Journal on Multimodal User Interfaces (2018) 12:273–283
https://doi.org/10.1007/s12193-018-0270-6

ORIGINAL PAPER

Three recent trends in Paralinguistics on the way to omniscient machine intelligence

Björn W. Schuller[1] · Yue Zhang[1] · Felix Weninger[1]

Received: 2 December 2017 / Accepted: 26 July 2018 / Published online: 12 September 2018

Abstract

A 2 year-old has approximately heard a 1000 h of speech—at the age of ten, around ten thousand. Similarly, automatic speech recognisers are often trained on data in these dimensions. In stark contrast, however, only few databases to train a speaker analysis system contain more than 10 h of speech and hardly ever more than 100 h. 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 best at human parity or even superhuman levels. While this is not reached at the time for many tasks such as speaker emotion recognition, deep learning—often described to lead to significant improvements—in combination with sufficient learning data, holds the promise to reach this goal. Luckily, every second, more than 5 h 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. A major effort could thus be invested in efficient labelling and sharing of these. In this contribution, first, benchmarks are given from the nine research challenges co-organised by the authors over the years at the annual Interspeech conference since 2009. Then, approaches to utmost efficient exploitation of the 'big' (unlabelled) data available are presented. Small-world modelling in combination with unsupervised learning help to rapidly identify potential target data of interest. Further, gamified crowdsourcing combined with human-machine cooperative learning turns the annotation process into an entertaining experience, while reducing the manual labelling effort to a minimum. Moreover, increasingly autonomous deep holistic end-to-end learning solutions are presented for the tasks at hand. The concluding discussion will contain some crystal ball gazing alongside practical hints not missing out on ethical aspects.

Keywords Computational Paralinguistics · Automatic speaker analysis · Big data · Deep learning

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 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 perceiver 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 automatic speaker analysis empowering humans beyond their natural skill-set in terms of listening such as when automatically estimating height or weight of a speaker [3,37] down to a few centimetres or kilograms of error, despite such tasks clearly being challenging [24] also for humans [56].

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, we have roughly listened to some 1000 h of speech. At the age of ten, this has already increased to around ten 1000 h of speech heard [33]. Obviously, these observations do not come with 'labels'—rather, we learn in a reinforced manner and from the situational context to recognise, understand, and analyse the speaker characteristics as conveyed in the speech signal.

✉ Björn W. Schuller
bjoern.schuller@imperial.ac.uk

[1] Department of Computing, Imperial College London, London, UK

At the same time, we synthesise speech and learn also from coupling analysis and synthesis efforts.

Considering speech recognition as related discipline, a technical system today is often trained on similar amounts of data as a human would hear in her lifetime. And in fact, also speech recognition engines increasingly learn in weakly supervised ways, exploiting also unlabelled speech data to go from some one or 2000 h of training material to the order of tens of thousands [58].

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 10 h of speech 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 sentiment or the perceived personality of a speaker—all subjective and therefore ambiguous tasks. At the same time, recognition should be independent of the person, i.e., reliable also for unknown speakers. Further independence requirements include phonetic content variation robustness, including varying language. This allows for not having to ask a user to speak prompted material with known phonetic content but enables to process arbitrary speech material.

Then, acoustic disturbances including complex cases such as multiple speakers cross-talking should not be in the way of reliable assessment—best at human parity or even superhuman levels such as when optimising automatic recognition of the human heart-beat from acoustics with only a few beats of error [22], or early diagnosis of diverse health conditions which at best a physician could ascertain from the voice [36].

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 abilities—an example being the above named emotion recognition from the voice acoustics [44,61]. However, the recent advances in processing power, and machine learning methods—most notably deep learning to which significant improvements and expectations are ascribed [12]—in combination with sufficient amount of learning data that can satisfy the increased requirements for data such models usually come with [7] hold the promise to reach the point of superhuman level on most or even all Computational Paralinguistics tasks likely already in the near future.

As for the amount of data available, luckily, every second, more than 5 h 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, human-alike or even beyond automatic speaker analysis could eventually be realised for improved human-computer interaction, mobile health applications, and many further fields of application.

In this context, the remainder of this paper is laid out as follows: first, the performance benchmarks of today's engines are given in Sect. 2. These stem from the nine research challenges dealing with Computational Paralinguistics held over the years at INTERSPEECH (leaving out the still ongoing tenth challenge). Following the belief 'there is no data than more data', approaches to utmost efficient exploitation of the 'big' (unlabelled) data available are presented in Sect. 3. Small-world modelling in combination with unsupervised learning help to rapidly identify potential target data of interest. Next presented, gamified dynamic cooperative crowdsourcing aims 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. 4 introduces increasingly autonomous deep holistic end-to-end learning solutions for the rich speaker analysis. The concluding discussion will contain some future perspectives alongside practical hints including ethical aspects.

