Introduction to the Issue on Automatic Assessment of Health Disorders Based on Voice, Speech, and Language Processing

I. Introduction

APPROXIMATELY one-fifth of the world's population suffer or have suffered from voice and speech production disorders due to diseases or some other dysfunction. Thus, there is a clear need for objective ways to evaluate the quality of voice and speech as well as its link to vocal fold activity, to evaluate the complex interaction between the larynx and voluntary movements of the articulators (i.e., lips, teeth, tongue, velum, jaw, etc), or to evaluate disfluencies at the language level. The underlying assumption is that deviations from patterns that might be considered normal can be correlated with many different symptoms and psychophysical situations. However, despite a large effort in this field in the last few years, useful services are still far fewer than those in other areas of speech technology (text-to-speech synthesis, speech recognition, speaker recognition and verification, etc.) and, as a result, many results have not been transferred to clinical settings.

In addition, the processing of the speech signal not only lets evaluate voice disorders, but also opens the door to contribute to the examination of other health disorders. The current state of the art is now opening the possibility of early detection, monitoring and evaluation of certain pathologies whose etiology is not directly related to the speech apparatus, such as Alzheimer's, Parkinson, other Parkinsonisms (amyotrophic lateral sclerosis, Huntington's disease …), dementia, autism, attention deficit hyperactivity disorder, depression, and obstructive sleep apnea, among others, which manifest certain alterations in speech at phonation, articulation, prosodic, or even linguistic levels.

The maturity of current speech technologies and earlier results reported in the specific field of this issue demonstrate the potential of addressing these challenges to provide new tools for clinicians. However, the application of speech technologies to the assessment of voice and health disorders is not restricted to the medical area alone, as it may also be of interest in forensic applications, the assessment of voice quality for voice professionals such as singers, the evaluation of stress and fatigue, the evaluation of surgical as well as pharmacological treatments and rehabilitation, etc.

Digital Object Identifier 10.1109/JSTSP.2020.2978566

II. Overviews

Motivated by the aforementioned observations, this special issue of the IEEE Journal of Selected Topics in Signal Processing puts together the most recent research about the *Automatic Assessment of Health Disorders based on voice, speech and language processing*. The special issue features 20 original articles out of the 53 submissions received worldwide, which can be roughly categorized in seven different categories: diagnosis and modelling of obstructive sleep apnea; evaluation and detection of cognitive disorders; assessment of neurological disorders; detection of speech and voice disorders; voice and speech assessment and detection; assessment and detection of mood disorders; and voice and speech modelling. In the following subsections of this Editorial, we briefly outline the contribution of each paper to the special issue.

A. Diagnosis and Modelling of Obstructive Sleep Apnea

In *"Modeling Obstructive Sleep Apnea voices using Deep Neural Network Embeddings and Domain-Adversarial Training,"* Espinoza-Cuadros *et al.* use state-of-the-art speaker recognition techniques based on acoustic subspace modeling (i-vectors), and deep neural network embeddings (x-vectors), and show a weak connection between speech and Obstructive Sleep Apnea (OSA). The authors hypothesize that this weak effect is due to undesired sources of variability as speakers' age, body mass index (BMI), or height, and introduce Domain-Adversarial Training to remove these sources of variability. Results show an increase of accuracy taking Body Mass Index (BMI) as adversarial domain.

The paper *"Diagnosis of Obstructive Sleep Apnea using Speech Signals from Awake Subjects,"* by Zigel *et al.*, introduces a system for the diagnosis of OSA fusing three different sub-systems fed with features extracted from breathing segments within continuous speech signals, information acquired from sustained vowels using a convolutional neural network, and inherent information in continuous speech signals using a recurrent neural network. Each sub-system provided an apnea-hypopnea index (AHI) estimation and was combined with age and BMI to train a single system that estimates AHI using a linear regression.

1932-4553 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

B. Evaluation and Detection of Cognitive Disorders

The paper *"Pragmatic Aspects of Discourse Production for the Automatic Identification of Alzheimer's Disease,"* by Pompilli *et al.*, characterizes language abilities for the automatic identification of Alzheimer Disease (AD) from narrative description tasks by also incorporating pragmatic aspects of speech production. The authors, investigate the relevance of a set of pragmatic features extracted from an automatically generated hierarchy graph in combination with a complementary set of state of the art features encoding lexical, syntactic and semantic cues. Experimental results suggest improvements identifying AD patients when pragmatic features are incorporated to the set of lexico-semantic features.

