*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Towards automatic assessment of voice disorders: A clinical approach

*Purva Barche, Krishna Gurugubelli, and Anil Kumar Vuppala*

Speech Processing Laboratory, LTRC, KCIS
International Institute of Information Technology, Hyderabad, India.
{purva.sharma, krishna.gurugubelli}@research.iiit.ac.in, and anil.vuppala@iiit.ac.in

## Abstract

Automatic detection and assessment of voice disorders is important in diagnosis and treatment planning of voice disorders. This work proposes an approach for automatic detection and assessment of voice disorders from a clinical perspective. To accomplish this, a multi-level classification approach was explored in which four binary classifiers were used for the assessment of voice disorders. The binary classifiers were trained using support vector machines with excitation source features, vocal-tract system features, and state-of-art OpenSMILE features. In this study source features namely, glottal parameters obtained from glottal flow waveform, perturbation measures obtained from epoch locations, and cepstral features obtained from linear prediction residual and zero frequency filtered signal were explored. The present study used the Saarbucken voice disorders database to evaluate the performance of proposed approach. The OpenSMILE features namely ComParE and eGEMAPS feature sets shown better performance in terms of classification accuracies of 82.8% and 76%, respectively for voice disorder detection. The combination of excitation source features with baseline feature sets further improved the performance of detection and assessment systems, that highlight the complimentary nature of exciting source features.

**Index Terms**: Clinical perspective, Detection and assessment, Excitation source features, Voice disorders.

## 1. Introduction

Speech is the natural mode of communication for human beings. Speech production requires airflow from the lungs to be phonated through vocal folds of the larynx and resonated in the vocal cavities shaped by the tongue, jaw, soft palate, lips, and other articulators. Phonation is a process by which the vocal folds produce certain sounds through quasi-periodic vibration, also known as voicing. Any abnormality in the larynx that affect voicing in speech production, refers to as voice disorder. From auditory-perceptual point of view, voice disorder affect voice quality, pitch, and loudness [1]. Compare to invasive methods of voice disorder detection, non-invasive methods which utilize acoustic information, received great attention. Non-invasive voice disorder detection uses perceptual and objective assessment approaches. Perceptual assessment is reliable but challenging, as speech language pathologists (SLPs) need to evaluate the abnormalities in voicing by listening to patients [2]. On the other hand, objective assessment or automatic detection of voice disorder relies on acoustic features extracted from speech using signal processing techniques [3]. Objective assessment methods are effective, and requires less time, moreover the acoustic features used in these methods are highly correlated to perceptual measure, so these methods are most widely explored for voice disorder detection [4, 5, 6].

Effect of voice disorders can be seen as irregularity in vocal fold vibration, so perturbation measures like jitter and shimmer [7, 8, 9], harmonic to noise ratio (HNR) [10, 11, 12], signal to noise ratio (SNR) [13, 14], and glottal to noise excitation (GNE) [4, 15] were used to capture irregularity characteristics of vocal fold vibration. The studies by Sudarsana et al. [16], found glottal source feature as good measure to differentiate the pathological voice from normal voice. In the literature, widely used system features such as the Mel frequency cepstral coefficient (MFCC), perceptual linear prediction (PLP), and linear prediction cepstral coefficient (LPCC) have been shown as reliable acoustic measures of voice impairment [17, 18]. The studies on filter bank based analysis of voice disorders [19, 20], have revealed that the characteristics of voice disorders can be found better in some frequency bands compare to other frequency bands. Recently, some authors used machine learning approaches for pathological voice detection [21, 22]. In all the above mentioned approaches, voice disorder detection was seen as two class problem which discriminate pathological voice from healthy voice. On the other hand, clinicians examine voice disorder in different way, first they detect the presence of voice disorder, later they perform differential diagnosis to identify the type of voice disorder such as structural, neurogenic, functional or psychogenic [23].

To the best of our knowledge, this is the first study, which assess the voice disorder from clinical point of view. In this study, firstly voice disorders are classified from the healthy voice. Then voice disorder problem is further classified as two class problem to know whether disorder is organic or non-organic. Organic disorders are further classified into structural or neurogenic. In the same way non-organic disorders are classified as functional or psychogenic. This study used the Saarbruecken voice disorder (SVD) database for assessment of voice disorders [24]. In this regard, this study explores different source features and compares them with state-of-art vocal-tract system features and OpenSMILE features.

