Journal of InterDisciplinary Research Applied to Medicine
JDReAM

JDREAM. Journal of interDisciplinary REsearch Applied to Medicine
JDREAM (2020), v.4 i. 2, 21-30
ISSN 2532-7518
DOI 10.1285/i25327518v4i2p21
http://siba-ese.unisalento.it, © 2020 Università del Salento

Automatic detection of Voice Disorders: recent literature advancements

Francesco Sigona[1]

[1]Department of Humanities & CRIL-DReAM, University of Salento, Italy

Corresponding author: Francesco Sigona
francesco.sigona@unisalento.it

Abstract

A short review of some recent findings in the field of automatic voice disorders detection and classification is provided in this article. The matter is getting more and more interest due to appealing non-invasiveness of the methods as well as the good achievable performances. An increasing role is played by Artificial Neural Networks (ANN), especially Deep ones, despite the need for large amounts of data for such networks, that are not always available for the task in question. The research in this field is directed in other directions too, including the investigation of new features and the capability to process running speech other than sustained sounds.

Keywords: voice disorders, machine learning, detection, classification

1. Introduction

The analysis of abnormal voice patterns or voice disorders plays an important role in diagnosing/treating several diseases and reducing the impact on the individual's communication skills. Voice disturbances include alterations in the quality, pitch, or amplitude, among other characteristics, that diverge from voices of similar age, gender, and social groups. Voice pathology detection and classification can be accomplished by means of automatic and a non-invasive method, by capturing patient's voice samples by a microphone, a smartphone, or any voice recorder, then submitting such samples to digital systems running specific software. Even though the terms detection and classification are used interchangeably very often, they actually refer to different tasks. Voice disorder detection refers to determining whether the given voice sample was produced by a person having a voice disorder or not; voice disorder classification task involves the ability to infer the type of disorder.

In this paper, some of the most recent studies in the field of automatic detection and classification of voice disorders are outlined. The found researches cover the use of sustained vowel as well as running speech, the investigation of novel features as well as novel classification methods to improve the accuracy of the outcome. Artificial Neural Networks (ANN) are more and more employed but traditional Machine Learning classifiers such as Support Vector Machines (SVM) still have their own advantages.

The remaining sections of the paper are organized as follows. The Section 2 reviews the most popular databases used to "learn" pathological and normal speech characteristics. Sections 3 and 4 respectively present some representative recent research papers on automatic voice disorders detection (including mobile apps solutions) and classification. Section 5 focuses on speech impairments by central nervous system disorders, while Section 6 present some articles about Specific language impairment (SLI), also known as development dysphasia. Section 7 present our conclusions.

2. Databases

2.1. Massachusetts Eye and Ear Infirmary (MEEI) database

The SVD database is publicly available via the Internet, (Barry and Putzer 2020) contains not only voice samples but also electroglottographic

(EGG) signals. The signals contain the information of the glottis movement during voice phonation. The materials include vowels /a/, /i/, /u / pronounced at different pitch (low, normal and high) for 1-3 s., the sentence "Guten Morgen, wie geht es Ihnen?" (Hello, how are you?), and EGG. The files have averages of around 1 and 3 s for sustained vowels and voice samples were sampled at 50 kHz with 16 bits of resolution.

2.3. Arabic voice pathology database (AVPD)

The AVPD (Mesallam et al. 2017) was recently developed at King Saud University, Riyadh. The database contains samples of sustained vowels, words, and paragraphs. All the speakers were native to Arabic language. Dysphonic patients suffering from five different types of organic voice disorders (cysts, nodules, polyps, paralysis and sulcus) were included in the database. The database contains repeated vowels, a running speech, Arabic digits and some common words. All subjects, including patients and normal persons, were recorded after clinical evaluation.

2.4. VOice ICarfEDerico II (VOICED)

The freely available VOICED database (Cesari et al. 2018) has been realized by the "Institute of High-Performance Computing and Networking of the National Research Council of Italy (IC-AR-CNR)" and the Hospital University of Naples "Federico II". It consists of 208 healthy and pathological voices collected during a clinical study performed following the guidelines of the medical SIFEL (Società Italiana di Foniatria e Logopedia) protocol and the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) 2013 Statement. For each subject, the database contains a recording of the vowel /a/ of five seconds in length, lifestyle information, the medical diagnosis, and the results of two specific medical questionnaires.

