Methods xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning

Nicholas Cummins[a,*], Alice Baird[a], Björn W. Schuller[a,b]

[a] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[b] GLAM – Group on Language, Audio & Music, Imperial College London, UK

ARTICLE INFO

Keywords:
Speech
Paralinguistics
Health
Deep learning
Challenges

ABSTRACT

Due to the complex and intricate nature associated with their production, the acoustic-prosodic properties of a speech signal are modulated with a range of health related effects. There is an active and growing area of machine learning research in this speech and health domain, focusing on developing paradigms to objectively extract and measure such effects. Concurrently, deep learning is transforming intelligent signal analysis, such that machines are now reaching near human capabilities in a range of recognition and analysis tasks. Herein, we review current state-of-the-art approaches with speech-based health detection, placing a particular focus on the impact of deep learning within this domain. Based on this overview, it is evident while that deep learning based solutions be become more present in the literature, it has not had the same overall dominating effect seen in other related fields. In this regard, we suggest some possible research directions aimed at fully leveraging the advantages that deep learning can offer speech-based health detection.

1. Introduction

Any given speech signal contains a rich array of information about the speaker. This information includes the linguistic content pertaining to the message the speaker wishes to communicate, as well as paralinguistic states and traits such as their current emotional states or their age and gender. Their are substantial and ongoing research efforts exploring the use of intelligent signal analysis and machine learning techniques to disengage these different facets with the aim of robustly and accurately recognising them, e.g., [1–4]. One particular aspect of speech processing research which is currently gaining in popularity, is the use of computational paralinguistic analysis to assess a multitude of different health conditions. Given the complexity of speech production and the importances of the physiological and cognitive systems involved to our (human) health and wellness – such as the respiratory system and the brain – slight changes in a speaker's physical and mental state can affect their ability to control their vocal apparatus, often at a subconscious level. Such changes can then alter the acoustic properties of the resulting speech in a manner that is measurable Moreover, as speech can be easily collected, transmitted and stored [2,5], speech-based analysis paradigms have the potential to herald a new form of active and passive remote sensing technology suitable for a broad range of health conditions.

Furthermore, speech-based health analysis is well placed at the intersection of arguably two of the most significant recent advances in computing and information systems; namely deep learning and ubiquitous computing. However, due in part to the small size of collected datasets, it is debatable if the full advantages of this positioning have even begun to be realised. The growing prevalence of deep learning can be exemplified by observing the shift in system topologies over the course of the popular *Computational Paralinguistics Challenge* (ComParE) series [6–11] and *Audio/Visual Emotion Challenge* (AVEC) workshops [12–15]. In the first health based ComParE challenge held in 2011, the recognition of speech affected by either intoxication or fatigue, not surprisingly as it was held before deep learning was considered state-of-the-art in speech processing, none of the contestants used deep learning system. On the other hand, by 2017 almost two-thirds of entrants in that year's ComParE challenge integrated some aspect of deep learning – for instance feature representation learning, classification or both – into their system.

Concurrent to the deep learning revolution, the advent of the *Internet-of-Things* (IoT) means there is currently a vast array of microphone enabled smart-devices and wearable technologies on the market, e.g., the Apple Watch™ series, Samsung Gear™ technology, or the Sony SmartWatch™ series. It has been predicted that the IoT will connect has many as 28 Billion devices by 2020 [16]. Voice-based applications for the remote monitor of speaker states and traits including health are becoming more conspicuous in the relevant

* Corresponding author.
E-mail addresses: nicholas.cummins@ieee.org (N. Cummins), bjoern.schuller@imperial.ac.uk (B.W. Schuller).

https://doi.org/10.1016/j.ymeth.2018.07.007
Received 28 March 2018; Received in revised form 11 July 2018; Accepted 19 July 2018

literature [17–19]. Smart monitoring technologies, based on deep learning and big data can potentially play a key role in helping the early and remote diagnosis of various health conditions.

Primarily, this review has been undertaken to compare the growing influence of deep learning approaches in speech analysis for health with current state-of-the-art (non-deep) machine learning approaches. It first overviews how different health states affect the various muscular and cognitive processes involved in speech production (Section 2), we then discuss the major breakthroughs that kick-started the current deep learning revolution (Section 3). The main contribution of this article is a review of the deep learning approaches conducted on the publicly available datasets associated with the ComParE and AVEC health related sub-challenges. Finally, the opportunities and challenges associated with advancing remote deep learning speech-based sensing technologies are then discussed (Section 5), and a brief conclusion is offered (Section 6).

2. Speech production and health

Human vocal anatomy is a unique and intricate anatomical structure (cf. Fig. 1.) which affords us the ability to vocalise a large variety of acoustic signals in a coordinated and meaningful manner [20]. It is the complexity of speech production that make it a suitable marker for a range of different health conditions; speech motor control, as well as requiring the coordination of the articulators; the mandible, lip, tongue, velum, jaw and pharyngeal constrictors, the respiratory muscles; including the diaphragm and the intercostal muscle, as well as the larynx [21,22], it is also the fastest discrete human motor activity [22]. There is also considerable cognitive aspect involved in speech production including message formation (what does the speaker wish to say) and storage in working memory [23], as well as the planning and implementation of the required articulatory movements to produce the required acoustic output [24].

The mechanisms of underlying vocalisation are commonly modelled as an independent (from each other) source-filter operation. The sound source in this model can either be periodic, in which air flow from the lungs produced during exhalation is modulated by the oscillating action of the vocal folds; or aperiodic, in which the vocal folds are lax and turbulent air exhaled from the lungs surges through the open glottis. The sound source then excites the vocal tract filter. The positing of the articulators, including the mandible, lip, tongue, velum, jaw and pharyngeal constrictors, alters the shape of the vocal tract determining its frequency response; the resonance frequencies of the vocal tract filter are known as the formant frequencies. Speech is considered a quasi-stationary process, i.e., while speech constantly changes the vocal tract properties is assumed to be stationary for small periods of time, roughly 25–40 ms, in order to permit short-time analysis.

In general, it can by hypothesised that cognitive related disorders can potentially disrupt message planning, and pre-articulatory functions including neurological processes relating to the implementation of the desired vocal tract movements. The fine motor control needed to produce speech also enhances its suitability as a health marker. Neuromuscular, muscular and movement disorders can effect the muscle and control systems needed to produce speech, affecting the quality of phonation, formant distribution, articulatory coordination, and diadochokinesis, to name but a few associated effects.

From here on in, we review the current state-of-the-art for detecting a range of different health conditions from the speech signal. Specifically, the health conditions reviewed include: *Intoxication* and *Fatigue* [6]; *Chemo-radiation therapy* [7], *Autistic Spectrum Conditions* [8]; *Physical and Cognitive Load* [9]; *Parkinson's Disease* and *Eating Condition* [10]; *Upper Respiratory Tract Infection* and *Snoring* [11]; as well as *Depression* [12–15]. We focus on these conditions as the corresponding datasets have been utilised in ComParE and AVEC challenges. Therefore, the datasets are publicly available, have commonly defined train, development and test partitions, and have a transparent baseline system and associated accuracy scores. Furthermore the nature of the challenges is such that they encourage participants to test the efficacy of newer approaches and compare performance not just with an established baseline, but with other state-of-the-art approaches. These properties allow for a straight comparison between different approaches and, as a core aim of the review process, effectively and systematically assess the impact of deep learning in this field of research.

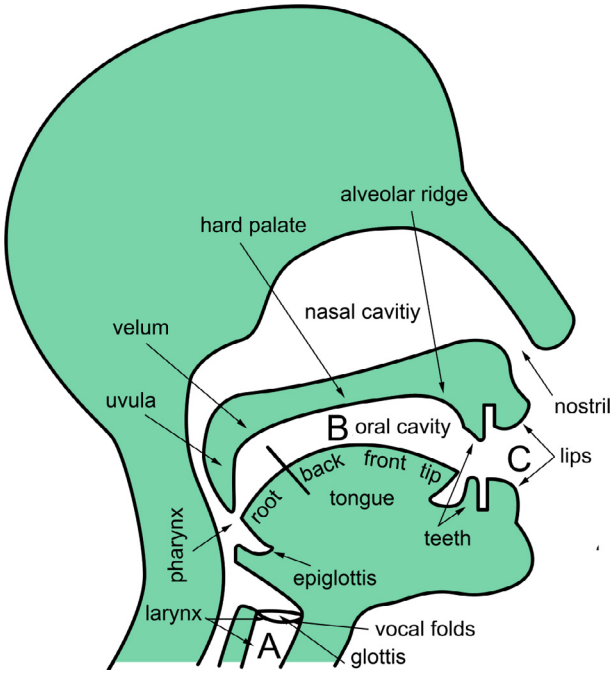

Fig. 1. An illustrative overview of the key muscle groups and anatomical structures used in speech production. The fine control required to coordinate these muscles and structures produce speech facilitate it's usefulness as a marker for a range of different health conditions. Figure adapted from [25].

3. Deep learning: the breakthroughs and current state-of-the-art

As a key objective of this review is to assess the impact of deep learning over the ComParE and AVEC challenges to date, the aim of this section is to briefly introduce deep learning from the prospective of the major breakthroughs related to the advent of the deep learning is the current state-of-the-art machine learning technique in a range of machine intelligence tasks. It should be noted that this section is intended to be an introductory, higher level overview to deep learning; for more in-depth information the interested reader is referred to the contained references, as well as to [26].

Deep learning is distinguished by artificial neural networks that, in general, have two or more layers. The core component of a *Deep Neural Network* (DNN) is the artificial neuron unit. Essentially, the role of these units is to either attenuate or amplify signals imputed from other neighbouring units. This is achieved by passing a weighted sum of inputs through a, typically non-linear, activation function thus creating the transformed output signal (cf. Fig. 2). The output equation of such as neuron is given by

$$y = \varphi(\omega^T x + b). \tag{1}$$

The concatenation of these neurons in a side-by-side manner forms a single layer network. However, the advantages of deep learning are realised by the stacking of single layer networks to create a multi-layered pipeline of non-linear transformations capable of learning feature representations, suitable for a given task, at various levels of abstraction.

Despite computational neural network research beginning in the

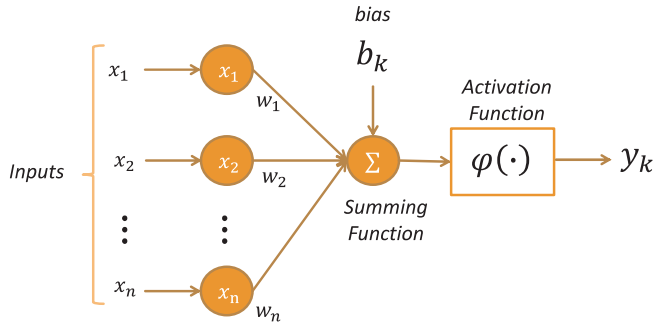

Fig. 2. Illustrative example of an artificial neuron unit. The neuron attenuates or amplifies signal a weighted sum of input signals by passing them through a (non-linear) activation function.

1960's and major advances in the subsequent years [27], the breakthrough that, arguably, kick-started the current deep learning revolution was the development of the structured layer-by-layer training paradigm of *Deep Belief Networks* (DBN) [28] and *Stacked Autoencoders* (SAE) [29]. These topologies are formed by stacking multiple layers of *Restricted Boltzmann Machines* (RBMs) or *feedforward autoencoders*, respectively. The deep nature of these topologies was realised by the unsupervised pre-training of the individual layers, followed by *back-propagation* to fine tuning of the entire network [30].

