Archives of Acoustics Vol. 48, No. 3, pp. 289–315 (2023), doi: 10.24425/aoa.2023.146640

Review Paper

Speech Analysis as a Tool for Detection and Monitoring of Medical Conditions: A review

Magdalena IGRAS-CYBULSKA[1],[3]*, Daria HEMMERLING[1],[3], Mariusz ZIÓŁKO[1],
Wojciech DATKA[2],[4], Ewa STOGOWSKA[2], Michał KUCHARSKI[1],
Rafał RZEPKA[5], Bartosz ZIÓŁKO[1],[5]

[1] *Techmo sp. z o.o.*
Kraków, Poland

[2] *Medical University of Bialystok*
Białystok, Poland

[3] *AGH University of Science and Technology*
Kraków, Poland

[4] *Faculty of Medicine, Jagiellonian University*
Kraków, Poland

[5] *Hokkaido University*
Kita Ward, Sapporo, Hokkaido, Japan

*Corresponding Author e-mail: migras@agh.edu.pl

(*received May 22, 2022; accepted March 21, 2023*)

The goal of this article is to present and compare recent approaches which use speech and voice analysis as biomarkers for screening tests and monitoring of some diseases. The article takes into account metabolic, respiratory, cardiovascular, endocrine, and nervous system disorders. A selection of articles was performed to identify studies that assess voice features quantitatively in selected disorders by acoustic and linguistic voice analysis. Information was extracted from each paper in order to compare various aspects of datasets, speech parameters, methods of applied analysis and obtained results. 110 research papers were reviewed and 47 databases were summarized. Speech analysis is a promising method for early diagnosis of certain disorders. Advanced computer voice analysis with machine learning algorithms combined with the widespread availability of smartphones allows diagnostic analysis to be conducted during the patient's visit to the doctor or at the patient's home during a telephone conversation. Speech analysis is a simple, low-cost, non-invasive and easy-to-provide method of medical diagnosis. These are remarkable advantages, but there are also disadvantages. The effectiveness of disease diagnoses varies from 65% up to 99%. For that reason it should be treated as a medical screening test and should be an indication of the need for classic medical tests.

Keywords: speech analysis; speech features; acoustic parameters; linguistic analysis; voice biomarkers; screening tests.

Acronyms

Acc – accuracy,
AD – Alzheimer's disease,
AI – artificial intelligence,
ALS – amyotrophic lateral sclerosis,
ALSFRS-R – sclerosis functional rating scale,
AMDF – average magnitude difference function,
APQ – amplitude perturbation quotient,
AR – average recall,
ASCVD – atherosclerotic cardiovascular disease,
ASR – automatic speech recognition,
AUC – area under a curve,
AUROC – area under the receiver operating characteristic,

AVEC	audio visual emotion challenges,
AVQI	acoustic voice quality index,
BDI	Beck depression inventory,
BFI	big five inventory,
BIDR	balanced inventory of desirable responding,
BMI	body mass index,
BPRS	brief psychiatric rating scale,
CAD	coronary artery disease,
CH	control healthy,
CHD	coronary heart disease,
CHR	clinical high-risk,
CNN	convolutional neural network,
CSL	computerized speech lab,
CVR	cockpit voice recorder,
DAIC	distress assessment interview corpus,
DCT	discrete cosine transform,
DDK	diadochokinetic,
DM	diabetes mellitus,
DNN	deep neural network,
DSM	diagnostic and statistical manual of mental disorders,
EM	expectation–maximization,
F0	fundamental frequency,
FBS	fetal bovine serum,
FT4	free thyroxine,
GC	NN gated convolutional neural network,
GFI	glottal function index,
GMM	the Gaussian mixture model,
GMM-UBM	the Gaussian mixture model-universal background model,
GPR	the Gaussian processes regression,
GRBAS	grade-roughness-breathiness-asthenia-strain scale,
H&Y	the Hoehn and Yahr scale,
HAMD	the Hamilton depression rating scale,
HbA1c	glycated hemoglobin A1c,
HC	healthy controls,
HLM	hierarchical linear modeling,
HMM	hidden Markov model,
HRSD	the Hamilton rating scale for depression,
HSC	hierarchical spectral clustering,
IQR	interquartile range,
IVR	interactive voice response,
JAD	Just Add Data,
K-SADS-PL	kiddie schedule for affective disorders and schizophrenia (present and lifetime version),
KNN	K-nearest neighbours,
LLD	low-level descriptors,
LR	logistic regression,
LSA	latent semantic analysis,
LSTM	multi-layer long short-term memory,
LTAS	long term average spectrum,
MAE	mean absolute error,
MAP	maximum a posteriori,
MDS-UPDRS	Movement Disorders Society UPDRS,
MDVP	Multi-Dimensional Voice Program,
MFCC	mel-frequency cepstral coefficients,
MHMC	Multimedia Human-Machine Communication,
MLP	multilayer perceptron,
MMSE	mini-mental state examination,
MPT	maximum phonation time,
NB	naive Bayes,
NN	neural networks,
NN LSTM	neural net multi-layer long short-term memory,
PANAS	positive and negative affect schedule,
PANSS	positive and negative syndrome scale,
PCL-C	post-traumatic stress disorder checklist,
PCOS	polycystic ovary syndrome,
PD	Parkinson's disease,
PDD	phase distortion deviation,
PHQ-9	patient health questionnaire-9,
PPQ	period perturbation quotient,
PTP	phonation threshold pressure,
PTSD	post-traumatic stress disorder,
PVRQoL	Pediatric Voice-Related Quality-of-Life,
QIDS	quick inventory of depressive, symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR),
RAP	relative average perturbation,
RBF	radial basis function,
RF	random forest,
RLR	randomized logistic regression,
RME	mental and emotional reinforcement,
RMSE	root mean square error,
RSI	reflux severity index,
RVM	relevance vector machines,
Sens	sensitivity,
SER	standard error of regression,
SIPS	semi-structured interview,
SIT	sentence intelligibility test,
SNR	signal-to-noise ratio,
SOPS	scale of prodromal symptoms,
Spec	specificity,
SPT	speech pause time,
STAI	state-trait anxiety inventory,
SVM	support vector machine,
T4	thyroxine, thyroid hormone,
TSH	thyroid-stimulating hormone,
TSST-C	trier social stress test for children,
UAR	unweighted average recall,
UBM	universal background model,
UD	unipolar depression,
UPDRS	Unified Parkinson's Disease Rating Scale,
Var	variance,
VC	vital capacity,
VHDAIC	virtual human distress assessment interview corpus,
VHI	voice handicap index,
YMRS	Young's rating scale for mania.

1. Introduction

Organs involved in the speech generation process are highly sensitive to both physical and mental ailments, hence the health of the speakers significantly affects their manner of speaking, voice emission, syntax, semantics and specific speech habits. Early detection and treatment of disorders can improve the effectiveness of treatment. In spite of this, speech analysis is currently rarely used in medical diagnostics of disorders other than those directly affecting the organs involved in speech generation and the respiratory system.

There is a high volume of publications on the diagnosis of specific disorders using speech analysis. The largest number of publications concerns the diagnosis of Parkinson's disease (PD). This high number of publications is reflected in this paper.

Speech analysis is a simple, low-cost, non-invasive and easy-to-provide preliminary test for disorders which affect speech, even marginally. Publications usually present problems related to the diagnosis of a single disease entity and state a high likelihood of diagnosis.

In this article we summarize different approaches which create systems to detect voice and speech impairments tackled by researchers around the world. Authors take into account different languages, dialects, types of speech (vowels, read text, monologue, etc.), algorithms, and different disorders to be analyzed. The current review aims to summarize the state-of-the-art of voice analysis as a biomarker of diseases. In particular, we want to compare properties of speech corpora and different approaches to speech processing paths.

Although some systematic reviews have been prepared recently, they are usually dedicated to certain groups of disorders (Cummins *et al.*, 2015a; Dogan *et al.*, 2017; Low *et al.*, 2020; Moro-Velazquez *et al.*, 2021; Stogowska *et al.*, 2022). To the best of our knowledge, there is a lack of a systematic review of voice analysis in connection with somatic disorders. In this study we did not include disorders which directly affect organs involved in speech generation, such as vocal cords or a vocal tract, as well as pulmonary disorders. The papers present a range of disorders that were analyzed by researchers and the results they obtained. This includes cardiovascular, metabolic, endocrine, COVID-19, schizophrenia, depression, amyotrophic lateral sclerosis (ALS), affective and neurodegenerative (Parkinson's, Alzheimer's, dementia) disorders. Creating a system that could monitor and reveal whether a patient has any voice/speech abnormalities and whether further diagnostics are required for a specific disease entity would be extremely useful in the medical environment. Another highly desirable tool would be a system allowing monitoring of treatment through voice analysis. In this article, we present scientific approaches to the problems of detecting disorders through voice analysis and a summary of the results, challenges, and problems.

2. Methods

2.1. Speech as an objective biomarker

For the purpose of this work, we have analyzed the PRISMA checklist which includes reports of reviews evaluating randomized trials. It is also a basis for reporting systematic reviews of different types of research. Articles with publication dates between January 2011 and November 2020 were selected from PubMed and ISCA Archive, using keywords 'voice' and 'speech' and respective disorder names. The records were screened for relevance to the topic of this review in order to identify studies that quantitatively assessed voice quality in the selected disorders by voice analysis. Studies which included only a perceptual assessment of voice were rejected.

For cardiovascular, metabolic and endocrine disorders, schizophrenia and ALS, all records were included as these disorders are less well documented (number of research papers less than 10). In contrast, affective disorders and neurodegenerative disorders (Parkinson's, Alzheimer's, dementia) are well investigated and the majority of publications describe automatic recognition with machine learning methods or even applications. Therefore only selected articles were included in our analysis, selected on the basis of publication in the ISCA archive and a high level of advancement. In this review we mentioned the research articles as well as three recent systematic reviews.

Determination of speech features we divided into several categories depending on the source of their origin (prosodic, spectral, voice source, linguistic) without going into further details. In some cases, standard feature vectors were selected. Machine learning approaches are summarized with information on the classifiers, the evaluation method and achieved best scores. If there was no attempt at automatic classification or regression, we report which statistical tests were used.

Finally, reviewed papers were flagged (Tables 10, 11, and 13) with one of the tags describing how advanced it is: 1 – basic research (investigating the statistical significance of acoustic parameters in the context of disorders); 2 – automated (using machine learning to classify regression); 3 – application (usually a smartphone app) for types of diagnostics (not simply collecting recordings).

In Sec. 3, each subsection starts with a medical description of how each disorder affects the voice and contains a summary of different processing approaches; they are supplemented by tables comparing cited studies. We reviewed a total of 110 papers, including disorders: endocrine, cardiovascular, metabolic, neurodegenerative, mental, and COVID-19.

Creating a computerized system of speech recognition-based diagnostic tools includes three steps. Voice recordings of a control group and individuals affected by the disorder in question are required first. Next, specific speech features are selected and calculated for both sets. In the training phase, the computer compares features of both sets of recordings and creates a classifier. This classifies the recordings into one of two sets: non affected individuals, or individuals suspected to be affected by the ailment. The structure of such a system is shown in Fig. 1.

Currently, the most difficult element is obtaining a sufficiently high number of recordings for the

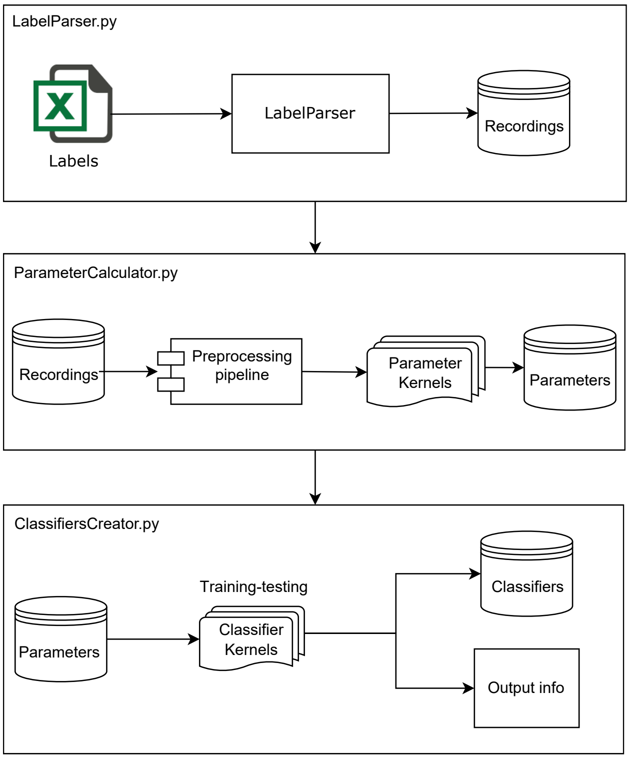

Fig. 1. Elements of a system for determining: speech features, training classifiers, and medical diagnosis.

training, validation, and testing phases. The recording databases are diverse and require a brief introduction. Software for determining speech features and creating classifiers is generally widely available, so the main task is selecting an effective algorithm. This means that the same algorithms are used for a variety of issues in speech technology, e.g., speech and speaker recognition, emotion detection and diagnosis of disease states.

We also performed a summary of available databases mentioned in the literature. Our analysis focused on the language and content of speech (with categories including sustained vowels, read speech, spontaneous monologue, dialogue with a human or virtual interviewer), recording protocol (how many times the individual was recorded, recording procedure, recording duration), number of speakers, their age and gender (in both the Control Healthy (CH) and affected persons), and other modalities (usually video, sometimes motion capture or biometric signals). We also included information on metadata: clinical evaluation of patients and perceptual evaluation of voice. Although the quality of recordings may be crucial for further processing, we did not compare technical details of recording procedures (equipment, acoustic conditions, sampling frequency, file parameters) because they are not usually systematically reported in the articles.