2.5. The Cantonese perceptual evaluation of voice (CanPEV) database

The CanPEV database (Law et al. 2010) was developed by the Division of Speech Therapy, the Chinese University of Hong Kong (CUHK). It consists of speech recordings from 232 native Cantonese speakers with either normal or pathological voices. The speech was recorded with a close-talking microphone in a quiet room at 44,100 Hz sampling rate. Each subject was required to produce repetitions of sustained vowels /a/, /i/ and /u/ (each one about 3 to 5 seconds long), 30–90 seconds of read speech, spontaneous speech from few seconds to few minutes.

2.6. The HUPA Database

This database was recorded at the Príncipe de Asturias hospital in Alcalá de Henares, Madrid, Spain (Moro-Velázquez et al. 2015; Arias-Londoño et al. 2011). The dataset contains sustained phonations of the vowel /a/ by 439 adult Spanish speakers (239 healthy and 200 pathological). Originally, the data was recorded with a sampling frequency of 50 kHz and later downsampled to 25 kHz. Pathological voices contain a wide variety of organic pathologies such as nodules, polyps, oedemas and carcinomas. More details of the database can be found also in Godino-Llorente et al. (2008).

2.7. The Advanced Voice Function Assessment Database (AVFAD)

The AVFAD database (Jesus et al. 2017) contains 363 healthy voices (253 females and 110 males) and 346 abnormal voices (247 females and 99 males). All clinical conditions were registered according to the Classification Manual of Voice Disorders-I (Verdolini, Rosen, and Branski 2006). Participants were audio-recorded, producing the following vocal tasks: sustaining vowels /a, i, u/; reading of six CAPE-V sentences; reading a phonetically balanced text; spontaneous speech (Behlau 2003).

2.8. The Voice disordered and Healthy Adults Speech Database

The recordings for this database (Tulics et al. 2019) were collected at the Outpatients' Department of the Head and Neck Surgery Department of the National Institute of Oncology, Budapest, Hungary, during consultations. The most common recordings are from patients with functional dysphonia and recurrent paresis. Samples from healthy people were recorded as well, these are used as comparison. All the participants had to read out loud the same eight sentence long text, titled 'The North Wind and the Sun'. This folk tale is frequently used in phoniatrics as a demonstration of continuous speech. The text was read in its Hungarian translation. The database contains a total of 450 recordings, 257 from patients with voice disorders (156 females and 101 males) and 193 people with a healthy voice (108 females and 85 males).

2.9. The LANNA children speech corpus

This database by the Laboratory of Artificial Neural Network Applications (Grill and Tučková 2016) contains speech samples of children suffering from SLI and healthy controls, i.e. normally developing children with no language or speech disorders diagnosed. The patients' group consists of 54 Czech children diagnosed with SLI in the 4 to 13 age group. Their speech was recorded in private speech and language therapist's office and doctor's office in Motol University Hospital. No information on the severity of the disorder is provided. The other (controls) group consists of 44 healthy Czech children in the 4 to 10 age group. The utterances from them were collected in school-rooms. The healthy controls subset comprises 1658 samples, while the SLI children subset comprises 2103 samples. All the recordings contain background noise as they were registered in situations simulating environment natural to the children to ensure their natural behaviour. The corpus consists of seven types of utterances: vowels, consonants, one-, two-, three-, and four-syllable words, and difficult words.

2.10. Parkinson's UI Machine Learning database

Acoustic data in this database (Little et al. 2009) consists of 195 sustained vocal phonations of 31 male and female subjects, of which 23 were diagnosed with Parkinson's disease. The age of the patients varies between 46 and 85 years (average of 65.8, standard deviation of 9.8). For each patient, averages of six phonations were recorded, with a length ranging from 1 to 36 s.

3. Some representative recent research papers on automatic voice disorders detection

AL-Dhief et al. (2020) presented a voice pathology detection system using Online Sequential Extreme Learning Machine (OSELM) to classify the voice signal into healthy or pathological, based on Mel-Frequency Cepstral Coefficients (MFCCs) as input feature. OSELM combines the advantage of good generalization performance at extremely fast learning speed of ELM (feedforward neural networks, "invented" in 2006 by G. Huang, 2006) with the capability to handle data samples obtained within packets over time (Abbas, Albadr, and Tiun 2017), instead of all at once. The voice samples for the vowel /a/ were collected equally from Saarbrücken voice database (SVD). The obtained results show that the maximum accuracy, sensitivity and specificity are 85%, 87% and 87%, respectively, showing that the proposed approach can differentiate healthy and pathological voices effectively.

