Automatic Speaker Analysis 2.0: Hearing the Bigger Picture

Björn W. Schuller
Chair of Complex & Intelligent Systems
University of Passau, Germany
&
Department of Computing
Imperial College London, U. K.
&
audEERING Gmbh, Gilching, Germany

Email: schuller@IEEE.org

***Abstract*—Automatic Speaker Analysis has largely focused on single aspects of a speaker such as her ID, gender, emotion, personality, or health state. This broadly ignores the interdependency of all the different states and traits impacting on the one single voice production mechanism available to a human speaker. In other words, sometimes we may sound depressed, but we simply have a flu, and hardly find the energy to put more vocal effort into our articulation and sound production. Recently, this lack gave rise to an increasingly holistic speaker analysis - assessing the 'larger picture' in one pass such as by multi-target learning. However, for a robust assessment, this requires large amount of speech and language resources labelled in rich ways to train such interdependency, and architectures able to cope with multi-target learning of massive amounts of speech data. In this light, this contribution will discuss efficient mechanisms such as large social-media pre-scanning with dynamic cooperative crowd-sourcing for rapid data collection, cross-task-labelling of these data in a wider range of attributes to reach 'big & rich' speech data, and efficient multi-target end-to-end and end-to-evolution deep learning paradigms to learn an accordingly rich representation of diverse target tasks in efficient ways. The ultimate goal behind is to enable machines to hear the 'entire' person and her condition and whereabouts behind the voice and words - rather than aiming at a single aspect blind to the overall individual and its state, thus leading to the next level of Automatic Speaker Analysis.**

I. Introduction

The automatic analysis of speech aiming at a rich characterisation of the speaker behind the sound of the voice and choice of her words has come to age by now. It offers an ever growing richness in speaker states and traits that can be assessed with increasing accuracy reaching from emotion, cognitive and physical load to eating condition, heart rate, deception and sincerity to sleepiness, intoxication, health state, age, gender, height, personality, pathologies, and whatnots as, for example, featured annually in the Interspeech Computational Paralinguistics Challenge (ComParE) competition series [1][1]. Reaching adulthood these days, it is increasingly used in a range of commercial and everyday applications. This often happens unnoticed by the larger public, such as in call centres for monitoring of quality or customer analysis, as results are – depending on the speaker characterisation task – still often not sufficiently robust to be used directly in an end-user application. However, when processing larger amounts of data, the automatic recognition of speaker states and traits such as emotional arousal, gender, or age group provides sufficiently meaningful results to be used for trend analyses and alikes.

A particular hope to increase robustness currently lies in the combined assessment of multiple speaker characteristics – the so called 'holistic' speaker or speech analysis. The idea is to become utmost independent of the co-influence of concurrent speaker states and traits – ultimately all impacting on the same voice production mechanism or the same cognitive processes responsible for the wording of one's phrases and grammar behind. In other words, we may sound depressed, but in fact perhaps simply are tired and exhausted or suffer from a flu. However, the more a technical system analyses not one state or trait of a speaker at a time in 'blind' isolation, but 'hears the larger picture' of what is going on in a speaker and what she is all about, the lower the risk of such confusions will likely be. Likewise, even if only interested in one aspect such as the speaker's emotion, it seems wise to grasp the overall state and traits of the speaker [2], [3], [4], [5]. This requires assessing a 'rich' variety of speaker characteristics simultaneously – ideally exploiting mutual dependencies between these.

On a related note in the domain of speech-to-text technology, NIST announced in 2002 the first Rich Transcription Evaluation (RT-02). The idea was to not only transcribe speech automatically, but include meta information on the speakers – at first mainly by diarisation of these. Later, this idea of targeting multiple aspects in the data was further extended, and off-springs organised such as in further "rich" transcription challenges. In [6], [7], for example, the task was alongside