2 Where are we on automatic speaker analysis?

The foundation for intelligent speech analysis is laid by the INTERSPEECH Computational Paralinguistic Challenges (ComParE).[2] 24 paralinguistic phenomena have been examined (in the ongoing 2018 challenge, further 4 were added), encompassing a speaker's transient states and more permanent traits, as well as speaking styles. The first INTERSPEECH 2009 Emotion Challenge (IS09EC) featured a binary (idle vs negative) and a five-way (anger, emphatic, neutral, positive, and rest) classification task on naturalistic children's speech. The follow-up INTERSPEECH 2010 Paralinguistic Challenge (IS10PC), evaluated the continuous-valued level of interest ([− 1,+ 1]) and the biometric primitives age (child, youth, adult, and senior) and gender/ age (female, male, and children). In the ensuing INTERSPEECH 2011 Speaker State Challenge (IS11SSC), intoxication (above or below .5 per mill blood alcohol concentration) and sleepiness (above or below 7.5 on the Karolinska sleepiness scale) had to be detected. Next, in the INTERSPEECH 2012 Speaker Trait Challenge (IS12STC), personality (openness, conscientiousness, extraversion, agreeableness, and neuroticism), likability, and intelligibility of pathological speakers were investigated, where all tasks were binarised to above or below average. Since 2013, the Challenge series has been consistently named to ComParE, subsuming all paralinguistic tasks under one

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

[2] http://compare.openaudio.eu/.

umbrella. The ComParE 2013 targeted for the first time the sensing of social signals such as laughter and fillers (as a localisation task), as well as conflict in dyadic group discussions. In addition, it addressed atypical speech patterns due to pervasive developmental disorders (autism), and enacted emotion. In ComParE 2014, the level of cognitive load (working memory) and physical load (based on heart rate and skin conductivity) were classified. The ComParE 2015 featured two regression tasks: the degree of nativeness (i.e., non-native English prosody) and Parkinson's condition on the unified Parkinsons disease rating scale (UPDRS); and the classification task of distinguishing six different food types/eating conditions while speaking. The ComParE 2016 posed new challenges in detecting deceptive vs non-deceptive speech, estimating the degree of sincerity, and identifying the native language out of eleven L1 classes of English L2 speakers. Finally, in the 2017 Addressee sub-challenge, it had to be determined whether speech produced by an adult was directed towards another adult or a child; in the Cold sub-challenge, speech under cold had to be told apart from 'healthy' speech; and in the Snoring sub-challenge, four different types of snoring had to be classified. To add the most recent still ongoing 2018 sub-challenges, these comprise the Atypical Affect sub-challenge, where emotion of individuals with disabilities has to be recognised in four classes; in the Self-Assessed Affect sub-challenge, three levels of valence as self-assessed have to be classified; in the Crying sub-challenge, infant's crying sounds in three groups have to recognised; and in the Heart Beats sub-challenge, three degrees of heart beat diseases are contained.

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 1582 (2010), 3996 (2011), 6125 (2012), to 6373 (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 to a fusion of methods. This was added by another deep-learning baseline in 2018.

Table 1 INTERSPEECH Computational Paralinguistics Challange (ComParE) benchmarks over the years following similar brute-force open-source computation by openSMILE and WEKA (in 2017, openXBOW and end-to-end deep learning have been used in addition). Given are the year the challenge was held, the name of the sub-challenge indicating the task targeted ("Pathology", however, deals with intelligibility of head and neck cancer patients before and after chemo-radiation treatment), the modelling scheme (column "Model") of the task either referring to the number of distinct classes to recognise, or the interval (marked by [· · ·]) 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

To provide an impression of what today's speaker analysis systems can reach, Table 1 shows the baseline results of the INTERSPEECH challenges centred on Computational Paralin-

[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/.

guistics. From the table, one can mainly see two things: an astonishing range of speaker characteristics can be automatically extracted significantly above chance level—sometimes already at superhuman 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., [41] for the latest edition), or the affective text [49], sentiment analysis [35], and other tasks in SemEval.

3 Big data, little labels: efficiency matters

While it was outlined above that there are sufficient data for most tasks of interest in Computational Paralinguistics owing to the rich amounts of data 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. 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 other tasks, semi-supervised or unsupervised learning approaches can exploit speech data available without labels such as found in large quantity on social media, TV, and broadcast [40]. In the ongoing, different ways of reaching utmost efficiency in exploiting big speech data are laid out.

3.1 Network analysis for pre-selection of social media data

It seems obvious that labelling social multimedia data 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 h, 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 are 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 featuring demographic aspects, number of likes/ dislikes given by viewers, and related search queries of the users [8]. 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.