While most approaches rely on feature extraction of the analysed signal with features subsequently fed into a classifier, Georgopoulos, 2020, investigated the direct use a time-frequency distribution (namely, the Wigner-Ville Distribution) of the voice signal and a deep learning classification method to automatically classify voice signals as normal or pathological. The time-frequency distribution is used as an image representation of the signal. The classification method is based on transfer learning of GoogleNet (Wu et al. 2018)) a well-trained Convolutional Neural Network (CNN) on large-scale natural images (unrelated to this problem) available in ImageNet. Voice data came from KAY Elemetrics (now Pentax Medical) database, developed by the MEEI Voice and Speech Lab. The samples used for analysis here is sustained phonation of vowel /a/. Achieved accuracy ranged from 69% to 74%.

Kadiri and Alku (2020), presented a systematic analysis of glottal source features in normal and pathological voice and investigated their effectiveness in voice pathology detection. Voice pathology detection experiments were carried out using the HUPA and the SVD databases. The glottal source features were derived from three signals: from the glottal flows estimated with the quasi-closed phase (QCP) glottal inverse filtering method (Airaksinen et al., 2014), from the approximate source signals computed with the zero frequency filtering (ZFF) method (Sri, Murty, and Yegnanarayana 2008) and directly from acoustic voice signals. The QCP method is based on the principles of closed phase (CP) analysis which estimates the vocal tract model from few speech samples located in the CP of the glottal cycle using linear prediction (LP) analysis. In contrast to the CP method, QCP takes advantage of all the speech samples of the analysis frame in computing the vocal tract model. ZFF is based on the fact that the effect of an impulse-like excitation (that occurs at the instant of glottal closure) is present throughout the spectrum including the zero frequency, while the vocal tract characteristics are mostly reflected in resonances at much higher frequencies. In this method, the acoustic speech signal is passed through a cascade of two zero frequency resonators and the resulting signal is equivalent to integration of the signal four times. Hence, the output grows or decays as a polynomial function of time. The trend is removed by subtracting the local mean computed over the average pitch period at each sample and the resulting output signal is referred as the zero-frequency filtered (ZFF) signal.

Analysis of features revealed that glottal source features help in discriminating normal voice from pathological voice. A Support Vector Machine (SVM) with a radial basis function (RBF) kernel has been used as classifier. The studied glottal source features provide better discrimination compared to spectral features such as MFCCs and perceptual linear prediction (PLPs) features. Further, the combination of the existing spectral features with the glottal source features resulted in improved detection performance, indicating the complementary nature of features. The best achieved accuracy was about 78%.

Oliveira et al. (2020), investigated the feasibility of combining sustained vowels for computer-based pathological voice characterization. The Authors conducted experiments on samples of sustained vowels /a/, /i/ and /u/ from SVD and AVFAD datasets, exploring the wavelet decomposition levels in the range of 4 to 18, revealing that wavelet coefficients extracted from the combination of vowels improved signal description and, hence, identification of subtle features of pathological voices, using Random Forest classifier (Breiman 2001). They also showed that the Haar wavelet-based features (Shia and Jayasree 2017) extracted from combined vowels achieved accurate voice classification with fewer decomposition levels. This approach enabled accuracy improvements of at least 15.61 and 2.61% for SVD and AVFAD datasets, respectively, regardless of the biological gender, achieving a final accuracy ranging from 78% to 83%.

Tulics et al. (2019) investigated two types of input vectors (acoustic features and Automatic Speech Recognition -ASR- posterior probabilities) with a SVM- and DNN-based classifiers, using read text materials from Hungarian-speaking patients suffering from multiple types of diseases from the Voice disordered and Healthy Adults Speech Database. They found that using acoustic parameters instead of the use phone-specific posteriors as input features increases the accuracy for the detection and classification of disordered voices. The most important parameters, as suggested by the employed Forward Feature Selection (FFS) algorithm, were the mean of MFCCs, the range of SPI on voiced plosives and affricates, the standard deviation of HNR, the range of IMF on nasals, the mean and standard deviation of jitter, the standard deviation of MFCCs, the mean of SPI on voiced plosives and affricates, the range of SPI on the vowel [E] and the standard deviation of SPI on voiced plosives and affricates. Also, the DNN approach outperformed the SVM classifier. Later, the same Authors (Tulics et al. 2020) examined the combination of the two input vectors can contribute to improve classification accuracy. They concluded that it is not worthwhile to calculate ASR phone posteriors, as it has no significant impact on classification outcome, but it can greatly complicate and slow down a diagnosis support system that models the cognitive decision-making process of an expert. ASR phone posterior derived features are less effective in the automatic classification of healthy and dysphonic voices, than using the acoustic feature set directly. Adding ASR phone posterior derived features to the acoustic features does not significantly improve the automatic classification accuracy of healthy and dysphonic voices.