Finally, we reviewed 47 databases: 10 corpora of endocrine diseases (three of diabetes, four of polycystic ovary syndrome (PCOS) and related disorders, three of thyroid disorders), two of cardiovascular disorders (one of CAD, one of CHD), six of metabolic disorders (obesity), 12 of neurodegenerative disorders (two of ALS, eight of Parkinson's disease, two of Alzheimer's disease), and 17 of mental disorders (three of bipolar, 11 depression and/or anxiety and/or PTSD, three of schizophrenia).

Additional diagrams were prepared to illustrate general tendencies and to compare advancements in the state-of-the-art in the analyzed groups of disorders. The main sources of problems were identified and some recommendations for future research were set.

2.2. Databases

Speech corpus development is generally time and cost consuming; however, good quality recordings are crucial for further processing. Figure 2 indicates that the majority of speech recording databases for the purposes of medical diagnostics emerged in 2016. This was likely in response to publications reporting a satisfactory effectiveness of speech analysis in an initial recognition of disease symptoms.

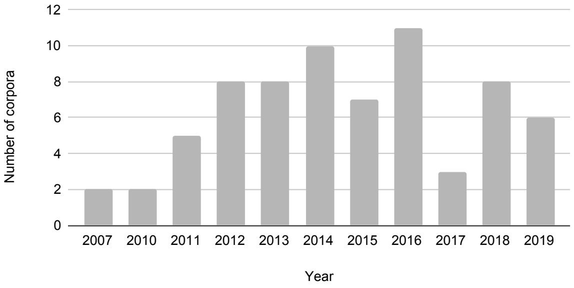

Fig. 2. Number of corpora created in 2007–2019.

There is a wide range of speech protocols, from the shortest (sustained phonation of vowels only) to the longest (interviews with a virtual agent). For read speech, there are standard text passages to be read, usually excerpts from stories or a short natural sentence. Counting one to 10 is also used. Sometimes the patient is asked for a short monologue. Just 21 of the 47 corpora contained more than two categories of speech (Fig. 3).

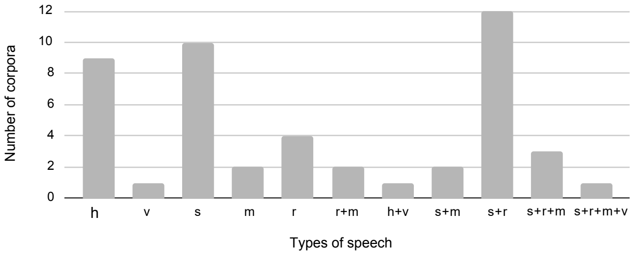

Fig. 3. Types of speech recorded in corpora 2009–2019: h – dialogue with a human; v – dialogue with a virtual agent; s – sustained vowels; m – monologue; r – read speech.

2.3. Features

Deviations of voice features are generally the result of anatomical changes or changes in the functioning of the nervous system. The cited publications do not concern analysis of sources of voice deviations, but only the ability to detect the effects of these changes is considered.

Three sets of speech parameters are usually distinguished: source parameters, vocal tract parameters, and prosody. The first and second groups are usually analyzed in the frequency domain of 20-30 milliseconds frames. Therefore, they are described as low-level descriptors (LLD).

Signal processing algorithms such as filtering or linear prediction allow for the extraction of acoustic features of the source and filter separately. The most popular source parameters include jitter and shimmer, which describe the stability of fundamental frequency production, and voice trembling. Examples of typical vocal tract features include formants and mel-frequency cepstral coefficients (MFCC). Prosodic features, such as syllables, phrases, and sentences, are observed in larger frames. For this reason, prosody parameters are also known as supra segmental or high-level features. Prosody describes intonation (modulation of fundamental frequency (F0) within the utterance), intensity (loudness, energy), and rhythm of speech (pauses, duration of speech segments, speech tempo).

Most of the authors of the reviewed articles used standard software tools to extract acoustic features. The most popular software tools for acoustic feature extraction are openSMILE, Praat, Multi-Dimensional Voice Program (MDVP), Kay Elemetrics-Computer Speech Lab, Dr. Speech, Snack Sound Toolkit and MATLAB (toolboxes such as Voice Sauce).

2.4. Classification models and evaluation metrics

In the reviewed papers focusing on the relationship between voice and the specified disorders, several study design scenarios can be found:

- ill/healthy comparison which leads to binary classification;
- comparison of subclasses of given disorders, which leads to multiclass classification;
- comparison between pre-treatment and post-treatment;
- monitoring the progression of disease severity;
- monitoring of treatment success;
- correlation with prodromal symptoms which leads to the measure of risk of the given disorder.

Depending on the case, classification or regression methods are applied. The most popular models are Gaussian mixture model (GMM) and *i*-vectors. Two type of classifiers are the most frequently used: support vector machine (SVM) and neural networks (NN). Different measures are used according to the classification of the regression model (Tables 2, 4, 6, 8, 10, 11, 13, 15, and 17).

3. Results

3.1. Cardiovascular diseases

Cardiovascular disease is the most common cause of death in both developed and developing countries (Kones, Rumana, 2017). There is currently little evidence on any association between cardiovascular disease and voice features (Table 1). It has been posited that the disease process may affect anatomical structures associated with voice generation, for example in atherosclerosis; as a systemic inflammatory process, it is associated with multiple pathological processes such as chronic kidney disease, cerebrovascular disease, vascular dementia, retinopathy and peripheral artery disease (Maor *et al.*, 2018).

Pareek and Sharma (2016) studied coronary heart disease (CHD) (Table 2). Their research reveals significant variations in spectrograms, the long-term average spectrum (LTAS) and other voice parameters such as jitter, shimmer and amplitude perturbation quotient (APQ), smoothed APQ, relative average perturbation (RAP), period perturbation quotient (PPQ),

Table 1. Comparison of speech databases in CAD and CHD
(v – sustained vowel; r – read speech; m – monologue; h – dialog with human; a – dialog with virtual agent).

Published	Speaker language	Number of patients	Number of HC	Age	Clinical evaluation	Voice evaluation	Type of speech v	r	m	h	a	Duration per speaker [h:min:s]
Cardiovascular: CHD												
Pareek, Sharma, 2016	N/A	80	80	53.4	NI	–	1	0	0	0	0	00:00:04
Cardiovascular: CAD												
Maor *et al.*, 2018	English	71	37	63	ASCVD risk score	–	0	1	1	0	0	NI

N/A – not applicable; NI – no information.

Table 2. Comparison of research methods in cardiovascular diseases
(pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Disease	Tool used for features extraction	Features					Classifier/Test
			pr	sp	vs	li	naf	
Pareek, Sharma, 2016	CHD	MDVP, CSL	0	1	1	0	18	–
Maor *et al.*, 2018	CAD	Beyond verbal communications	0	1	0	0	81	Logistic regression

smoothed PPQ in comparison with the control group ($P < 0.05$).

In research conducted by (Maor *et al.*, 2018), coronary artery disease (CAD) patients were compared to CH (Table 2). MFCC parameters were extracted. Univariate binary logistic regression analysis identified five voice features that were associated with CAD. Multivariate binary logistic regression with adjustment for atherosclerotic cardiovascular disease (ASCVD) risk scores identified two voice features that were independently associated with CAD (odds ratio OD = 0.37; 95% CI, interquartile range IQR = 0.18–0.79; OD = 4.01; 95% CI, IQR = 1.25–12.84; p-value = 0.009 and p-value = 0.02, respectively). Both features were more strongly associated with CAD when patients were asked to describe an emotionally significant experience.

3.2. COVID-19

Acute respiratory disease caused by the SARS-CoV-2 locates and multiplies mainly in the cytoplasm of lung cells. Thanks to the techniques enabling precise voice analysis combined with artificial intelligence (AI), it is possible to effectively and, above all early, diagnose COVID-19. Diagnostic scenarios can be conducted based on voice samples sent over a telephone.

Han *et al.* (2020) focused on developing some potential use-cases of intelligent speech analysis for COVID-19 diagnosed patients. By analysing speech recordings they constructed audio-only-based models to automatically categorise the health state of patients from four aspects: severity of illness, sleep quality, fatigue, and anxiety.

The prominent symptoms of COVID-19 include cough and breathing difficulties. Sharma *et al.* (2020) claim that respiratory sounds (cough, breath, and voice) can provide useful insights, enabling the design of a diagnostic tool. In their research, they determined 9 sound categories describing the voice, breath and cough. The acoustic analysis included the spectral analysis, energy description and zero-crossing rate. The accuracy on test data was 67%.

In order to better evaluate the COVID-19 infection, Wei *et al.* (2020) proposed an end-to-end method for cough detection and classification. It is based on real human-robot conversation data, which processes speech signals to detect cough and classifies it if detected. They find that the weighted sum can generate a 76% top-1 accuracy.

Pinkas *et al.* (2020) studied the harnessed deep machine learning and speech processing to detect the SARS-CoV-2 positives. Their dataset of cellular phone recordings included vocal utterances, speech, and coughs that were self-recorded by the subjects in either hospitals or isolation sites. They achieved the following diagnostic efficiency: a recall of 78% and a probability of false alarm (PFA) of 41%.

The papers of (Lechien *et al.*, 2020; Stasak *et al.*, 2021) presented voice analysis to classify the severity of COVID-19, from mild to moderate. They have reported the severity of COVID-19 might have influenced abnormally high rates of vocal dysphonia likely due to glottic (e.g., vocal folds) edema and tissue inflammation. In the study, the scientists used glottal, prosodic and spectral acoustic features from short-duration speech segments and applied them to machine learning algorithms. Experimental results indicate that certain feature-task combinations can produce COVID-19 classification accuracy of up to 80% as compared with using the all-acoustic feature baseline (68%).

Despotovic *et al.* (2021) presents the experiments with cough patterns using standard acoustic features sets, wavelet scattering features and deep audio embeddings extracted from low-level feature representations. The models achieve accuracy of 89% confirming the applicability of audio signatures to identify the COVID-19 symptoms.

The authors of (Hassan *et al.*, 2020) applied six speech features from a collected dataset and deep neural network (DNN) to create system for COVID-19 detection. The results show the classification accuracy for breathing sound reaching up to 98%, for cough sounds an accuracy of 97% was attained, while the voice accuracy of the system was only 88%. Their analysis shows that in the first place collecting cough and breathing sounds should make a COVID-19 detection system.

Subirana *et al.* (2020) showed that AI transfer learning algorithms trained on cough phone recordings results in diagnostic tests for COVID-19. They suggest a novel open collective approach to large-scale real-time health care AI. They evaluated the performance of four shallow machine learning classification algorithms: SVM, K-nearest neighbors, random forest, logistic regression. The presented graphs show that, depending on the methods used, the accuracy of COVID-19 diagnostics ranged from less than 80% (KNN and DenseNet201) to almost 100% (logistic regression and DenseNet201).

Laguarta *et al.* (2020) noticed that COVID-19 subjects, especially including asymptomatic, could be accurately discriminated from a forced-cough cell phone recording using AI. When validated with subjects diagnosed using an official test, the model achieved COVID-19 sensitivity of 99% with a specificity of 94%.

Deshpande and Schuller (2020) summarised efforts taken by the research community towards helping the individuals and the society in the fight against COVID-19 using speech signal processing.

3.3. Obesity and metabolic syndrome

Obesity is a growing health problem in many parts of the world. Excessive body fat is associated with multiple disorders such as diabetes, heart disease, hypertension, and stroke. Obesity itself is characterized by chronic low grade inflammation with permanently increased oxidative stress (Kopp, 2019). Given the potential influence of obesity (body mass index (BMI) of 30 or above) on the size and configuration of upper airway structures, it follows that other structures involved in voice production may be affected by body mass (Kopp, 2019).

The effect of obesity on voice changes is scarcely analyzed in the literature. The results of the conducted research suggest that there is a link between vocal tract morphology and obesity. Research requires vast databases and selecting patients with specific comorbidities. In this case, more effort to obtain data must be made. The summary of databases used so far in the literature is shown in Table 3.

Solomon *et al.* (2011) conducted a longitudinal analysis over a period of six months on eight obese and eight non-obese adults who underwent bariatric surgical procedures. No significant differences were detected between the groups during the preoperative assessment for acoustic parameters, maximum phonation time, laryngeal airway resistance and airflow during a sustained vowel. The only minor differences were detected for strain, pitch and loudness perception of voice over time, but not between groups. Phonation threshold pressure (PTP), at comfortable and high pitches (30% and 80% of the F0 range), changed significantly over time, but not between groups. Analysis of individual data revealed a trend for PTP at 30% F0 to decrease as BMI decreased.

da Cunha *et al.* (2011) posited that obese individuals' voices are more aperiodic than non-obese individuals' voices, as jitter and shimmer were increased and harmonic-to-noise ratio was decreased in the former group in their study.

Table 3. Comparison of speech databases used for obesity analysis
(v – sustained vowel; r – read speech; m – monologue; h – dialog with human; a – dialog with virtual agent).

Published	Speaker language	Number of patients	Number of HC	Age	Clinical evaluation	Voice evaluation	Type of speech					Duration per speaker [h:min:s]
							v	r	m	h	a	
Solomon *et al.*, 2011	English	8	8	53.1	laryngeal imaging	severity, roughness, breathiness, strain, pitch, loudness	1	1	0	0	0	NI
Celebi *et al.*, 2013	Turkish	20	20	38.8	BMI, laryngoscopic en	VHI, GRBAS	1	1	0	0	0	NI
de Souza *et al.*, 2014	N/A	44	30	42.45	–	vocal complaint	1	0	0	0	0	00:00:10
Hamdan *et al.*, 2014	English	9	0	35.56	BMI, laryngeal examination	simplified GRBAS	1	1	0	0	0	NI
Barsties *et al.*, 2013	German	22	7	21.4	BMI, body fat volume	roughness, hoarseness, breathiness, AVQI	1	1	0	0	0	NI
de Souza, Santos, 2018	N/A	42	42	26.83	BMI		1	0	0	0	0	NI

N/A – not applicable; NI – no information.

Hamdan *et al.* (2014) investigated 15 subjects undergoing bariatric surgery. They also found no significant difference in any of the acoustic features or in the laryngeal findings before and after surgery.