3.2 Game's on!: making crowdsourcing fun: seriously

Whether freshly recorded or retrieved from social media, the speech and audio or language data have 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 [13], or "ethical norms of privacy" could be violated—potentially even knowingly by the crowd workers [20]. In addition, unreliable raters can be a severe problem adding noise to the labels [53]. 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 [34]. 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 [18]. More interestingly, crowd workers could experience how their work empowers Artificial Intelligence by having a gamified crowdsourcing 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. 2.

3.3 Cooperative learning: human + machine

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 active learning. This could be shown successful in Computational Paralinguistics tasks starting with the recognition of emotion [68] or the confidence estimation in emotion recognition results [11] exploiting unlabelled data and even earlier on in textual cues' exploitation [16] in sentiment analysis. Purely self-learning seems unsuited, as the risk to run into stagnation of improvement can be high, despite adding exponentially more unlabelled data. 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' [31], 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 [70], but only more recently in Computational Paralinguistics [67].

Active learning, i.e., pre-selecting the most informative instances for labelling by humans, has thereby mostly been shown to work well in simulations with ground-truth labels. This means, experiments were simulated 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 [19]. In future solutions, learning the labellers, i.e., 'being careful whom to trust when' [53] can play an increasingly important role when it comes to crowdsourcing-based annotation in an active learning manner [62]. This can also help increase efficiency when learning profiles of inter-rater reliability to determine the optimal grouping of crowd-workers for reducing human labelling effort.

3.4 Using synthesised speech for training

A promising approach that disposes of the need of data labelling at all is to synthesise speech according to a specification of speaker states and traits, and then use the resulting audio as training data, along with the ground truth labels that are already known by definition. As for other emerging techniques in speaker analysis, the field of emotion recognition pioneered this paradigm in early works on using emotional speech synthesis to complement human natural emotional speech in training recognition models [46]. However, the synthesis approach used in this work suffers from a dependency on expert crafted rules to vary the speech synthesis parameters according to specific conceptualisations of emotion, making it difficult to generalise the method to other speaker states and traits, for which such mappings would have to be obtained by laborious manual research. For instance, we do not know of an implementation of speech synthesis according to a specified personality vector in terms of the OCEAN personality dimensions (openness, conscientiousness, extraversion, agreeableness, neuroticism)—although first studies exist on parameters of speech synthesis engines that contribute to personality perception, e.g., [4]—not to mention the large variety of speaker states and traits that can currently be analysed by machines (see for example Table 1). In this vein, deep learning based generative models for speech synthesis such as WaveNet [57] could be highly promising, as these allow the learning of generative models conditioned on—in principle—arbitrary speaker profile vectors as well as linguistic and prosodic features, automatically learning the relation from input profiles to synthesised waveforms in an end-to-end fashion (cf. Sect. 4.2). In the experiments done in [57], these vectors comprised encodings of speaker ID as well as textual features and fundamental frequency. In the context of data scarcity for Computational Paralinguistics, it is particularly interesting that the training material for speech synthesis per speaker ID includes several dozen hours; thus, it is conceivable to be able to train speaker trait/ state conditioned WaveNet synthesis models on the set of ComParE databases mentioned in this article, where a similar amount of data per class would be available in many cases.

4 Deep learning: machine intelligence matters

4.1 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 10 years [59], the first to use a deep architecture based on restricted Boltzmann machines—again for speech emotion recognition—appeared some 3 years later [50]. More recently, first works on convolutional neural networks (CNNs) for speech emotion recognition appeared [29]. However, only 2 years ago, the first true end-to-end Computational Paralinguistics system using convolutional layers ahead of LSTM layers [55] 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 (GANs) [5], as well as for attention mechanisms [21,30]. Both GANs and attention can be seen as mechanisms to learn features from data – while GANs can be seen as representation learning (cf. Sect. 4.4) and also allow for simulating training data, the use of attention is related to learning statistical functionals, i.e., features that summarise lower level feature contours over time. This is particularly important for Computational Paralinguistics, as many tasks in this field are formulated as sequence classification or regression tasks (mapping a sequence of features, e.g., a spectrogram, to a single target value, e.g., arousal of a speaker), and not all areas of the feature sequence are equally relevant to the problem at hand (e.g., emotional utterances may contain a significant amount of silence or non-emotional words). While these first works on attention learning were—again—restricted to emotion recognition, it is conceivable that many more works will follow this highly promising paradigm for holistic Computational Paraliguistics in the near future, given the high flexibility of the general approach.

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 [42,69]. Alternatively, gated recurrent units have been considered to the same task in [52]. CNNs are for example applied for personality analysis [28,39], computation of sentiment [39,51,69], and emotion features [39], or dialect and variety recognition [17]. Adversarial network inspirations can be found on sentiment tasks as well in [25,32].