[1] cf. also http://www.compare.openaudio.eu

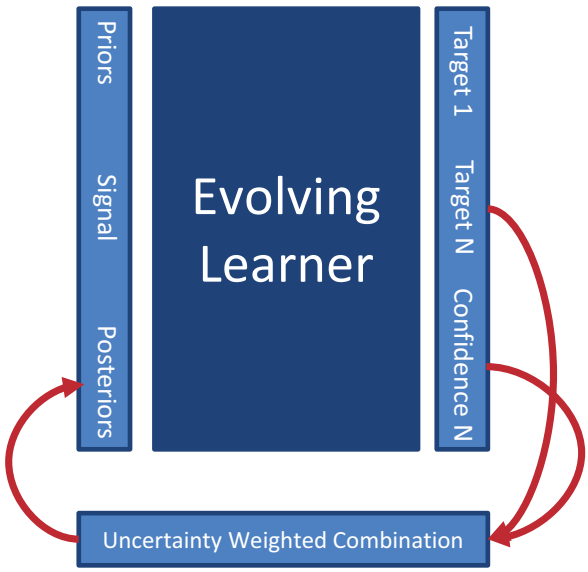

Fig. 1. Rich multiple-target speech analysis by an evolving learner that learns confidences per tasks alongside the tasks. These are iteratively fed back as input to refine the model. Further explanations are given in the text.

automatic orthographic transcription of speech to provide event detection and tracking such as for example speech vs music, and "speaker tracking" as well as "information extraction", namely named entity detection. One motivation behind is similar: the more a system understands what is overall ongoing in the data, the more likely it will assess the parts of interest correctly.

II. Automatic Speaker Analysis 2.0 – What is missing?

In the following, let us first broadly consider where we stand at in Computational Paralinguistics, or Automatic Speaker Analysis, and what is missing when doing a coarse comparison with human speaker analysis abilities.

A. Superhuman, yet? The "first encounter phone test"

An interesting way of looking at how good humans – i. e., we – are at 'speaker analysis' is a setting in everyday life, where we hear an unknown voice for the first time without seeing the person. Such a setting is given, for example, when hearing a conversational partner on a conventional (i. e., non-video enhanced) phone. We quickly assess the gender, age, likability, personality [8], social status, emotion and many further characteristics and refine our impression as the conversation goes. In other words, we do not focus on one aspect, but in fact draw the 'larger picture'. When it comes to the 'tone of the voice' we relate it carefully to single words or phrases as is needed in the context of the conversation. For example, when hearing a sequence of names uttered by a superior or someone we are interested in, we carefully relate her or his voice exactly to *our* name to sense their appreciation or sentiments towards us.

So where is Automatic Speaker Analysis in relation to these human skills? Above, it was already mentioned that most current technical systems are usually only targeting one aspect at a time or a very few at best. In addition, the temporal resolution is often somewhat arbitrary either related to the database (when processing of pre-chunked material) or to the framing or windowing of a technical process, but hardly to the word or semantic level.

As to reliability of the assessment, looking again at the related field of automatic speech recognition, first papers there claim human-level [9] or even 'superhuman' levels in accuracy. In other words, the computer has – according to these claims – exceeded human perceptive and cognitive ability in certain tasks such as speech-to-text transcriptions in particular test conditions such as adverse acoustic settings. Other examples exist, such as in image processing, where such claims are similarly made in nowadays deep learning era [10]. But where is Automatic Speaker Analysis in such terms?

Most certainly, the computer has exceeded human ability of laymen when considering the domain of health state or pathology assessment from the voice and words such as when automatically diagnosing Autism Spectrum Condition [11], Alzheimer's [12] or Parkinson's disease [13]. Other examples exist such as predicting height [14] or heart rate [15] from voice acoustics down to some centimetres or beats per minute, where automatic approaches are likely a nodge ahead, albeit human perception tests for comparison are largely missing. Mostly in the psychological and phonetic literature, some do exist such as for exemptions for human age perception in speech such as [16], [17]) or speaker height such as [**?**]. However, these studies are mostly executed on other data than the studies working on machine 'perception'. One exemplary task where both human [18] and machine perception studies exist on the same data set, and there also exists a solid ground truth rather than a subjective fuzzy point of relation such as in the case of emotion is the recognition of alcohol intoxication at comparably lower level (0.5 per mill blood alochol concentration): In a perception study [19], the participants seemed to have a tendency of lower accuracy than the top systems of the Interspeech 2011 Speaker State Challenge [20].