In the study conducted by Celebi *et al.* (2013), 20 obese and 20 non-obese volunteers underwent voice evaluation by laryngoscopy, acoustic analysis, aerodynamic measurement and perceptual analysis, using the grade-roughness-breathiness-asthenia-strain (GRBAS) scale and the 10 scales voice handicap index (VHI). No differences were found in acoustic analysis parameters between the two groups ($P > 0.05$). Maximum phonation time in the obese group (mean ± standard deviation, 19.6 ± 4.9 seconds) was significantly shorter than in the control group (26.4 ± 4.1 seconds) ($P < 0.001$), although the S/Z ratio was similar between the two groups.

de Souza *et al.* (2014) verified the presence of vocal complaints and a correlation between the auditory-perceptual analysis of voice and vocal self-assessment in a group of women with morbid obesity before and after bariatric surgery. There were no statistically significant differences regarding the mean fundamental frequency of the voice in both groups; however, there was a significant difference between the two groups regarding maximum phonation.

Barsties *et al.* (2013) analyzed the impact of body mass and body fat volume on selected parameters of vocal quality, a phonatory range and aerodynamics in women. Significant differences between three weight groups were found across several measures of intensity: vital capacity (VC), maximum phonation time (MPT), and shimmer. As compared to other groups, significantly higher values of maximum and minimum intensity levels, as well as sound pressure level during habitual running speech, were observed for the obese group. In contrast, the underweight group had significantly lower values for VC and the ratio of expected to measured VC. Furthermore, underweight subjects differed significantly as compared to normal weight subjects with lower MPT and the higher lowest F0. Finally, the obese group showed significantly lower shimmer values than normal-range weight subjects.

de Souza and Santos (2018) investigated the relationship between BMI and average acoustic voice features. The subjects were grouped according to BMI: 19 underweight, 23 in the normal range, 20 overweight, and 22 obese. Regarding the average F0, there was a statistically significant difference between underweight and overweight and obese groups, and the normal range and overweight and obese groups. The average MPT revealed a statistically significant difference between underweight and obese, the normal range and obese, and overweight and obese individuals. Obese women showed lower MPT.

3.4. PCOS

Polycystic ovary syndrome (PCOS) is the most common cause of hyperandrogenism in women of reproductive age, with the prevalence of 10–15%. The main characteristics of this endocrinopathy are menstrual disorders, clinical and/or laboratory hyperandrogenism and polycystic ovary morphology on ultrasonography. The elevated serum concentration of testosterone can account for symptoms such as hirsutism, acne and androgenic alopecia, as well as a deep, low voice. Moreover, other common conditions in women with PCOS are insulin resistance, disturbances of glucose metabolism, and dyslipidemia. Recent studies have shown that insulin resistance is associated with poorer verbal fluency in women (Ekblad *et al.* 2015; Sirmans, Pate, 2014). The significant delay in diagnosing this endocrinopathy still remains a worldwide issue. The use of speech analysis, as an easily accessible, a convenient screening test, could possibly expedite establishing a proper diagnosis and, in consequence, help provide the women with a proper treatment and an early monitoring for metabolic complications of PCOS. The speech databases and methods applied for classification to analyze PCOS diseases by speech are shown in Tables 5 and 6.

Hannoun *et al.* (2011) found that there was no statistically significant difference in the acoustic parameters except for an increase in the relative average perturbation ($P < 0.035$) and a decrease in the maximum phonation time ($P < 0.001$) in patients with PCOS.

Table 4. Comparison of research methods used for obesity analysis
(pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Disease	Tool used for features extraction	Features					Classifier/Test
			pr	sp	vs	li	naf	
Solomon *et al.* (2011)	obesity	MDVP	0	0	1	0	7	ANOVA
Barsties *et al.* (2013)		Voice Profiler 4.2, Speech Tool, Praat	0	1	1	0	20	Mann-Whitney U-test
Celebi *et al.* (2013)		Praat	0	0	1	0	7	Mann–Whitney U
de Souza *et al.* (2014)		ANAGRAF	0	0	1	0	4	–
Hamdan *et al.* (2014)		Visi-Pitch IV	0	0	1	0	7	Wilcoxon
de Souza, Santos (2018)		Praat	0	0	1	0	2	Mann-Whitne

Table 5. Comparison of speech databases in diabetes, thyroid and PCOS diseases
(v – sustained vowel; r – read speech; m – monologue; h – dialog with human; a – dialog with virtual agent).

Published	Speaker language	Number of patients	Number of HC	Age	Clinical evaluation	Voice evaluation	Type of speech					Duration per speaker [h:min:s]
							v	r	m	h	a	
Endocrine: diabetes												
Chitkara, Sharma, 2016	N/A	NI	NI	40–60	–	GRBAS	1	0	0	0	0	00:00:04
Pinyopodjanard *et al.*, 2019	N/A	83	70	54	HbA1c, FBS	–	1	0	0	0	0	00:00:05
Hamdan *et al.*, 2012	NI	82	29	52.83	HbA1c, FBS	GRBAS	1	1	0	0	0	NI
Endocrine: thyroid												
Dassie-Leite *et al.*, 2018	NI	100	100	3–12	altered T4, FT4, TSH	PVRQoL	1	1	0	0	0	NI
Mohammadzadeh *et al.*, 2011	N/A	120	88	35.9	T4, serum TSH	GRBAS	1	NI	NI	NI	NI	NI
Junozović-Žunić *et al.*, 2019	N/A	47	0	45	–	GRBAS	1	1	0	0	0	00:00:02
Endocrine: PCOS												
Huang *et al.*, 2015	English	48	0	41–62	free T concentrations	VHI	1	1	0	0	0	NI
Hannoun *et al.*, 2011	N/A	17	21	26	testosterone level	–	1	0	0	0	0	00:00:02
Gugatschka *et al.*, 2013	German	24	10	29	endocrinologic	VHI	1	1	0	0	0	NI
Aydin *et al.*, 2016	N/A	30	22	23.8	endocrinologic, laryngeal	VHI, GFI, RSI	1	0	0	0	0	NI

N/A – not applicable; NI – no information.

Table 6. Comparison of research methods in endocrine diseases
(pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Disease	Tool used for features extraction	Features					Classifier/Test
			pr	sp	vs	li	naf	
Dassie-Leite *et al.*, 2018	thyroid	VOXMETRIA	0	0	1	0	4	Student's t-test, Mann-Whitney
Hamdan *et al.*, 2012	diabetes	VISI-PITCH IV	0	0	1	0	7	Wilcoxon Mann–Whitney rank sum, Pearson's Chi-square
Chitkara, Sharma, 2016		MDVP, CSL	0	0	1	0	22	–
Pinyopodjanard *et al.*, 2019		MDVP, CSL	0	0	1	0	7	Logistic regression
Mohammadzadeh *et al.*, 2011	thyroid	Visipitch III, MDVP	0	0	1	0	13	Student t-test, chi-square, Mann-Whitney
Junozović-Žunić *et al.*, 2019		Speech Training for Windows, Dr. Speech, EZ Voice Plus	0	0	1	0	4	Paired-samples t-test
Huang *et al.*, 2015	hysterectomy	CSL	0	0	1	0	5	Linear regression
Hannoun *et al.*, 2011	PCOS	VISI Pitch (Model 3300)	0	0	1	0	6	Chi-square, Mann-Whitney tests
Gugatschka *et al.*, 2013		MDVP	0	0	1	0	9	Student t-test
Aydin *et al.*, 2016		Dr. Speech	0	0	1	0	8	Pearson chi-square, Fisher's exact test, Student t-test, Mann-Whitney U

Gugatschka *et al.* (2013) observed a trend towards a lower mean fundamental frequency, although it was not statistically significant. Elevated serum levels of androgens, as found in women with PCOS, were shown not to have an impact on the subjective and objective voice parameters.

Aydin *et al.* (2016) claimed that abnormal muscle tension patterns and impaired vocal fold vibration are common in patients with PCOS, although they are not accompanied by increased vocal symptoms or deteriorated acoustic voice parameters.

Huang *et al.* (2015) showed that testosterone administration in women with low T levels over 24 weeks was associated with dose- and concentration-dependent decreases in average pitch in the higher dose groups. These changes were seen in spite of an absence of self-reported changes in voice. The participants were healthy women, 41–62 years of age, who had undergone hysterectomy with or without partial or total oophorectomy.

3.5. Diabetes

Diabetes mellitus (DM) is a group of metabolic diseases characterized by chronic hyperglycemia which is a result of defective secretion and/or action of insulin. Type 2 DM is the most prevalent type of diabetes and concerns about 90% of diabetic patients worldwide. In the pathogenesis of Type 2 DM both mechanisms – impaired insulin action (insulin resistance) and impaired insulin secretion – play a role. Chronic hyperglycemia leads to development of diabetic complications and affects among others neurological, vascular, and muscular systems, all of which are essential components of the phonatory apparatus (Hamdan *et al.*, 2012) and methods applied for classification used by the researchers to analyze diabetes diseases are shown in Tables 5 and 6, respectively.

Hamdan *et al.* (2012) measured fundamental frequency, shimmer, relative average perturbation, harmonic-to-noise ratio and voice turbulence index, and reported no significant differences in any of the acoustic variables between diabetic patients and CH. There was no significant difference in the mean score of any of the perceptual evaluation parameters between diabetic patients and CH, despite the fact that mean scores were all higher in the diabetic group except for roughness. Patients with type 2 DM and poor glycemic control or neuropathy showed a significant difference in the grade GRBAS classification of their voice compared to CH.

In research conducted by Chitkara and Sharma (2016), the goal was to distinguish between vocal characteristics of patients with type 2 DM and control group. All the voice parameters that were investigated (jitter, shimmer, smoothed amplitude perturbation quotient, noise to harmonic ratio, relative average perturbation, amplitude perturbation quotient) show a significant difference in their values for the diabetic group versus CH.

Pinyopodjanard *et al.* (2019) found that F0 in female diabetic patients was significantly lower than controls (222.23 ± 27.89 Hz versus 241.08 ± 28.21 Hz, $P < 0.01$). In female diabetic subgroups with disease duration of over 10 years, poor glycemic control or neuropathy, F0 remained significantly lower. Multivariate analysis showed that F0 was significantly associated with diabetes after controlling for age, BMI, presence of hypertension, and dyslipidemia. However, F0 was not able to predict the presence of diabetes as shown by the logistic regression analysis ($P = 0.243$).

3.6. Hypothyroidism and hyperthyroidism

Hypothyroidism is the state of insufficient hormone production by the thyroid gland. Commonly reported symptoms in patients with this condition are hoarseness, deep or weak voice, vocal fatigue and tension while speaking as a result of vagus nerve edema, laryngeal muscle weakness and vocal cord paresis caused by an enlarged thyroid gland. Hyperthyroidism, as a state of increased thyroid hormone secretion, can also significantly reduce voice intensity and deepen its timbre. Hoarseness, roughness and trembling voice are also observed (Junozović-Žunić *et al.*, 2019). The research databases used in literature are shown in Table 5 and methods for further speech analysis are shown in Table 6.

Mohammadzadeh *et al.* (2011) found that F0, voice turbulence index and soft phonation index were significantly different from control values. There was positive correlation between thyroid-stimulating hormone (TSH) concentration and variation in F0 and prevalence of voice disorders.

Dessie-Leite *et al.* (2018) led an observational, analytical, cross-sectional study including 200 prepubertal children, of whom 100 had congenital hypothyroidism. The following parameters were evaluated: 1) history (identification, complaints, and interfering variables); 2) auditory-perceptual and acoustic evaluation; 3) self-assessment scores in the Pediatric Voice-Related Quality-of-Life (PVRQoL) survey; 4) laryngological evaluation; 5) medical records (congenital hypothyroidism etiology, age at treatment initiation, disease severity at diagnosis, treatment quality, and thyroid function tests on the day of the examination). Both groups had mean/median acoustic measurements within normal limits. There was no association between voice/larynx characteristics and endocrinological data.

Junozović-Žunić *et al.* (2019) reported that patients with hypothyroidism displayed significant differences in amplitude perturbation, jitter and noise-to-harmonics ratio between pre-treatment and post-treatment periods. In the group of patients with hyperthyroidism, significant differences were noted in the aerodynamic parameter maximum phonation time only. There were significant differences in all perceptual parameters in both groups of patients ($P < 0.05$) in pre- and post-treatment, except in the grade and asthenia parameter in the group of patients with hypothyroidism. The parameter grade was a border line insignificant in the group of patients with hyperthyroidism.

3.7. Mental and neurodegenerative disorders

There is a high volume of publications on the diagnosis of neurodegenerative and mental disorders using speech analysis. Most of the works concern the acoustic analysis of speech, but there are publications informing about the high effectiveness of linguistic analysis, an example is (Stasak *et al.*, 2017).

The application described in (Kiss *et al.*, 2021) is capable of estimating the probability of three types of voice disorders in English and Hungarian: depression, dysphonia, and Parkinson's disease.

Villatoro-Tello *et al.* (2021) used the available lexicon for a mentally ill and control subjects in a classification process to detect depression and dementia.

3.7.1. Schizophrenia

Schizophrenia is a chronic psychiatric disorder that affects 1% of the world's adult population. Language, thought and communication dysfunction characterize all its symptoms. They can be broadly divided into two groups: positive and negative. Positive thought disorder leads to a discourse that is difficult to understand (derailment, contact, neologisms, etc.). The research databases and conducted research are summarized in Tables 7 and 8, respectively.

Mota *et al.* (2012) compared patients with schizophrenia and mania. Global speech graph measures were not significantly different for the groups; however, patients with schizophrenia produced significantly fewer words per report than patients with mania. The authors observed poor speech, logorrhea and flight of thoughts.

Bedi *et al.* (2015) reported a proof-of-principle study which aimed to test automated speech analysis combined with machine learning to predict later psychosis onset in youths at clinical high-risk (CHR) for psychosis. Thirty-four CHR youths had baseline interviews and were assessed quarterly for up to two and a half years; five transitioned to psychosis. Speech features included a latent semantic analysis (LSA) measure of semantic coherence and two syntactic markers of speech complexity: maximum phrase length and the use of determiners (e.g., which). These speech features predicted later psychosis development with 100% accuracy, outperforming classification from clinical interviews. Speech features were significantly correlated with prodromal symptoms.

