013) features on 'functional' level – partially, however, also directly (lower numbers of) low-level-descriptors on frame level were used), and WEKA[4] (mostly using Support Vector Machines). In 2017, openXBOW[5] and end-to-end learning based on TensorFlow[6], were used in addition in a fusion of methods.

From the table, one can mainly see two things: an astonishing range of speaker characteristics can be automatically extracted significantly above chance level –

[2] See http://compare.openaudio.eu/ for details on these events.
[3] http://audeering.com/technology/opensmile/.
[4] http://www.cs.waikato.ac.nz/ml/weka/.
[5] http://github.com/openXBOW/openXBOW/.
[6] http://www.tensorflow.org/.

Table 1. Interspeech Computational Paralinguistics Challange series (ComParE) baseline results over the years following similar brute-force open-source focussed computation by openSMILE and WEKA (in 2017, openXBOW and end-to-end deep learning have been used in addition) as seen in the challenge by sub-challenge. Given are the year the challenge was held, the name of the sub-challenge usually clearly representing the task targeted ("Pathology", however, deals with intelligibility of head and neck cancer patients before and after chemo-radiation treatment), the modelling (column "Model") of the task either naming the number of distinct classes to recognise, or the interval (marked by $[\cdots]$) in case of a regression task, or "x" in case several (classification) tasks had to be addressed, and the baseline results (column "Base"). Different evaluation measures were used for competition depending on the type of task and modelling of it as *classification* (result given in terms of percentage of unweighted accuracy (% UA), i. e., added recall per class divided by the number of classes to cope with imbalance across classes in the sense of chance-normalisation), *regression* (shown is the correlation coefficient (CC (2010)/ρ (else)) – marked by $^+$) or *detection* task (given is the percentage of unweighted average area under the curve (% UAAUC) – marked by *)

Year	Sub-challenge	Model	Base
2017	Addressee	2	70.2
	Cold	2	71.0
	Snoring	4	58.5
2016	Deception	2	68.3
	Sincerity	[0,1]	$.602^+$
	Native Language	11	47.5
2015	Degree of Nativeness	[0,1]	$.425^+$
	Parkinson's Condition	[0,100]	$.390^+$
	Eating Condition	7	65.9
2014	Cognitive Load	3	61.6
	Physical Load	2	71.9
2013	Social Signals	2×2	83.3^*
	Conflict	2	80.8
	Emotion	12	40.9
	Autism	4	67.1
2012	Personality	5×2	68.3
	Likability	2	59.0
	Pathology	2	68.9
2011	Intoxication	2	65.9
	Sleepiness	2	70.3
2010	Age	4	48.91
	Gender	3	81.21
	Interest	$[-1,1]$	$.421^+$
2009	Emotion	5	38.2
		2	67.7

sometimes already at super-human level such as in the case of intoxication or some pathologies –, yet leaving head room for improvement for several others if not all.

Note that in this series, both, acoustic and textual cues can mostly be exploited unless – in rare cases – the data of a sub-challenge features prompted speech. However, other challenges exist focussing on textual cues such as the annual author profiling task at PAN within the CLEF framework (cf. e. g., [37] for the latest edition), or the affective text [44], sentiment analysis [31], and other tasks in SemEval.

5 Conclusions and Perspectives

Concluding this contribution, a short summary is given followed by some perspectives.

5.1 Conclusions

Current-state performances based on the Interspeech challenge series on Computational Paralinguistics over nine years have been shown that demonstrated the richness of speaker characteristics that can be automatically accessed already by today. At the same time, these results showed the room left over for future improvements. To address this issue, an argument was made to go 'broader' in automatic speaker analysis in terms of assessment of multiple characteristics of a speaker in full parallel to avoid confusion due to co-influence of these. Further, deep learning has been named as current promising solution for modelling in terms of machine learning. As particular advantage, this allows the learning of the feature representation directly from the data – an interesting and valuable aspect in a field that is ever-since marked by major efforts going into the design of optimal feature representations. As such going 'deep and broad' requires 'big' training data, avenues towards efficient exploitation of 'big' social multimedia data in combination with gamified crowd-sourcing were shown. These included efficiency-optimising measures by smart pre-selection of instances and combined active and semi-supervised learning mechanisms to avoid human involvement in labelling as much as possible. Alternatively, exploitation of pre-trained networks on 'big' image data was named to analyse speech data based on image-related representations such as spectograms or scalograms and alike in potential future efforts. However, for some under-resourced special types of data, such as of vulnerable parts of the population [24], 'conventional' collection of data will still be required.