They explained this finding by the fact that ASR acoustic models are trained with the objective of being robust to variation, which is likely to shade the differences between dysphonic and non-dysphonic voices. If the training corpora used to train the ASR models were controlled and labelled w.r.t. dysphonia, phone posteriors could become useful for detecting dysphonia, however, as ASR training relies on very large datasets (ideally several hundred, rather thousands of hours of speech), this requirement is quasi hard to fulfil in practice.

3.1. Mobile apps for automatic detection of voice disorders.

Ilapakurti et al. (2019), aiming at developing mobile diagnostic voice disorder app, investigated Mel-Scale Spectrogram and MFCCs as input features for several NN architectures: a 5-layer plain network, 5-layer CNN and a Recurrent Neural Network (RNN). Voice samples were obtained from a voice clinic in a tertiary teaching hospital (Far Eastern Memorial Hospital, FEMH), which included 50 normal voice samples and 150 samples of com-mon voice disorders, including vocal nodules, polyps, and cysts (collectively referred to Phono trauma); glottis neoplasm; unilateral vocal paralysis. Voice samples of a 3-second sustained vowel sound /a:/ were recorded at a comfortable level of loudness. The best model was a 5-layer CNN trained with MFCC and Mel-Spectrogram. It had a Sensitivity: 96% & Specificity: 18% on the test data.

Verde et al. (2019) propose a machine learning (ML)-based mobile voice disorder detection system. A trained model was directly embedded in a mobile application, allowing the user to evaluate the health of his/her own voice anywhere and at any time, without the necessity of trans-mitting user data to or storing user data on any server. This constitutes, at the time of writing, a significant innovation on account of the fact that most of the existing studies in literature limit the use of the mobile device to the tasks of acquiring the useful signal, transmitting it to an external server to be analysed and visualizing and communicating the results obtained to the users (Alhussein and Muhammad 2018; Muhammad et al. 2018). Unfortunately, the transmission of these patient data can be subject to a security attacks on security or privacy violations. The proposed mobile system, instead, has no need to transmit any data, so limiting the probability of any security attack.

4. Automatic classification of voice disorders

Liu et al. (2019) investigated phone posterior probabilities from a large-vocabulary ASR system trained with normal speech (Cantonese) to classify spoken utterances of 80 subjects extracted from the CanPEV database, which were already rated and divided into several categories: normal+mild, moderate, and severe. In addition to the proposed ASR voice features, the effectiveness of a set of conventional voice features that can be extracted from the utterance without using acoustical model has been investigated. The Authors adopted a minimalistic acoustic parameter set for voice analysis, known as eGeMAPS, which is implemented with the OpenSMILE toolkit (Eyben et al. 2016). Given the sequence of posterior vectors, the Authors proposed to use it to compute four types of feature parameters (a total of 18-dimension features), which are used to locate and quantify irregular posterior variations at specific speech sounds: PPV (Phone Posterior Variation), GOP (Goodness of Pronunciation), GOPV (GOP Variation) and BFR (Blurred Frame Ratio). By combining the contributions from the ASR voice features and conventional voice features, a subject-level prediction accuracy of over 80% on three severity classes has been achieved. Subjects with mild disorder and those with severe disorder could be perfectly distinguished by the proposed method.

(Miliaresi, Poutos, and Pikrakis 2021) addressed the task to classify functional dysphonia, phonotrauma, laryngeal neoplasm and vocal paralysis and showed that it is possible to treat MFCC derived features and data from medical records as two different input sources to a single neural network architecture consisting of two subnetworks. The first one, a CNN is used to treat the acoustic signal as an image, that captures spectral shape by operating on MFFC derived features and simple filterbank outputs. The second (feed-forward) network analyses an enhanced input vector, consisting of the demographic parameters and mid-term signal features. The outputs of the aforementioned subnetworks are concatenated and fed to a dense layers with 1024 nodes, with each node's output being processed by a Rectified Linear Unit (ReLu) activation function. Finally, a softmax output layer of four units is used to produce posterior probability estimations of the four classes of the problem under study.