B. Paralinguistic granularity

From the above discussion, it seems obvious that humans are better at assessing a common analysis of a speaker, but in fact, they are also doing this in much more nuanced ways than today's technical systems do: in [21], a dozen taxonomies were taken as basis to extend the analysis beyond coarse states such as a speaker's degree of pain as perceived by others, by adding aspects such as the (degree of) acting of the felt pain (as perceived by others), the (degree of) intentionality of this acting (as perceived by others), and the (degree of) prototypicality of this acting (as perceived by others). It becomes clear that likewise, a certain depth can be established as related to each speaker characteristic by considering suited taxonomies such as degree of acting, etc. This in combination with extending the width results in a sheer endless richness – potentially also

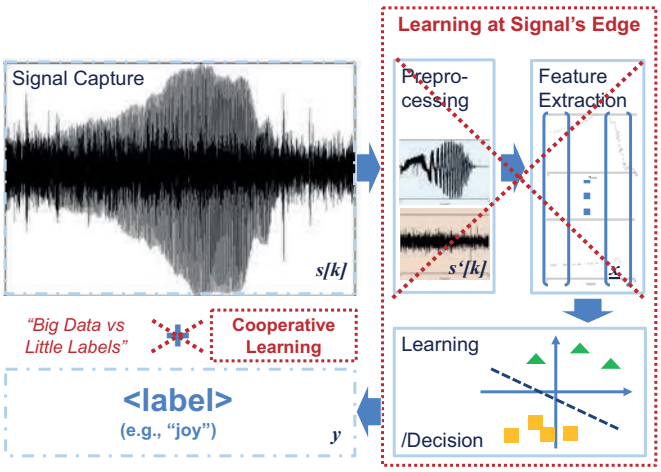

Fig. 2. Seamless learning from the audio data 'end-to-end'. Rather than dealing with pre-processing and feature extraction individually before-hand learning of a decision model, one model learns 'through' from the raw signal (i. e., at 'signal's edge') to the final label. To learn the accordingly higher number of free parameters needed to not only learn separation functions in non-linear space, but also feature representations and pre-processing, large(r) amounts of learning data are usually required. To provide such, 'cooperative learning' putting together active and semi-supervised learning [23] shall help to efficiently label the 'big' available unlabelled speech data such as on the Internet, on the radio, in television, etc. Further explanations are given in the text.

in tree structures – which is certainly not at all times followed upon by humans who rather focus depth if related to a certain purpose or interest. However, humans *are* capable of assessing such nuances, and the technical systems would yet have to follow.

C. Confidence

Furthermore, we as humans do usually have a somewhat reliable feeling for the reliability of our assessments: For example, when negotiating for a raise in salary, we listen carefully to the reaction, and not only analyse whether "let me think about it" is more likely positive or negative, but also are able to attach a confidence such as "I am quite sure it was positive" to the assessment. In Automatic Speaker Analysis, comparably little effort has been spent on provision of independent confidence measures, such as based on automatic estimation of human agreement on a paralinguistic phenomenon [22]. However, in an application context, such information is particularly useful, and more efforts have thus to follow into this direction.

III. Automatic Speaker Analysis 2.0 – Getting There

The above discussion makes it obvious that next generation speaker analysis systems should assess multiple tasks in one pass – potentially in a broad and deep fine-grained manner and relating to semantically meaningful units such as words.

A. Holism by multi-target evolving learning

Catering the concept of holism, an according technical scheme is shown in Figure 1. There, one can see multiple targets on the output side of a (machine) learning algorithm such as a neural network. Each target thereby has its own confidence information provided. This could be an additional output per target. An example would be co-learning of the labeller agreement level alongside the target task as additional output [22]. This information can be fed back to the input side as 'posteriors'. In principle, frontiers between features and target label could be washing away in such an architecture, which could be fed on the input side in the first place by the raw signal (cf. also below). Likewise, such an approach could consider co-learning of features and target labels alongside confidence measures for each of these attributes. As an example, consider co-learning of pitch – a psycho-acoustically highly complex perceptual phenomenon [24] – with speaker emotion. The learner could likewise simultaneously refine its modelling of pitch and emotion – two clearly correlated phenomena. In such way, pitch as perceived by humans could be approached more closely rather than in todays' engines which mostly use the physical fundamental frequency and some rule-based approaches towards human perception modelling such as based on frequency-dependant scaling. In fact, this can be also of particular use to aim towards better understanding of such interdependencies if the learning algorithm allows for sufficient according interpretation. This can be of particular help in coaching applications such as when giving feedback on acoustic features in relation to paralinguistic phenomena. As an example, consider the case of automatic recognition of atypical emotion such as by individuals on the Autism spectrum. A system that co-learnt feature relations alongside atypical and typical emotion could potentially give richer feedback on how to change one's vocalisations to change the perception of a certain state – as an example, pitch in order to convey emotion in less atypical manners. Similar coaching could target apps for likability, etc.