Other breakthroughs which have, arguably, also played a considerable role in the deep learning revolution are the use of *dropout*, the use of *Rectified linear units* (ReLU) as activation functions, and improved *Graphics Processing Units* (GPU). Dropout is an exceptionally simple yet powerful regularisation technique to alleviate overfitting when training deep networks. Overfitting is the effect observed when a model is said to have high variance in relation to the training data; essentially the network accounts for a maximal amount of variation in the training data. This effect results in a model that does not generalise too additional, held-out, testing data. Dropout is the random removal of units from a network during training to help improve generalisation [31]. The advent of using ReLUs [32] was of particular importance for task such as speech-based health analysis which often have a limited amount of training data [33]. The output of the ReLU is given as:

$$y = max(0, \omega^T x + b), \tag{2}$$

which allows the unit to have a constant gradient when the summation of the inputs is positive, while the desired non-linearity is achieved by having a zero output when the summation of the inputs non-positive. The nature of this output results in a network which consists of ReLU's, being easier to train as it does not suffer from the 'vanishing gradient' issues associated with other activation functions [33]. Finally, the ongoing improvements in GPU technology has enabled researchers to train networks at speeds considerably faster than those realistically achievable using a standard computer processing units [34,35].

Although DBNs and SAEs marked a major breakthrough in deep learning, it is disputable if they are still considered state-of-the-art. More recent deep learning advances in speech processing and computational paralinguistics have been focused around the development of *End-to-End* (E2E) learning paradigms [36]. Such systems generally consist of a *Convolutional Neural Network* (CNN) to learn robust feature representations directly from a raw waveform, followed by a *Recurrent Neural Network* (RNN) to leverage the temporal dynamics associated with time series data such as speech.

CNNs originated in image processing applications [37] and are considered a biological inspired variant of the *Multilayer Perceptron* (MLP) [27]. They generally consist of a combination of three main building blocks; conventional filters, pooling operations and ReLU activations. The Convolutional layers perform a of set filtering (kernel) operations; each neuron of the filter is connected with a local receptive field of previous layers and extracts a local feature map. The role of the pooling or sub-sampling operations is then to reduce the dimensionality of each feature map while retaining the most important information. Finally, non-linearities are introduced to the feature extraction processes the use of the ReLu activations. Most contemporary CNNs involve several combinations these blocks concatenated in a deep structure. When used in isolation, i.e., not in conjunction with a RNN, this structure is followed by fully connected layers to achieved the required prediction output.

A drawback of the DBN, SAE and CNN topologies is that connections are only available between two adjacent layers; this means they typically operate on inputs of fixed dimensionality and do not take into account any temporal dependencies that may exist between processing blocks. RNN's on the other hand allow for cyclical connections endowing them with the capability of accessing previously processed inputs. To cope with the vanishing or exploding gradient problems caused by the backpropagated error when training either blows up or decays over time [38], recurrent neurons contain a gating mechanism, such as *Long-Short-Term-Memory* (LSTM) [39] or *Gated Recurrent Units* (GRU) [40], to explicitly model long-term dependencies. The role of a gating mechanism is to control the flow of information inside each neuron, essentially allowing the network to learn how much past information is retained or forgotten. As well as being a core component in E2E networks, RNNs are widely used as a stand-alone classifier in a range of computational paralinguistic tasks, e.g., [17,41–43].

4. Speech-health challenges

The aim of this section is to highlight the current state-of-the-art approaches associated with the publicly available datasets released through the aforementioned ComParE and AVEC speech-health challenges. An overview of the core characteristics of these datasets is given in Table 1. To better understand the increasing impact of deep learning in this research field, this review is conducted in chronological order – from the first challenge to include a speech health conditions in 2011, speech affected by *sleepiness* or *intoxication*,through to the *cold and flu* and *snore sound* challenges held in 2017. However, before beginning these reviews, the set-up and evaluation metrics used to set the baseline scores in these challenges are first introduced.

4.1. Baseline systems

As already mentioned, a particular reason for focusing this review on the ComParE and AVEC challenges is the presence of a transparent baseline system and associated scoring metrics. This property enables the use of straightforward comparisons to assess improvements, if any, offered through new approaches. However, aside from outperforming the baseline, it is worth noting that there is currently no strict rule as to what constitutes a 'good' system performance. Furthermore, it is worth noting here that larger scale studies, potentially undertaken as part of a clinical trail, are required to establish what constitutes system performances suitable to deploy speech-based systems into clinical practices.

The baseline system for the majority of these works is set using a *brute-forced* feature extraction paradigm to create a single, yet high dimensional, acoustic representation of an utterance. First, various frame-level features, commonly refereed to as *Low Level Descriptors* (LLDs), are extracted from a speech utterance typically at frame rates of 25–40 ms, as at a frame rate of 10 ms. Exemplar speech LLDs's include pitch, energy and spectral descriptors. Next, numerous functionals, such as moments, extremes, percentiles and regression parameters, are applied to each LLD to produce a set of utterance levels summaries. These statistical summaries are then concatenated together to form the single, but often high dimensional, feature representation of an utterance. In all the ComParE and AVEC challenges the baseline sets are extracted using the openSMILE toolkit [44]. openSMILE provides software solutions to enable users to extract large audio feature spaces in (near) real-time. Furthermore to ensure reproducibility a standard set of scripts is

Table 1
Available (speech health vocalisation) corpora, from the previous AVEC and ComParE challenges. Where possible information on language, participants, gender, and age is given.

Corpora (Domain)	Challenge	Participant information
Sleep Language Corpus (SLC) (Fatigue)	ComParE-2011 [6]	German, 99 (56 f), mean age 24.9 years (20–52), ± 4.2
Alcohol Language Corpus (ALC) [45] (Intoxication)	ComParE-2011 [6]	German, 162 (78 f), mean age 31.0 years (21–75) ± 9.5
Netherlands Cancer Institute Concomitant ChemoRadioTherapy Speech Corpus (NKI CCRT) [46] (Pathology)	ComParE-2012 [7]	Dutch, 55 (10 f), mean age 57.0 year
Child Pathological Dataset (CPD) [47]	ComParE-2013 [8]	French, 99, (gender distribution not given), 6–18 years
Audio-Visual Depressive Language Corpus (AViD-Corpus) (Depression)	AVEC-2013 [12] & AVEC-2014 [13]	German, 292, (gender distribution not given), mean age 31.5 years
Munich Bio-voice Corpus (MBC) [48] (Physical Load)	ComParE-2014 [9]	German, 19, (4 f)
Cognitive Load with Speech and EGG (CSLE) [49] (Cognitive Load)	ComParE-2014 [9]	English, 26 (6 f)
Parkinson's Corpus (PC) [50] (Parkinson's)	ComParE-2015 [10]	Spanish, 100 (50 PD and 50 healthy, 25M 25F each), Men 33–77 mean 62.2 sd 11.2, Women 44–75, m 60, sd, 7.8
iHEARu-EAT (EC) (Eating Condition) [51]	ComParE-2015 [10]	German, 30 subjects (15 f, 15 m; 26.12.7 years)
Upper Respiratory Tract Infection Dataset (URTI) [52] (Cold)	ComParE-2017 [11]	630, (248 f), mean age 29.5 years, standard deviation 12.1
Munich-Passau Snore Sound Corpus (MPSSC) (Snore)	ComParE-2017 [11]	843 snore events, from 224 subjects (no native language, gender, or age is given)
Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) (Depression) [53]	AVEC-2016 [14] &AVEC-2017 [15]	(multimodal) English, 621, 18 and 60 years old

provided with openSMILE to enable users to replicate the feature representation used in the challenges.

Classification is, mainly, achieved in the baseline systems using linear kernel *Support Vector Machines* (SVM)/*Support Vector Regression* (SVR). These techniques are used as they are considered to be robust against overfitting, which is a common issue when utilising small and unbalanced datasets. Again, to ensure reproducibility, these systems are implemented in the open-source machine learning toolkit Weka [54].

4.2. Evaluation metrics

Within Machine Learning there are many metrics in which accuracy and efficiency can be measured against. Within this review we mainly refer to *Unweighted Average Recall* (UAR) for class-based classification challenges and the *Root Mean Square Error* (RMSE) for the regression tasks. UAR is commonly used as the vast majority of datasets have a (highly) imbalanced class ratio. Therefore, as this metric gives equal weight to all classes it is better suited than a weighted metric e.g., accuracy, which could give an artificially high reading caused by constantly picking the majority class. UAR is given by:

$$UAR = \frac{TP}{(TP + FN)} \tag{3}$$

where *TP* denotes the number of true positives and, *FN* the number of false negatives as classified by a model, UAR is expressed as a percentage value between 0 and 100. RMSE is a error metric that measures the spread of the predicted values around a regression line. RMSE is calculated as

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)} \tag{4}$$

where $\hat{y}$ denotes the predicted value, y_n, the actual score and N the total number of test instances. RMSE has the same units as the response variable, with lower values of RMSE indicating better system performances.

The remainder of Section 4 reviews the state-of-art in the speech-based health challenges; namely: *Sleepiness* (Section 4.3); *Intoxication* (Section 4.4); *Pathologic Speech* (Section 4.5); *Autistic Spectrum Conditions* (Section 4.6); *Depression*, specifically work carried out for AVEC-2013 and −2014 (Section 4.7); *Cognitive and Physical Load* (Section 4.8); *Parkinson's Disease* (Section 4.9); *Eating Condition* (Section 4.10); *Depression*, specifically work carried out for AVEC-2016 and −2017 (Section 4.11); *Cold and Flu* (Section 4.12); and, *Snoring* (Section 4.13). Each challenge is presented as a stand-alone subsection and also offers a comparisons between state-of-the-art and deep learning approaches undertaken on the various corpora.

4.3. Sleepiness (2011)

The third edition of the ComParE challenge series, held in 2011, was the first to feature sub-challenges on speech affected by a health condition, namely sleepiness and intoxication [6]. Together with the physical and cognitive load challenges in 2014 (cf. 4.8), sleepiness can be grouped under the umbrella terminology 'speech affected by fatigue'. Whilst fatigue has no strict medical condition definition, it can generally be regarded as a state of which the core symptoms include a reduction in efficiency and motivation. Fatigue degrades an individual's performances and can lead to errors and, potentially fatal, accidents in many settings, such as on the road [55,56], or in the workplace [57]. The early and remote diagnosis of fatigue is therefore potentially life-saving on the roads and in related transport industries such as aeronautical and freight [58]. Fatigue contains both cognitive and physical aspects, which are expected to have notable effects on speech. Indeed, speech affected by sleepiness has been associated with effects relating to impaired speech planning such as flattened prosody; effects include reduced intelligibility and tense and breathy qualities associated with irregular vocal fold actions [59].

The challenge dataset was the *Sleepy Language Corpus* (cf. Table 2). The official feature set was a 4368 dimensional acoustic feature representation, and was used in conjunction with *Synthetic Minority Oversampling Technique* (SMOTE) resampling [60], to account for class imbalances and a SVM to set a 2-class, 'not-sleepy' and 'sleepy', development set UAR of 67.3% and test set UAR of 70.3% [6]. As the challenge was set in 2011 and, deep learning technologies arguably did not begin to replace *Gaussian Mixture Models* (GMMs) and *Hidden Markov*

Table 2
Partitioning of the Sleepy Language Corpus (SLC) into the Train, (Devel) opment and Test partitions used for the ComParE-2011 2-class – Non Sleepy Language (NSL) and Sleepy Language (SL) – classification task. Displayed are the number of utterances (#) per class, per partition. Table reproduced from [6].

SLC #	Train	Devel.	Test	Σ
NSL	2125	1836	1957	5918
SL	1241	1079	851	3171
Σ	3366	2915	2808	9089

Table 3
Partitioning of the Alcohol Language Corpus (ALC) into the Train, (Devel) opment and Test partitions used for the ComParE-2011 2-class – Non Alcoholized Language (NAL) and Alchoholized Language (AL) – classification task. Displayed are the number of utterances (#) per class, per partition. Table reproduced from [6].