5. Speech impairments by central nervous system disorders.

Lauraitis et al. (2020) adopted Bidirectional Long Short-Term Memory (BiLSTM) neural network and Wavelet Scattering Transform with Support Vector Machine (WST-SVM) classifier for detecting speech impairments of patients at the early stage of central nervous system disorders (CNSD). The study includes 339 voice samples collected from 15 subjects: 7 patients with early stage CNSD (3 Huntington, 1 Parkinson, 1 cerebral palsy, 1 post stroke, 1 early dementia), other 8 subjects were healthy. Speech data is collected using voice recorder from Neural Impairment Test Suite (NITS) mobile app. Features are extracted from pitch contours, Mel-frequency cepstral coefficients (MFCC), Gammatone cepstral coefficients (GTCC), Gabor (analytic Morlet) wavelet and auditory spectrograms. 94.50% (BiLSTM) and 96.3% (WST-SVM) accuracy is achieved for solving healthy vs. impaired classification problem. The developed method can be applied for automated CNSD patient health state monitoring and clinical decision support systems as well as a part of Internet of Medical Things (IoMT).

5.1. Parkinons' desease (PD)

Kodrasi and Bourlard (2020), proposed to use the spectro-temporal sparsity characterization as a robust feature for dysarthric speech detection, based on the motivation that since dysarthric speech of patients suffering from PD is breathy, semi-whispery, and is characterized by abnormal pauses and imprecise articulation, it can be expected that its spectro-temporal sparsity differs from the spectro-temporal sparsity of healthy speech. The Authors first provided a numerical analysis of the suitability of different non-parametric and parametric measures (i.e., l1-norm, kurtosis, Shannon entropy, Gini index, shape parameter of a Chi distribution, and shape parameter of a Weibull distribution) for sparsity characterization. It is shown that kurtosis, the Gini index, and the parametric sparsity measures are advantageous sparsity measures, whereas the l1-norm and entropy measures fail to robustly characterize the temporal sparsity of signals with a different number of time frames. Second, they proposed to characterize the spectral sparsity of an utterance by initially time-aligning it to the same utterance uttered by a (arbitrarily selected) reference speaker using dynamic time warping. Experimental results on a Spanish database of healthy and dysarthric speech showed that estimating the spectro-temporal sparsity using the Gini index or the parametric sparsity measures and using it as a feature in a support vector machine results in a high classification accuracy of 83.3%.

Asmae et al. (2020), used ANN and K-Nearest Neighbours (KNN) algorithms, in the purpose of distinguishing between PD patient and healthy individual. Voice data in the Parkinson's UI Machine Learning has been used. Standard features derived from fundamental frequency, jitter and shimmer, have been used, as well as non-standard features such as Correlation Dimension (Kantz and Schreiber 2003), Recurrence Period Density Entropy and Detrended Fluctuation Analysis (Dixit 1988): 22 features in total. The ANN has two hidden layers and the Levenberg-Marquardt (LM) has been used as training optimization algorithm (Hagan and Menhaj 1994). Experimental results have showed that the ANN classifier achieved higher average performance than the KNN classifier in term of accuracy. The established system can distinguish healthy people from an acceptable range of people with PD with an accuracy rate of 96.7% by using ANN, and 79.3% by using KNN when the number of neighbours taken was k=1, by using the cosine distance.

6. Specific language impairment (SLI)

Several very recent researches regarded Specific language impairment (SLI), also known as development dysphasia.

Kotarba and Kotarba (2020), proposed an efficient approach to automatic detection of SLI based on log-power spectrograms of speech samples. The utterances from the LANNA children speech corpus were used to calculate the normalized log-power spectrograms. Deep neural network algorithm based on ResNet architecture (He et al. 2016) was used to perform the classification task. The accuracy rate of proposed SLI detection method exceeds 99% in the speaker independent scenario.

Reddy, Alku, and Rao (2020), proposed a method for SLI detection in children that utilizes time- and frequency-domain glottal parameters, which are extracted from the voice source signal obtained using quasi-closed phase (QCP). In addition, 12 MFCCs and openSMILE based acoustic features are also extracted from speech utterances, including min (or max) value and its relative position, standard deviation, range, median, skewness, kurtosis, 2 linear regression coefficients, and quadratic error of the following features: root mean square (RMS) energy, zero crossing rate, pitch and voicing probability. SVMs with RBF kernel and feed-forward neural network (FFNN), are trained separately for the MFCCs, openSMILE and glottal features. A leave-fourteen-speakers-out cross-validation strategy is used for evaluating the classifiers. The experiments are conducted using the LANNA corpus. Experimental results show that the glottal parameters contain significant discriminative information required for identifying children with SLI. Furthermore, the complementary nature of glottal parameters is investigated by independently combining these features with the MFCCs and openSMILE acoustic features. The overall results indicate that the glottal features when used in combination with MFCCs feature set provides the best performance with the FFNN classifier in the speaker-independent scenario (98.82%).