Gosztolya *et al.* (2018) matched 10 subjects with schizophrenia and eight HC with age and gender. The speakers performed spontaneous speech about their previous day for 5 min. The automatic speech recognition (ASR) system was trained to recognize the phones

Table 7. Comparison of speech databases used for schizophrenia analysis
(v – sustained vowel; r – read speech; m – monologue; h – dialog with human; a – dialog with virtual agent).

Published	Speaker language	Number of patients	Number of HC	Age	Clinical evaluation	Voice evaluation	Type of speech					Duration per speaker [h:min:s]
							v	r	m	h	a	
Mota *et al.*, 2012	Spanish	16	8	35.75	DSM-IV, BPRS, PANSS		0	0	0	1	0	00:20:00 –00:40:00
Bedi *et al.*, 2015	English	34	0	14–27	SIPS/SOPS	–	0	0	0	1	0	01:00:00
Gosztolya *et al.*, 2018	Hungarian	10	8	39.9	MMSE	–	0	0	1	0	0	NI

NI – no information.

Table 8. Comparison of research methods in schizophrenia examination
(pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Disease	Tool used for features extraction	Features					Classifier/Test
			pr	sp	vs	li	naf	
Mental								
Mota *et al.*, 2012	schizophrenia	Network Analysis Toolkit, MATLAB	0	0	0	1	11	NB, RBF, MLP, SVM, DT
Bedi *et al.*, 2015		Natural Language Toolkit (NLTK), POS-Tag Pen Tree Bank	1	0	0	1	7	Convexhull classifier
Gosztolya *et al.*, 2018		ASR with DNN,	1	0	0	0	8	SVM

in the utterances. For acoustic modeling, a standard deep neural network (DNN) with feed-forward topology was applied. A detailed examination revealed that, among the pause-related temporal parameters, those which took into account both the silent and filled pauses were the most useful in distinguishing the two speaker groups.

In studies of links between voice and schizophrenia, more research has been conducted in the area of voice perception (e.g., Pinheiro, Nizinkiewicz, 2019). Their findings support the hypothesis that higher-order operations reflected in amplitude modulations are abnormal in schizophrenia in a valence-dependent manner. The altered detection of vocal changes with a positive quality may lead to deficits in the comprehension of emotional states and intentions of social partners during vocal communication.

3.7.2. Depression and bipolar disorder

Reduced speech activity in patients with depression, especially with psychomotor impairment, is confirmed by many systematic studies. A number of clinical observations suggest that changes in voice features, such as pitch, may be important measures in diagnosing the early-stages of depression. They can also assess the progress of treatment for a depressive episode. Speech pause times (SPT), a silent interval between phonations during automatic speech, can be useful as an objective pathophysiological marker in depression. In clinical remission, depressive patients had comparable SPT values to the CH group. Studies show that in bipolar disorder, increased speech activity can predict a switch to hypomania. Depression is characterized by psychomotor retardation; in speech,

Table 9. Comparison of speech databases in bipolar and depression disorder
(v – sustained vowel; r – read speech; m – monologue; h – dialog with human; a – dialog with virtual agent).

Published	Speaker language	Number of patients	Number of HC	Age	Clinical evaluation	Voice evaluation	Type of speech					Duration per speaker [h:min:s]
							v	r	m	h	a	
Mental: bipolar												
Grünerbl *et al.*, 2014	German	NI	NI	18–65	HAMD, YMRS	–	0	0	0	1	0	NI
Guidi *et al.*, 2015	NI	11	18	NI	QID, YMRS	–	0	1	0	0	0	NI
Faurholt-Jepsen *et al.*, 2016	Danish	28	0	30.3 ± 9.3	HAMD, YMRS	–	0	0	0	1	0	NI
Alghowinem *et al.*, 2016	English	30	30	21–75	DSM-IV	–	0	0	0	1	0	00:08:33
Yang *et al.*, 2012	English	57	0	39.7(19–65)	HRSD	–	0	0	0	1	0	NI
Mundt *et al.*, 2012	English	165	0	37.8 ± 12.5	HAM-D, DSM-IV, QIDS-C, QIDS-SR	–	1	1	1	0	0	NI
Valstar *et al.*, 2013	German	NI	NI	31.5 ± 12.3	BDI-II	Valence, Arousal	1	1	1	0	1	00:25:00
Hönig *et al.*, 2014	German	219	0	31.5 ± 12.9	BDI	–	0	1	1	0	0	00:08:08
DeVault *et al.*, 2014	English	351	0	45.6 ± 12.2	–	–	0	0	0	1	1	00:17:30
Afshan *et al.*, 2018	Mandarin	735	953	NI	DSM-IV	–	0	0	0	1	0	00:01:52
McGinnis *et al.*, 2019	English	71	0	3–8	TSST-C, K-SADS-PL	–	0	0	1	0	0	00:03:00
Scherer *et al.*, 2013b	English	43	0	41.2 ± 11.6	PHQ-9, PCL-C	–	0	0	0	0	1	00:60:00
Gratch *et al.*, 2014	English	110	NI	18–65	PCL-C, PHQ-9, STAI-T, BIDR, BFI, RME, PANAS	–	0	0	0	1	0	00:38:11

NI – no information.

this shows up in reduction of pitch (variation, range), loudness, tempo and in voice qualities different from those of typical modal speech (Mundt *et al.*, 2012; Faurholt-Jepsen *et al.*, 2016). The comparison of speech databases and conducted research summary are presented in Tables 9, 10, and 11.

One of the most popular corpora used for depression research is the Mundt database. Thirty five physician-referred patients beginning treatment for depression were assessed weekly, using standard depression severity measures during a six-week observational study. Speech samples were also obtained over the telephone each week using an interactive voice response (IVR) system to automate data collection (Faurholt-Jepsen *et al.*, 2016).

Helfer *et al.* (2013) measured articulatory precision manifested through formant frequency tracking. GMM and SVM were applied using the Mundt database. They showed that a depression state can be classified with only formant frequencies and their dynamics given by the velocity and acceleration.

Cummins *et al.* (2015a) also used the Mundt corpus and the GMM model, although they focused on the spectral and energy-based properties of speech. They stated that depression-induced changes are in the laryngeal coordination and the vocal tract behavior. They show that depression is associated with a decreases in the pitch variability, changes in formant frequencies and decreases in the sub-band energy variability.

Hönig *et al.* (2014) observed a similar speech disorder in depression and sleepiness. They employed a small group of acoustic features, modeling prosody and spectrum, enriched with voice quality. The dataset comprises 1122 recordings from 219 German subjects (66 male); mean age 31.5 years, total duration of all files 29.7h. The database consists of read and spontaneous speech. The length of the speech tasks is between 5.8sec and 5.3min (mean 1.6min).

Scherer *et al.* (2013a; 2013b) examined another group of features – voice quality related, especially in the breathy-tense dimension – and used SVM as a classifier. The authors introduced a new dataset: virtual human distress assessment interview corpus with patients with depression and PTSD.

Experiments presented by Cummins *et al.* (2015b) support the hypothesis that reduction in acoustic variation of speech is produced under more severe depression. Their dataset comprises the Mundt corpus and AVEC2013. MFCC features were extracted. Universal

Table 10. Comparison of research methods and results in mental disorders
(pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Database	Tool used for features extraction	Features					Classifier/Test	Evaluation metrics	Best score	Tag
			pr	sp	vs	li	naf				
Mental: depression											
Alghowinem *et al.*, 2012	Black Dog	openSMILE	1	1	1	0	NI	HMM, GMM	VAR, F1	VAR = 71%	2
Mundt *et al.*, 2012	Mundt 2	Praat	1	1	1	0	12	logistic regression, t-test	Log Odds Response, *p*-value	–	1
Cummins *et al.*, 2013	Mundt 1, Black Dog	openSMILE	0	1	0	0	39	GMM-UBM	Acc	~65%	2
Scherer *et al.*, 2013a	VHDAIC	NI	1	1	1	0	4	SVM	Acc, F1	75%	2
Valstar *et al.*, 2013	AViD	openSMILE	1	1	1	0	2268	CVR	RMSE, MAE	10.35, 14.12	2
Hönig *et al.*, 2014	AVEC2014	openSMILE	1	1	1	0	3805	RLR	spearman's ρ, pearson's r	−0.46	1
Bozkurt *et al.*, 2014	Mundt 1	openSMILE, Praat	0	1	0	0	2860	SVM	UAR	69.48%	2
Cummins *et al.*, 2015c	Mundt, AVEC2013	openSMILE	0	1	1	0	2268	RVM	RMSE	10.89	2
Zhao *et al.*, 2020	AVEC2013, AVEC2014	openSMILE	0	1	1	0		DCNN	RMSE, MAE	9.57 7.9	2
Seneviratne *et al.*, 2020	Mundt	Aperiodicity, Periodicity, Pitch (APP) detector	0	1	1	0	20	SVM	Acc	81.77%	2
Alghowinem *et al.*, 2016	Black Dog, Pitt, AVEC	openSMILE	1	1	1	0	504	SVM	AR	96.90%	2

NI – no information.

Table 11. Comparison of research methods and results in mental disorders
(pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Database	Tool used for features extraction	Features					Classifier/Test	Evaluation metrics	Best score	Tag
			pr	sp	vs	li	naf				
Stasak *et al.*, 2016	AVEC2014	openSMILE	1	1	1	0	2155	SVM, RVM	RMSE	11.20%	2
Shau, Espy-Wilson, 2016	Mundt 1	NI	0	0	1	0	NI	SVM	Acc	62–87%	2
Simantiraki *et al.*, 2017	AVEC2014	openSMILE	0	0	1	0	247	JAD	AUC	0.88	2
Stasak *et al.*, 2017	DAIC	COVAREP speech toolkit	0	1	0	1	74	DT	Acc	82%	2
Afshan *et al.*, 2018	CONVERGE	openSMILE, VoiceSauce	0	1	1	0	>6300	GMM + *i*-vectors	P, R, F1, A	F1 = 0.95	2
Al Hanai *et al.*, 2018	DAIC	NI	1	1	1	1	553	NN LSTM	MAE, RMSE	MAE = 4.97 RMSE = 6.27	2
Xezonaki *et al.*, 2020	DAIC, General Psychotherapy Corpus (GPC)	NI	1	1	1	1		SVM, hierarchical attention-based Network	UAR, F1-macro	UAR = 0.72 F1-macro = 0.69	2
McGinnis *et al.*, 2019	own	MATLAB	0	1	1	0	NI	LR, SVM	Acc, Sens, Spec	80%, 54%, 93%	2
Mental: bipolar disease											
Grünerbl *et al.*, 2014	own	openSMILE	1	1	1	1	17	NB	Acc, recall, precision	70%	2
Guidi *et al.*, 2015	own	–	1	0	1	0	2	non-parametric Friedman test for paired data	$P < 0.05$	–	1
Faurholt-Jepsen *et al.*, 2016	own	openSMILE	0	1	1	0	6552	RF	Acc, Sens, Spec	0.74, 0.97, 0.52	2
Guidi *et al.*, 2017	own	NI	1	0	1	0	12	Friedman's test for paired data, Mann–Whitney U-test	$P < 0.05$	–	1

NI – no information.

background models (UBMs) were trained with the expectation-maximization (EM) algorithm. Speaker-specific GMMs were formed using full adaptation, with five iterations of the maximum a posteriori (MAP) algorithm.

The goal of research conducted by Bozkurt *et al.* (2014) was to find speech features that can distinguish speaking patterns of individuals with a diagnosis of clinical depression on a speaker-independent basis. Speech parameters were obtained from spectral analysis. Experiments for two-class depressed vs. non-depressed subjects were performed using SVM classifiers implemented on free speech recordings. Features advanced recognition rates up to 69% of the arithmetic average of individual class accuracies.

Stasak *et al.* (2016) showed performance boosting for the depression classification by using speech affect ratings in combination with low level features vs using these descriptors alone. They also showed the importance of setting thresholds for data selection for a specific affect. An automatic emotion-rating system derived from GeMAPS may positively contribute to performance in combination with low level features.

Khorram *et al.* (2016) took an approach for depression detection in bipolar patients. They combined two systems. The first was patient-specific and used unlabeled personal calls along with assessment calls to develop a unified background model and *i*-vectors, respectively. The second system was cohort-general and based on rhythm features. Their results showed improvement from the baseline with the unweighted average recall increasing from 0.66 ± 0.11 to 0.73 ± 0.09 and the area under the receiver operating curve increasing from 0.69 ± 0.15 to 0.78 ± 0.12.

Huang *et al.* (2016) collected six speech responses from 30 subjects – 15 with unipolar depression (UD) and 15 with bipolar disorder – and used the hierarchical spectral clustering (HSC) algorithm to adapt the larger Multimedia Human-Machine Communication (MHMC) emotion database to the obtained mood database CHI-MEI. Experimental results show that the proposed method achieved 73%, improving the detection accuracy by 13% compared to the commonly used SVM-based classifiers.

Sahu and Espy-Wilson (2016) explored features which are important in detecting depression. They used short audio samples of sustained vowels (5–6 s) and longer (30 s – 2 min) of free speech from the Mundt database and explored breathiness, jitter, and shimmer-based features using the average magnitude difference function (AMDF) to quantify them. Using the AMDF based feature they got 62–87% frame-wise accuracy for 5 out of 6 speakers.

Alghowinem *et al.* (2016) investigated the feasibility of cross-cultural UD detection from z-score normalized prosody features. They used German and English datasets (AVEC, BlackDog, and Pitt, respectively). Authors showed that binary (depressed/mild or not depressed) classification trained on SVM performs very well on each individual dataset.

Lopez-Otero *et al.* (2017) presented a study on how speaker de-identification affects the performance of a depression-detection system based on speech transcriptions. For depression detection, it is necessary to know which words are related to depression.

Stasak *et al.* (2017) proposed a novel measure for quantifying articulation effort and demonstrated that depression classification can be achieved by selecting speech with the higher articulation effort, linguistic complexity, or word-based arousal/valence.