#	Train	Devel.	Test	Σ
NAL	3750	2790	1620	8160
AL	1650	1170	1380	4200
Σ	5400	3960	3000	12360

Table 4
Partitioning of the Netherlands Cancer Institute Concomitant ChemoRadioTherapy Speech Corpus (NKI CCRT) into the Train, (Devel) opment and Test partitions used for the ComParE-2012 2-class – Intelligible (I) and Non-Intelligible (NL) – classification task. Displayed are the number of utterances (#) per class, per partition. Table reproduced from [7].

#	Train	Devel.	Test	Σ
I	384	341	475	1200
NI	517	405	264	1186
Σ	901	746	739	2386

Models (HMMs) as state-of-the-art classification techniques in speech processing tasks until 2012 [61] none of the challenge entrants were based on deep, or for that matter shallow, neural networks. Indeed, the winner of the sleepiness sub challenge used a *Mel Frequency Cepstral Coefficent* (MFCC) based on a GMM-*Universal Background Model-Maximum a Posteriori* (UBM-MAP) *supervector* approach utilised in a 3-state left-to-right HMMs, with the resulting supervectors being classified using a SVM [62]. For information on supervector formation, the interested reader is referred to [63,64]. This approach achieved a UAR of 71.7%, representing a 3.1% percentage point improvement over the baseline system. For an in-depth review of other approaches in the sleepiness challenge the interested reader is referred to [59]. Somewhat surprisingly, the authors were unable to identify any works that had revisited this corpus and attempted to improve on the challenge wining score using deep learning.

4.4. Intoxication (2011)

The second health challenge held in 2011, was the detection of speech affected by alcohol. The dataset was the *Alcohol and Language Corpus* (cf. Table 3) in which speech affected by alcohol was determined by a *Blood Alcohol Concentration* (BAC) reading between 0.5 and 1.75 per mill, and speech not affected by alcohol, a BAC in the interval 0.0–0.5 [45]. Excessive alcohol consumption is a major public health concern. A *World Health Organisation* (WHO) report released in 2014, estimated that there are at least 200 disease associated with harmful levels of alcohol use and that in 2012, with approximately 3.3 million deaths worldwide (5.9% of the worldwide total number of deaths) which are attributable to alcohol consumption [65]. It is widely accepted that alcohol affects speech, indeed slurred speech is a hallmark effect associated with intoxication [66,67]. Corroborating evidence for alcohol induced alterations on the phonetic structure of speech can been seen in studies showing the difficulties of performing accurate *Automatic Speech Recognition* (ASR) on speech affected by alcohol [68].

The Challenge baseline, set using the same set-up as the sleepiness challenge (cf. Table 4.3), was a development set UAR of 65.3% and test set UAR of 65.9% [6]. Again, no entrants to this challenge used deep learning, furthermore only one entrant [69], utilised a ANN, in the form of a *Multilayer Perceptron* (MLP). Reflecting the then state-of-the-art, GMM and HMM based system were also popular within this sub-challenge; 5 out of 7 entrants utilised such a technique. The winning system leveraged a GMM-UBM supervector system, the challenge feature set and an iterative speaker normalisation technique to achieve a test set UAR of 70.5% [70]. Bone et al. [71], refined different aspects of this approach outside of the challenge and were able to achieve a small improvement to the test set UAR of 71.4%.

To the best of the authors knowledge, only one deep learning approach has been tested on the ACL corpus, namely [72]. In this work, 40 dimensional mel filter bank features were fed into a two layer bi-directional RNN utilising GRU's and dropout. This set-up was able to achieve best development set UAR of 69.2% and test set UAR of 71.0%. This score almost matches state-of-the-art for this dataset and notably, this was achieved *without* compensation for data class imbalances, or performing speaker normalisation.

4.5. Pathologic speech (2012)

The first pathological speech challenge, held in 2012, was based on assessing the level of intelligibility of speech samples taken from individuals pre and post *Concomitant Chemo-Radiation Treatment* (CCRT) for inoperable tumours of the head and neck [7]. All patients in the *Netherlands Cancer Institute Concomitant ChemoRadioTherapy Speech Corpus* (NKI CCRT) (cf. Table 4) had a primary tumour located on a physical structure associated with articulation i.e., the oral cavity, oropharynx, hypopharynx, larynx and nasopharynx [73].

For the challenge baseline, the feature space was expanded on from the 2011 challenge and consisted of a 6125 dimensional feature representation [7]. The challenge organisers tested both SVM and Random Forest classifiers to set the baseline and observed a slight, but not significant, advantage in using Random forest which yielded a development set UAR of 64.8% and test set UAR of 68.9%.

As in the sleepiness and alcohol challenges, none of the entrants to the Pathology Sub-Challenge utilised a neural network based classifier. The winners of the sub-challenge developed a knowledge based approach that combined multiple subsystems covering different characteristics of pathological speech; namely, phoneme probability, prosodic and intonational features, Voice quality and pronunciation features; fused using a Naive Bayes framework [74]. Combining this approach with speaker clustering the winning test set UAR was 76.8%

It is worth noting that a DNN approach was proposed and explored in the 2-class speaker likeability challenge [75]. The presented results indicated that while a DNN could improve upon the baseline, the authors where unable to formulate a network structure that could improve upon a more standard MLP approach. Moreover, DNN topologies have been used in other related speech intelligibility tasks such as improving ASR accuracy for individuals with dysarthria [76,77]. Recently, on a database containing 323 Chinese participants speaking Cantonese with normal or pathological voices [78]. The authors proposed the use of acoustic posterior probabilities of phones computed using a DNN-HMM ASR for distinguishing between mild, moderate and severe categories of voice disorder severity.

4.6. Autism spectrum conditions (2013)

Autism is a spectrum of conditions which can be defined by limitations and irregularities in socialisation and communication [79]. The early diagnosis of an *Autism Spectrum Condition* (ASC) is important for increased positive outcomes from therapy, as well as for reducing parental stress [80,81]. Linguistic peculiarities, associated with ASCs, include echolalia, out of context phrasing, as well as pronoun and role reversal [82,83]. However, language skills vary considerably across subtypes within the spectrum [84,85] reducing the reliable of linguistics cues to aid the diagnosis of ASC. Paralinguistics, on the other hand, appear better suited as an ASC diagnostic aid; acoustic features relating to articulation, loudness, pitch, and rhythm have consistently shown promising results in this regard [71,86,87].

Table 5
Partitioning of the Child Pathological Speech Database (CPSD) into the Train, (Devel) opment and Test partitions used for Autism ComParE-2013 sub-challenge. This challenge consisted of a 2-class task Typically Developing, and Atypically Developing classification task, as well as a 4-class task, Typically developing (TYP), Pervasive Developmental Disorders (PDD), pervasive developmental disorders Non-Otherwise Specified (NOS), and specific language impairment such as Dysphasia (DYS) classification task. The number of utterance (#) per partitioning is shown. Table reproduced from [8].

#	Train	Devel	Test	Σ
TYP	566	543	542	1651
PDD	104	104	99	307
NOW	104	68	75	247
DYS	129	104	104	337
Σ	903	819	820	2542

Participants the 2013 Autism Sub-Challenge had to classify the type of pathology of a child [8]. The Child Pathological Speech Database (CPSD) (cf. Table 5) was used to provided speech recording of children who are either: (i) *Typically Developing*; (ii) diagnosed as having a *Pervasive Developmental Disorders* (PDD) such as Autism; (iii) diagnosed as having a *Pervasive Developmental Disorders Non-Otherwise Specified* (NOS); or, (iv) diagnosed as having a specific language impairment such as *Dysphasia* (DYS). Two evaluation tasks were defined: a binary 'typical' vs. 'atypical' classification task and a four-class 'diagnosis' task. The challenge baseline was set using a new feature set containing 6373 statistical functions derived from a set of 65 low level descriptors [8]. Challenge baseline development partitions 2- and 4-class UAR's were 92.8% and 52.4% respectively, with the corresponding test set UAR's being 90.7% and 67.1%.

For the first time in a health challenge, we see a participant utilising a DNN [88]. The system consisted of two hidden layers of stacked RBMs. As the system was used as part of an ensemble of classifiers, the authors only reported development set UARs for the DNN, 94.4% and 57.5% for the 2- and 4-class tasks, which were both above baseline. It is worth noting that the proposed ensemble method was able to beat the baseline score for the 2-class task, UAR of 92.2%, but not for the four class problem in which it achieved a UAR of 64.8%. The challenge was won by a knowledge driven system which leveraged a harmonic model of the voiced speech to extracted features including *Fundamental Frequency* (f0), *Harmonic to-Noise Ratio* (HNR), shimmer, and jitter [89]. Combined with a SVM classifier this approach achieved 2- and 4-class test set UARs of 93.6% and 69.4%.

Outside of the challenge Huang and Hori [90] compared a range of different DNN structures. Their system also utilised feature normalisation and dimensionality reduction approaches including *Principal Component Analysis* (PCA) and *Linear Discriminant Analysis* (LDA). The authors state test set UARs of 92.9% and 66.0% for the 2- and 4-class classification tasks, speculating that these results could have been improved through a consideration of the class imbalances.

4.7. Depression (2013 & 2014)

Major Depressive Disorders (MDD) are a growing global health concern. It has been estimated that there are approximately 322 million individuals worldwide living with depression [91] and that the total number of individuals estimated to be living with depression has increased by 18.4% between 2005 and 2015 [92]. Speech, as well as other behavioural and physiological signals, has the potential to provide a set of objective markers to aid depression detection [3,93].

Both the 2013 and 2014 Audio-Visual Emotion Challenges (AVEC) required participants to predict the level of self-reported depression, as scored by the *Beck Depression Index II* (BDI-II) [94], from a given audiovisual clip [12,13]. Both challenges corpora are based on a subset 150 files taken from the *Audio-Visual Depressive Language Corpus* (AViD-Corpus)[1]. The audio baseline feature set for both challenges consisted of 2 268 features and the classifier was a SVR. The metric for both challenges was the RMSE over all sequences. The audio RMSE development and test baselines for AVEC-2013 were 10.75 and 14.12; the corresponding scores for AVEC-2014 were 11.52 and 12.57.

None of the participants, in either of the challenges, utilised a deep learning approach, this includes both audio-only, visual-only and audiovisual systems. Both challenges were won by teams from Lincoln Laboratory, Massachusetts Institute of Technology [95,96]. These systems were highly knowledge driven, exploiting a specifically designed feature space which captured coordination across articulators and a purpose built *Gaussian staircase* regressor. Their lowest AVEC-2013 RMSE was 8.50, and the lowest test set RSME of their AVEC-2014, which also included visual information, was 8.12. For a more in depth review of both challenges the interested reader is referred to [3].

To the best of the authors knowledge there has been no speech and deep learning approaches on either corpora. However, there has been recent research interest in predicting depression scores using deep learning vision systems. Zhu et al. [97], proposed a dual CNN structure to exploit both facial appearance and dynamic information. The authors report a lowest AVEC-2014 test set RMSE of 9.55. Similarly, Kang et al. [98], used the AVEC-2014 data to re-tune the *VIPLFaceNet* network, a 10-layer CNN with 7 convolutional layers and 3 fully-connected layers, for the task of depression detection. This approach yielded a RMSE of 9.43.