Sharma and Singh (2020) used sustained phonation of vowel /a/ uttered by children, from the LANNA database, to detect and classify control (healthy) and experimental (SLI) group using linear predictive coding (LPC) feature set. LPC order was set to 8, and a set of 408 features was build using 17 statistical function applied to the 8 coefficients, their delta and delta-delta ("delta" refers to the difference between two consecutive feature frames). A standard non-parametric Mann-Whitney non-parametric U-test was applied to filter the significant features for 95% level of confidence. The top-20 and top-10 features were then selected by computing the Spearman's rank correlation coefficients. Naïve-Bayes (NB) and SVM were employed for machine learning task. The best accuracies were obtained from NB classifiers i.e. 97.9% (for top-20 LPC features) and 97.8% (for top-10 LPC features) with 5-fold cross-validation protocol.

7. Conclusions

From this unexhaustive review of the most recent attempts to improve automatic voice disorders detection and classification, it is clear that the role of ANN is going to increase in the near future. More and more complex networks are investigated such as Online Sequential Extreme Learning Machine (OSELM) and GoogleNet and transfer learning is applied to face the availability of limited amount of data compared to the amount that would be required to train from scratch the most complex networks.

Accuracy achievable with ANN can be higher than 90%, with peaks close to 99% in some specific tasks.

One of the most promising approach is to embed the capability to detect voice disorders in mobile devices as software app, allowing a portable and usable solution to monitor the quality of voice in real time.

An intense research activity is also devoted to find and select new features to augment or replace the classical one to improve the recognition accuracy, including a trend to use directly spectrographic or time-frequency representations of voice samples as images to feed ANN good at image-recognition.

Finally, running speech is increasingly considered instead of sustained vowels to take into account more realistic speech scenarios.

8. References

- Abbas, Musatafa, Abbood Albadr, and Sabrina Tiun. 2017. "Extreme Learning Machine: A Review." *International Journal of Applied Engineering Research*. 12: 4610-4623.
- Airaksinen, Manu, Tuomo Raitio, Brad Story, and Paavo Alku. 2014. "Quasi Closed Phase Glottal Inverse Filtering Analysis with Weighted Linear Prediction." *IEEE Transactions on Audio, Speech and Language Processing* 22 (3): 596–607. https://doi.org/10.1109/TASLP.2013.2294585.
- AL-Dhief, Fahad Taha, Nurul Mu'azzah Abdul Latiff,

Nik Noordini Nik Abd. Malik, Naseer Sabri, Marina Mat Baki, Musatafa Abbas Abbood Albadr, Aymen Fadhil Abbas, Yaqdhan Mahmood Hussein, and Mazin Abed Mohammed. 2020. "Voice Pathology Detection Using Machine Learning Technique." In *2020 IEEE 5th International Symposium on Telecommunication Technologies (ISTT)*, 99–104. IEEE. https://doi.org/10.1109/ISTT50966.2020.9279346.