Coming back to Figure 1, on the input side, one further finds (optional) knowledge on priors. Optimal decisions usually exploit such knowledge on the a-priori distribution or expectancy of phenomena such as in (optimal) Bayesian decisions. These priors could obviously also be learnt as more data is seen gradually.

The learner as such is described as 'evolving' learner in the figure. This lends space to the idea of having the learning algorithm change itself over time if either receiving more and more data thus increasing the number of free parameters for learning, or by evolving over the output layer such as when identifying novel features or target tasks to add during seeing novel data. In simple forms, this could also be simply evolving over self-learnt feature representations such as in [25].

B. End-to-end learning

In order to cope with a huge variety of speaker analysis target tasks, self-learning of feature representations from the raw (speech signal) data has recently appeared as convenient alternative option such as by convolutional neural networks (CNNs) [26], [27], [28]. This principle is shown in Figure 2. There, one can see the raw signal as captured by a microphone

as input to further processing. Traditionally, these would be different individually tweaked blocks of processing often operating with quite diverse and heterogeneous approaches for pre-processing (depicted as source separation of speech and noise components – e. g., by non-negative matrix factorisation or other suited means), and feature extraction (shown as a series of feature vectors over time given the time series character of an audio stream). In fact, one could find many more according individual building blocks in a real system such as for hierarchical feature extraction from low-level descriptors to functional level or even histogram level in the case of bags-of-(audio)-words, or feature space optimisation, etc. Then follows the actual learning for decision making. In end-to-end learning, however, the idea would be to learn as seaminglessly as possible in order to avoid 'quantisation'-based information loss along the chain of information reduction from a several kbit/sec speech signal to a few bits of label information after the decision process.

Below the signal capture, one sees in the figure the label that is needed in order to learn (unless, of course, unsupervised clustering would be sufficient). In fact, this would rather be a vector of rich label information following the principles described above to target not one task at a time, but multiple such. Speech data for learning is usually available, yet unfortunately mostly without the needed label information. This leads to the 'big data vs little labels' paradox requiring efficient ways of labelling with low involvement of (cost-intensive) human labelling efforts. This will be considered next.

C. Data – the final frontier?

Advances in machine learning – in particular 'deep learning' recently increasingly changed the challenges of this field which just recently had been massively pre-occupied with the choice of ideal features. Nowadays, this can increasingly be tackled by learning feature representations directly from the data – even learning 'end-to-end' from the raw audio signal – as outlined above. While this solves problems, it emphasises another ever present bottleneck in the field: data scarcity. To cater the sheer endless hunger for data that comes with learning representations and models of a rich variety of speaker characteristics in parallel, one has little labelled resources at hand these days. This holds in particular for such that are labelled with a multitude of speaker characteristics rather than just one. While even some of the very early speech databases such as the 'classic' TIMIT database [29] already provide a range of speaker attributes in one database, such as age, gender, being a native speaker or not, eight "major dialects of American English", race, and even height and education level of the speaker, this information was not included for the sake of multi-target 'holistic' speaker analysis in the first place, but rather as rich information on the subjects of a corpus mainly intended for speech recognition. Such data is unfortunately also mostly based on lab recordings – potentially of prompted speech – rather than conversational speech recorded 'in the wild' such as in [30], which is much more desirable to train with in order to prepare an application for real-world conditions.