In *"An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech,"* Luz *el al*. present a study of the predictive value of purely acoustic features automatically extracted from spontaneous speech for AD dementia detection, from a computational paralinguistics perspective. The effectiveness of several state-of-the-art paralinguistic feature sets for AD detection were assessed on spontaneous speech dataset. The feature sets assessed were the extended Geneva minimalistic acoustic parameter set, the emobase feature set, the ComParE 2013 feature set, and new Multi-Resolution Cochleagram features. The authors introduce a new active data representation method for feature extraction in AD dementia recognition. Results show that classification models based solely on acoustic speech features extracted with the method proposed can achieve accuracy levels comparable to those achieved by models that employ higher-level language features.

With an emphasis on cognitive and thought disorders, the manuscript *"A Review of Automated Speech and Language Features for Assessment of Cognition and Thought Disorders,"* by Voleti *et al.*, reviews a set of features from recorded and transcribed speech to objectively assess the speech and language, for the early diagnosis of the disease, and for tracking it after diagnosis. The study reviews existing speech and language features used in this domain, discuss their clinical application, and highlight their advantages and disadvantages. The authors review features to measure complementary dimensions of cognitive-linguistics, including syntactic complexity, language diversity, semantic coherence, and timing. The review concludes with a proposal of new research directions to further advance the field.

The paper *"A Multimodal Interlocutor-Modulated Attentional BLSTM for Classifying Autism Subgroups during Clinical Interviews,"* by Yun-Shao *et al.*, proposes a computational framework for differentiating three Autism Spectrum Disorder (ASD) subgroups: Autistic Disorder vs. High Functioning Autism vs. Asperger. The authors used two different Bidirectional Long-Short Term Memory (BLSTM) networks to model the speech and gestural movement separately and a deep fusion network to combine the information for the final classification. Data were obtained during interviews based on the Autism Diagnostic Observation Schedule (ADOS) protocol, which is considered the gold standard in the clinic to assess the severity of ASD. The multimodal approach embeds the interlocutors behavior coordination using the interlocutor modulation attention mechanism, where it automatically learns to emphasize the important part during the interaction progression.

In *"Transitive Entropy - A Rank Ordered Approach for Natural Sequences,"* Back *et al.* propose a new probabilistic measure, termed Transitive Entropy (TE) to tackle with the characterization of changes in language due to diseases such as dementia. The authors show that degenerative solutions can occur when using current entropy measures, being them less suitable for classifying disorders in natural language. The TE is proposed to overcome this problem. The authors examine the properties of the proposed entropy measure and demonstrate its effectiveness on successfully classifying patient dementia by application to a probabilistic model of pause length in the speech.

C. Assessment of Neurological Disorders

In *"Automatic Assessment of Sentence-Level Dysarthria Intelligibility using BLSTM,"* Bhat *et al.* propose a machine learning-based method to automatically classify dysarthric speech into intelligible and unintelligible using BLSTM networks. Additionally, the paper presents a method to use the available pre-trained acoustic models for transfer-learning, showing that this method was able to handle channel noise, providing a significant improvement in comparison to traditional machine learning methods.

The paper *"Automatic Assessment of Speech Impairment in Cantonese-speaking People with Apha*sia," by Qin *et al.*, describes a fully automated system for speech assessment of Cantonese-speaking people with aphasia (PWA). The authors used a deep neural network based automatic speech recognition system for aphasic speech trained with both in-domain and out-of-domain speech data. Story-level embedding and a siamese network are applied to derive text features, which are used to quantify the difference between aphasic and unimpaired speech. The proposed text features are combined with conventional acoustic features to cover different aspects of speech and language impairment in PWA. Results show a high correlation between predicted scores and subject assessment scores. The siamese network significantly outperforms story-level embedding in generating text features for automatic assessment.

Mariya *et al.*, in "*Data Augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition*," propose the use of data augmentation techniques to create new corpora of dysarthric speech. The authors, developed a two-level data augmentation from dysarthric speech using virtual microphone array synthesis and multi-resolution feature extraction. With the augmented speech data, an isolated word Automatic Speech Recognition (ASR) system was trained and tested with two additional corpora of dysarthric speech. Performance of the ASR system showed a reduced word error rate for low and very low intelligible speakers with dysarthria compared to recent works on data augmentation reported for dysarthric speech recognition.