5.2 Some Crystal-Ball Gazing

Putting the above together in a 'life-long learning' [42] Computational Paralinguistics system supported by the crowd during 24/7 learning efforts based on big social media and contributed data, we may soon witness passing the 'edge'

of 'x-ray speaker analysis', i. e., soon see super-human level automatic speaker analysis for an astonishingly broad range of speaker characteristics.

Further supporting approaches not mentioned here include transfer learning [23], reinforcement learning [49], and tighter coupling of generative and discriminative approaches [51] or synthesis and analysis of speaker states and traits, to name but three of the most promising aspects.

Once reaching such abilities, ethical, legal, and societal implications (ELSI) will play an important role [8] if such technology is increasingly used in human-decision support such as in automatic job interviews, tele-diagnosis in health care, or monitoring of customers, and employees, to name again but three use-cases. It will be of crucial importance to invest efforts into privacy protection, reliable and meaningful automatic confidence measure provision to explain the certainty and trust one should have in the automatic assessments, and accountable communication of the 'possible' to the general public such as in down-toning trust in deception recognition, if it only works at – say – some 70 % accuracy as shown in the table above. This will require organisation of future challenges in the research community as well as ensuring widest possible spread of the word.

May we soon experience powerful and reliable automatic speaker analysis and Computational Paralinguistics applied in the best possible ways only to benefit society at large in everyday problem solving and increase of wellbeing.

Acknowledgment. The author acknowledges funding from the European Research Council within the European Union's 7th Framework Programme under grant agreement no. 338164 (Starting Grant Intelligent systems' Holistic Evolving Analysis of Real-life Universal speaker characteristics (iHEARu)) and the European Union's Horizon 2020 Framework Programme under grant agreement no. 645378 (Research Innovation Action Artificial Retrieval of Information Assistants - Virtual Agents with Linguistic Understanding, Social skills, and Personalised Aspects (ARIA-VALUSPA)). The responsobility lies with the author. The author would further like to thank his team colleague Anton Batliner at University of Passau/Germany as well as Stefan Steidl at FAU Erlangen/Germany and all other co-organisers and participants over the years for running the Interspeech Computational Paralinguistics related challenge events and turning them into a meaningful benchmark.

References

1. Adda, G., Besacier, L., Couillault, A., Fort, K., Mariani, J., De Mazancourt, H.: "Where the data are coming from?" ethics, crowdsourcing and traceability for big data in human language technology. In: Proceedings Crowdsourcing and Human Computation Multidisciplinary Workshop, Paris, France (2014)
2. Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Schuller, B.: Snore sound classification using image-based deep spectrum features. In: Proceedings INTERSPEECH, 5 p. ISCA, Stockholm (2017)
3. Arsikere, H., Lulich, S.M., Alwan, A.: Estimating speaker height and subglottal resonances using mfccs and gmms. IEEE Signal Process. Lett. **21**(2), 159–162 (2014)