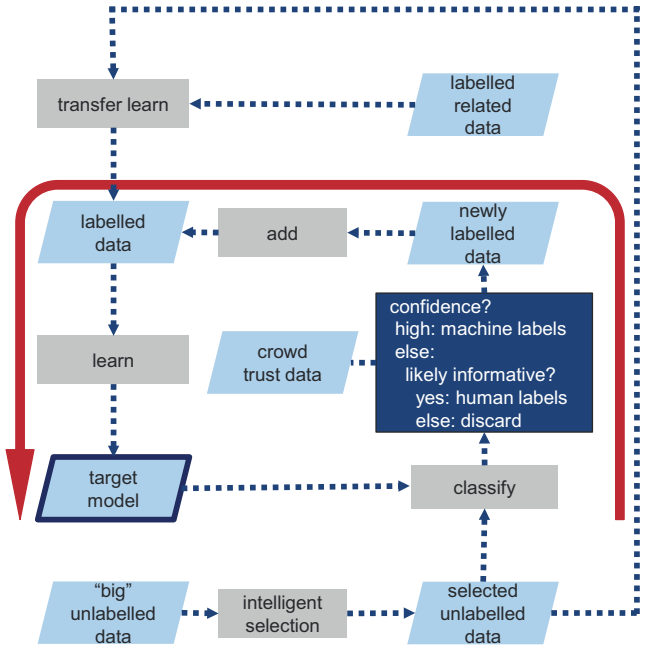

Fig. 3. In order to quickly obtain sufficient amounts of (labelled) speech data to train models for a rich variety of speaker characteristics, novel approaches of 'cooperative learning' appear as an option. In an initial step, suited (speech) data is pre-filtered from sources of 'big' (speech) data such as social media (e. g., YouTube). Suited means of pre-selection can base on social media network analyses via links across media and content descriptions, etc. Content-based potentially unsupervised verification mechanisms can be added (not shown as extra block). Then, one can optionally transfer learn from related data to produce an initial amount of labelled data. From this, a first model is learnt on assessing the target task (the 'target model'). With this model, a decision is made on novel unlabelled data. Decisions on whether a machine label (in case of high machine confidence in its decision) or human label (in case the machine cannot label itself, but the data point seems informative, i. e., relevant) is added or the data is discarded (the machine cannot label with sufficient certainty, but deems the data not of sufficient interest to ask for human aid) is chosen. In addition, the trust in the individual human annotators (the crowd) is learnt (i. e., 'whom to trust when') to further increase efficiency by asking the right annotator or optimal combination of annotators at the 'right time'. After adding the newly added data – potentially in batches – an iterative re-training and further processing of data is executed (following the red arrow). Further explanations are given in the text.

However, one is not faced with a lack of speech data, as the Internet, television, radio, and many other resources are loaded with endless amounts of data. Likewise, the real task is to cope with sparsely labelled or unlabelled data. A rich variety of solutions exists to transfer knowledge across tasks if similar tasks have been labelled previously, to label data by the machine itself, or together with the human, albeit utmost efficiently pre-selecting such data points of highest information to the machine rather than having the human label all data. For a comprehensive survey on these options and further such as using synthesised speech suiting target speaker characteristics, the reader is referred to [31].

Here, rather than giving details on transfer, active, semi-supervised, cooperative learning, and alike, I want to highlight a particularly promising avenue for increasing efficiency when aiming to 'get that data in' and most importantly, to 'get that *richly labelled* data in'. This is shown in Figure 3. Starting

with large-data social-media pre-scanning by suited means (bottom left in the figure) such as by network analysis (e. g., by small world models) and/or based on semantic tag information, one pre-selects a set of likely suited, yet unlabelled data. If related labelled data is at hand, transfer learning, e. .g, by neural network-based approaches [32] or 'cross-task-labelling' [33] where databases with different partially overlapping labels are used to label each other with all available labels in a semi-supervised iterative manner can follow to provide an initial model for decision making. This model needs to produce above chance level decisions – ideally of course comparably higher ones already.

Then, based on the initial model, decisions on the pre-selected unlabelled data are made. Dynamic cooperative learning sharing labelling efforts among machine and human while learning trust in human raters – potentially with crowd-sourcing – can then follow. Indeed, this has already been successfully shown to be efficient in real data annotation tasks for Computational Paralinguistics [34]. The idea is thereby to reduce the human labelling effort to a minimum by letting the machine annotate the data whenever it is sufficiently confident it can do so, and ask for human help only in other cases when at the same time there seems to be sufficient interest in knowing the label of the current data. According measures of informativeness can base on (high expected) novelty, scarceness of the data, or expected (significant) change of model parameters if the machine would know the label. Ideally, one also learns the trust in the raters per task and label and while obtaining labels, and their optimal combination. As an example, consider the machine being uncertain, but believing the data point is of sufficient interest. Based on its own assumption, it forwards the data to the rater who is best suited in this particular case. If his label deviates from the expected, the machine can decide to ask another rater who in this case might be best suited to ask next, etc. Note that the cooperative learning process is iterative, as the models can be retrained to increase the amount of machine annotated data with usually gradually improving models and likewise increasing confidence of the machine in its predictions.