Simantiraki *et al.* (2017) tested the hypothesis that UD could be detected from glottal source signals by using discrete cosine transform (DCT) coefficients of phase distortion deviation (PDD) preceded by the feature selection using Just Add Data (JAD). The models (SVM, ridge logistic regression and random forest) were evaluated on read and spontaneous speech from the AVEC2014 dataset. The researchers concluded that the lack of harmonic clarity is mainly disruptive in the region above F1 where harmonic amplitudes are relatively low.

Afshan *et al.* (2018) stressed that depression can be characterized by prosodic abnormalities and/or articulatory and phonetic errors. They verified various aspects of speech signals: cepstral features, difference in harmonic amplitudes, formant amplitudes, and i-vector-based systems using MFCCs ad score-level fusion to combine the two systems.

3.7.3. Parkinson's disease

PD is the second-most prevalent neurodegenerative disease in the world. It is caused by a loss of dopaminergic neurons and it causes severe motor and cognitive dysfunctions. It is characterized by hypophonia (reduced voice volume) and dysphonia (breathiness, hoarseness or creakiness in the voice), typically preceding more generalized speech disorders. About 90% of PD patients develop speech impairments such as monopitch, monoloudness, imprecise articulation and other symptoms (Orozco-Arroyave *et al.*, 2014a; 2014b). The summary of speech databases and research methods are presented in Tables 12 and 13.

In comprehensive review Moro-Velazquez *et al.* (2021) identified the most common features and ma-

Table 12. Comparison of speech databases in PD analysis
(v – sustained vowel; r – read speech; m – monologue; h – dialog with human; a – dialog with virtual agent).

Published	Speaker language	Number of patients	Number of HC	Age	Clinical evaluation	Voice evaluation	Type of speech					Duration per speaker [h:min:s]
							v	r	m	h	a	
Pinto *et al.*, 2016	French, Portugese	139	65	65–70	NI	–	0	1	1	1	0	14:54:36
Skodda *et al.*, 2011	German	138	50	66.74(8.48)	UPDRS III, H&Y	–	1	1	0	0	0	NI
Rusz *et al.*, 2018	Czech	78	30	62.3(11.3)	MDS-UPDRS III	MDS-UPDRS III	1	0	1	0	0	NI
Orozco-Arroyave *et al.*, 2014a	Spanish	50	50	62.2(11.2)	UPDRS-III and H&Y	–	1	1	1	0	0	00:01:34
Pettorino *et al.*, 2017	Mandarin	13	12	62.1(52–72)	NI	–	0	1	0	0	0	NI
Zhan *et al.*, 2016	English	121	105	57.6(9.1)	UPDRS	–	1	0	0	0	0	NI
Hemmerling *et al.*, 2016	Polish	27	0	65(7.9)	UPDRS, H&Y	–	1	1	0	0	0	00:00:59
Anatolík, Fougeron, 2013	French	79	26	32–89	NI	intelligibility and articulatory imprecision	0	1	0	0	0	NI

NI – no information.

Table 13. Comparison of research methods and results in Parkinson's disease (pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Database	Tool used for features extraction	Features					Classifier/Test	Evaluation metrics	Best score	Tag
			pr	sp	vs	li	naf				
Orozco--Arroyave *et al.*, 2014b	German, Spanish, Czech	Praat	0	1	1	0	68	SVM	Acc Sens, Spec, AUC	Acc 97.6%	2
Mallela *et al.*, 2020	Indian	NI	0	1	0	0	NI	CNN + Bidirectional Long Short-Term Memory	Acc	97.41%	2
Pompili *et al.*, 2020	Portugese	HTK	1	1	1	0	52	DNN TensorFlow	Acc	95.27%	2
Vásquez--Correa *et al.*, 2015	PC-GITA, German, Czech	NI	0	1	0	0	NI	CNN	Acc	89%	2
Villa-Canas *et al.*, 2015	PC-GITA	NI	0	1	1	0	NI	GMM, GMM-UBM, SVM	Acc	77%	2
Pettorino *et al.*, 2017	Mandarin, Polish, Italian	NI	1	0	0	0	2	t-test	$P < 0.05$	–	1
Zhan *et al.*, 2016	own	NI	1	0	1	0	9	RF	Acc	71.0(± 0.4)%	3
Klumpp *et al.*, 2017	NI	NI	1	0	0	0	NI	–	SER	1.34	3
Rusz *et al.*, 2018	own	NI	1	0	1	0	NI	binary logistic regression	AUC	0.85	2
Wodzinski *et al.*, 2019	PC-GITA	NI	0	1	0	0	NI	LSTM	Acc	0.917	2

NI – no information.

chine learning techniques employed in automatically detecting and assessing the severity of PD using phonatory and articulatory aspects of speech and voice.

Orozco-Arroyave *et al.* (2014b) investigated the detection of PD by analyzing speech recordings in German, Spanish, and Czech. The Spanish database contains speech recordings of 50 PD patients (mean age 61), the set of German patients includes 88 individuals (mean age 66.5), and in the group of Czech native speakers 21 were diagnosed with idiopathic PD (mean age 62.2). For each language a CH group with the same number of individuals was selected. The databases contain recordings of sustained phonation of vowels, diadochokinetic (DDK) evaluation, sentences, and monologues. Three features in the phonation process were taken into account: harmonics-to-noise ratio, normalized noise energy, and glottal-to-noise excitation ratio. Additionally, the first and second formant and 12 MFCC were extracted. Authors suggest that it is possible to detect PD using the same method in different languages. Additionally, Orozco-Arroyave *et al.* (2014a) presented further tests using jitter and shimmer.

Villa-Canas *et al.* (2015) analyzed low-frequency components of speech signals by using three different time-frequency techniques. Their results showed that the changes in the low frequency components are able to discriminate between people with Parkinson's and healthy speakers with an accuracy of 77%, using a single sentence.

Zlotnik *et al.* (2015) considered four groups of features: phonation, articulation, prosody, and intelligibility. Many techniques were tested for predicting the stage of PD using the patient's voice exclusively. Finally, an ensemble of classifiers obtained the best results (0.609), combining the output of the best Random Forest, with intelligibility features.

Vásquez-Correa *et al.* (2015; 2017; 2018) found that voice impairments appear in about 90% of speech samples as reduced loudness, monopitch, monoloudness, reduced stress, breathy, hoarse voice quality, and imprecise articulation. Speech samples were subjected to frequency analysis: F0, its variability and MFCC (Vásquez-Correa *et al.*, 2015), wavelet transform and short-time Fourier transform (Vásquez-Correa *et al.*, 2017; 2018). Physical parameters of voice were used for modeling the SVM and the convolutional neural network (CNN). In the publication from 2015, they declared the highest accuracies with the voiced frames a range from 64% to 86%, while the results with unvoiced frames range from 78% to 99%. In the paper published in 2017, they mentioned accuracies of up to 89% for the classification of Parkinson's patients vs. control group when a convolutional neural network was used to extract features from the time-frequency representations. Their study published in 2018 proposes a multitask learning approach based on CNNs to assess at the same time eleven speech aspects, including difficulties of the patients to move articulators such as lips, palate, tongue, and larynx. The input to the CNNs are time-frequency representations obtained from transitions between voiced and unvoiced segments.

Hemmerling *et al.* (2016) explored phonatory and articulatory features for modulated vowels in PD detection. They used jitter, shimmer, F0, energy and MFCC statistical values along side instantaneous energy and its range coming from the Hilbert-Huang transformation. The data used contained sustained and modulated vowels. As the result, the sustained vowels covered higher binary accuracy classification.

Pettorino *et al.* (2017) investigated whether speech of PD patients presents rhythmic abnormalities. Twenty-five Mandarin speakers (13 PD and 12 HC matched on age) and thirty-one Polish speakers (18 PD and 13 HC matched on age) read aloud a passage of story. The vowel percentage and the interval between two consecutive vowel onset points were calculated. They segmented the recorded speech into vocalic and consonantal intervals, and then calculated the vocalic portion in the utterance and the duration of the interval between two consecutive vowels. The effectiveness of the rhythmic metric appears to be language-dependent. For Polish was distinctly higher while for Mandarin there was no significant difference. They concluded that the analyzed method could be used for automatic diagnosis of PD for Polish and Italian, but not for Mandarin.

Klumpp *et al.* (2017) introduced Apkinson – a smartphone application providing a mobile monitoring solution for PD patients. The severity and progression of PD can be tracked. The patient has to perform a speech exercise in which the person has to constantly produce syllables of subsequent consonant-vowel combinations. The Levenshtein distance evaluates the similarity by computing the required insertions, deletions and substitutions to transform one string to the other. For the recordings of the HC, the recognizer correctly counted the number of keywords in 76% of the cases. For the group of PD patients, the number of correctly assessed records was slightly lower with 72%.

The most common scale for assessing the severity of PD is the Unified Parkinson's Disease Rating Scale (UPDRS) and Movement Disorders Society UPDRS (MDS-UPDRS), which evaluates non-motor and motor experiences of daily living and motor complications.

The Computational Paralinguistics Challenge Special Session of Interspeech 2015 was dedicated, alongside two other topics, to the degree to which PD can be detected by speech analysis. Recordings of 50 PD patients were provided. The dataset was divided into 42 tasks per speaker, yielding 1470 recordings in the training set (3 speakers), 630 recordings in the development set (15 speakers), and 462 recordings (11 speakers) in the test set. The duration of recordings ranges from 0.24 seconds to 154 seconds and consist of a monologue, a read text and sentence recordings.

Grósz *et al.* (2015) applied two state-of-the-art machine learning methods in the regression Sub-Challenges of the Interspeech 2015 Computational Paralinguistics Challenge. They showed that both DNN and Gaussian process regression (GPR) are competitive with the baseline SVM, and the results can be improved by combining the classifiers. They trained DNNs with five hidden layers and 1000 neurons in each hidden layer. The best results (0.671) they obtained by using a feature selection and by averaging out the scores of multiple recordings clustered to the same person.

Sztahó *et al.* (2015) presented the method of linear regression models on a set of extracted acoustic features from the middle of vowels in words, sentences and continuous speech, and the partitioning of speech samples according to their total length into parts with long, medium and short duration. Jitter, shimmer, articulation rate, intensity and its variation, rate of transients and MFCC were extracted. They notice that in terms of the final results on the test set that was uploaded for the challenge, many conclusions cannnot be deduced due to the variation in the data. They emphasized that correlations (and also the baseline results) count as weak. They experienced high intra-variation of the extracted features.

3.7.4. Alzheimer's disease and dementia

Alzheimer's disease (AD) is a progressive neurodegenerative disorder clinically defined as an impairment of certain cognitive and functional abilities. As the result of the aging society, there are growing number of people affected by AD.

Gosztolya *et al.* (2019) took an automatic approach focusing on features describing the number of pauses in spontaneous speech, specifically filled gaps. The authors constructed a large set of descriptors and used correlation and the sequential forward selection algorithm to find the most promising ones. Based on only the acoustic features, they were able to separate the various groups with accuracy scores between 74–82%. They attained similar accuracy scores using only the linguistic features. With the combination of the two types of features, the accuracy scores between 80–86%.

Sadeghian *et al.* (2017) presented empirical evidence that AD patients can be reliably distinguished from HC through a combination of acoustic features from speech and linguistic features extracted from an automatically determined transcription of speech.

Wankerl *et al.* (2017) proposed a purely statistical approach towards automatic diagnosis of AD, solely based on n-gram models with subsequent evaluation of the perplexity. The system works independently in a concrete language. AD patients show emotional prosodic impairment.

The Pitt corpus used by Warnita *et al.* (2018) consists of speech samples and their transcriptions from 244 HC and 309 dementia patients. They used a gated convolutional neural network (GCNN) for speech data. The presented study is the non-linguistic approach for detecting AD by utilizing only the speech audio data. Since it does not utilize linguistic information, authors can apply it to low resource languages. The proposed method achieved the accuracy of 74%.

Weiner *et al.* (2016; 2018) investigated a way of automatic classification of AD using conversational speech. They derived 10 prosodic and textual features from 98 voice samples from the interdisciplinary longitudinal study on adult development and the aging (ILSE) dataset. They compared two pipelines of feature extraction for dementia detection: manual transcription and ASR using samples from the ILSE corpus. The acoustic and linguistic features were extracted and several models were built with the Gaussian classifier as the top performer. Weiner *et al.* (2018) stated that early detection of dementia is possible by automatically processing conversational speech. They tested a group of more than 200 subjects. Conversational speech (12 min of interview) was chosen since it is a natural form of communication that can be recorded without causing stress to subjects. The best results were obtained through a combination of acoustic and linguistic features. Finally, a Gaussian classifier was trained to discriminate three cognitive diagnoses. The authors declare that it is possible to detect dementia using speech of duration 2.5 min, although the most reliable results require between 10 and 15 min. In the publication of 2016, the authors declared the F-score of 0.8 for the detection of AD. In the paper from 2018, they stated detected dementia with an UAR of 0.64 using acoustic features extracted from speech segments.

Rohanian *et al.* (2021) present two multimodal fusion-based deep learning models that consume ASR transcribed speech and acoustic data simultaneously to classify whether a speaker in a structured diagnostic task has AD. They achieved an accuracy of 84% using words, word probabilities, disfluency features, pause information, and a variety of acoustic features.

The summary of speech databases and research methods applied for AD and dementia analysis are shown in Tables 14 and 15.

Table 14. Comparison of speech databases in Alzheimer's disease and dementia (v – sustained vowel; r – read speech; m – monologue; h – dialog with human; a – dialog with virtual agent).

Published	Speaker language	Number of patients	Number of HC	Age	Clinical evaluation	Voice evaluation	Type of speech					Duration per speaker [h:min:s]
							v	r	m	h	a	
Sattler *et al.*, 2017	German	NI	NI	>40	medical, psychological, cognitive, physical, dental examinations, semi-standardized biographic interview	–	0	0	0	1	0	NI
Ujiro *et al.*, 2018	Japanese	12	12	74.5 ± 4.3	DSM-IV-T	–	0	1	0	0	1	NI

NI – no information.