4. Chang, J., Scherer, S.: Learning representations of emotional speech with deep convolutional generative adversarial networks. arXiv preprint (2017). arXiv:1705.02394
5. Chen, N., Qian, Y., Yu, K.: Multi-task learning for text-dependent speaker verification. In: Proceedings INTERSPEECH, 5 p. ISCA, Dresden, Germany (2015)
6. Chen, X.W., Lin, X.: Big data deep learning: challenges and perspectives. IEEE Access **2**, 514–525 (2014)
7. Covington, P., Adams, J., Sargin, E.: Deep neural networks for youtube recommendations. In: Proceedings 10th ACM Conference on Recommender Systems (RecSys), pp. 191–198. ACM, Boston (2016)
8. Davis, K.: Ethics of Big Data: Balancing risk and innovation. O'Reilly Media Inc., Newton (2012)
9. Deng, J., Schuller, B.: Confidence measures in speech emotion recognition based on semi-supervised learning. In: Proceedings of INTERSPEECH, 5 p. ISCA, Portland (2012)
10. Deng, L., Li, J., Huang, J.T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., et al.: Recent advances in deep learning for speech research at microsoft. In: Proceedings ICASSP, pp. 8604–8608. IEEE, Vancouver (2013)
11. Deng, X.N., Joshi, K.: Is crowdsourcing a source of worker empowerment or exploitation? understanding crowd workers perceptions of crowdsourcing career (2013)
12. Eyben, F., Wöllmer, M., Schuller, B.: A Multi-task approach to continuous five-dimensional affect sensing in natural speech. ACM Trans. Interact. Intell. Syst. Spec. Issue Affect. Interact. Nat. Environ. **2**(1), 6 (2012)
13. Freitag, M., Amiriparian, S., Cummins, N., Gerczuk, M., Schuller, B.: An 'end-to-evolution' hybrid approach for snore sound classification. In: Proceedings INTERSPEECH, 5 p. ISCA, Stockholm (2017)
14. Goldberg, A.B., Zhu, X.: Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In: Proceedings 1st Workshop on Graph Based Methods for Natural Language Processing, pp. 45–52. ACL, Stroudsburg (2006)
15. Guggilla, C.: Discrimination between similar languages, varieties and dialects using cnn-and lstm-based deep neural networks. VarDial **3**, 185 (2016)
16. Hantke, S., Eyben, F., Appel, T., Schuller, B.: ihearu-play: Introducing a game for crowdsourced data collection for affective computing. In: Proceedings 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII), pp. 891–897. aaac/IEEE, Xi'An (2015)
17. Hantke, S., Zhang, Z., Schuller, B.: Towards intelligent crowdsourcing for audio data annotation: integrating active learning in the real world. In: Proceedings INTERSPEECH, 5 p. ISCA, Stockholm, Sweden (2017)
18. Harris, C.G., Srinivasan, P.: Crowdsourcing and ethics. In: Altshuler, Y., Elovici, Y., Cremers, A.B., Aharony, N., Pentland, A. (eds.) Security and Privacy in Social Networks, pp. 67–83. Springer, Heidelberg (2013)
19. Kranjec, J., Beguš, S., Geršak, G., Drnovšek, J.: Non-contact heart rate and heart rate variability measurements: a review. Biomed. Signal Process. Control **13**, 102–112 (2014)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
21. Künzel, H.J.: How well does average fundamental frequency correlate with speaker height and weight? Phonetica **46**(1–3), 117–125 (1989)