IV. Automatic Speaker Analysis 2.0 – A brief on Responsibility

Clearly, with growing richness and fine-granularity or depth of automatically assessed speaker characteristics at increasing robustness comes an increasing ethical responsibility. This obviously holds especially in areas of 'super-human' assessment performance, as the machine may reveal aspects that humans would not notice. Likewise, once such systems start to be used on a broad scale in decision support or decision making such as in automated phone-based job interviews, computational tele-diagnosis of health state, or machine monitoring of drivers' or pilots' states (such as in case of insurance cases) to name but three delicate examples, a range of aspects need to be carefully addressed by a responsible empowering technology. These mainly include data privacy, honest and transparent communication of confidence levels and reliability to the user of such technology, but also to society at large. In other words, the limitations of such systems need to be clearly outlined in order to avoid over-expectancy. To this end, research competitions with well-defined test-beds such as the Interspeech Computational Paralinguistics Challenge mentioned above form a basis. Yet, further efforts will need to address the broader society such as a current effort by the World Economic Forum's Young Scientists'[2] recommendation on best ethical practices on a more general note. Also, with increasing big data exploitation, further ethical challenges may arise due to potential cross-correlation and connection of data points [35], [36].

V. Conclusion

In this contribution, a holistic view on Automatic Speaker Analysis was suggested that aims at assessment of fine-grained speaker characteristics in a maximal width and depth alongside confidence levels for each aspect. As a learning approach to this end, data-driven learning directly from the raw signal was suggested as one solution. The advantage being is that likewise no expert knowledge about peculiarities of each nuance of speaker characteristics are needed. Obviously, however, more 'traditional' feature brute-forcing such as by the openSMILE toolkit is an alternative. To cater the increased data requirement that arises I) from seamlessly learning from signal's edge thus largely increasing the number of free parameters to be learnt and II) the opening up towards a rich multitude of fine-grained speaker characteristics, avenues based on efficient big data exploitation were further suggested. These base on pre-filtering the large amount of available speech data such as by semantic content descriptions or network features on social media platforms before training an initial model, such as by transfer learning. Then, an iterative loop is entered where efficiency optimisation is in the foreground of efforts. The human is reasonably kept in the loop – such as by gamified crowd-sourcing, e. g., via the iHEARu-PLAY platform, but dynamic active learning helps to minimise human efforts as the machine labels itself whenever sufficiently confident.

In future architectures, an evolving element was further suggested. This may I) change the learner configuration as more data comes in. Likewise, with more data gradually available, the number of free parameters in the learner could be self-adapted. As an example, more layers may be added in a deep neural network, or more neurons in a broad neural network.

In addition, future speaker analysis engines could identify novelty to self-broaden up on the diversity of speaker characteristics or increase depth such as by known or even novel taxonomies.

In the longer run with further evolving Automatic Speaker Analysis systems, one will notice increasing impact on society once our technical systems understand our state and ad-hoc make meaningful assessments of new speakers – may these be used for the best such as in health care and wellbeing, human-machine interaction, entertainment, coaching, and many more exciting applications to be soon expected.

[2]The author of this contribution is a member.

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)). The responsobility lies with the author.

REFERENCES

[1] B. Schuller, "The Computational Paralinguistics Challenge," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, July 2012.

[2] A. H. Poorjam, M. H. Bahari *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)*. Mashhad, Iran: IEEE, 2014, pp. 7–12.

[3] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, 2015.

[4] S. B. Kalluri, A. Vijayakumar, D. Vijayasenan, and R. Singh, "Estimating multiple physical parameters from speech data," in *Proceedings IEEE 26th International Workshop onMachine Learning for Signal Processing (MLSP)*. Vietri sul Mare, Italy: IEEE, 2016, pp. 1–5.

[5] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, 2017.

[6] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news." in *Proceedings Interspeech*. Lisbon, Portugal: ISCA, 2005, pp. 1149–1152.

[7] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *Proceedings Interspeech*. Brighton, UK: ISCA, 2009, 4 pages.

[8] P. McAleer, A. Todorov, and P. Belin, "How do you say hello? personality impressions from brief novel voices," *PloS one*, vol. 9, no. 3, p. e90779, 2014.

[9] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, 2015, pp. 1026–1034.

[11] F. Pokorny, B. Schuller, P. Marschik, R. Brückner, P. Nyström, N. Cummins, S. Bölte, C. Einspieler, and T. Falck-Ytter, "Earlier Identification of Children with Autism Spectrum Disorder: An Automatic Vocalisation-based Approach," in *Proceedings Interspeech*. Stockholm, Sweden: ISCA, 2017, 5 pages.