4.2 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, in 2016 first efforts in doing so were successfully reported [55]. In that 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 [43], 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.

4.3 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 networks such as AlexNet [23] or VGG19 [48] exist. In [2], these are for the first time exploited for Computational Paralinguistics showing the power of the approach on the INTERSPEECH 2017 ComParE's [43] 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 [15], 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.

4.4 Coupling analysis and synthesis

The usage of generative adversarial networks (GANs) in combination with discriminative training is a recent trend that seems highly promising due to offering a novel paradigm of combining unlabelled with labelled data. The generic principle of GANs is as follows: a generator network transforms random noise into a waveform or feature representation of a signal (such as speech). A second network, called discriminator, is trained to distinguish outputs of the generator from real-world training examples, while the objective of the generator is to have its outputs classified as real. However, recently GANs are increasingly used as part of a discriminatively trained model, where the output of the generator is additionally fed into a neural network classifier (or regressor) predicting a target value such as emotion (see Fig. 1). For instance, this paradigm has recently been proposed for automatic valence recognition from speech [5]. In particular, this work used a large set of meeting speech, which is not labelled at all in emotional dimensions, to train the generator and discriminator—valence labels were only required for training the weights of the final classification layers [10]. Applied a similar technique to the classification of autism disorders from children's speech. It seems easy to extend this

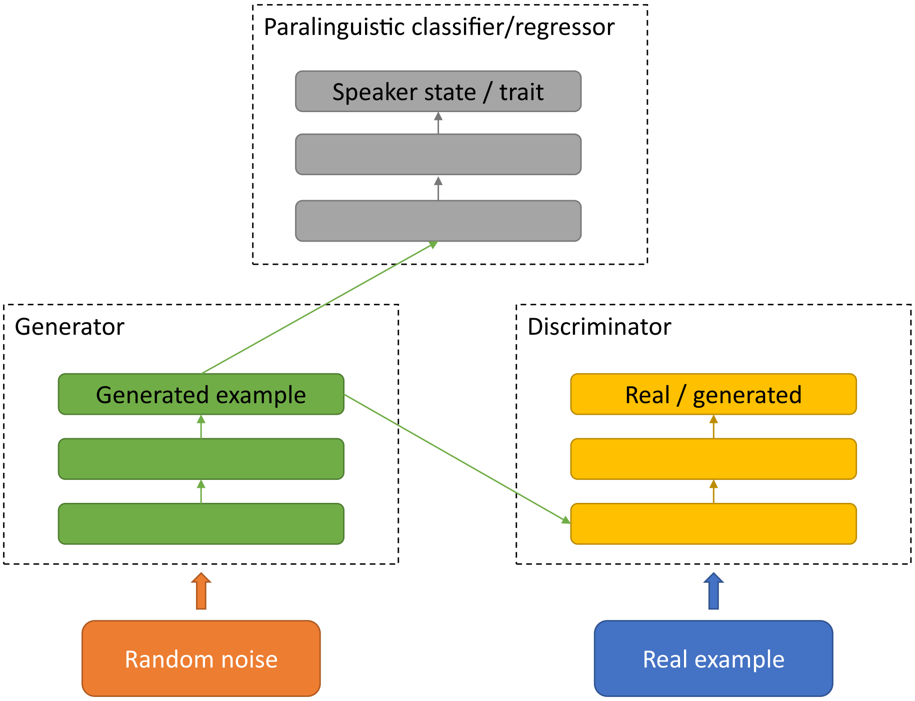

Fig. 1 Generative adversarial network (GAN) applied to feature generation for Computational Paralinguistics

to a variety of other speaker states and traits in the future, offering a new concept of leveraging unlabelled, yet diverse speech data, and, as opposed to the training with synthesised speech mentioned in Sect. 3.4, does not rely on synthesising high-quality speech in an intermediate step—instead, the system directly learns to synthesise meaningful feature representations.

5 Broad tasks: holism matters

From a methodological point of view, contemporary machine intelligence lacks holistic sensing and analysis ability in two aspects: *(1) At the front-end,* the bulk of studies treat human communication channels separately based on single-modal analysis. *(2) On the output side*, there is currently a wealth of loosely connected studies on affect recognition and machine analysis of social signals and human user characteristics; however, there is no holistic concept considering all these contiguous ontological phenomena jointly and in an associative context.

To make the dream of omniscient machine intelligence come true, we envision an end-to-end unified sensing and analysis framework based on multi-modal data processing and multi-task learning for holistic machine perception, using a novel "deep fusion" approach that explores the space between traditional early and late fusion approaches in a continuous way, where modalities can be processed with modality-dependent or shared hidden layers (see Fig. 2).