22. Liu, P., Qiu, X., Huang, X.: Adversarial multi-task learning for text classification. arXiv preprint (2017). arXiv:1704.05742
23. Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., Zhang, G.: Transfer learning using computational intelligence: a survey. Knowl. Based Syst. **80**, 14–23 (2015)
24. Lyakso, E., Frolova, O., Dmitrieva, E., Grigorev, A., Kaya, H., Salah, A.A., Karpov, A.: EmoChildRu: emotional child russian speech corpus. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS, vol. 9319, pp. 144–152. Springer, Cham (2015). doi:10.1007/978-3-319-23132-7_18
25. Majumder, N., Poria, S., Gelbukh, A., Cambria, E.: Deep learning-based document modeling for personality detection from text. IEEE Intell. Syst. **32**(2), 74–79 (2017)
26. Mao, Q., Dong, M., Huang, Z., Zhan, Y.: Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimedia **16**(8), 2203–2213 (2014)
27. Mitchell, T.M., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Mishra, B.D., Gardner, M., Kisiel, B., Krishnamurthy, J., et al.: Never-ending learning. In: Proceedings 29th AAAI Conference on Artificial Intelligence. AAAI, Austin (2015)
28. Miyato, T., Dai, A.M., Goodfellow, I.: Virtual adversarial training for semi-supervised text classification. Stat **1050**, 25 (2016)
29. Moore, R.K.: A comparison of the data requirements of automatic speech recognition systems and human listeners. In: Proceedings INTERSPEECH, pp. 2582–2584, Geneva, Switzerland (2003)
30. Morschheuser, B., Hamari, J., Koivisto, J.: Gamification in crowdsourcing: A review. In: Proceedings 49th Hawaii International Conference on System Sciences (HICSS). pp. 4375–4384. IEEE (2016)
31. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: Semeval-2016 task 4: sentiment analysis in twitter. In: Proceedings International Workshop on Semantic Evaluations (SemEval), pp. 1–18 (2016)
32. Pokorny, F., Schuller, B., Marschik, P., Brückner, R., Nyström, P., Cummins, N., Bölte, S., Einspieler, C., Falck-Ytter, T.: Earlier identification of children with autism spectrum disorder: an automatic vocalisation-based approach. In: Proceedings INTERSPEECH, 5 p. ISCA, Stockholm (2017)
33. Poorjam, A.H., Bahari, M.H., Vasilakakis, V., et al.: Height estimation from speech signals using i-vectors and least-squares support vector regression. In: Proceedings 38th International Conference on Telecommunications and Signal Processing (TSP), pp. 1–5. IEEE, Prague (2015)
34. Poorjam, A.H., Bahari, M.H., et al.: Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In: Proceedings 4th International eConference on Computer and Knowledge Engineering (ICCKE). pp. 7–12. IEEE, Mashhad (2014)
35. Poria, S., Cambria, E., Hazarika, D., Vij, P.: A deeper look into sarcastic tweets using deep convolutional neural networks. arXiv preprint (2016). arXiv:1610.08815
36. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings 24th International Conference on Machine learning. pp. 759–766. ACM, Corvallis, OR (2007)
37. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF (2016)
38. Schuller, B., Mousa, A.E.D., Vryniotis, V.: Sentiment analysis and opinion mining: on optimal parameters and performances. Wiley Interdisc. Rev. Data Min. Knowl. Disc. **5**(5), 255–263 (2015)

39. Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., Trigeorgis, G., Tzirakis, P., Zafeiriou, S.: The INTERSPEECH 2017 computational paralinguistics challenge: addressee, Cold and Snoring.. In: Proceedings INTERSPEECH, 5 p. ISCA, Stockholm (2017)
40. Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. **1**(2), 119–131 (2010)
41. Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G., Arsić, D.: Semantic speech tagging: towards combined analysis of speaker traits. In: Proceedings AES 42nd International Conference, pp. 89–97. AES, Ilmenau (2011)
42. Silver, D.L., Yang, Q., Li, L.: Lifelong machine learning systems: Beyond learning algorithms. In: Proceedings AAAI spring symposium series. AAAI, Palo Alto (2013)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint (2014). arXiv:1409.1556
44. Strapparava, C., Mihalcea, R.: Semeval-2007 task 14: Affective text. In: Proceedings 4th International Workshop on Semantic Evaluations (SemEval), pp. 70–74. ACL, Swarthmore (2007)
45. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., Schuller, B.: Deep neural networks for acoustic emotion recognition: raising the benchmarks. In: Proceedings ICASSP, pp. 5688–5691. IEEE, Prague (2011)
46. Sun, X., Gao, F., Li, C., Ren, F.: Chinese microblog sentiment classification based on convolution neural network with content extension method. In: Proceedings 6th Biannual Conference on Affective Computing and Intelligent Interaction (ACII), pp. 408–414. aaac/IEEE, Xi'an (2015)
47. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1422–1432. ACL, Lisbon, Portugal (2015)
48. Tarasov, A., Delany, S.J., Mac Namee, B.: Dynamic estimation of worker reliability in crowdsourcing for regression tasks: making it work. Expert Syst. Appl. **41**(14), 6190–6210 (2014)
49. Taylor, M.E., Stone, P.: Transfer learning for reinforcement learning domains: a survey. J. Mach. Learn. Res. **10**, 1633–1685 (2009)
50. Trigeorgis, G., Ringeval, F., Brückner, R., Marchi, E., Nicolaou, M., Schuller, B., Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: Proceedings ICASSP, pp. 5200–5204. IEEE, Shanghai (2016)
51. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation (workshop extended abstract) (2017)
52. Van Dommelen, W.A., Moxness, B.H.: Acoustic parameters in speaker height and weight identification: sex-specific behaviour. Lang. Speech **38**(3), 267–287 (1995)
53. Walker, S., Pedersen, M., Orife, I., Flaks, J.: Semi-supervised model training for unbounded conversational speech recognition. arXiv preprint (2017). arXiv:1705.09724
54. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R.: Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In: Proceedings INTERSPEECH, pp. 597–600. ISCA, Brisbane (2008)