D. Detection of Speech and Voice Disorders

In *"Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness,"* Villegas *et al.* developed a psychoacoustic roughness model as a predictor of creaky voice. The authors used a simple bi-directional Recurrent Neural Network to predict the presence of creakiness in vocalic segments from only roughness traces. The proposed roughness-based predictor eases interpretation and comparison of creakiness among corpora and suggests that roughness prediction models could be successfully used for classification of creaky intervals in speech.

The paper *"Analysis and Detection of Pathological Voice using Glottal Source Features,"* by Sudarsana *et al.*, provides a systematic analysis of glottal source features and investigates their effectiveness in voice pathology detection. Glottal source features are extracted using glottal flows estimated with the quasi-closed phase (QCP) glottal inverse filtering method, using approximate glottal source signals computed with the zero frequency filtering (ZFF) method, and using acoustic voice signals directly. In addition, the authors propose to derive mel-frequency cepstral coefficients (MFCCs) from the glottal source waveforms computed by QCP and ZFF to effectively capture the variations in glottal source spectra of pathological voice. The analysis revealed that the glottal source contains significant information that discriminates normal and pathological voice, being complementary to the information provided by the conventional MFCCs and Perceptual Linear Prediction features.

E. Voice and Speech Assessment and Detection

In *"Automated Speech Production Assessment of Hard of Hearing Children,"* Czap presents a method for the automated speech production assessment (ASPA) of hearing impaired children, providing feedback about the pronunciation quality of words and sentences uttered during unsupervised practice in the course of speech development. The essence of the ASPA method is the joint assessment of sound and rhythm errors. The method uses the output activity of the neural networks trained to classify speech sounds to assess sound correctness. A dynamic time warping adapted to the speech of the hearing impaired was used to determine rhythm errors. The novelty of the procedure is that it provides a method for the assessment of non-typifiable pronunciation errors.

Chandrashekar *et al.*, in *"Spectro-Temporal Representation of Speech for Intelligibility Assessment of Dysarthria,"* explore the use of spectro-temporal representations for speech intelligibility assessment of dysarthric speech. The authors investigate the use of spectro-temporal representations to evaluate intelligibility levels using artificial neural network (ANN) and convolutional neural network (CNN), concluding that the CNN classifier performs better than the one based on ANN. The authors also tested time-frequency CNN configurations, which improved their performance in comparison with time or frequency CNN configurations.

In *"The Automatic Detection of Speech Disorders in Children: Challenges, Opportunities and Preliminary Results,"* Shahin *et al.* discuss three key challenges in processing child disordered speech. First, authors investigate the effectiveness of high-level paralinguistic features in disordered speech detection to reduce the dependency on annotated data, training a binary classifier using paralinguistic features extracted from both typically developing children and those suffering from Speech Sound Disorders. Second, the authors tackle the speech disorder detection as an anomaly detection problem where models were trained on typically developing speech, reducing the need for disordered training data. An anomaly detection-based system was trained with speech attribute features to classify between typical and atypical phoneme pronunciations of children with speech disorder. Finally, they test the efficiency of an x-vector based speaker diarization technique in pediatric therapy session recordings, successfully distinguishing between therapist and child speech.

In *"Multimodal and multi-output deep learning architectures for the automatic assessment of voice quality using the GRB scale,"* Arias-Londoño *et al.* propose a technique based on deep learning to objectively evaluate the speech according to the perceptual criteria commonly used in the clinic by phoniatricians and speech therapists. The ultimate goal is contributing to remove the subjectivity of an evaluation process widely affected by intra- and inter-rater variability.

F. Assessment and Detection of Mood Disorders

The paper *"Automatic Assessment of Depression from Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders,"* by Zhao *et al*. contributes to the field of quantitative mental health research with a deep learning approach that combines unsupervised learning, knowledge transfer and hierarchical attention for the task of speech-based depression severity measurement. The authors propose an attention transfer process that transfers attentions to measure the depression severity in both frame and sentence levels across tasks. The paper also proposes a novel hierarchical attention autoencoder paradigm that applies attention mechanisms to train hierarchical autoencoders capable of generating representations of depressed speech in an unsupervised manner.