5.1 Going multi-modal: holistic machine perception

Research efforts towards multi-modal data processing have been mainly dedicated to affect recognition by means of bimodal audio-visual fusion on data-level, feature-level or decision-level. As for the machine analysis of speaker characteristics in general, the core machine perception research fields, i.e., computer vision, computer audition, and computer touch, are still widely decoupled and there exist only few works on multi-sensory integration of all perceptual modalities, including auditory, visual and tactile sensing (e.g., physiological signals, or textual input). In order to exploit joint training with databases featuring different modalities, stacked auto-encoders can be applied to raw signals as well as intermediate features extracted by the lower hidden layers of the neural network.

A number of multi-modal databases exist, mostly containing annotations with human affect, for instance, the RECOLA (REmote COLlaborative and Affective interactions) corpus, which provides audio, video and biosignals (ECG, EDA) and serves as the standard dataset in the ACM Multimedia Audio/Visual Emotion Challenge (AVEC). Another commonly used database is MAHNOB-HCI, comprising a large collection of modalities (multicamera video of face, head, speech, eye gaze, pupil size, ECG, GSR, respiration amplitude, and skin temperature). Similarly, the HUMAINE database provides naturalistic clips which record pervasive emotion (forms of feeling, expression and action that colour human life).

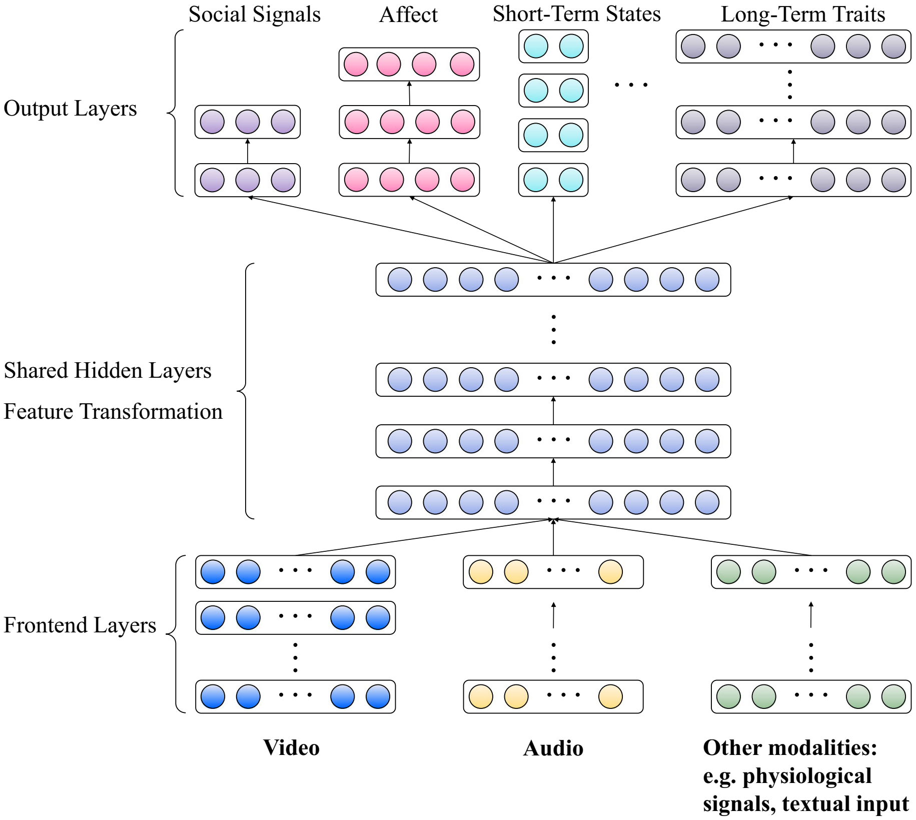

Fig. 2 End-to-end unified multi-modal multi-task deep neural network

5.2 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 [45], age, height, weight, and smoking habits recognition at the same time [38], emotion, likability, and personality assessment in one pass [66], commonly targeting deception and sincerity [64] or drowsiness and alcohol intoxication [65] in the recognition, as well as assessment of several emotion dimensions or representations in parallel [14,60,61,63], and aiming at speaker verification [6] co-learning other aspects. Similar approaches can be found in text-based information exploitation [25].

6 Conclusions and perspectives

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

6.1 Conclusions

In this article, we showed the results of the INTERSPEECH challenge series on Computational Paralinguistics over the last 9 years since their beginning (the tenth edition is currently still ongoing). The results from this series clearly indicate that a broad choice of speaker states and traits such as emotions, health state, age, personality, or gender—naming but a few—can be recognised from the voice significantly above chance level and often already quite reliably.