4.8. Cognitive and physical load (2014)

The 2014 iteration of the ComParE challenge focused on the recognition of speakers cognitive and physical load in speech [9]. The datasets utilised for these sub-challenges were the *Cognitive Load with Speech and EGG* (CLSE) database (cf. Table 6) and the*Munich Bio-voice Corpus* (MBC) (cf. Table 7). High cognitive load and mental fatigue is strongly associated increasing demands on working memory and impaired mental performances [56,99]. Effects such as increased articulation rate, an increase in the number of filled pauses, and a reduction in formant *vowel space area* are commonly reported for speech produced at high cognitive load [99]. Physical fatigue on the other hand, is a reduction in muscle power and movement with a key symptom being impaired co-ordination [56]. Further there are strong links between changes in heart-rate and changes in prosodic and voice quality (glottal) features [48,100].

The baseline was again set with the 6373 dimensional ComParE-2013 feature set. The official baseline scores for the cognitive load sub-challenge were 63.2% for the development partition and 61.6% for the test set. While for the physical load challenge development and test set UAR's were 67.2% and 71.9% respectively [9]. Both challenges had two entrants using DNN based approaches [101,102], as well as a DNN based representation learning paradigm [103].

Gosztolya et al. [101,102] proposed and developed a *Deep Rectifier Network* (DRN) approach which utilised rectified linear units (ReLU) The reported test set UAR of the proposed DRN is 63.05%, representing a slight improvement on the baseline. However, this accuracy was well below the challenge winning UAR of 73.9%, achieved by fusing four *i-vector* based systems which utilise different low-level feature representations [104]. For details of the i-vector paradigm the interested reader is referred to [105,106]. It can be speculated that the gulf in system performance between the DNN system presented in [101] and the challenge winner could be due in part to the small amount of data in the cognitive load corpora.

[1] The AVEC-2014 corpora is a reduced file length (time) version of the AVEC-2013 corpora. Further, 5 files were replaced from the AVEC-2013 corpora when for the 2014 challenge due to unsuitability.

Table 6
Partitioning of the Cognitive Load with Speech and EGG (CLSE) database into the Train, (Devel) opment and Test partitions used for the ComParE-2014 3-class – (L) oad level from 1–3 – classification task. Displayed are the number of utterances (#) per class, per partition. Table reproduced from [9].

#	Train	Devel	Test	Σ
L1	297	189	216	702
L2	297	189	216	702
L3	429	273	312	1014
Σ	1023	651	744	2418

Table 7
Partitioning of the Munich Bio-voice Corpus (MBC) into the Train, (Devel) opment and Test partitions used for the ComParE-2014 2-class – low or level of physical load – classification task. Displayed are the number of utterances (#) per class, per partition. Table reproduced from [9].

#	Train	Devel	Test	Σ
Low	199	199	154	552
High	186	185	165	532
Σ	385	384	319	1088

The DRN was also used in physical load sub-challenge achieving a test set UAR of 73.03% [101]. Again, this is represented only a slight improvement on the baseline approach, 71.9% UAR, and was only slightly below the winning physical load UAR of 75.4%, achieved using a multi-view Canonical Correlation Analysis feature selection paradigm [107].

Development partition results are also reported for a DNN-based approach for both the cognitive and physical load sub-challenges in [102]. Specifically the authors tested both *Conditional Restricted Boltzmann Machines* (CRBM) and a DNN topology, again built with ReLU (ReLUNet). The CRBM achieved a development set UAR of 58.6% and 69.2% for the cognitive and physical load challenges respectively, while the ReLUNet achieved scores of 61.9% and 67.7%. While these approaches are above the development set baseline performances, such as result was not replicated on the test sets. By the authors own admission their approaches overfitted to the development set.

Finally, Nwe et al. [103], used, as part of a wider fusion system, a DNN bottleneck paradigm to extract a higher level feature representation from the baseline features for cognitive load classification. Presented result indicated that these features could outperform a GMM supervector approach as well as the baseline features on the development partition, and that they could aid GMM supervectors in a late fusion system.

4.9. Parkinson's (2015)

Parkinson's disease (PD) is a neurodegenerative disease characterised by motor deficits including bradykinesia, e.g., slow movement, rigidity and tremors [108]. PD is estimated to be the second most common degenerative disorder, after Alzheimers, and effects approximately 12 individuals per 1000 of the population [108], with an increased prevalence in persons aged over 65 [109]. The links between PD and the effects on speech motor control such as decreases in speech rate; monotonic pitch; increases in articulatory and phonetic errors; and, breathy and tense voice qualities, are well established in the relevant literature [110–113]. Moreover, the associated effects are capable of differentiating between speech either affected by or speech not affected by PD, e.g., [114].

In the Parkinson's Condition Sub-Challenge of ComParE-2015 [10], the participants had to estimate a patients Parkinson's state according to the *Unified Parkinsons Disease Rating Scale, motor subscale* (UPDRS-III) [115]. Unlike the previous ComParE health related sub-challenges this was a regression task, with the performance metric being the *Spearmans Correlation Coefficient* (ρ). Three scores were offered for the baseline which was set using the ComParE-2013 feature set and a SVR. First a cross-fold validation of the training set $\rho = 0.434$; second a standalone development set score of $\rho = 0.492$; and finally the official test set baseline $\rho = 0.390$. For full details on the baseline set-up, the interested reader is referred to [10].

In the challenge, Hahm and Wang [116], utilised two DNN systems in their approach. Firstly, to extract quasi-articulatory features, the authors utilised a DNN-based transfer learning approach to estimate an inverse mapping between acoustic features and articulatory features. This mapping was learnt on the *Multi-Channel Articulatory* (MOCHA)-TIMIT corpora, a data set consisting of simultaneous recordings of speech and articulatory data – as determined by *Electromagnetic Articulograph* (EMA) sensors – from 2 British English speakers [117]. This information was then fused with baseline features and processed using a 3-layer DNN regressor, yielding a test set score of $\rho = 0.485$.

For the first time, a health challenge was won by a system utilised a DNN based approach. Grósz et al. [118] used a 5 layer DRN trained on a subset of baseline features, as selected by a Pearsons correlation coefficient feature selection methodology, to achieve a test set score of $\rho = 0.306$. It is worth noting here that the test set contained multiple instances from a smaller number of speakers [10]. Thereby, as with other challenge participants e.g., [119], the Grósz et al. approach gained substantial improves in system performance by apply a post-processing speaker clustering method. The final system – utilising both a DNN and speaker clustering – yielded a test set score of $\rho = 0.649$.

4.10. Eating (2015)

A second health related sub-challenge was also held in the ComParE-2015 challenge, namely the Eating condition sub-challenge in which participants had to classify the type of food, or not, a speaker was eating [10]. For this sub-challenge the iHEARu-EAT dataset was utilised (cf. Table 8). Such a system represents an acoustic means of objectively monitoring of ingestive behaviour (MIB) as well as the prevention of obesity and eating disorders [51]. The WHO has estimated that obesity levels have triple globally since 1975 [120]; furthermore, this increase has had a considerable economic burden on individuals and societies. Therefore research into techniques to automatically monitor food intake is increasing e.g., [51,121,122]. Speech is a potential information channel for automatically monitoring food intake; results presented by Hantke et al. [51], indicated that eating whilst speaking has a statistically significant effect on the accuracy of an ASR system. Furthermore, such errors differ across different food-types.

Participants in the Eating condition sub-challenge had to classify classes speech which were either 'not affected' by food or were affected by an Apple, a Nectarine, a Bannana, Crisps, Biscuits or Gummi bears [10]. The baseline was set again using the ComParE-2013 feature set and a SVM, the baseline training partition UAR, set using leave-one-out

Table 8
Partitioning of the iHEARu-EAT dataset into the Train and Test partitions used for the ComParE-2015 7-class eating condition classification task. Displayed are the number of utterances (#) per class, per partition. Table reproduced from [10].

#	Train	Test	Σ
No Food	140	70	210
Apple	140	56	196
Nectarine	133	63	196
Banana	140	70	210
Crisp	140	70	210
Biscuit	133	70	203
Gummi bear	119	70	189
Σ	945	469	1414

speaker cross fold validation, was 61.3% and the test set UAR 65.9%. In the challenge a *shallow* neural network approach was proposed in [123]. More specifically, Pellegrini tested a variety of shallow and deep DNN based approaches, however results indicated, that a shallow network with a single hidden layer with ReLu activations, and trained with the momentum update rule, performed the strongest for the task, achieving a UAR of 68.4%.

A CNN based representation learning approach, designed to leverage information from out-of-domain data, was proposed for the Eating sub-challenge in [124]. This system trained a CNN comprised of three convolutional and maxpool layers, followed by two fully connected layers. When training the network, the authors utilised dropout, data augmentation via pitch shifting, and a transfer learning approach in which they trained their CNN on data from the Voxforge[2] In their final approach, the authors trained a logistic regression classifiers on the CNN features and obtained a test set UAR of 75.9%. However, this score was still below the challenge winning UAR of 83.1% [125]. Kaya et al. [125] approach, used a novel framework based on Fisher vector encoding of extracted features and cascaded normalisation to account for variability due to differing speakers characteristics and spoken content.

Recently, DRN approaches were shown to produce equivalent performance as conventional linear SVM and AdaBoost systems for the eating task [126]. However, this comparison was not the main aim of the authors. Instead, this paper focused on the advantages gained by performing *speaker clustering* with different classifier approaches. In this regard, the choice of classifier prior to speaker clustering had little effect i.e., all system accuracies after clustering were highly similar regardless of their underlying classification paradigm. Such a result is due to multiple instances from the same speaker being used in the test partition, and highlights the need for larger and more diverse datasets.

4.11. Depression (2016 & 2017)

The Audio-Visual Emotion recognition Challenge (AVEC) series revisited the task of depression detection (cf. 4.7), in 2016 and 2017 [14,15]. Both challenges utilised the *Distress Analysis Interview Corpus – Wizard of Oz* (DAIC-WOZ) database of 193 clinical interviews designed to support the diagnosis of psychological distress conditions such as depression. For AVEC-2016 [14], the challenge was a 2-class multimodal level of severity (high/low) task, based on the patients *Patient Health Questionnaire* (PHQ)-8 depression index score [127], see Fig. 3 for score distribution of the DAIC-WOZ data. The challenge organisers provided extracted video features (for ethical reasons the raw video could not be shared), audio files and transcription of the interviews. The audio baseline feature representation was realised using a set of prosodic, voice quality and spectral LLDs extracted from the *Collaborative Voice Analysis REPository* (COVAREP) Matlab toolbox [128]. The COVAREP LLDs were combined with a SVM classification, with majority voting used to produce one single depression label per file, with the official metric being the *F1 score* for the 'depressed' class. Using this method, the development score was 0.46 (0.68, for 'not-depressed') and the test set score was 0.41 (0.58, for 'not-depressed').

Deep learning based approaches were not prominent in the 2016 challenge, with the only deep learning approach being the DEPAUDIONET E2E system proposed in [129]. This system feed Mel Spectrum features into an E2E network comprising a feature extraction topology of; a one-dimensional convolution layer, a batch normalization layer, a ReLu layer, a one-dimensional max-pooling layer and a dropout layer; this is then followed by a LSTM layer and two fully connected layers. To help overcome issues pertaining to imbalance in training data (4:1 'not-depressed' to 'depressed'), the authors performed a random down sampling of the non depression class when creating their mini-batch samples for network training. The DEPAUDIONET system outperformed the baseline approach on development set 0.52 (0.70, for 'not-depressed'); however, no test set score was given in the paper [129]. The winning entrant utilised a gender-dependant multi-modal decision tree approach, gaining a test set score of 0.57 (0.88, for 'not-depressed') [130].