Table 15. Comparison of research methods and results in Alzheimer's and dementia disorder (pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Database	Tool used for features extraction	Features					Classifier/Test	Evaluation metrics	Best score	Tag
			pr	sp	vs	li	naf				
Warnita *et al.*, 2018	Pittcorpus	openSMILE	0	1	1	0	12459	GCNN	Acc	73.6%	2
Weiner *et al.*, 2018	ILSE	NI	1	1	0	0	40	GMM	UAR	0.645	2
Ujiro *et al.*, 2018	own	Snack Sound Toolkit	0	0	1	1	21	SVM and logistic	AUROC	0.95	2
Pan *et al.*, 2020	DementiaBank dataset	NI	1	0	1	1		bi-LSTM	F-score	78.34%	2
Mirheidari *et al.*, 2018	own	Praat	1	0	1	1	12	HMM-GMM	Acc	91%	2

NI – no information.

Table 16. Comparison of speech databases in ALS
(v – sustained vowel; r – read speech; m – monologue; h – dialog with human; a – dialog with virtual agent).

Published	Speaker language	Number of patients	Number of HC	Age	Clinical evaluation	Voice evaluation	Type of speech					Duration per speaker [h:min:s]
							v	r	m	h	a	
Horwitz-Martin *et al.*, 2016	English	34	0	NI	–	SIT	1	1	0	0	0	NI
Wang *et al.*, 2016	English	11	11	60	ALSFRS-R	SIT	1	1	0	0	0	NI

NI – no information.

Table 17. Comparison of research methods in ALS
(pr – prosodic; sp – spectral; vs – voice source; li – linguistic; naf – number of acoustic features).

Published	Disease	Tool used for features extraction	Features					Classifier/Test
			pr	sp	vs	li	naf	
Neurodegenerative								
Horwitz-Martin *et al.*, 2016	ALS	MATLAB	1	1	0	0	36	Spearman correlations, regression
Wang *et al.*, 2016		openSMILE	1	1	1	0	6373	RLR, *i*-vectors, SVM, DNN

3.7.5. Amyotrophic lateral sclerosis

ALS is a rapidly progressing disorder that causes the death of neurons controlling voluntary movements. Symptoms include difficulty in speaking or swallowing.

Antolík and Fougeron (2013) found that among consonant distortions, the most frequent type of distortion in ALS is an incomplete closure of stops.

Wang *et al.* (2016) explored the option to diagnose ALS from short speech acoustic and articulatory samples. They examined 11 affected patients and 11 HC, and constructed a large dimensionality dataset of acoustic features from audio samples and articulatory features derived from tongue and lip sensors. Furthermore, randomized logistic regression (RLR) was used as a feature selection method, and the *i*-vector was calculated for each speaker and concatenated to feature the vector for speaker normalization. Next, they trained two classes of models: SVM with a radial basis function (RBF) kernel, and DNN achieving maximum performance when all acoustic and sensor features were combined and provided to train the DNN model.

Horwitz-Martin *et al.* (2016) presented research into an objective and automatic assessment of speech loss with features extracted from the first and second formant. They found that acceleration features derived from F2 were the most informative for speech predicting the rate of speech decline and assessing the intelligibility decline.

The speech databases and research methods applied for ALS analysis are shown in Tables 16 and 17.

4. Discussion

4.1. Recruitment process

To create an effective computer system it is necessary to have the right number of recordings of individuals affected by each given disorder. Each patient must have an official diagnosis from a physician. Recordings taken at early disease stages are the most desirable since the main purpose of computer systems is to recognize the onset of the disease. Sometimes it is difficult to convince patients of the desirability of collecting recordings. Patients must consent to participate in research. Thus, the recruitment process is difficult and frequently time-consuming. Collecting recordings requires approval from the Ethics Committee.

4.2. Standardization of database descriptions

We commonly encountered insufficient descriptions of the collected speech databases. As shown in Tables 1, 3, 5, 7, 9, 12, 14, 16 only some of the parameters of recordings are reported systematically across the corpora. The descriptions frequently omit the duration of the recordings or detailed technical information on the conditions of the recording process. We propose to standardize the descriptions as a table or an annex, which would clearly outline the properties of the corpus, especially when the database is not publicly available. Important parameters assessing the usefulness of recording databases are the number of recorded speakers (Fig. 4), the duration of each speech and total recording times representing the size of the database speaker (Tables 1, 3, 5, 7, 9, 12, 14, 16). In addition, it may be useful to divide the recordings by gender. Generally, patients' age ranges are given, although some authors present the average age of patients and standard deviation. Similar information is presented regarding the age of CH. Comparing these data sets shows that frequently the patients are more advanced in age than the CH group. The correlation between the patient age and their health may affect the functionality of the classifiers. An older patient's voice may be incorrectly classified as a voice of an ill patient. Conversely, a young person's voice may be incorrectly classified as a voice of a healthy individual.

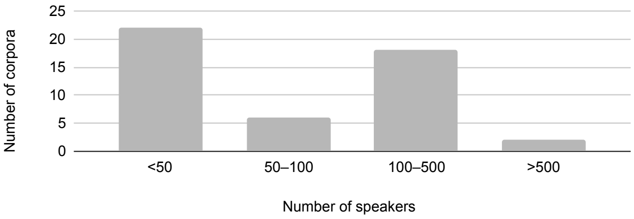

Fig. 4. Corporas from 2009–2019.

The recordings were divided into three groups. The first includes read texts or, far less frequently, individual words or chosen vowels. Such recordings can be used to analyze the same statements for all respondents, which is a valuable advantage when analyzing the acoustic features of voices. In this case, there is no dependency on the content of the statements. The second group includes recordings of spontaneous statements. These enable linguistic analysis, which is especially useful when diagnosing mental illness (Stasak *et al.*, 2017; Lopez-Otero *et al.*, 2017). The third group includes recordings from dialogues between subjects and interviewers (Weiner *et al.*, 2016). Conversational speech is the most natural form of communication that can be recorded without putting subjects under stress. Moreover, acoustic and language information can be obtained from conversations between patients and virtual agents. Generally, subjects receive instruction on how to speak. This applies to duration, avoiding emotions, monologue topics and sometimes loudness. For example, in sporadic cases, patients are asked to give a prolonged statement of the desired vowel.

The most frequently recorded statements are in one language only (sporadically dialect). The main reason is that such recordings are the easiest to obtain. Their effective analysis is also easier, especially linguistic analysis.

Recordings can be obtained using a variety of equipment, such as computers or mobile phones. Specialist recording equipment can be used to obtain a wider frequency range. However, such equipment is less widely available, and recordings must be taken at specific locations. Acoustic conditions for recordings using mobile phones are poorer than those in recording studios. If possible, the signal-to-noise ratio (SNR) parameter should be determined.

There is a general view that speech does not contain frequencies above 8 kHz. Determining the threshold frequency of recorded speech determines the sampling frequency of the digital signals. The size of sampling and the resolution (usually 16 bits per sample) determines the size of the acoustic files.

It is extremely important to label recordings with accurate information on the health of the subjects. Information about any medication taken by the patient is also useful, as it can have a significant impact on the features of the recorded speech.

4.3. Standardization of recording protocols

There is no consistent scenario for the content of recorded speech. Some corpora contain recordings of sustained vowels only, while others use read speech or interviews (Weiner *et al.*, 2018). Each form of speech has its own advantages: sustained vowels provide material which is standard in phoniatric investigations, read speech gives the same content for each speaker, while dialogues are usually the most natural.

4.4. Controlled clinical laryngeal examination

Only a few of the studies included phoniatric/ laryngeal imaging (i.e., laryngoscopic, laryngostroboscopic, videokymography, and high-speed digital imaging). In the absence of such evaluation, the condition of the vocal folds and surrounding structures cannot be determined accurately, and potential laryngeal disorders could interfere with the disorder being investigated and have an impact on acoustic parameters of the voice. Such findings would have been bolstered by including additional information regarding thew vocal fold structure and physiology. Future studies should include endoscopic methods to place these results in the diagnostic context, especially in case of somatic disorders. This recommendation is especially relevant to disorders for which the state-of-the-art is less advanced or where voice sample sizes are small. The issue is that such examinations are costly.

4.5. Exclusion criteria

There is a lack of consistency concerning excluding voice recording of certain subjects. In some cases, researchers only excluded recordings which did not follow the recording protocol; however, in most cases specific exclusion criteria were chosen, e.g., excluding subjects who were under the influence of alcohol of drugs, smokers (de Souza, Santos, 2018) or individuals whose jobs put an increased strain on their voice. In some cases exclusions were made on the basis of medical records (e.g., the disorder was too advanced). In general, the criteria should be standardized. Additionally, the excluded recordings can be valuable for case studies investigating factors influencing the human voice.

4.6. Selection criteria for the HC group

In the majority of studies, control groups were matched by age and gender only. In some cases, additional factors were also considered, e.g., BMI. In general, there is no standardized approach.

4.7. Single disease

While some of the disorders may be intercorrelated (e.g., obesity and CAD, depression, and diabetes), the reviewed articles rarely took into account more than one or two disorders at a time.

4.8. Sample size

The number of speakers in the corpora varies from around 10 to several thousand. If the number is low, their medical records are generally more detailed. In order to apply certain modern machine learning methods, e.g., DNN, large datasets are required. This could be achieved through crowdsourcing, which has been effective in other fields of speech technology.

4.9. Legal issues of voice recording

While state-of-the-art algorithms are able to identify individuals using their voice prints, new problems are emerging. When speech is recorded on smartphones (Dogan *et al.*, 2017), user data should be treated as sensitive and protected with appropriate privacy policy security procedures for transfer and storage.

4.10. Numbers of recordings

In most cases, just one recording is taken for each individual and there are no reference points from the onset of disease or from its progression. There is a need for more frequent monitoring of the patient voice status.

4.11. Availability of corpora

The majority of datasets were prepared for specific studies and were not available publicly to other researchers (Low *et al.*, 2020). This makes results less comparable and reproducible while new methods are being developed. However, several corpora have become international benchmark standards on which novel methods can be validated. This trend should also be adopted for other disorders.

4.12. Feature sets

Tables 2 and 4 clearly show that there are no standards in speech parameterization or, if they exist, that they are rarely used. Only a few results were obtained using standard feature vectors, such as those offered by openSMILE. The number of extracted features varies from several to several thousand features (e.g., openSMILE vectors such as AVEC or ComPare) (Fig. 5). In most cases, researchers choose one or two parameters from four categories (prosodic, spectral, voice source, linguistic), on the basis of their previous experience or their tools. Technically, there are no limitations on verifying features from all the categories, which could improve our understanding of the correlations of voice features with disorders (Sadeghian *et al.*, 2017). When novel parameterizations are introduced, their implementations are usually not made public.

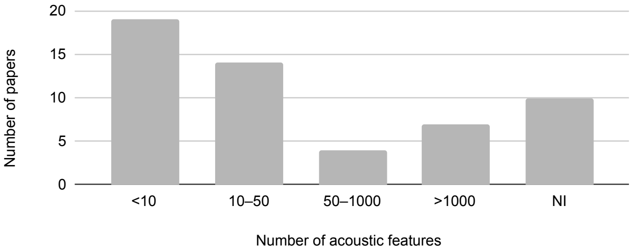

Fig. 5. Acoustic features investigated in the articles 2009–2019.

4.13. Lack of cross-cultural and cross-language comparisons

The majority of articles focus on a single corpus or language. Gathering recordings to create databases is time consuming, and requires a high level of cooperation between engineers and medical professionals from different specializations. It is necessary to identify patients who meet the criteria set out in the study (e.g., no specific comorbidities, addictions, etc.). Being able to share databases (including speech and voice signals) among researchers usually requires formal consent and the willingness to cooperate. In the literature, there are still few studies that take into account more than one language (Orozco-Arroyave *et al.*, 2014b), even though there is a need for systematic comparisons of findings between languages (Maor *et al.*, 2018; Vásquez-Correa *et al.*, 2017; Pettorino *et al.*, 2017). The importance of heterogeneous training sets in machine learning should also be emphasized. Acoustic features of speech are affected by the diagnosed disorder, as well as by the language spoken by the subject. In many cases the same algorithms can be used effectively regardless of the language, and only the acoustic feature values will change. The situation is significantly more complicated in linguistic testing of speech. Transferring research methods from one language to another may not be sufficiently effective (Warnita *et al.*, 2018).

5. Conclusions

Current disease diagnostic methods are adapted from those which have been verified as useful in speech or speaker recognition and investigating emotions. However, it should be remembered that speech analysis for medical purposes should follow different rules than speech recognition analysis. As a result of evolution, the human voice has adapted to the perceptual capabilities of the hearing system. This has greatly improved the efficiency of communication by speech. The human ear is a frequency analyzer with nonlinear characteristics. Imitating these properties in speech technology has been shown to be highly effective, e.g., MFCC. We suppose that in the case of speech analysis for medical purposes there are no indications to take the perceptual properties of the hearing system into account. Authors of the studies included in our paper an attempt to detect speech features which can be indicative of specific disorders. To achieve this, they mainly use acoustic analysis and sometimes linguistic analysis.

Determining the reasons (e.g., anatomical or neurological) for voice changes in specific disease states is a separate issue and not the subject of the cited publications. Characteristic features of speech result from the anatomical structure of the vocal tract and the way it is stimulated by the nervous system. Therefore, medical speech diagnostics are aimed at neurodegenerative and mental disorders as well as disorders affecting the physical structure of the vocal tract. The majority of publications refer to speech deviations caused by disorders of the nervous system (e.g., depression, dementia, Parkinson's and Alzheimer's diseases).

Both acoustic and linguistic properties of speech are taken into account. The methods of acoustic analysis are more useful and were tested for all the diseases presented above. Linguistic methods have been tested for mental and neurodegenerative diseases. The conducted experiments show that the combination of both methods improves the efficiency of diagnosis. The relatively large number of publications in this field is testimony to the influence of the nervous system on the generation of speech. This is particularly important for psychiatry, which usually lacks objective clinical measurements used in other specializations. While nervous system disorders result in both acoustic and linguistic features of speech, somatic disorders are diagnosed by analyzing acoustic parameters of the voice.