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 crowdsourcing 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 [27], 'conventional' collection of data will still be required.

6.2 Some crystal-ball gazing

Putting the above together in a 'life-long learning' [47] Computational Paralinguistics system supported by the crowd during 24/7 learning efforts based on big social media and contributed data, we may soon see superhuman level automatic speaker analysis for an astonishingly broad range of speaker characteristics.

Further supporting approaches not mentioned here include transfer learning [26] and reinforcement learning [54], to name but two of the most promising aspects.

Once reaching such abilities, ethical, legal, and societal implications (ELSI) will play an important role [9] 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.

Acknowledgements The authors acknowledge 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 responsibility lies with the author. The author would further like to thank his team colleague Anton Batliner at University of Augsburg/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 (2014) "Where the data are coming from?" ethics, crowdsourcing and traceability for big data in human language technology. In: Proceedings of crowdsourcing and human computation multidisciplinary workshop, Paris, France
2. Amiriparian S, Gerczuk M, Ottl S, Cummins N, Freitag M, Pugachevskiy S, Schuller B (2017) Snore sound classification using image-based deep spectrum features. In: Proceedings of INTERSPEECH. ISCA, Stockholm, Sweden
3. Arsikere H, Lulich SM, Alwan A (2014) Estimating speaker height and subglottal resonances using MFCCs and GMMs. IEEE Signal Process Lett 21(2):159–162
4. Aylett MP, Vinciarelli A, Wester M (2017) Speech synthesis for the generation of artificial personality. IEEE Trans Affect Comput
5. Chang J, Scherer S (2017) Learning representations of emotional speech with deep convolutional generative adversarial networks. In: Proceedings of ICASSP. New Orleans, LA, USA, pp 2746–2750
6. Chen N, Qian Y, Yu K (2015) Multi-task learning for text-dependent speaker verification. In: Proceedings of INTERSPEECH. ISCA, Dresden, Germany, 5 p
7. Chen XW, Lin X (2014) Big data deep learning: challenges and perspectives. IEEE Access 2:514–525
8. Covington P, Adams J, Sargin E (2016) Deep neural networks for youtube recommendations. In: Proceedings of 10th ACM conference on recommender systems (RecSys). ACM, Boston, MA, pp 191–198
9. Davis K (2012) Ethics of big data: balancing risk and innovation. O'Reilly Media Inc, Newton
10. Deng J, Cummins N, Schmitt M, Qian K, Ringeval F, Schuller B (2017) Speech-based diagnosis of autism spectrum condition by generative adversarial network representations. In: Proceedings of of the 2017 international conference on digital health. ACM, New York, NY, USA, pp 53–57
11. Deng J, Schuller B (2012) Confidence measures in speech emotion recognition based on semi-supervised learning. In: Proceedings of INTERSPEECH. ISCA, Portland, OR
12. Deng L, Li J, Huang JT, Yao K, Yu D, Seide F, Seltzer M, Zweig G, He X, Williams J, et al (2013) Recent advances in deep learning for speech research at microsoft. In: Proceedings of ICASSP. IEEE, Vancouver, BC, pp 8604–8608
13. Deng XN, Joshi K (2013) Is crowdsourcing a source of worker empowerment or exploitation? understanding crowd workers perceptions of crowdsourcing career. In: 34th International conference on information systems, Milan 2013, pp 1–10. https://pdfs.semanticscholar.org/73ef/ab88621309fdf3d39ac2aff8c70b193c0606.pdf