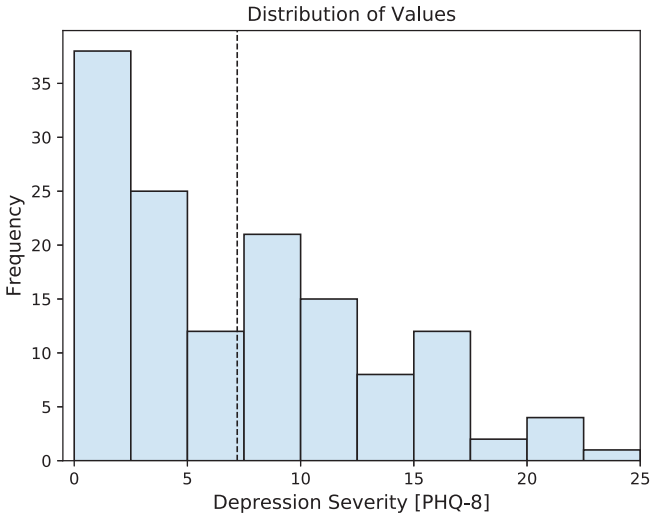

Fig. 3. Frequency Distribution of Development + Training partitions used for the AVEC-2016 and AVEC-2017, Depression sub-challenges. Figure reproduced from [14,15].

The AVEC-2017 depression sub-challenge required participants to predict – again from multimodal audio, visual, and text data – the PHQ-8 score of each patient in the DAIC-WOZ corpus [15]. The baseline audio system was again based on COVAREP features, however this was in conjunction with a Random Forest regressor. As in AVEC-2013 and AVEC-2014 (cf. 4.7), the metric is the RMSE, with the audio development set and test set RMSE's being 6.74 and 7.78 respectively.

In the challenge, two papers, both lead by Le Yang of Northwestern Polytechnical University, explored the suitability of CNNs for the task. The first [131] system fed COMPARE-2013 features into a CCN trained to predict PHQ8 score. After training the CNN, the weights were frozen and the last layer removed. The output of the remaining fully connected layer was then fed into a new DNN which was trained using the CNN features. The authors ultimately trained 4 systems; a combination of gender specific models for 'depressed' and 'not-depressed' classes. The outputs of these four systems were then fused (also with outputs of similar systems from the video and text modalities) via multivariate linear regression. To alleviate issues relating to training data imbalance, the authors also proposed a data expansion technique in which the 'depressed' class by cutting longer segments into multiple chunks to artificially create more depression data instance, albeit with the same labels. Their audio system achieved development set RMSE of 6.62, which was below the provided baseline; only audio-visual fusion RMSE were given for test set, with the lowest reported value being 5.40.

In the second paper [132], the same CNN-DNN set up was used. However the authors utilised the downsampling methodology during training presented in [129] for the 'not-depressed' class. Further, only gender dependent models (not class specific as well) were trained. No audio only scores directly comparable with the baseline are given in the paper. The audio-visual approach achieved a RMSE of 5.97. For the interested reader, a unification of both CNN approaches is presented in [133].

The winning approach [134] exploited the provided patient interview transcripts to identify salient topics and segments. They performed random forest regression on the baseline audio and visual features corresponding to these segments. This method was augmented with

[2] http://www.voxforge.org/home corpus. The aim of this network was to extract a high level feature representation from a 40 dimensional Mel-filter bank representation.

Table 9
Partitioning of the Upper Respiratory Tract Infection Dataset (URTID) into the Train, (Devel) opment and Test partitions used for the ComParE-2017 2-class – cold (c) and not cold (NC) – classification task. Displayed are the number of utterances (#) per class, per partition. Table reproduced from [11]. Note, at time of publishing the number of utterance in the cold and non-cold URTID test partition had not been publicly released by the challenge organisers, overall the test set contains 9551 utterances.

#	Train	Devel	Σ
C	970	1011	1981
NC	8535	8585	17120
Σ	9505	9596	19101

semantic features, which were manually extracted using the *Linguistic Inquiry and Word Count* (LIWC) text analysis program [135]. The combined approach utilising acoustic, visual and semantic features produced the winning RMSE of 4.99.

4.12. Cold and Flu (2017)

Upper Respiratory Tract Infections (URTIs) such as the Common Cold and Influenza (flu) are another serious public health concern, which cause approximately 3–5 million cases of severe illness, and about 250 000 to 500 000 thousand deaths per year [136]. One of the best method for slowing the spread of these illnesses through a population is early diagnosis and social isolation. In this regard, the Cold sub-challenge in ComParE-2017 [11], required participants to classify speech as affected by cold and flu or speech under 'normal' health conditions.

Utilising the *Upper Respiratory Tract Infection Dataset* (cf. Table 9), four separate systems were presented as a baseline: (i) the ComParE-2013 feature set together with a SVM which achieved development and test UAR's of 64.0% and 70.2% respectively; (ii) a *Bag-of-Audio-Words* (BoAW) approach, for details see [137], which quantised the ComParE-2013 LLD's into a sparse histogram-of-occurrences representation, 64.2% development and 67.3% for test; (iii) an E2E topology of two convolutional layers and one LSTM layer, operating directly on the raw audio waveform 59.1% development and 60.0%; and (iv), a majority vote fusion of these approaches which returned the official baseline test set score of 71.0%. The challenge organisers speculate that the weaker performance of the E2E model was due to the imbalance in the training data towards the healthy condition, approximately 8:1 [11]. E2E models rely purely on the statistics of the available data to learn optimal features, therefore it is possible that the available data, due to the imbalance, did not contain a sufficient variation for the cold class.

An E2E approach was also presented by one of challenge participants [138]. In this system, *Constant Q Transform* (CQT) spectrum and Gammatone spectrum features were feed into 5 CNN layers followed by a single gated recurrent unit layer. The paper does not give a test set result for the E2E system; however, their it outperforms the challenge baseline E2E system on the development set 68.4% to 58.6 %. For the test set submission, the E2E system was fused with other more conventional approaches and a best UAR of 71.4% was reported.

A more conventional DNN classifier approach was proposed in [139]. The authors explored the suitability of different feature representations with either a SVM or a range of different DNN topologies. The authors best development performance, UAR 80.0%, was found using a DNN classifier, trained with a z-score normalised *Spectral Modulation* features using sigmoid nodes in a 3 layer topology. However, the strong development set results did not translate to strong test, UAR 62.1%, with the authors speculating that overfitting was the reason.

This sub-challenge was won by a DNN-based feature extraction system [140]. The approach extracted frame-level posterior probabilities as given by a Deep Neural Network with 3 hidden layers, each containing 256 ReLU neurons and a softmax function in the output layer. These probabilities were combined into an utterance level representation then classified with a SVM. This approach did not outperform any of the baseline scores, but when fused with ComParE–2013 features achieved the winning test set UAR of 72.0%.

Table 10
Partitioning of the Munich-Passau Snore Sound Corpus (MPSSC) into the Train, (Devel)opment and Test partitions used for the ComParE-2017 4-class – snore originating from the (V)elum, (O)ropharyngeal lateral walls, (T) ongue base or (E) piglottis – classification task. Displayed are the number of utterances (#) per class, per partition. Table reproduced from [11]. Note, at time of publishing the number of utterance per snore type in the MPSSC test partition had not been publicly released by the challenge organisers, overall the test set contains 263 snoring instances.

#	Train	Devel	Σ
V	168	161	329
O	76	75	151
T	8	15	23
E	30	32	62
Σ	282	283	565

4.13. Snore (2017)

A second health task was run in ComParE-2017, although not strictly a speech task. Snoring, can be a marker of *Obstructive Sleep Apnea* (OSA) a highly prevalent sleep disorders, affecting approximately 37% of middle-aged men and 25% of middle-aged women in the general population [141–143]. An integral part of successful treatment of OSA is the location of the site of obstruction and vibration for targeted surgical intervention [144]. In this regard, participants in the snoring sub-challenge had to classify four different snore types [11]. The four classes related to the site of obstruction and vibration in the corresponding snore sound include, the *velum* (V), the *oropharyngeal* area including the palatine tonsils (O), the *tongue* base (T) and the *epiglottis* (E).

The same four baseline approaches as per the cold sub-challenge (cf. 4.12), however, this time using the *Munich-Passau Snore Sound Corpus* (cf. Table 10): (i) ComParE-2013 feature set and SVM, 40.6% development and 58.5% test (this UAR was the official baseline; (ii) BoAW a, 44.2% development and 51.2% for test; (iii) E2E system, however with 3 convolutional layers, 40.3% development and 40.3%; and (iv), a majority vote fusion of these approaches 43.4% development and 55.6% test. Again, the challenge data was not balanced, approximately half the training and development samples were from the V classes. As already mentioned (cf. 4.12), this imbalance could have reduced the effectiveness of the E2E approach.

The DNN-based posterior probabilities that won the cold sub-challenge was also tested on the snore data [140]. Again, the DNN features did not outperform the conventional baseline on development set and produced a test UAR of 64.0% when fused with the baseline features. The sub-challenge winners [145] utilised a shallow network in their approach. Using the same fisher vector based features space introduced in [125] in combination with weighted kernel *Extreme Learning Machines* (ELM) to achieve a test set UAR of 64.2%. ELMs are a single hidden layer feedforward networks, in which the parameters of the hidden nodes are randomly assigned and the output weights of hidden nodes are learned in a single pass [146]. They can be implemented to include the kernel trick to handle non-linearities in the feature space, while the weighting function used [145] can help account for class imbalances in the training data.

Outside of the official sub-challenge, however still on the challenge data, Amiriparian et al. [147], proposed and explored a novel deep spectrum features approach. These features are derived by forwarding spectrograms through very deep task-independent pre-trained image

classification CNNs such as *AlexNet* [35] and *VGG19* [148] and the activations of a fully connected layer is then used as a feature. In [147], the authors demonstrated that the activations of a second fully connected layer of AlexNet using a viridis colour map are well suited to the task. When combined with a SVM classifier, these deep spectrum features achieved a test set UAR of 67.0%. This approach was extended in [149], by using performing evolutionary feature selection based on competitive swarm optimisation on the extracted features. This approach yielded strong development set performances, but results indicate it was potentially susceptible to overfitting. The highest test set UAR achieved by the 'End-to-Evolution' system was 66.5%.

5. Opportunities for leveraging deep learning

A common reoccurring theme, not only limited to deep learning paradigms, that has continually come up throughout this review has been that of model overfitting. This effect is most likely due to the small and imbalanced datasets often use in the challenges (cf. Table 1). The following paragraphs highlight some potential deep learning based research avenues to help alleviate the issues associated with such operating conditions.

Inclusion of intelligent labelling paradigms: Techniques such as semi-supervised learning, active learning and cooperative learning have been shown to enhance recognition models in a range of speech tasks [150]. These approaches leverage a smaller set of labelled data to annotate a larger dataset with minimal human involvement. Such approaches have been shown to aid a range of speech based classification tasks including emotion recognition; see [150] for a recent review. Recent image classification research shows that the combination of semi-supervised learning in combination with *few-shot networks* can efficiently solve new learning tasks using only a small number of training samples [151,152]. To the best of the authors knowledge, such approaches have yet to be tested in the speech-health domain.

Data augmentation: Deep learning techniques such as *Generative Adversarial Networks* (GANS) [153,154] can be used to generate new training data samples [155,156]. This is a particularly promising development due to the high costs associated with obtaining high quality clinical data, from vulnerable populations. GANS consist of a generative model (generator) which is set to compete against a discriminative model (discriminator) in an adversarial setting. The discriminator is trained to accurately distinguish whether a given sample has been produced by the generator or drawn from a training data distribution. The concurrent objective of the generator is to fool the discriminator into misclassifying the generated samples [153,154]. The overall objective of a GAN network is to compel both models to continuously improve their methods until the generator is able to perfectly synthesise the training data, and the discriminator is unable to find a difference between the samples synthesised by the generator and real samples from the dataset.