55. Xia, R., Liu, Y.: Leveraging valence and activation information via multi-task learning for categorical emotion recognition. In: Proceedings ICASSP, pp. 5301–5305. IEEE, Brisbane (2015)
56. Zhang, B., Provost, E.M., Essi, G.: Cross-corpus acoustic emotion recognition from singing and speaking: a multi-task learning approach. In: Proceedings ICASSP, pp. 5805–5809. IEEE, Shanghai (2016)
57. Zhang, B., Provost, E.M., Essl, G.: Cross-corpus acoustic emotion recognition with multi-task learning: seeking common ground while preserving differences. IEEE Trans. Affect. Comput. (2017)
58. Zhang, Y., Coutinho, E., Zhang, Z., Adam, M., Schuller, B.: On rater reliability and agreement based dynamic active learning. In: Proceedings 6th Biannual Conference on Affective Computing and Intelligent Interaction (ACII), pp. 70–76. aaac/IEEE, Xi'an (2015)
59. Zhang, Y., Liu, Y., Weninger, F., Schuller, B.: Multi-task deep neural network with shared hidden layers: breaking down the wall between emotion representations. In: Proceedings ICASSP, pp. 4990–4994. IEEE, New Orleans (2017)
60. Zhang, Y., Weninger, F., Ren, Z., Schuller, B.: Sincerity and deception in speech: two sides of the same coin? a transfer- and multi-task learning perspective. In: Proceedings INTERSPEECH, pp. 2041–2045. ISCA, San Francisco (2016)
61. Zhang, Y., Weninger, F., Schuller, B.: Cross-domain classification of drowsiness in speech: the case of alcohol intoxication and sleep deprivation. In: Proceedings INTERSPEECH, 5 p. ISCA, Stockholm (2017)
62. Zhang, Y., Zhou, Y., Shen, J., Schuller, B.: Semi-autonomous data enrichment based on cross-task labelling of missing targets for holistic speech analysis. In: Proceedings ICASSP, pp. 6090–6094. IEEE, Shanghai (2016)
63. Zhang, Z., Coutinho, E., Deng, J., Schuller, B.: Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 115–126 (2015)
64. Zhang, Z., Weninger, F., Wöllmer, M., Schuller, B.: Unsupervised learning in cross-corpus acoustic emotion recognition. In: Proceedings ASRU, pp. 523–528. IEEE, Big Island (2011)
65. Zhou, C., Sun, C., Liu, Z., Lau, F.: A c-lstm neural network for text classification. arXiv preprint (2015). arXiv:1511.08630
66. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings ICML 2003 Workshop on the Continuum From Labeled to Unlabeled Data in Machine Learning and Data Mining, vol. 3, Washington, DC (2003)