14. Eyben F, Wöllmer M, Schuller B (2012) A multi-task approach to continuous five-dimensional affect sensing in natural speech. ACM Trans Interact Intell Syst 2(1). https://doi.org/10.1145/2133366.2133372
15. Freitag M, Amiriparian S, Cummins N, Gerczuk M, Schuller B (2017) An 'end-to-evolution' hybrid approach for snore sound classification. In: Proceedings of INTERSPEECH. ISCA, Stockholm, Sweden
16. Goldberg AB, Zhu X (2006) Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In: Proceedings of 1st workshop on graph based methods for natural language processing. ACL, Stroudsburg, PA, pp 45–52
17. Guggilla C (2016) Discrimination between similar languages, varieties and dialects using cnn-and lstm-based deep neural networks. VarDial 3:185
18. Hantke S, Eyben F, Appel T, Schuller B (2015) ihearu-play: Introducing a game for crowdsourced data collection for affective computing. In: Proceedings of 6th biannual conference on affective computing and intelligent interaction (ACII). AAAC/IEEE, Xi'An, P. R. China, pp 891–897
19. Hantke S, Zhang Z, Schuller B (2017) Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world. In: Proceedings of INTERSPEECH. ISCA, Stockholm, Sweden
20. Harris CG, Srinivasan P (2013) Crowdsourcing and ethics. In: Altshuler Y, Elovici Y, Cremers AB, Aharony N, Pentland A (eds) Security and privacy in social networks. Springer, Berlin, pp 67–83
21. Huang CW, Narayanan SS (2016) Attention assisted discovery of sub-utterance structure in speech emotion recognition. In: Proceedings of INTERSPEECH. San Francisco, CA, USA, pp 1387–1391
22. Kranjec J, Beguš S, Geršak G, Drnovšek J (2014) Non-contact heart rate and heart rate variability measurements: a review. Biomed Signal Process Control 13:102–112
23. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Solla SA, Leen TK, Müller K-R (eds) Advances in neural information processing systems. NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems, vol 1. Lake Tahoe, Nevada, pp 1097–1105. https://dl.acm.org/citation.cfm?id=2999257
24. Künzel HJ (1989) How well does average fundamental frequency correlate with speaker height and weight? Phonetica 46(1–3):117–125
25. Liu P, Qiu X, Huang X (2017) Adversarial multi-task learning for text classification. arXiv preprint arXiv:1704.05742
26. Lu J, Behbood V, Hao P, Zuo H, Xue S, Zhang G (2015) Transfer learning using computational intelligence: a survey. Knowl-Based Syst 80:14–23
27. Lyakso E, Frolova O, Dmitrieva E, Grigorev A, Kaya H, Salah AA, Karpov A (2015) Emochildru: emotional child russian speech corpus. In: International conference on speech and computer. Springer, Athens, Greece, pp 144–152
28. Majumder N, Poria S, Gelbukh A, Cambria E (2017) Deep learning-based document modeling for personality detection from text. IEEE Intell Syst 32(2):74–79
29. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans Multimed 16(8):2203–2213
30. Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. In: Proceedings of ICASSP. New Orleans, LA, USA, p 5
31. Mitchell TM, Cohen W, Hruschka E, Talukdar P, Betteridge J, Carlson A, Mishra BD, Gardner M, Kisiel B, Krishnamurthy J, et al (2015) Never-ending learning. In: Proceedings of 29th AAAI conference on artificial intelligence. AAAI, Austin, TX
32. Miyato T, Dai AM, Goodfellow I (2016) Virtual adversarial training for semi-supervised text classification. Statistics 1050:25
33. Moore RK (2003) A comparison of the data requirements of automatic speech recognition systems and human listeners. In: Proceedings of INTERSPEECH. Geneva, Switzerland, pp 2582–2584
34. Morschheuser B, Hamari J, Koivisto J (2016) Gamification in crowdsourcing: a review. In: IEEE proceedings of 49th Hawaii international conference on system sciences (HICSS). pp 4375–4384
35. Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V (2016) Semeval-2016 task 4: sentiment analysis in twitter. In: Proceedings of international workshop on semantic evaluations (SemEval), pp 1–18
36. Pokorny F, Schuller B, Marschik P, Brückner R, Nyström P, Cummins N, Bölte S, Einspieler C, Falck-Ytter T (2017) Earlier identification of children with autism spectrum disorder: an automatic vocalisation-based approach. In: Proceedings of INTERSPEECH. ISCA, Stockholm, Sweden
37. Poorjam AH, Bahari MH, Vasilakakis V, et al (2015) Height estimation from speech signals using i-vectors and least-squares support vector regression. In: IEEE Proceedings of 38th international conference on telecommunications and signal processing (TSP). Prague, Czech Republic, pp 1–5
38. Poorjam AH, Bahari MH, et al (2014) Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In: IEEE proceedings of 4th international conference on computer and knowledge engineering (ICCKE). Mashhad, Iran, pp 7–12
39. Poria S, Cambria E, Hazarika D, Vij P (2016) A deeper look into sarcastic tweets using deep convolutional neural networks. arXiv preprint arXiv:1610.08815
40. Raina R, Battle A, Lee H, Packer B, Ng AY (2007) Self-taught learning: transfer learning from unlabeled data. In: Proceedings of 24th international conference on machine learning. ACM, Corvallis, OR, pp 759–766
41. Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B (2016) Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF
42. Schuller B, Mousa AED, Vryniotis V (2015) Sentiment analysis and opinion mining: on optimal parameters and performances. Wiley Interdiscip Rev: Data Min Knowl Discov 5(5):255–263
43. 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 (2017) The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In: Proceedings of INTERSPEECH. ISCA, Stockholm, Sweden
44. Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G (2010) Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans Affect Comput 1(2):119–131
45. Schuller B, Wöllmer M, Eyben F, Rigoll G, Arsić D (2011) Semantic speech tagging: towards combined analysis of speaker traits. In: Proceedings of AES 42nd international conference. AES, Ilmenau, Germany, pp 89–97
46. Schuller B, Zhang Z, Weninger F, Burkhardt F (2012) Synthesized speech for model training in cross-corpus recognition of human emotion. Int J Speech Technol 15(3):313–323
47. Silver DL, Yang Q, Li L (2013) Lifelong machine learning systems: beyond learning algorithms. In: Proceedings of AAAI spring symposium series. AAAI, Palo Alto, CA
48. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