- Alhussein, Musaed, and Ghulam Muhammad. 2018. "Voice Pathology Detection Using Deep Learning on Mobile Healthcare Framework." *IEEE Access* 6: 41034–41. https://doi.org/10.1109/ACCESS.2018.2856238.
- Arias-Londoño, Julián David, Juan I. Godino-Llorente, Maria Markaki, and Yannis Stylianou. 2011. "On Combining Information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for Automatic Detection of Pathological Voices." *Logopedics Phoniatrics Vocology* 36 (2): 60–69. https://doi.org/10.3109/14015439.2010.528788.
- Asmae, Ouhmida, Raihani Abdelhadi, Cherradi Bouchaib, Sandabad Sara, and Khalili Tajeddine. 2020. "Parkinson's Disease Identification Using KNN and ANN Algorithms Based on Voice Disorder." In *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 1–6. IEEE. https://doi.org/10.1109/IRASET48871.2020.9092228.
- Barry, WJ, and M Putzer. 2020. "Voice Database." 2020. Accessed: December 10, 2020. http://www.stimmdatenbank.coli.uni-saarland.de/.
- Behlau, M. 2003. "Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)." *ASHA* 9: 187–89.
- Breiman, L. 2001. "Random Forest." *Machine Learning* 45 (1): 5–32.
- Cesari, Ugo, Giuseppe De Pietro, Elio Marciano, Ciro Niri, Giovanna Sannino, and Laura Verde. 2018. "A New Database of Healthy and Pathological Voices." *Computers and Electrical Engineering* 68: 310–21. https://doi.org/10.1016/j.compeleceng.2018.04.008.
- Dixit, RP. 1988. "On Defining Aspiration." In *Proc. 12th Int. Conf. Linguistics*, 606–10. Tokyo, Japan.
- Elemetrics, K. 1994. "Kay Elemetrics Corp. Disordered Voice Data-Base." *Model* 4337.
- Eyben, Florian, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, et al. 2016. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing." *IEEE Transactions on Affective Computing* 7 (2): 190–202. https://doi.org/10.1109/TAFFC.2015.2457417.
- Georgopoulos, Voula C. 2020. "Advanced Time-Frequency Analysis and Machine Learning for Pathological Voice Detection." In *2020 12th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, 1–5. IEEE. https://doi.org/10.1109/CSNDSP49049.2020.9249603.
- Godino-Llorente, Juan Ignacio, Víctor Osma-Ruiz, Nicolás Sáenz-Lechón, Ignacio Cobeta-Marco, Ramón González-Herranz, and Carlos Ramírez-Calvo. 2008. "Acoustic Analysis of Voice Using WPCVox: A Comparative Study with Multi Dimensional Voice Program." *European Archives of Oto-Rhino-Laryngology* 265 (4): 465–76. https://doi.org/10.1007/s00405-007-0467-x.
- Grill, Pavel, and Jana Tučková. 2016. "Speech Databases of Typical Children and Children with SLI." Edited by Frederic Dick. *PLOS ONE* 11 (3): e0150365. https://doi.org/10.1371/journal.pone.0150365.
- Hagan, Martin T., and Mohammad B. Menhaj. 1994. "Training Feedforward Networks with the Marquardt Algorithm." *IEEE Transactions on Neural Networks* 5 (6): 989–93. https://doi.org/10.1109/72.329697.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *2016 IEEE Conference on Computer Vi-Sion and Pattern Recognition (CVPR)*, 770–78.
- Huang, Guang-Bin. 2015. "What Are Extreme Learning Machines? Filling the Gap Between Frank Rosenblatt's Dream and John von Neumann's Puzzle." *Cognitive Computation* 7: 263–78. https://doi.org/10.1007/s12559-015-9333-0.
- Ilapakurti, Anitha, Sharat Kedari, Jaya Shankar Vuppalapati, Santosh Kedari, and Chandrasekar Vuppalapati. 2019. "Artificial Intelligent (AI) Clinical Edge for Voice Disorder Detection." In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, 340–45. IEEE. https://doi.org/10.1109/BigDataService.2019.00060.
- Jesus, Luis M.T., Inês Belo, Jessica Machado, and Andreia Hall. 2017. "The Advanced Voice Function Assessment Databases (AVFAD): Tools for Voice Clinicians and Speech Research." *Advances in Speech-Language Pathology* 9: 237255. https://doi.org/10.5772/intechopen.69643.
- Kadiri, Sudarsana Reddy, and Paavo Alku. 2020. "Analysis and Detection of Pathological Voice Using Glottal Source Features." *IEEE Journal of Selected Topics in Signal Processing* 14 (2): 367–79. https://doi.org/10.1109/JSTSP.2019.2957988.
- Kantz, Holger, and Thomas Schreiber. 2003. *Nonlinear Time Series Analysis* . Edited by Cambridge University Press. 2nd ed. Cambridge, UK.
- Kodrasi, Ina, and Herve Bourlard. 2020. "Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28: 1210–22. https://doi.org/10.1109/TASLP.2020.2985066.
- Kotarba, Katarzyna, and Michal Kotarba. 2020. "Efficient Detection of Specific Language Impairment in Children Using ResNet Classifier." In *2020 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 169–73. IEEE. https://doi.org/10.23919/SPA50552.2020.9241289.
- Lauraitis, Andrius, Rytis Maskeliunas, Robertas Damasevicius, and Tomas Krilavicius. 2020. "Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering

Domain Features." *IEEE Access* 8: 96162–72. https://doi.org/10.1109/ACCESS.2020.2995737.