Huang *et al.*, in *"Natural Language Processing Methods for Acoustic and Landmark Event-based Features in Speech-based Depression Detection,"* propose a framework for analyzing speech as a sequence of acoustic events, and investigates its application to depression detection. Acoustic space regions are tokenized to 'words' representing speech events at fixed or irregular intervals. This tokenization allows the exploitation of acoustic word features using natural language processing methods. An advantage of this framework is its ability to accommodate heterogeneous event types, combining acoustic words and speech landmarks, which are articulation-related speech events.

G. Voice and Speech Modelling

The paper *"Improved subglottal pressure estimation from neck-surface vibration in healthy speakers producing non-modal phonation,"* by Lin *et al*., describes a non invasive methodology to improve the estimation of subglottal pressure during natural

speech from neck surface vibration during non-modal phonation by incorporating accelerometer-based measures of cepstral peak prominence, fundamental frequency, and of the subglottal impedance-based inverse filtering waveform. The method has the potential to be used for the clinical assessment of voice disorders, particularly in ambulatory monitoring and biofeedback.

Pandei *et al*., in *"Epoch Detection Using Hilbert Envelope for Glottal Excitation Enhancement and Maximum-Sum Subarray for Epoch Marking,"* present a novel technique for detecting glottal excitation epochs with an application to the challenging scenario of the analysis of pathological voices, in which periodicity is significantly affected by different types of perturbations.

III. Summary

This special issue is expected to contribute to make this field more open and visible to speech and signal processing experts, presenting the state of the art and the potentiality of the current speech and language technologies to contribute to solve important challenges that were partially hidden to the scientific community. Moreover, it is expected to contribute to stimulate the multi and interdisciplinary work that has evolved the speech technologies from its very early stages, but with a turn to the biomedical field.

The emphases are on both basic and applied research related to the monitoring of voice and speech production status, as well as in the clinical evaluation of new developments. From these papers, we hope that the interested reader will find useful suggestions and further stimulation to carry on research in this field.

This special issue is expected to elucidate not only the technical, but also some of the clinical difficulties inherent to the field, providing a forum for sharing thoughts on how to overcome them.

Although there has been significant progress in the field, there are many future challenges, including the interpretation of learned models, adversarial examples, the lack of reproducibility due to the non-existence of open corpora, the existence of comorbidities affecting the different disorders, and problems associated with data-poor domains.

Acknowledgment

The Guest Editors acknowledge all the authors for their valuable contributions to this special issue, thanking them for their patience during the always hard and long reviewing process, especially to those that unfortunately had no opportunity to see their work published. They encourage the authors to continue with their research in this field.

They would also like to thank all the reviewers who took time and consideration to assess the submitted manuscripts. Their diligence and their constructive criticisms and remarks significantly contributed to to the high quality of the papers included in this special issue. Finally, they would like to thank the Editorial Board, Editor-in-Chief Prof. Lina Karam and the Journal Manager Rebecca Wollman for their valuable and necessary support throughout the entire process.

Juan I. Godino-Llorente, *Lead Guest Editor*
Theory and Communications Department
Universidad Politécnica de Madrid
28031 Madrid, Spain

Douglas O'Shaughnessy, *Guest Editor*
Centre Énergie Matériaux Télécommunications
Institut National de la Recherche Scientifique
Montréal, QC H5A 1K6, Canada

Tan Lee, *Guest Editor*
Department of Electronic Engineering
The Chinese University of Hong Kong
Hong Kong

Najim Dehak, *Guest Editor*
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218-2680 USA

Claudia Manfredi, *Guest Editor*
Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Firenze
50121 Florence, Italy

Juan I. Godino-Llorente (Senior Member, IEEE) was born in Madrid, Spain, in 1969. He received the B.Sc. and M.Sc. degrees in telecommunications engineering, and the PhD. degree in computer science from Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 1992, 1996, and 2002, respectively. From 1996 to 2003, he was an Assistant Professor with the Circuits and Systems Engineering Department, UPM. From 2003 to 2005, he joined the Signal Theory and Communications Department, University of Alcala. From 2005, he joined again UPM, being the Head of the Circuits and Systems Engineering Department from 2006 till 2010. Since 2011, he has been a Full Professor with Signal Theory and Communications Department, UPM. He has been the Spanish Coordinator of the 2103 COST Action funded by the European Science Foundation, and the General Chairman of the 3rd Advanced Voice Function Assessment Workshop. He is a member of ISCA. During his career, he has lead more than 20 research projects funded by national or international public bodies and by the industry. During the academic term 2003–2004, he was a Visiting Professor with Salford University, Manchester, U.K; and in 2016, he was a Visiting Researcher with the Massachusetts Institute of Technology, USA funded by a Fulbright grant. He has served as an Editor for the *Speech Communication Journal* and for the *EURASIP Journal of Advances in Signal Processing*, and has also been a member of the scientific committee of INTERSPEECH, IEEE ICASSP, EUSIPCO, BIOSIGNALS, and other events.