Promising results presented in [157] highlight the need for continued research in this direction. The authors demonstrated that GAN based methods can be used to synthesis new training instances to aid classification. This was particular important as it was demonstrated for a pathology classification task similar to the 2013 Autism sub-challenge (cf. 4.6), for which there was a relatively small amount of associated training data. Furthermore, their GAN system produced competitive performances when compared with more conventional classification paradigms [157].

Data representation learning: Conventional hand crafted features may not adequately capture the required variability to differentiate between acoustic spaces associated with different health conditions. Deep representation and E2E learning have the potential to determine highly abstract, thus more robust, representations specific for the task at hand. Non-deep approaches, such as BoAW [137], highlight the potential of representation learning; BoAW has achieved the state-of-the-art for emotion prediction in particular [158] and is now a ComParE baseline system [11,159]. This assertion is also further supported by the state-of-the-art performance gained using the deep spectrum features for snore sounds [147]. The upcoming ComParE-2018 challenge, which includes 'state-of-mind' and 'abnormal heart sound' sub-challenges, features deep representation learning as a baseline system [159]. This will be set using the AUDEEP toolkit [160] which utilises recurrent sequence to sequence autoencoders to learn representations which take into account the temporal dynamics of time series data such as speech.

More multitask and transfer learning: Many of the different health conditions discussed in this review have either a direct, i.e., they are co-morbid, or have an indirect relationship to each other. For instance increased fatigue is a core symptom of depressive disorders [94]. Mutlitask learning and transfer learning approaches have the potential to help exploit commonalities to create a more robust system [150]. Interrelationships between the sleepiness (cf. 4.3) and intoxication (cf. 4.4) datasets were explored in [161]. The authors show that an effective, SVM based, classifier can be obtained by aggregating the training data from both corpora. Similarly results presented in [162], indicate that the combined learning of depression and affects in a multitask LSTM-RNN paradigm aids depression prediction. Moreover, it has also been shown that the inclusion OF depression information aids the accuracy of deep regressors performing affect detection on AVEC-2014 data [163,164]. Autoencoder and RNN based transfer learning system have been shown to aid speech-based emotion recognition [165,166]. However, to the best of the authors knowledge such approaches are yet to be tested in the speech-health domain. Within image processing techniques GANs are also being explored for domain adaptation solutions, e.g., [167]. The core idea being, the use if an adversarial network to learn an effective mapping between two domains.

6. Conclusion

Deep learning has unquestionably improved on more conventional machine learning paradigms in terms of system accuracy and robustness in many speech-based applications. With this in mind, we reviewed the impact of deep learning paradigms in the speech and health domain. By reviewing the health based *Computational Paralinguistics Challenge* (ComParE) series and *Audio/Visual Emotion Challenge* (AVEC) workshops we observed a steady increase in the percentage of deep learning based entries from none in 2011 through to two-thirds in ComParE-2017 However, it is clear from this review that deep learning systems are still not yet a dominant force in speech and health. While two challenges have been won using a deep learning based approach: 2015's Parkinson's condition sub-challenge [118] and 2017's cold and flu sub-challenge [140], it is debatable given the results in [126] if the winning approach in 2015 was due to the deep learning approach or the post-processing speaker clustering method employed [118]. Furthermore, the majority of challenge participants still use very conventional feature extraction and classification paradigms.

A major reason for this is most likely the database, compared to other speech-tasks such as speaker and speech recognition the databases used for speech-health tasks are often small, containing multiple samples from a single speaker, and can be unbalanced in terms of samples per class. To further facilitate deep learning in such conditions, we identified future research topics in intelligent labelling, data augmentation, representation learning as well as multitask and transfer learning. However to fully overcome these aforementioned issues, considerable investment is needed in collecting truly large datasets. The need for this data is twofold, firstly to facilitate the true impact that deep learning solutions could possibly make, and secondly to enable large scale clinical studies. Furthermore, this second aspect would enable investigations into a major outstanding research question; which recognition accuracies correspond to positive impacts in real-world healthcare situations.

Looking into the future, ongoing developments and advancements in ubiquitous computing devices for example, have the potential to

provide truly big data for researchers in the speech-health domain [168]. The combination of such data with current and next generation deep learning paradigms can foster a new generation of patient-driven healthcare devices. Such devices could offer a range of benefits such as the improvement of diagnostics, triggering of earlier interventions and discovery of more effective treatments [169,170].

Acknowledgements

This research has received funding from the EU's 7th Framework Programme ERC Starting Grant No. 338164 (iHEARu), the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B), and the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902, and the European Unions Horizon 2020 research and innovation programme EFPIA.

References

[1] D. Bone, C.C. Lee, T. Chaspari, J. Gibson, S. Narayanan, Signal processing and machine learning for mental health research and clinical applications, IEEE Signal Process. Mag. 34 (2017) 189–196.
[2] S. Cunningham, P. Green, H. Christensen, J. Atria, A. Coy, M. Malavasi, L. Desideri, F. Rudzicz, Cloud-based speech technology for assistive technology applications (CloudCAST), Stud. Health Technol. Inf. 242 (2017) 322–329.
[3] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T.F. Quatieri, A review of depression and suicide risk assessment using speech analysis, Speech Commun. 71 (2015) 10–49.
[4] B. Schuller, Can affective computing save lives? Meet mobile health, IEEE Comput Mag 50 (2017) 40.
[5] M.S. Hossain, G. Muhammad, Cloud-assisted industrial internet of things (IIOT) enabled framework for health monitoring, Comput. Netw. 101 (2016) 192–202.
[6] B. Schuller, A. Batliner, S. Steidl, F. Schiel, J. Krajewski, The INTERSPEECH 2011 speaker state challenge, Proceedings INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, ISCA, Florence, Italy, 2011, pp. 3201–3204.
[7] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, B. Weiss, The INTERSPEECH 2012 Speaker Trait Challenge, in: Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, ISCA, Portland, OR, 2012, pp. 254–257.
[8] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Mar.i, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, S. Kim, The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism, in: Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, ISCA, Lyon, France, 2013, pp. 148–152.
[9] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Mar.i, Y. Zhang, The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load, in: Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, ISCA, Singapore, Singapore, 2014, pp. 427–431.
[10] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J.R. Orozco-Arroyave, E. Nöth, Y. Zhang, F. Weninger, The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition, in: Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, ISCA, Dresden, Germany, 2015, pp. 478–482.
[11] 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, S. Zafeiriou, The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring, in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3442–3446.
[12] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge, in: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC '13, ACM, Barcelona, Spain, 2013, pp. 3–10.
[13] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, ACM, Orlando, FL, 2014, pp. 3–10.
[14] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, Avec 2016: Depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, Amsterdam, Netherlands, 2016, pp. 3–10.
[15] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, M. Pantic, Avec 2017: Real-life depression, and affect recognition workshop and challenge, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17, ACM, Mountain View, CA, 2017, pp. 3–9.
[16] S. Jankowski, J. Covello, H. Bellini, J. Ritchie, D. Costa, The Internet of Things: Making sense of the next mega-trend, http://www.goldmansachs.com/our-thinking/outlook/internet-of-things/iot-report.pdf, 2014 (accessed: 25-06-2018).
[17] G. Hagerer, N. Cummins, F. Eyben, B. Schuller, Did you laugh enough today? – Deep Neural Networks for Mobile and Wearable Laughter Trackers, in: Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 2044–2045.
[18] E. Marchi, F. Eyben, G. Hagerer, B.W. Schuller, Real-time Tracking of Speakers' Emotions, States, and Traits on Mobile Platforms, in: Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, ISCA, San Francisco, CA, 2016, pp. 1182–1183.
[19] A. Tsiartas, C. Albright, N. Bassiou, M. Frandsen, I. Miller, E. Shriberg, J. Smith, L. Voss, V. Wagner, Sensay analyticstm: A real-time speaker-state platform, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '17, IEEE, New Orleans, LA, 2017, pp. 6582–6483.
[20] W. Fitch, The evolution of speech: a comparative review, Trends Cognit. Sci. 4 (2000) 258–267.
[21] P. Mermelstein, Articulatory model for the study of speech production, J. Acoust. Soc. Am. 53 (1973) 1070–1082.
[22] R.D. Kent, Research on speech motor control and its disorders: A review and prospective, J. Commun. Disord. 33 (2000) 391–428.
[23] A. Baddeley, Working memory and language: an overview, J. Commun. Disord. 36 (2003) 189–208.
[24] W.J.M. Levelt, A. Roelofs, A.S. Meyer, A theory of lexical access in speech production, Behav. Brain Sci. 22 (1999) 1–38.
[25] D. O'Shaughnessy, Speech Communications: Human and Machine, 2nd edition, Wiley-IEEE Press, Piscataway, NJ, 1999.
[26] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, MA, 2016 URL http://www.deeplearningbook.org.
[27] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Networks 61 (2015) 85–117.
[28] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.
[29] P. Vincent, H. Larochelle, Y. Bengio, P.A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th International Conference on Machine Learning, ICML '08, ACM, Helsinki, Finland, 2008, pp. 1096–1103.
[30] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[32] G.E. Nair, V. and Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, Omnipress, Haifa, Israel, 2010, pp. 807–814.
[33] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: G. Gordon, D. Dunson, M. Dudk (Eds.), Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 of Proceedings of Machine Learning Research, PMLR, Fort Lauderdale, FL, 2011, pp. 315–323.
[34] R. Raina, A. Madhavan, A.Y. Ng, Large-scale deep unsupervised learning using graphics processors, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, Montreal, Canada, pp. 873–880.
[35] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 25, Curran Associates Inc, 2012, pp. 1097–1105.
[36] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, S. Zafeiriou, Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network, in: Proceedings 41st IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2016, IEEE, Shanghai, PR. China, 2016, pp. 5200–5204.
[37] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (1989) 541–551.
[38] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Networks 5 (1994) 157–166.
[39] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[40] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Gated feedback recurrent neural networks, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, PMLR, Lille, France, vol. 37, 2015, pp. 2067–2075.
[41] R. Brueckner, B. Schulter, Social signal classification using deep BLSTM recurrent neural networks, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '14, IEEE, Florence, Italy, 2014, pp. 4823–4827.
[42] J. Han, Z. Zhang, N. Cummins, F. Ringeval, B. Schuller, Strength modelling for real-world automatic continuous affect recognition from audiovisual signals, Image Vision Comput. 65 (2016) 76–86 Special Issue on Multimodal Sentiment Analysis and Mining in the Wild.
[43] D. Le, Z. Aldeneh, E.M. Provost, Discretized continuous speech emotion recognition with multi-task deep recurrent neural network, in: Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech

Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 1108–1112.
[44] F. Eyben, F. Weninger, F. Gro, B. Schuller, Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor, in: Proceedings of the 21st ACM International Conference on Multimedia, MM '13, ACM, Barcelona, Spain, 2013, pp. 835–838.
[45] F. Schiel, C. Heinrich, S. Barfüsser, Alcohol language corpus: the first public corpus of alcoholized german speech, Lang. Resour. Eval. 46 (2012) 503–521.
[46] R.P. Clapham, L. van der Molen, R.J.J.H. van Son, M.W.M. van den Brekel, F.J.M. Hilgers, NKI-CERT Corpus – speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC '12, ELRA, Istanbul, Turkey, 2012, pp. 23–25.
[47] F. Ringeval, J. Demouy, G. Szaszak, M. Chetouani, L. Robel, J. Xavier, D. Cohen, M. Plaza, Automatic intonation recognition for the prosodic assessment of language-impaired children, IEEE Trans. Audio Speech Lang. Process. 19 (2011) 1328–1342.
[48] B. Schuller, F. Friedmann, F. Eyben, The munich BioVoice corpus: effects of physical exercising, heart rate, and skin conductance on human speech production, in: Proceedings 9th Language Resources and Evaluation Conference, LREC '14, ELRA, Reykjavik, Iceland, 2014, pp. 1506–1510.
[49] T.F. Yap, Speech production under cognitive load: effects and classification (Ph.D. thesis), Electrical Engineering & Telecommunications, Faculty of Engineering, University of New South Wales, Australia, 2012.
[50] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesus Francisco Vargas Bonilla, Maŕa Claudia Gonzalez-Rátiva, Elmar Nöth, New spanish speech corpus database for the analysis of people suffering from parkinson's disease, in: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC '14, ELRA, Reykjavik, Iceland, 2014, pp. 342–347.
[51] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. El-Desoky Mousa, B. Schuller, I hear you eat and speak: automatic recognition of eating condition and food types, use-cases, and impact on ASR performance, PLoS ONE 11 (2016) 1–24.
[52] J. Krajewski, S. Schnieder, A. Batliner, Description of the Upper Respiratory Tract Infection Corpus (URTIC), in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017. No pagination.
[53] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, L.-P. Morency, The Distress Analysis Interview Corpus of human and computer interviews, in: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC '14, ELRA, Reykjavik, Iceland, 2014, pp. 3123–3128.
[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explorations Newsl. 11 (2009) 10–18.
[55] P. Philip, P. Sagaspe, N. Moore, J. Taillard, A. Charles, C. Guilleminault, B. Bioulac, Fatigue, sleep restriction and driving performance, Accid. Anal. Prev. 37 (2005) 473–478.
[56] S.K.L. Lal, A. Craig, A critical review of the psychophysiology of driver fatigue, Biol. Psychol. 55 (2001) 173–194.
[57] A. Williamson, R. Friswell, Fatigue in the workplace: causes and countermeasures, Fatigue: Biomed. Health Behav. 1 (2013) 81–98.
[58] G. Belenky, A. Lamp, A. Hemp, J.L. Zaslona, Fatigue in the Workplace, in: M. Bianchi (Ed.), Sleep Deprivation and Disease: Effects on the Body, Brain and Behavior, Springer, New York, NY, 2014, pp. 243–268.
[59] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, F. Eyben, Medium-term speaker states – a review on intoxication, sleepiness and the first challenge, Comput. Speech Lang - Special Issue on Broadening the View on Speaker Analysis 28 (2014) 346–374.
[60] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[61] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, IEEE Signal Process. Mag. 29 (2012) 82–97.
[62] D.-Y. Huang, S.S. Ge, Z. Zhang, Speaker state classification based on fusion of asymmetric simpls and support vector machines, in: Proceedings INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, ISCA, Florence, Italy, 2011, pp. 3301–3304.
[63] W.M. Campbell, D.E. Sturim, D.A. Reynolds, Support vector machines using GMM supervectors for speaker verification, IEEE Signal Process. Lett. 13 (2006) 308–311.
[64] W.M. Campbell, D.E. Sturim, D.A. Reynolds, A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, ICASSP '06, IEEE, Toulouse, France, 2006, 4 pages.
[65] World Health Organization, Global status report on alcohol and health, 2014, http://www.who.int/substanceabuse/publications/globalalcoholreport/en/, 2014 (accessed: 26-06-2018).
[66] L.C. Sobell, M.B. Sobell, Effects of alcohol on the speech of alcoholics, J. Speech Lang. Hearing Res. 15 (1972) 861–868.
[67] F. Klingholz, R. Penning, E. Liebhardt, Recognition of low level alcohol intoxication from speech signal, J. Acoust. Soc. Am. 84 (1988) 929–935.
[68] Z. Zhang, F. Weninger, M. Wllmer, J. Han, B. Schuller, Towards intoxicated speech recognition, in: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, Anchorage, AK, 2017, pp. 1555–1559.
[69] C. Montacié, M.J. Caraty, Combining multiple phoneme-based classifiers with audio feature-based classifier for the detection of alcohol intoxication, in: Proceedings INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, ISCA, Florence, Italy, 2011, pp. 3205–3208.
[70] D. Bone, M.P. Black, M. Li, A. Metallinou, S. Lee, S. Narayanan, Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors, in: Proceedings INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, ISCA, Florence, Italy, 2011, pp. 3217–3220.
[71] D. Bone, M. Li, M.P. Black, S.S. Narayanan, Intoxicated speech detection: a fusion framework with speaker-normalized hierarchical functionals and GMM supervectors, Comput. Speech Lang. 28 (2014) 375–391.
[72] K. Berninger, J. Hoppe, B. Milde, Classification of Speaker Intoxication Using a Bidirectional Recurrent Neural Network, in: P. Sojka, A. Horák, I. Kopeček, K. Pala (Eds.), Text, Speech, and Dialogue, Springer International Publishing, Brno, Czech Republic, 2016, pp. 435–442.
[73] L. van der Molen, M.A. van Rossum, A.H. Ackerstaff, L.E. Smeele, C.R. Rasch, F.J. Hilgers, Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views, BMC Ear Nose Throat Disord. 9 (2009) 10.
[74] J. Kim, N. Kumar, A. Tsiartas, M. Li, S. Narayanan, Intelligibility classification of pathological speech using fusion of multiple high level descriptors, in: Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, ISCA, Portland, OR, 2012, pp. 534–537.
[75] R. Brückner, B. Schuller, Likability classification – a not so deep neural network approach, in: Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, ISCA, Portland, OR, 2012, pp. 290–293.
[76] E. Ylmaz, M. Ganzeboom, C. Cucchiarini, H. Strik, Combining non-pathological data of different language varieties to improve DNN-HMM performance on pathological speech, in: Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, ISCA, San Francisco, CA, 2016, pp. 218–222.
[77] S. Chandrakala, N. Rajeswari, Representation learning based speech assistive system for persons with dysarthria, IEEE Trans. Neural Syst. Rehabil. Eng. 25 (2017) 1510–1517.
[78] T. Lee, Y. Liu, Y.T. Yeung, T.K. Law, K.Y. Lee, Predicting severity of voice disorder from DNN-HMM acoustic posteriors, in: Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, ISCA, San Francisco, CA, 2016, pp. 97–101.
[79] American Psychiatric Association, Diagnostic and statistical manual of mental disorders: DSM-4, Washington, D.C., fourth ed., 2000.
[80] C. Dover, A.L. Couteur, The prevalence of anxiety and mood problems among children with autism and asperger syndrome, Arch. Dis. Child. 92 (2000) 540–545.
[81] J. Kim, P. Szatmari, S. Bryson, D. Streiner, F. Wilson, The prevalence of anxiety and mood problems among children with autism and asperger syndrome, SAGE publications and The National Autistic Society 4 (2000) 117–132.
[82] M. Carpenter, M. Tomasello, T. Striano, Role reversal imitation and language in typically developing infants and children with autism, Infancy 8 (2005) 253–278.
[83] A. Le Couteur, G. Haden, D. Hammal, H. McConachie, Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: the ADI-R and the ADOS, J. Autism Dev. Disord. 38 (2008) 362–372.
[84] M. Kjelgaard, H. Tager-Flusberg, An investigation of language impairment in autism: Implications for genetic subgroups, Lang. Cognit. Process. 16 (2001) 287–308.
[85] M. Kjelgaard, H. Tager-Flusberg, Update on the language disorders of individuals on the autistic spectrum, Brain Dev. 25 (2003) 166–172.
[86] D. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, S. Warren, Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development, Proc. Natl. Acad. Sci. 107 (2010) 13354–13359.
[87] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, M. Chetouani, D. Cohen, B. Schuller, Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children, in: Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, ISCA, San Francisco, CA, 2016, pp. 1210–1214.
[88] H.-Y. Lee, T.-Y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao, T.-L. Pao, Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition, in: Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, ISCA, Lyon, France, 2013, pp. 215–219.
[89] M. Asgari, A. Bayestehtashk, I. Shafran, Robust and accurate features for detecting and diagnosing autism spectrum disorders, in: Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, ISCA, Lyon, France, 2013, pp. 191–194.
[90] C.L. Huang, C. Hori, Classification of children with voice impairments using deep neural networks, in: 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE, Kaohsiung, Taiwan, 2013. 5 pages.
[91] World Health Organization, Depression and other common mental disorders: global health estimates, http://www.who.int/mentalhealth/management/depression/prevalenceglobalhealthestimates/en/, 2017 (accessed: 26-06-2018).
[92] T. Vos, C. Allen, M. Arora, R.M. Barber, Z.A. Bhutta, A. Brown, A. Carter, D.C. Casey, F.J. Charlson, A.Z. Chen, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 19902015: a systematic analysis for the global burden of disease study 2015, Lancet 388 (2016) 1545–1602.

[93] A. Pampouchidou, P. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, M. Tsiknakis, Automatic assessment of depression based on visual cues: a systematic review, IEEE Trans. Affective Comput. (2017) 27 Pages, in press.

[94] A.T. Beck, R.A. Steer, R. Ball, W.F. Ranieri, Comparison of beck depression inventories-ia and-ii in psychiatric outpatients, J. Pers. Assess. 67 (1996) 588–597.

[95] J.R. Williamson, T.F. Quatieri, B.S. Helfer, R. Horwitz, B. Yu, D.D. Mehta, Vocal biomarkers of depression based on motor incoordination, in: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC '13, ACM, Barcelona, Spain, 2013, pp. 41–48.

[96] J.R. Williamson, T.F. Quatieri, B.S. Helfer, G. Ciccarelli, D.D. Mehta, Vocal and facial biomarkers of depression based on motor incoordination and timing, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, ACM, Orlando, FL, 2014, pp. 65–72.

[97] Y. Zhu, Y. Shang, Z. Shao, G. Guo, Automated depression diagnosis based on deep networks to encode facial appearance and dynamics, IEEE Trans. Affective Comput. (2017) 8 pages, in press.

[98] Y. Kang, X. Jiang, Y. Yin, Y. Shang, X. Zhou, Deep transformation learning for depression diagnosis from facial images, in: J. Zhou, Y. Wang, Z. Sun, Y. Xu, L. Shen, J. Feng, S. Shan, Y. Qiao, Z. Guo, S. Yu (Eds.), Biometric Recognition, Springer International Publishing, 2017, pp. 13–22.

[99] T.F. Yap, J. Epps, E. Ambikairajah, E.H.C. Choi, Formant frequencies under cognitive load: effects and classification, EURASIP J. Adv. Signal Process. 2011 (2011) 219253.

[100] B. Schuller, F. Friedmann, F. Eyben, Automatic recognition of physiological parameters in the human voice: heart rate and skin conductance, in: Proceedings 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '13, IEEE, Vancouver, Canada, 2013, pp. 7219–7223.

[101] G. Gosztolya, T. Grósz, R. Busa-Fekete, L. Tóth, Detecting the intensity of cognitive and physical load using adaboost and deep rectifier neural networks, in: Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, ISCA, Singapore, Singapore, 2014, pp. 452–456.

[102] H. Jing, T.-Y. Hu, H.-S. Lee, W.-C. Chen, C.-C. Lee, Y. Tsao, H.-M. Wang, Ensemble of machine learning algorithms for cognitive and physical speaker load detection, in: Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, ISCA, Singapore, Singapore, 2014, pp. 447–451.

[103] T.L. Nwe, T.H. Nguyen, B. Ma, On the use of bhattacharyya based GMM distance and neural net features for identification of cognitive load levels, in: Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, ISCA, Singapore, Singapore, 2014, pp. 736–740.

[104] M.V. Segbroeck, R. Travadi, C. Vaz, J. Kim, M.P. Black, A. Potamianos, S.S. Narayanan, Classification of cognitive load from speech using an i-vector framework, in: Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, ISCA, Singapore, Singapore, 2014, pp. 751–755.

[105] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, A study of interspeaker variability in speaker verification, IEEE Trans. Audio Speech Lang. Process. 16 (2008) 980–988.