Speech analysis systems provide a promising approach for creating low-cost, non-invasive and remote diagnostic tools for automatic assessment or monitoring of certain disorders. Early and sensitive disease detection can support medical intervention and treatment. Speech corpora can be collected conveniently in a clinical environment or at home using smartphones.

Various methods of assessing the effectiveness of diagnosing diseases through acoustic voice analysis and linguistic speech analysis were used. This diversity makes it difficult to answer the question for which diseases the analyzed methods are by far the most effective and for which the least reliable. Effectiveness of disease diagnosis varies from 65% up to 99%. From a medical point of view, such results should be treated as a screening tests only and should be an indication of the need for standard medical tests. According to literature reports, the highest effectiveness was obtained for: COVID-19, schizophrenia, and Parkinson's disease. Worse results were obtained for depression, bipolar disorder and Alzheimer's disease. Relatively weaker results were obtained for PCOS, diabetes, hypothyroidism, hyperthyroidism, and amyotrophic lateral sclerosis. The weakest accuracies were obtained for the diagnosis of obesity and metabolic syndrome.

Acknowledgments

This paper was supported by the National Centre for Research and Development [grant VAMP on http://voicebiometrics.pl/ number POIR.04.01.04-00-0052/18].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Afshan A., Guo J., Park S.J., Ravi V., Flint J., Alwan A. (2018), Effectiveness of voice quality features in detecting depression, [in:] *Interspeech*, pp. 1676–1680, doi: 10.21437/Interspeech.2018-1399.
2. Al Hanai T., Ghassemi M.M., Glass J.R. (2018), Detecting depression with audio/text sequence modeling of interviews, [in:] *Interspeech*, pp. 1716–1720, doi: 10.21437/Interspeech.2018-2522.
3. Alghowinem S., Goecke R., Epps J., Wagner M., Cohn J.F. (2016), Cross-cultural depression recognition from vocal biomarkers, [in:] Interspeech, pp. 1943–1947, doi: 10.21437/Interspeech.2016-1339.
4. Alghowinem S., Goecke R., Wagner M., Epps J., Breakspear M., Parker G. (2012), From joyous to clinically depressed: Mood detection using spontaneous speech, [in:] *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, Youngblood G.M., McCarthy P.M. [Eds.], pp. 141–146, https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS12/paper/view/ 4478/4782.
5. Antolík T.K., Fougeron C. (2013), Consonant distortions in dysarthria due to Parkinson's disease, amyotrophic lateral sclerosis and cerebellar ataxia, [in:] *Interspeech*, pp. 2152–2156, doi: 10.21437/Interspeech.2013-509.
6. Aydin K. *et al.* (2016), Voice characteristics associated with polycystic ovary syndrome, *The Laryngoscope*, **126**(9): 2067–2072, doi: 10.1002/lary.25818.
7. Barsties B., Verfaillie R., Roy N., Maryn Y. (2013), Do body mass index and fat volume influence vocal quality, phonatory range, and aerodynamics in females?, *CoDAS*, **25**(4): 310–318, doi: 10.1590/s2317-17822013000400003.

8. Bedi G. *et al.* (2015), Automated analysis of free speech predicts psychosis onset in high-risk youths, *npj Schizophrenia*, **1**(1): 15030, doi: 10.1038/npjschz.2015.30.

9. Bozkurt E., Toledo-Ronen O., Sorin A., Hoory R. (2014), Exploring modulation spectrum features for speech-based depression level classification, [in:] *Interspeech*, doi: 10.21437/Interspeech.2014-312.

10. Celebi S. *et al.* (2013). Acoustic, perceptual and aerodynamic voice evaluationin an obese population, *The Journal of Laryngology and Otology*, **127**(10): 987–990, doi: 10.1017/s0022215113001916.

11. Chitkara D., Sharma R.K. (2016), Voice based detection of type 2 diabetes mellitus, [in:] *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 83–87, doi: 10.1109/AEEICB.2016.7538402.

12. Cummins N., Epps J., Sethu V., Breakspear M., Goecke R. (2013), Modeling spectral variability for the classification of depressed speech, [in:] *Interspeech*, pp. 857–861, doi: 10.21437/Interspeech.2013-242.

13. Cummins N., Scherer S., Krajewski J., Schnieder S., Epps J., Quatieri T.F. (2015a), A review of depression and suicide risk assessment using speech analysis, *Speech Communication*, **71**: 10–49, doi: 10.1016/j.specom.2015.03.004.

14. Cummins N., Sethu V., Epps J., Krajewski J. (2015b), Relevance vector machine for depression prediction, [in:] *Interspeech*, pp. 110–114, doi: 10.21437/Interspeech.2015-37.

15. Cummins N., Sethu V., Epps J., Schnieder S., Krajewski J. (2015c), Analysis of acoustic space variability in speech affected by depression, *Speech Communication*, **75**: 27–49, doi: 10.1016/j.specom.2015.09.003.

16. Da Cunha M.G.B., Passerotti G.H., Weber R., Zilberstein B., Cecconello I. (2011), Voice feature characteristic in morbid obese population, *Obesity Surgery*, **21**(3): 340–344, doi: 10.1007/s11695-009-9959-7.

17. Dassie-Leite A.P., Behlau M., Nesi-França S., Lima M.N., de Lacerda L. (2018), Vocal evaluation of children with congenital hypothyroidism, *Journal of Voice*, **32**(6): 11–19, doi: 10.1016/j.jvoice.2017.08.006.

18. Deshpande G., Schuller B. (2020), An overview on audio, signal, speech, & language processing for COVID-19, *arXic preprint*, doi: 10.48550/arXiv.2005.08579.

19. Despotovic V., Ismael M., Cornil M., Mc Call R., Fagherazzi G. (2021), Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results, *Computers in Biology and Medicine*, **138**: 104944, doi: 10.1016/j.compbiomed.2021.104944.

20. DeVault D. *et al.* (2014), SimSensei kiosk: A virtual human interviewer for healthcare decision support, [in:] *AAMAS '14: Proceedings of the 2014 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1061–1068.

21. Dogan E., Sander C., Wagner X., Hegerl U., Kohls E. (2017), Smartphone-based monitoring of objective and subjective data in affective disorders: Where are we and where are we going? Systematic review, *Journal of Medical Internet Research*, **19**(7): e262, doi: 10.2196/jmir.7006.

22. Ekblad L.L. *et al.* (2015), Insulin resistance is associated with poorer verbal fluency performance in women, *Diabetologia*, **58**(11): 2545–2553, doi: 10.1007/s00125-015-3715-4.

23. Faurholt-Jepsen M. *et al.* (2016), Voice analysis as an objective state marker in bipolar disorder, *Translational psychiatry*, **6**(7): e856–e856, doi: 10.1038/tp.2016.123.

24. Gosztolya G., Bagi A., Szalóki S., Szendi I., Hoffmann I. (2018), Identifying schizophrenia based on temporal parameters in spontaneous speech, doi: 10.13140/RG.2.2.10884.78721.

25. Gosztolya G., Vincze V., Tóth L., Pákáski M., Kálmán J., Hoffmann I. (2019), Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using ASR and linguistic features, *Computer Speech & Language*, **53**: 181–197, doi: 10.1016/j.csl.2018.07.007.

26. Gratch J. *et al.* (2014), The distress analysis interview corpus of human and computer interviews, [in:] *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3123–3128, http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf.

27. Grósz T., Busa-Fekete R., Gosztolya G., Tóth L. (2015), Assessing the degree of nativeness and Parkinson's condition using Gaussian processes and deep rectifier neural networks, [in:] *Interspeech*, pp. 919–923, doi: 10.21437/Interspeech.2015-195.

28. Grünerbl A. *et al.* (2014), Smartphone-based recognition of states and state changes in bipolar disorder patients, *IEEE Journal of Biomedical and Health Informatics*, **19**(1): 140–148, doi: 10.1109/jbhi.2014.2343154.

29. Gugatschka M. *et al.* (2013), Subjective and objective vocal parameters in women with polycystic ovary syndrome, *Journal of Voice*, **27**(1): 98–100, doi: 10.1016/j.jvoice.2012.07.007.

30. Guidi A., Schoentgen J., Bertschy G., Gentili C., Scilingo E.P., Vanello N. (2017), Features of vocal frequency contour and speech rhythm in 1250 bipolar disorder, *Biomedical Signal Processing and Control*, **37**: 23–31, doi: 10.1016/j.bspc.2017.01.017.

31. Guidi A., Scilingo E. P., Gentili C., Bertschy G., Landini L., Vanello N. (2015), Analysis of running speech for the characterization of mood state in bipolar patients, [in:] *2015 AEIT International Annual Conference (AEIT)*, pp. 1–6, doi: 10.1109/AEIT.2015.7415275.

32. Hamdan A.-l., Jabbour J., Nassar J., Dahouk I., Azar S.T. (2012), Vocal characteristics in patients with type 2 diabetes mellitus, *European Archives of Oto-Rhino-Laryngology*, **269**(5): 1489–1495, doi: 10.1016/j.amjoto.2012.03.008.
33. Hamdan A.-L., Safadi B., Chamseddine G., Kasty M., Turfe Z.A., Ziade G. (2014), Effect of weight loss on voice after bariatric surgery, *Journal of Voice*, **28**(5): 618–623, doi: 10.1016/j.jvoice.2014.03.004.
34. Han J. *et al.* (2020), An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety, *arXiv preprint*, doi: 10.48550/arXiv.2005.00096.
35. Hannoun A., Zreik T., Husseini S.T., Mahfoud L., Sibai A., Hamdan A.-l. (2011), Vocal changes in patients with polycystic ovary syndrome, *Journal of Voice*, **25**(4): 501–504, doi: 10.1016/j.jvoice.2009.12.005.
36. Hassan A., Shahin I., Alsabek M.B. (2020), COVID-19 detection system using recurrent neural networks, [in:] *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, pp. 1–5, doi: 10.1109/CCCI49893.2020.9256562.
37. Helfer B.S., Quatieri T.F., Williamson J.R., Mehta D.D., Horwitz R., Yu B. (2013), Classification of depression state based on articulatory precision, [in:] *Interspeech*, pp. 2172–2176, doi: 10.21437/Interspeech.2013-513.
38. Hemmerling, D., Orozco-Arroyave J.R., Skalski A., Gajda J., Nöth E. (2016), Automatic detection of Parkinson's disease based on modulated vowels, [in:] *Interspeech*, pp. 1190–1194, doi: 10.21437/Interspeech.2016-1062.
39. Hönig F., Batliner A., Nöth E., Schnieder S., Krajewski J. (2014), Automatic modelling of depressed speech: relevant features and relevance of gender, [in:] *Interspeech*, pp. 1248–1252, doi: 10.21437/Interspeech.2014-313.
40. Horwitz-Martin R.L. *et al.* (2016), Relation of automatically extracted formant trajectories with intelligibility loss and speaking rate decline in amyotrophic lateral sclerosis, [in:] *Interspeech*, pp. 1205–1209, doi: 10.21437/Interspeech.2016-403.
41. Huang G., Pencina K.M., Coady J.A., Beleva Y.M., Bhasin S., Basaria S. (2015), Functional voice testing detects early changes in vocal pitch in women during testosterone administration, *The Journal of Clinical Endocrinology Metabolism*, **100**(6): 2254–2260, doi: 10.1210/jc.2015-1669.
42. Huang K.-Y., Wu C.-H., Kuo Y.-T., Jang F.-L. (2016), Unipolar depression vs. bipolar disorder: An elicitation-based approach to short-term detection of mood disorder, [in:] *Interspeech*, pp. 1452–1456, doi: 10.21437/Interspeech.2016-620.
43. Junuzović-Žunić L., Ibrahimagić A., Altumbabić S. (2019), Voice characteristics in patients with thyroid disorders, *The Eurasian Journal of Medicine*, **51**(2): 101–105, doi: 10.5152/eurasianjmed.2018.18331.
44. Khorram S., Gideon J., McInnis M.G., Provost E.M. (2016), Recognition of depression in bipolar disorder: Leveraging cohort and person specific knowledge, [in:] *Interspeech*, pp. 1215–1219, doi: 10.21437/Interspeech.2016-837.
45. Kiss G., Sztahó D., Tulics M.G. (2021), Application for detecting depression, Parkinson's disease and dysphonic speech, [in:] *Interspeech*, pp. 956–957.
46. Klumpp P., Janu T., Arias-Vergara T., Vásquez-Correa J.C., Orozco-Arroyave J.R., Nöth E. (2017), Apkinson – A mobile monitoring solutionfor Parkinson's disease, [in:] *Interspeech*, pp. 1839–1843, doi: 10.21437/Interspeech.2017-416.
47. Kones R., Rumana U. (2017), Cardiometabolic diseases of civilization: History and maturation of an evolving global threat. An update and call to action, *Annals of Medicine*, **49**(3): 260–274, doi: 10.1080/07853890.2016.1271957.
48. Kopp W. (2019), How western diet and lifestyle drive the pandemic of obesity and civilization diseases, *Diabetes, Metabolic Ayndrome and Obesity: Targets and Therapy*, **12**: 2221–2236, doi: 10.2147/DMSO.S216791.
49. Laguarta J., Hueto F., Subirana B. (2020), COVID-19 artificial intelligence diagnosis using only cough recordings, *IEEE Open Journal of Engineering in Medicine and Biology*, **1**: 275–281, doi: 10.1109/OJEMB.2020.3026928.
50. Lechien J. *et al.* (2020), Features of mild-to-moderate COVID-19 patients with dysphonia, *Journal of Voice*, doi: 10.1016/j.jvoice.2020.05.012.
51. Lopez-Otero P., Docio-Fernandez L.D., Abad A., Garcia-Mateo C. (2017), Depression detection using automatic transcriptions of de-identified speech, [in:] *Interspeech*, pp. 3157–3161, doi: 10.21437/Interspeech.2017-1201.
52. Low D.M., Bentley K.H., Ghosh S.S. (2020), Automated assessment of psychiatric disorders using speech: A systematic review, *Laryngoscope Investigative Otolaryngology*, **5**(1): 96–116, doi: 10.1002/lio2.354
53. Mallela J. *et al.* (2020), Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM, [in:] *Interspeech*, pp. 4586–4590, doi: 10.21437/Interspeech.2020-2221.
54. Maor E., Sara J.D., Orbelo D.M., Lerman L.O., Levanon Y., Lerman A. (2018), Voice signal characteristics are independently associated with coronary artery disease, *Mayo Clinic Proceedings*, pp. 840–847, doi: 10.1016/j.mayocp.2017.12.025.
55. McGinnis E.W. *et al.* (2019), Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood, *IEEE Journal of Biomedical and Health Informatics*, **23**(6): 2294–2301, doi: 10.1109/JBHI.2019.2913590.