Douglas O'Shaughnessy (Fellow, IEEE) received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1976. He has been a Professor with INRS (University of Quebec) and an Adjunct Professor with McGill University since 1977. He is the author of the textbook *Speech Communications: Human and Machine* (1986 Addison-Wesley; revised in 2000, IEEE Press). He is a Fellow of the Acoustical Society of America (1992) and of the International Speech Communication Association (2019). He was the founding Editor-in-Chief of *the EURASIP Journal on Audio, Speech, and Music Processing*. He was the Secretary of the International Speech Communication Association. He has presented tutorials on automatic speech recognition at ICASSP-1996, ICASSP-2001, ICC-2003, and ICASSP-2009. He is also an Associate Editor for *JASA Express Letters* (2008-present). He is currently a member of the IEEE Signal Processing Society Board of Governors, as well as the IEEE Fellow Committee.

Tan Lee (Member, IEEE) was a Postdoctoral Researcher with the Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden, in 1997–1998. He is currently an Associate Professor with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong and the Director of the DSP and Speech Technology Laboratory. He has been working on speech and language related research since early 90s. In 2001, he was a Visiting Researcher with AT&T Bell Laboratories, Murray Hill, NJ, USA. In recent years, he has been collaborating closely with medical doctors, and speech and hearing professionals, to apply advanced signal processing methods in dealing with human communication disorder problems. His recent research interests include automatic assessment of voice disorder, speech disorder and language impairment, analysis of child and elderly speech, and speech under pressure. He is a member of the International Speech Communication Association. He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, and an Editorial Board Member of the *EURASIP Journal on Advances in Signal Processing*. He served as an Area Chair for INTERSPEECH 2014, 2016, and 2018, and the General Co-Chair of the 2018 International Symposium on Chinese Spoken Language Processing.

Najim Dehak (Senior Member, IEEE) received the Ph.D. degree from the School of Advanced Technology, Montreal in 2009. During his Ph.D. studies, he worked with the Computer Research Institute of Montreal, Canada. He is well known as a leading developer of the I-vector representation for speaker recognition. He first introduced this method, which has become the State-of-the-Art in this field, during the 2008 summer Center for Language and Speech Processing workshop with Johns Hopkins University. This approach has become one of the most known speech representations in the entire speech community. He is currently a Faculty Member with the Department of Electrical & Computer Engineering, Johns Hopkins University, Baltimore, MD, USA. Prior to joining Johns Hopkins, he was a Research Scientist with the Spoken Language Systems Group, the MIT Computer Science and Artificial Intelligence Laboratory. His research interests are in machine learning approaches applied to speech processing, audio classification, and health applications. He is a member of the IEEE Speech and Language Technical Committee.

Claudia Manfredi (Member, IEEE) is with the Department of Information Engineering, Universita' degli Studi di Firenze, Firenze, Italy, where she has been working in the field of control systems and identification, linear and nonlinear system analysis. Her main research activity concerns voice analysis under a biomedical perspective. She is a member of the International Speech and Communication Association, the Italian Biomedical Engineering Group, the Pacific Voice and Speech Foundation, the Pan European Voice Conference, the Collegium Medicorum Theatri. She is member of the Editorial Board of *Biomedical Signal Processing and Control Journal* (Elsevier Ltd.). She was the Italian representative in the Management Committee of the European COST Action 2103 "Advances in voice quality assessment" (2006–2011) and WG member in the COST 1101 action "European network for the studies of dystonia syndromes," 2011–2015. Since 1999, she has been organizes the series of biennial International Workshops: Models and Analysis of Vocal Emissions for Biomedical Applications, Firenze, Italy. She has been a member of the organizing committee of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE-EMBS) Milano, 2015.