49. Strapparava C, Mihalcea R (2007) Semeval-2007 task 14: affective text. In: Proceedings of 4th international workshop on semantic evaluations (SemEval). ACL, Swarthmore, PY, pp 70–74
50. Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B (2011) Deep neural networks for acoustic emotion recognition: raising the benchmarks. In: Proceedings of ICASSP. IEEE, Prague, Czech Republic, pp 5688–5691
51. Sun X, Gao F, Li C, Ren F (2015) Chinese microblog sentiment classification based on convolution neural network with content extension method. In: Proceedings of 6th biannual conference on affective computing and intelligent interaction (ACII). AAAC/IEEE, Xi'An, P. R. China, pp 408–414
52. Tang D, Qin B, Liu T (2015) Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of conference on empirical methods in natural language processing (EMNLP). ACL, Lisbon, Portugal, pp 1422–1432
53. Tarasov A, Delany SJ, Mac Namee B (2014) Dynamic estimation of worker reliability in crowdsourcing for regression tasks: making it work. Exp Syst Appl 41(14):6190–6210
54. Taylor ME, Stone P (2009) Transfer learning for reinforcement learning domains: a survey. J Mach Learn Res 10:1633–1685
55. Trigeorgis G, Ringeval F, Brückner R, Marchi E, Nicolaou M, Schuller B, Zafeiriou S (2016) Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: Proceedings of ICASSP. IEEE, Shanghai, P. R. China, pp 5200–5204
56. Van Dommelen WA, Moxness BH (1995) Acoustic parameters in speaker height and weight identification: sex-specific behaviour. Lang Speech 38(3):267–287
57. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499
58. Walker S, Pedersen M, Orife I, Flaks J (2017) Semi-supervised model training for unbounded conversational speech recognition. arXiv preprint arXiv:1705.09724
59. Wöllmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas-Cowie E, Cowie R (2008) Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. In: Proceedings of INTERSPEECH. ISCA, Brisbane, Australia, pp 597–600
60. Xia R, Liu Y (2015) Leveraging valence and activation information via multi-task learning for categorical emotion recognition. In: Proceedings of ICASSP. IEEE, Brisbane, Australia, pp 5301–5305
61. Zhang B, Provost EM, Essl G (2017) Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences. IEEE Trans Affect Comput
62. Zhang Y, Coutinho E, Zhang Z, Adam M, Schuller B (2015) On rater reliability and agreement based dynamic active learning. In: Proceedings of 6th biannual conference on affective computing and intelligent interaction (ACII). AAAC/IEEE, Xi'An, P. R. China, pp 70–76
63. Zhang Y, Liu Y, Weninger F, Schuller B (2017) Multi-task deep neural network with shared hidden layers: breaking down the wall between emotion representations. In: Proceedings of ICASSP. IEEE, New Orleans, LA, pp 4990–4994
64. Zhang Y, Weninger F, Ren Z, Schuller B (2016) Sincerity and deception in speech: two sides of the same coin? a transfer- and multi-task learning perspective. In: Proceedings of INTERSPEECH. ISCA, San Francisco, CA, pp 2041–2045
65. Zhang Y, Weninger F, Schuller B (2017) Cross-domain classification of drowsiness in speech: the case of alcohol intoxication and sleep deprivation. In: Proceedings of INTERSPEECH. ISCA, Stockholm, Sweden
66. Zhang Y, Zhou Y, Shen J, Schuller B (2016) Semi-autonomous data enrichment based on cross-task labelling of missing targets for holistic speech analysis. In: Proceedings of ICASSP. IEEE, Shanghai, P. R. China, pp 6090–6094
67. Zhang Z, Coutinho E, Deng J, Schuller B (2015) Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Trans Audio Speech Lang Process 23(1):115–126
68. Zhang Z, Weninger F, Wöllmer M, Schuller B (2011) Unsupervised learning in cross-corpus acoustic emotion recognition. In: Proceedings of ASRU. IEEE, Big Island, HI, pp 523–528
69. Zhou C, Sun C, Liu Z, Lau F (2015) A c-lstm neural network for text classification. arXiv preprint arXiv:1511.08630
70. Zhu X, Lafferty J, Ghahramani Z (2003) Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: Proc. of ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining. vol. 3. Washington, DC