[106] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (2011) 788–798.

[107] H. Kaya, T. Özkaptan, A.A. Salah, S.F. Gürgen, Canonical correlation analysis and local fisher discriminant analysis based multi-view acoustic feature reduction for physical load prediction, in: Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, ISCA, Singapore, Singapore, 2014, pp. 442–446.

[108] O.-B. Tysnes, A. Storstein, Epidemiology of parkinson's disease, J. Neural Transm. 124 (2017) 901–905.

[109] T. Pringsheim, N. Jette, A. Frolkis, T.D. Steeves, The prevalence of parkinson's disease: a systematic review and metaanalysis, Mov. Disord. 29 (2014) 1583–1590.

[110] G.J. Canter, Speech characteristics of patients with Parkinsons disease: intensity, pitch, and duration, J. Speech Hearing Disord. 28 (1963) 221–229.

[111] J.A. Logemann, H.B. Fisher, B. Boshes, E.R. Blonsky, Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients, J. Speech Hearing Disord. 43 (1978) 47–57.

[112] L. Hartelius, P. Svensson, Speech and swallowing symptoms associated with parkinsons disease and multiple sclerosis: a survey, Folia Phoniatrica et Logopaedica 46 (1994) 9–17.

[113] S. Skodda, W. Visser, U. Schlegel, Vowel articulation in Parkinson's disease, J. Voice 25 (2011) 467–472.

[114] J.R. Orozco-Arroyave, F. Hönig, J.D. Arias-Londoño, J.F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, E. Nöth, Automatic detection of parkinson's disease in running speech spoken in three different languages, J. Acoust. Soc. Am. 139 (2016) 481–500.

[115] G.T. Stebbins, C.G. Goetz, Factor structure of the unified parkinson's disease rating scale: motor examination section, Mov. Disord. 13 (1998) 633–636.

[116] S. Hahm, J. Wang, Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data, in: Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, ISCA, Dresden, Germany, 2015, pp. 513–517.

[117] A. Wrench, K. Richmond, Continuous speech recognition using articulatory data, 2000, 145–148.

[118] T. Grósz, R. Busa-Fekete, G. Gosztolya, L. Tóth, Assessing the degree of nativeness and parkinson's condition using gaussian processes and deep rectifier neural networks, in: Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, ISCA, Dresden, Germany, 2015, pp. 919–923.

[119] J.R. Williamson, T.F. Quatieri, B.S. Helfer, J. Perricone, S.S. Ghosh, G. Ciccarelli, D.D. Mehta, Segment-dependent dynamics in predicting parkinsons disease, in: Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, ISCA, Dresden, Germany, 2015, pp. 518–522.

[120] World Health Organization, Obesity and overweight, http://www.who.int/mediacentre/factsheets/fs311/en/, 2018 (accessed: 26-03-2018).

[121] J.M. Fontana, M. Farooq, E. Sazonov, Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior, IEEE Trans. Biomed. Eng. 61 (2014) 1772–1779.

[122] E.S. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E.L. Melanson, M.R. Neuman, Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behaviour, IEEE Trans. Biomed. Eng. 57 (2010) 626–633.

[123] T. Pellegrini, Comparing svm, softmax, and shallow neural networks for eating condition classification, in: Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, ISCA, Dresden, Germany, 2015, pp. 899–903.

[124] B. Milde, C. Biemann, Using representation learning and out-of-domain data for a paralinguistic speech task, in: Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, ISCA, Dresden, Germany, 2015, pp. 904–908.

[125] H. Kaya, A.A. Karpov, A.A. Salah, Fisher vectors with cascaded normalization for paralinguistic analysis, in: Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, ISCA, Dresden, Germany, 2015, pp. 909–913.

[126] G. Gosztolya, L. Tóth, A feature selection-based speaker clustering method for paralinguistic tasks, Pattern Anal. Appl. 21 (2018) 193–204.

[127] K. Kroenke, T.W. Strine, R.L. Spitzer, J.B. Williams, J.T. Berry, A.H. Mokdad, The PHQ-8 as a measure of current depression in the general population, J. Affective Disord. 114 (2009) 163–173.

[128] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP – A collaborative voice analysis repository for speech technologies, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '14, Florence, Italy, pp. 960–964.

[129] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, Depaudionet: an efficient deep model for audio based depression classification, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, Amsterdam, Netherlands, 2016, pp. 35–42.

[130] L. Yang, D. Jiang, L. He, E. Pei, M.C. Oveneke, H. Sahli, Decision tree based depression classification from audio video and language information, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, Amsterdam, Netherlands, 2016, pp. 89–96.

[131] L. Yang, H. Sahli, X. Xia, E. Pei, M.C. Oveneke, D. Jiang, Hybrid depression classification and estimation from audio video and text information, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17, ACM, Mountain View, CA, 2017, pp. 45–51.

[132] L. Yang, D. Jiang, X. Xia, E. Pei, M.C. Oveneke, H. Sahli, Multimodal measurement of depression using deep learning models, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17, ACM, Mountain View, CA, 2017, pp. 53–59.

[133] L. Yang, D. Jiang, W. Han, H. Sahli, DCNN and DNN based multi-modal depression recognition, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction, ACII '17, IEEE, San Antonio, TX, 2017, pp. 484–489.

[134] Y. Gong, C. Poellabauer, Topic modeling based multi-modal depression detection, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17, ACM, Mountain View, CA, 2017, pp. 69–76.

[135] J.W. Pennebaker, M.E. Francis, R.J. Booth, Linguistic inquiry and word count: LIWC2001, Lawrence Erlbaum Associates, Mahwah, NJ, 2001 http://liwc.wpengine.com/.

[136] World Health Organization, Influenza (Seasonal), http://www.who.int/mediacentre/factsheets/fs211/en/, 2018. (accessed: 28-01-2018).

[137] M. Schmitt, B. Schuller, openXBOW – Introducing the Passau open-source crossmodal bag-of-words toolkit, J. Mach. Learn. Res. 18 (2017) 5 pages.

[138] D. Cai, Z. Ni, W. Liu, W. Cai, G. Li, M. Li, End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum, in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3452–3456.

[139] M. Huckvale, A. Beke, It sounds like you have a cold! Testing voice features for the Interspeech 2017 Computational Paralinguistics Cold Challenge, in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3442–3446.

[140] G. Gosztolya, R. Busa-Fekete, T. Grósz, L. Tóth, Dnn-based feature extraction and classifier combination for child-directed speech, cold and snoring identification, in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3522–3526.

[141] P. Jennum, R.L. Riha, Epidemiology of sleep apnoea/hypopnoea syndrome and sleep-disordered breathing, Eur. Respir. J. 33 (2009) 907–914.

[142] T. Young, L. Evans, L. Finn, M. Palta, et al., Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women, Sleep 20 (1997) 705–706.

[143] I. Fietze, T. Penzel, A. Alonderis, F. Barbe, M. Bonsignore, P. Calverly, W. De

Backer, K. Diefenbach, V. Donic, M. Eijsvogel, et al., Management of obstructive sleep apnea in Europe, Sleep Med. 12 (2011) 190–197.
[144] C.B. Croft, M. Pringle, Sleep nasendoscopy: a technique of assessment in snoring and obstructive sleep apnoea, Clin. Otolaryngol. 16 (1991) 504–509.
[145] H. Kaya, A.A. Karpov, Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold, in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3527–3531.
[146] G.-B. Huang, D.H. Wang, Y. Lan, Extreme learning machines: a survey, Int. J. Mach. Learn. Cybern. 2 (2011) 107–122.
[147] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, B. Schuller, Snore Sound Classification Using Image-based Deep Spectrum Features, in: Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3512–3516.
[148] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv abs/1409.1556, 2014.
[149] M. Freitag, S. Amiriparian, N. Cummins, M. Gerczuk, B. Schuller, An 'End-to-Evolution' Hybrid Approach for Snore Sound Classification, in: Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3507–3511.
[150] Z. Zhang, N. Cummins, B. Schuller, Advanced data exploitation in speech analysis – an overview, IEEE Signal Process. Mag. 34 (2017) 107–129.
[151] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates Inc, 2017, pp. 4077–4087.
[152] E. Triantafillou, H. Larochelle, J. Snell, J. Tenenbaum, K.J. Swersky, M. Ren, R. Zemel, S. Ravi, Meta-learning for semi-supervised few-shot classification, 2018. arXiv abs/1803.00676.
[153] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates Inc, 2014, pp. 2672–2680.
[154] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, X. Chen, Improved techniques for training GANs, in: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, Curran Associates Inc, 2016, pp. 2234–2242.
[155] Y. Saito, S. Takamichi, H. Saruwatari, Statistical parametric speech synthesis incorporating generative adversarial networks, IEEE/ACM Trans. Audio Speech Lang. 26 (2018) 84–96.
[156] C. Donahue, J. McAuley, M. Puckette, Synthesizing audio with generative adversarial networks, 2018. arXiv abs/1802.04208.
[157] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, B. Schuller, Speech-based diagnosis of autism spectrum condition by generative adversarial network representations, in: Proceedings of the 7th International Digital Health Conference, DH '17, ACM, London, U.K., 2017, pp. 53–57.
[158] M. Schmitt, F. Ringeval, B. Schuller, At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech, in: Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, ISCA, San Francisco, CA, 2016, pp. 495–499.
[159] B. Schuller, S. Steidl, P. Marschik, H. Baumeister, F. Dong, F.B. Pokorny, E.-M. Rathner, K.D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, S. Zafeiriou, the INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats, in: Proceedings INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, ISCA, Hyderabad, India, 2018, 5 pages.
[160] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, B. Schuller, auDeep: unsupervised learning of representations from audio with deep recurrent neural networks, J. Mach. Learn. Res. 18 (2018) 1–5.
[161] Y. Zhang, F. Weninger, B. Schuller, Cross-domain classification of drowsiness in speech: the case of alcohol intoxication and sleep deprivation, in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3152–3156.
[162] L. Chao, J. Tao, M. Yang, Y. Li, J. Tao, Multi task sequence learning for depression scale prediction from video, in: 2015 International Conference on Affective Computing and Intelligent Interaction ACII '15, IEEE, Xi'an, P.R.China, 2015, pp. 526–531.
[163] R. Gupta, S.S. Narayanan, Predicting affective dimensions based on self assessed depression severity, in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2016, pp. 1427–1431.
[164] R. Gupta, S. Sahu, C. Espy-Wilson, S.S. Narayanan, An affect prediction approach through depression severity parameter incorporation in neural networks, in: Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, ISCA, Stockholm, Sweden, 2017, pp. 3122–3126.
[165] E. Coutinho, B. Schuller, Shared acoustic codes underlie emotional communication in music and speech – evidence from deep transfer learning, PLoS ONE 12 (2017).
[166] J. Deng, S. Frühholz, Z. Zhang, B. Schuller, Recognizing emotions from whispered speech based on acoustic feature transfer learning, IEEE Access 5 (2017) 5235–5246.
[167] S. Sankaranarayanan, Y. Balaji, C.D. Castillo, R. Chellappa, Generate to adapt: aligning domains using generative adversarial networks, 2017, arXiv abs/1704.01705.
[168] R.S. Istepanian, T. Al-Anzi, m-health 2.0: new perspectives on mobile health, machine learning and big data analytics, Methods (2018) (in press).
[169] D. Metcalf, S.T.J. Milliard, M. Gomez, M. Schwartz, Wearables and the internet of things for health: Wearable, interconnected devices promise more efficient and comprehensive health care, IEEE Pulse 7 (2016) 35–39.
[170] L. Piwek, D.A. Ellis, S. Andrews, A. Joinson, The rise of consumer health wearables: promises and barriers, PLOS Medicine 13 (2016) 1–9.