[12] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, "On automatic diagnosis of alzheimers disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[13] T. Arias-Vergara, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Noeth, "Gender-dependent gmm-ubm for tracking parkinson's disease progression from speech," in *Proceedings Speech Communication; 12. ITG Symposium*. Paderborn, Germany: VDE/IEEE, 2016, pp. 1–5.

[14] H. Arsikere, S. M. Lulich, and A. Alwan, "Estimating speaker height and subglottal resonances using mfccs and gmms," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 159–162, 2014.

[15] B. Schuller, F. Friedmann, and F. Eyben, "Automatic Recognition of Physiological Parameters in the Human Voice: Heart Rate and Skin Conductance," in *Proceedings ICASSP*. Vancouver, Canada: IEEE, 2013, pp. 7219–7223.

[16] P. Assmann, S. Barreda, and T. Nearey, "Perception of speaker age in children's voices," in *Proceedings of Meetings on Acoustics ICA*, vol. 19, no. 1. ASA, 2013, p. 060059.

[17] S. S. Waller, M. Eriksson, and P. Sörqvist, "Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age," *Frontiers in Psychology*, vol. 6, 2015.

[18] B. Baumeister and F. Schiel, "Fundamental frequency and human perception of alcoholic intoxication in speech," in *Proceedings 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, 2015, 5 pages.

[19] ——, "Human perception of alcoholic intoxication in speech." in *Proceedings Interspeech*. Lyon, France: ISCA, 2013, pp. 1419–1423.

[20] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-Term Speaker States – A Review on Intoxication, Sleepiness and the First Challenge," *Computer Speech and Language*, vol. 28, no. 2, pp. 346–374, 2014.

[21] B. Schuller, "Reading the Author and Speaker: Towards a Holistic Approach on Automatic Assessment of What is in One's Words," in *Proceedings 18th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*, ser. LNCS. Budapest/Hungary: Springer, 2017, 12 pages.

[22] J. Deng, W. Han, and B. Schuller, "Confidence Measures for Speech Emotion Recognition: a Start," in *Proceedings Speech Communication; 10. ITG Symposium*, T. Fingscheidt and W. Kellermann, Eds., ITG. Braunschweig, Germany: IEEE, September 2012, pp. 1–4.

[23] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433–444, 2010.

[24] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[25] M. Freitag, S. Amiriparian, N. Cummins, M. Gerczuk, and B. Schuller, "An 'End-to-Evolution' Hybrid Approach for Snore Sound Classification," in *Proceedings Interspeech*. Stockholm, Sweden: ISCA, 2017, 5 pages.

[26] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network," in *Proceedings ICASSP*. Shanghai, P. R. China: IEEE, 2016, pp. 5200–5204.

[27] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Proceedings Interspeech*. Stockholm, Sweden: ISCA, 2017, 5 pages.

[28] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1705.02394*, 2017.

[29] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[30] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "Cosine-a corpus of multi-party conversational speech in noisy environments," in *Proceedings ICASSP*. Taipei, Taiwan: IEEE, 2009, pp. 4153–4156.

[31] Z. Zhang, N. Cummins, and B. Schuller, "Advanced Data Exploitation in Speech Analysis – An Overview," *IEEE Signal Processing Magazine*, vol. 34, July 2017.

[32] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum Autoencoder-based Domain Adaptation for Speech Emotion Recognition," *IEEE Signal Processing Letters*, vol. 24, 2017.

[33] Y. Zhang, Y. Zhou, J. Shen, and B. Schuller, "Semi-autonomous Data Enrichment Based on Cross-task Labelling of Missing Targets for Holistic Speech Analysis," in *Proceedings ICASSP, year = 2016, editor = , volume = , series = , pages = 6090–6094, address = Shanghai, P. R. China, publisher = IEEE,*.

[34] S. Hantke, Z. Zhang, and B. Schuller, "Towards Intelligent Crowdsourcing for Audio Data Annotation: Integrating Active Learning in the Real World," in *Proceedings Interspeech*. Stockholm, Sweden: ISCA, 2017, 5 pages.

[35] A. Abbasi, S. Sarker, and R. H. Chiang, "Big data research in information systems: Toward an inclusive research agenda," *Journal of the Association for Information Systems*, vol. 17, no. 2, p. 3, 2016.

[36] D. E. O'Leary, "Ethics for big data and analytics," *IEEE Intelligent Systems*, vol. 31, no. 4, pp. 81–84, 2016.