- Law, T, K Lee, JH Lam, AC van Hasselt, and MCF Tong. 2010. "The Construction of the Cantonese Percep-Tual Evaluation of Voice (CanPEV): The Content Validation Process." In *Proc. 4th World Voice Congr. World Voice Consortium*, 159. Seoul, Korea.
- Little, Max A., Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig. 2009. "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease." *IEEE Transactions on Biomedical Engineering* 56 (4): 1015–22. https://doi.org/10.1109/TBME.2008.2005954.
- Liu, Yuanyuan, Tan Lee, Thomas Law, and Kathy Yuet-Sheung Lee. 2019. "Acoustical Assessment of Voice Disorder With Continuous Speech Using ASR Posterior Features." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27 (6): 1047–59. https://doi.org/10.1109/TASLP.2019.2905778.
- Mesallam, Tamer A., Mohamed Farahat, Khalid H. Malki, Mansour Alsulaiman, Zulfiqar Ali, Ahmed Al-nasheri, and Ghulam Muhammad. 2017. "Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms." *Journal of Healthcare Engineering* 2017: 1–13. https://doi.org/10.1155/2017/8783751.
- Miliaresi, Ioanna, Kyriakos Poutos, and Aggelos Pikrakis. 2021. "Combining Acoustic Features and Medical Data in Deep Learning Networks for Voice Pathology Classification." In *2020 28th European Signal Processing Conference (EUSIPCO)*, 1190–94. IEEE. https://doi.org/10.23919/Eusipco47968.2020.9287333.
- Moro-Velázquez, Laureano, Jorge Andrés Gómez-García, Juan Ignacio Godino-Llorente, and Gustavo Andrade-Miranda. 2015. "Modulation Spectra Morphological Parameters: A New Method to Assess Voice Pathologies According to the GRBAS Scale." *BioMed Research International* 2015. https://doi.org/10.1155/2015/259239.
- Muhammad, G, MF Alhamid, M Al-sulaiman, and B Gupta. 2018. "Edge Computing with Cloud for Voice Disor-Der Assessment and Treatment." *IEEE Commun. Mag* 56 (4): 6065.
- Oliveira, Brigada F. C., Deborah M. V. Magalhaes, Daniel S. Ferreira, and Fatima N. S. Medeiros. 2020. "Combined Sustained Vowels Improve the Performance of the Haar Wavelet for Pathological Voice Characterization." In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 381–86. IEEE. https://doi.org/10.1109/IWSSIP48289.2020.9145258.
- Reddy, Mittapalle Kiran, Paavo Alku, and Krothapalli Sreenivasa Rao. 2020. "Detection of Specific Language Impairment in Children Using Glottal Source Features." *IEEE Access* 8: 15273–79. https://doi.org/10.1109/ACCESS.2020.2967224.
- Sharma, Yogesh, and Bikesh Kumar Singh. 2020. "Prediction of Specific Language Impairment in Children Using Speech Linear Predictive Coding Coefficients." In *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, 305–10. IEEE. https://doi.org/10.1109/ICPC2T48082.2020.9071510.
- Shia, S. Emerald, and T. Jayasree. 2017. "Detection of Pathological Voices Using Discrete Wavelet Transform and Artificial Neural Networks." In *Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017*, 1–6. https://doi.org/10.1109/ITCOSP.2017.8303086.
- Sri, K, Rama Murty, and B Yegnanarayana. 2008. "Epoch Extraction From Speech Signals." *IEEE Trans. Audio, Speech, Lang Process* 16 (8). https://doi.org/10.1109/TASL.2008.2004526.
- Tulics, Miklos Gabriel, Gyorgy Szaszak, Krisztina Meszaros, and Klara Vicsi. 2019. "Artificial Neural Network and SVM Based Voice Disorder Classification." In *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 307–12. IEEE. https://doi.org/10.1109/CogInfoCom47531.2019.9089908.
- Tulics, Miklos Gabriel, Szaszak, Gyorgy, Meszaros Krisztina, Vicsi Klara. 2020. "Using ASR Posterior Probability and Acoustic Features for Voice Disorder Classification." In *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 000155–60. IEEE. https://doi.org/10.1109/CogInfoCom50765.2020.9237866.
- Verde Laura, Giuseppe De Pietro, Mubarak Alrashoud, Ahmed Ghoneim, Khaled N. Al-Mutib, and Giovanna Sannino. 2019. "Leveraging Artificial Intelligence to Improve Voice Disorder Identification Through the Use of a Reliable Mobile App." *IEEE Access* 7: 124048–54. https://doi.org/10.1109/ACCESS.2019.2938265.
- Verdolini, K, C Rosen, and R Branski. 2006. *Classification Manual for Voice Disorders*. Edited by I. Mahwah. Lawrence Erlbaum.
- Wu, Huiyi, John Soraghan, Anja Lowit, and Gaetano Di Caterina. 2018. "A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks," September, 446–50.