56. Mirheidari B., Blackburn D., Walker T., Venneri A., Reuber M., Christensen H. (2018), Detecting signs of dementia using word vector representations, [in:] *Interspeech*, pp. 1893–1897, doi: 10.21437/Interspeech.2018-1764.

57. Mohammadzadeh A., Heydari E., Azizi F. (2011), Speech impairment in primary hypothyroidism, *Journal of Endocrinological Investigation*, **34**(6): 431–433, doi: 10.1007/BF03346708.

58. Moro-Velazquez L., Gomez-Garcia J.A., Arias-Londoño J.D., Dehak N., Godino-Llorente J.I. (2021), Advances in Parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects, *Biomedical Signal Processing and Control*, **66**: 102418, doi: 10.1016/j.bspc.2021.102418.

59. Mota N.B. *et al.* (2012), Speech graphs provide a quantitative measure of thought disorder in psychosis, *PLOS ONE*, **7**(4): e34928. doi: 10.1371/journal.pone.0034928.

60. Mundt J.C., Vogel A.P., Feltner D.E., Lenderking W.R. (2012), Vocal acoustic biomarkers of depression severity and treatment response, *Biological psychiatry*, **72**(7): 580–587, doi: 10.1016/j.biopsych.2012.03.015.

61. Orozco-Arroyave J.R., Arias-Londoño J.F., Vargas-Bonilla J.F., Gonzalez-Rativa M.C., Nöth E. (2014a), New spanish speech corpus database for the analysis of people suffering from Parkinson's disease, [in:] *LREC*, pp. 342–347.

62. Orozco-Arroyave J.R. *et al.* (2014b), Automatic detection of Parkinson's disease from words uttered in three different languages, [in:] *Interspeech*, doi: 10.21437/Interspeech.2014-375.

63. Pan Y., Mirheidari B., Reuber M., Venneri A., Blackburn D., Christensen H. (2020), Improving detection of Alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction, [in:] *Interspeech*, pp. 4961–4965, doi: 10.21437/Interspeech.2020-2698.

64. Pareek V., Sharma R.K. (2016), Coronary heart disease detection from voice analysis, [in:] *2016 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1–6, doi: 10.1109/SCEECS.2016.7509344.

65. Pettorino M., Gu W., Półrola P., Fan P. (2017), Rhythmic characteristics of Parkinsonian speech: A study on Mandarin and Polish, [in:] *Interspeech*, pp. 3172–3176, doi: 10.21437/Interspeech.2017-850.

66. Pinheiro A.P., Niznikiewicz M. (2019), Altered attentional processing 1395 of happy prosody in schizophrenia, *Schizophrenia Research*, **206**: 217–224, doi: 10.1016/j.schres.2018.11.024.

67. Pinkas G., Karny Y., Malachi A., Barkai G., Bachar G., Aharonson V. (2020), SARS-CoV-2 detection from voice, *IEEE Open Journal of Engineering in Medicine and Biology*, **1**: 268–274, doi: 10.1109/ojemb.2020.3026468.

68. Pinto S. *et al.* (2016), Dysarthria in individuals with Parkinson's disease: A protocol for a binational, cross-sectional, case-controlled study in French and European Portuguese (FraLusoPark), *BMJ Open*, **6**(11): doi: 10.1136/bmjopen-2016-012885.

69. Pinyopodjanard S., Suppakitjanusant P., Lomprew P., Kasemkosin N., Chailurkit L., Ongphiphadhanakul B. (2019), Instrumental acoustic voice characteristics in adults with type 2 diabetes, *Journal of Voice*, **35**(1): 116–121, doi: 10.1016/j.jvoice.2019.07.003.

70. Pompili A. *et al.* (2020), Assessment of Parkinson's disease medication state through automatic speech analysis, *arXiv preprint*, doi: 10.48550/arXiv.2005.14647.

71. Rohanian M., Hough J., Purver M. (2021), Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs, [in:] *Interspeech*, pp. 3820–3824, doi: 10.21437/Interspeech.2021-1633.

72. Rusz J. *et al.* (2018), Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **26**(8): 1495–1507, doi: 10.1109/TNSRE.2018.2851787.

73. Sadeghian R., Schaffer J.D., Zahorian S.A. (2017), Speech processing approach for diagnosing dementia in an early stage, [in:] *Interspeech*, pp. 2705–2709, doi: 10.21437/Interspeech.2017-1712.

74. Sahu S., Espy-Wilson C.Y. (2016), Speech features for depression detection, [in:] *Interspeech*, pp. 1928–1932, doi: 10.21437/Interspeech.2016-1566.

75. Sattler C. *et al.* (2017), Interdisciplinary longitudinal study on adult development and aging (ILSE), [in:] *Encyclopedia of Geropsychology*, Pachana N.A. [Ed.], pp. 1213–1222, Springer, doi: 10.1007/978-981-287-082-7_238.

76. Scherer S., Stratou G., Gratch J., Morency L.-P. (2013a), Investigating 1435 voice quality as a speaker-independent indicator of depression and PTSD, [in:] *Interspeech*, pp. 847–851, doi: 10.21437/Interspeech.2013-240.

77. Scherer S. *et al.* (2013b), Automatic behavior descriptors for psychological disorder analysis, [in:] *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, doi: 10.1109/FG.2013.6553789.

78. Seneviratne N., Williamson J.R., Lammert A.C., Quatieri T.F., Espy-Wilson C. (2020), Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression, [in:] *Interspeech*, pp. 4551–4555, doi: 10.21437/Interspeech.2020-2758.

79. Sharma N. *et al.* (2020), Coswara – A database of breathing, cough, and voice sounds for COVID-19 diagnosis, *arXiv preprint*, pp. 4811–4815, doi: 10.21437/Interspeech.2020-2768.

80. Simantiraki O., Charonyktakis P., Pampouchidou A., Tsiknakis M., Cooke M. (2017), Glottal source features for automatic speech-based depression assessment, [in:] *Interspeech*, pp. 2700–2704, doi: 10.21437/Interspeech.2017-1251.

81. Sirmans S.M., Pate K.A. (2014), Epidemiology, diagnosis, and management of polycystic ovary syndrome, *Clinical Epidemiology*, **6**: 1–13, doi: 10.2147/clep.s37559.

82. Skodda S., Grönheit W., Schlegel U. (2011), Intonation and speech rate in Parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission, *Journal of Voice*, **25**(4): e199–e205, doi: 10.1016/j.jvoice.2010.04.007.

83. Solomon N.P., Helou L.B., Dietrich-Burns K., Stojadinovic A. (2011), Do obesity and weight loss affect vocal function?, [in:] *Seminars in Speech and Language*, **31**(1): 31–42, doi: 10.1055/s-0031-1271973.

84. de Souza L.B.R., Pereira R.M., dos Santos M.M., Godoy C.M.A. (2014), Fundamental frequency, phonation maximum time and vocal complaints in morbidly obese women, *ABCD. Arquivos Brasileiros de Cirurgia Digestiva*, **27**(1): 43–46. doi: 10.1590/ s0102-67202014000100011.

85. de Souza L.B.R., dos Santos M.M. (2018), Body mass index and acoustic voice parameters: Is there a relationship?, *Brazilian Journal of Otorhinolaryngology*, **84**(4): 410–415, doi: 10.1016/j.bjorl.2017.04.003.

86. Stasak B., Epps J., Cummins N., Goecke R. (2016), An investigation of emotional speech in depression classification, [in:] *Interspeech*, pp. 485–489, doi: 10.21437/Interspeech.2016-867.

87. Stasak B., Epps J., Goecke R. (2017), Elicitation design for acoustic depression classification: An investigation of articulation effort, linguistic complexity, and word affect, [in:] *Interspeech*, pp. 834–838, doi: 10.21437/Interspeech.2017-1223.

88. Stasak B., Huang Z., Razavi S., Joachim D., Epps J. (2021). Automatic detection of COVID-19 based on short-duration acoustic smartphone speech analysis, *Journal of Healthcare Informatics Research*, **5**(2): 201–217, doi: 10.1007/s41666-020-00090-4.

89. Stogowska E., Kamński K.A., Ziółko B., Kowalska I. (2022), Voice changes in reproductive disorders, thyroid disorders and diabetes: A review, *Endocrine Connections*, **11**(3): e201505, doi: 10.1530/EC-21-0505.

90. Subirana B. *et al.* (2020), Hi sigma, do I have the Coronavirus?: Call for a new artificial intelligence approach to support health care professionals dealing with the COVID-19 pandemic, *arXiv preprint*, doi: 10.48550/arXiv.2004.06510.

91. Sztahó D., Kiss G., Vicsi K. (2015), Estimating the severity of Parkinson's disease from speech using linear regression and database partitioning, [in:] *Interspeech*, pp. 498–502, doi: 10.21437/Interspeech.2015-183.

92. Ujiro T. *et al.* (2018), Detection of dementia from responses to atypical questions asked by embodied conversational agents, [in:] *Interspeech*, pp. 1691–1695, doi: 10.21437/Interspeech.2018-1514.

93. Valstar M. *et al.* (2013), Avec 2013: The continuous audio/visual emotion and depression recognition challenge, [in:] *AVEC'13 Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, doi: 10.1145/2512530.2512533.

94. Vásquez-Correa J.C., Arias-Vergara T., Orozco-Arroyave J.R., Nöth E. (2018), A multitask learning approach to assess the dysarthria severity in patients with Parkinson's disease, [in:] *Interspeech*, pp. 456–460, doi: 10.21437/interspeech.2018-1988.

95. Vásquez-Correa J.C., Arias-Vergara T., Orozco-Arroyave J.R., Vargas-Bonilla J.F., Arias-Londoño J.D., Nöth E. (2015), Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions, [in:] *Interspeech*, pp. 105–109, doi: 10.21437/Interspeech.2015-36.

96. Vásquez-Correa J.C., Orozco-Arroyave J.R., Nöth E. (2017), Convolutional neural network to model articulation impairments in patients with Parkinson's disease, [in:] *Interspeech*, pp. 314–318, doi: 10.21437/Interspeech.2017-1078.

97. Villa-Cañas T., Arias-Londoño J.D., Orozco-Arroyave J.R., Vargas-Bonilla J.F., Nöth E. (2015), Low-frequency components analysis in running speech for the automatic detection of Parkinson's disease, [in:] *Interspeech*, pp. 100–104, doi: 10.21437/Interspeech.2015-35.

98. Villatoro-Tello E., Dubagunta P., Fritsch J., Ramírez-de-la Rosa G., Motlicek P., Magimai-Doss M. (2021), Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition, [in:] *Interspeech*, pp. 1927-1931, doi: 10.21437/Interspeech.2021-1288.

99. Wang J., Kothalkar P.V., Cao B., Heitzman D. (2016), Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples, [in:] *Interspeech*, pp. 1195–1199, doi: 10.21437/Interspeech.2016-1542.

100. Wankerl S., Nöth E., Evert S. (2017), An n-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language, [in:] *Interspeech*, pp. 3162–3166, doi: 10.21437/Interspeech.2017-1572.

101. Warnita T., Inoue N., Shinoda K. (2018), Detecting Alzheimer's disease using gated convolutional neural network from audio data, [in:] *Interspeech*, pp. 1706–1710, doi: 10.21437/Interspeech.2018-1713.

102. Wei W., Wang J., Ma J., Cheng N., Xiao J. (2020), A real-time robot-based auxiliary system for risk evaluation of COVID-19 infection, *arXiv preprint*, doi: 10.48550/arXiv.2008.07695.

103. Weiner J., Angrick M., Umesh S., Schultz T. (2018), Investigating the effect of audio duration on dementia detection using acoustic features, [in:] *Interspeech*, pp. 2324–2328, doi: 10.21437/Interspeech.2018-57.

104. Weiner J., Herff C., Schultz T. (2016), Speech-based detection of Alzheimer's disease in conversational German, [in:] *Interspeech*, pp. 1938–1942, doi: 10.21437/Interspeech.2016-100.

105. Wodzinski M., Skalski A., Hemmerling D., Orozco-Arroyave J.R., Nöth E. (2019), Deep learning approach to Parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification, [in:] *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 717–720, doi: 10.1109/embc.2019.8856972.

106. Xezonaki D., Paraskevopoulos G., Potamianos A., Narayanan S. (2020), Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews, [in:] *Interspeech*, pp. 4556–4560, doi: 10.21437/Interspeech.2020-2819.

107. Yang Y., Fairbairn C., Cohn J.F. (2012), Detecting depression severity from vocal prosody, *IEEE Transactions on Affective Computing*, **4**(2): 142–150, doi: 10.1109/T-AFFC.2012.38.

108. Zhan A. *et al.* (2016), High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection, *arXiv preprint*, doi: 10.48550/arXiv.1601.00960.

109. Zhao Z. *et al.* (2020), Hybrid network feature extraction for depression assessment from speech, [in:] *Interspeech*, pp. 4956–4960, doi: 10.21437/Interspeech.2020-2396.

110. Zlotnik A., Montero J.M., San-Segundo R., Gallardo-Antolín A. (2015), Random forest-based prediction of Parkinson's disease progression using acoustic, ASR and intelligibility features, [in:] *Interspeech*, pp. 503–507, doi: 10.21437/Interspeech.2015-184.