Rest of the paper is organised as follows. Section 2 describes clinical perspective of voice disorder classification, Section 3 presents experimental setup with details of database, extraction of excitation source evidence, feature extraction and classifier. Results and discussion of the voice disorder detection and assessment system are presented in Section 4. Finally, summary and conclusion of this work are discussed in Section 5.

## 2. Clinical perspective of voice disorder classification

This section discuss about classification of voice disorders from a clinical point of view, which is used to plan the experiments. According to American Speech-Language-Hearing Association, voice disorders are broadly characterized into organic and non-organic [25]. Organic voice disorders are physiological in nature, result from vocal cord paralysis, lesion of the



http://dx.doi.org/10.21437/Interspeech.2020-2160

larynx, problems with nervous system innervation to the larynx [26]. Organic voice disorders such as cyst, polyp, laryngitis and vocal nodules are due to the physical abnormality of the larynx, referred to as structural voice disorders [27]. On the other hand, voice disorders like, Parkinson's, recurrent laryngeal nerve palsy, spasmodic dysphonia, and asymptomatic lateral sclerosis are caused by the damage of recurrent laryngeal nerve or the problems in the central nervous system, referred to as neurological voice disorders [27]. In contrast to the organic voice disorders, the non-organic voice disorders are caused by ineffective use of the vocal mechanism or poor muscle control in subjects with normal physical structure, called as functional voice disorders such as muscle tension dysphonia, ventricular phonation, and vocal fold bowing [28]. Sometimes voice quality may degrade due to psychological factors, which in turn leads to psychological disorders such as psychogenic dysphonia (PD) and aphonia [28].

# 3. Experimental setup

This section describes details of database, extraction of excitation source features along with state-of-art openSMILE and vocal-tract-system features for voice disorder detection and assessment system. It also describes classifier and it's parameters used in our study.

## 3.1. Database

The database used in this experiment is saarbruecken voice disorder dataset which is freely available on `http://www.stimmdatenbank.coli.uni-saarland.de/` [24]. It contains more than 2000 voice recordings sampled at 50 kHz, out of which, 687 are collected from healthy subjects (428 females and 259 males) and 1356 are collected from subjects (629 males and 727 females) with voice disorders. This database contains 71 different voice disorders. Each recording session consists of a German sentence ("Guten Morgen, wie geht es Ihnen?") and vowels of /a/, /i/, and /u/ in normal, high, low and rising-falling pitch. In this work, the voice disorders which have more than 30 voice recordings are considered for the assessment. Based on our discussion in Section 2, voice disorders are categorized into different categories, as shown in Table 1. In this study, for the assessment of voice disorders, the vowels /a/, /i/, and /u/ in normal, high, low and rising-falling pitch are considered.

Table 1: *Details of the voice disorders used in our experiment from SVD database. Here, SD: Spasmodic dysphonia, RLNP: Recurrent laryngeal nerve palsy, FD: Functional dysphonia, and PD: Psychogenic dysphonia.*

| Disorder Type | | Disorder name | #Samples |
|---|---|---|---|
| Organic | Structural | Laryngitis | 30 |
| | | Leukoplakia | 41 |
| | | Polyp | 45 |
| | Neurogenic | SD | 30 |
| | | RLNP | 188 |
| Non-organic | Functional | FD | 254 |
| | Psychogenic | PD | 91 |

## 3.2. Extraction of excitation source evidences

The present study explored the glottal flow waveform, zero frequency filtered signal and linear prediction residual to estimate the source information that can characterize the functioning of vocal mechanism. In this regard, two state-of-art signal processing techniques, called quasi-closed-phase (QCP) analysis and zero frequency filtering techniques, have been used to extract the glottal flow waveform, and zero frequency filtered signal, respectively. Extraction of excitation source evidences is discussed as follows:

- **Quasi-Closed-Phase analysis** is a state-of-art technique to estimate the glottal flow waveform [29]. The QCP method is based on closed phase analysis in which vocal tract model was estimated from speech samples in closed phase of glottal cycle [30]. On the other hand, QCP estimates vocal tract model from all speech samples by using a weighted linear prediction analysis. The attenuated main excitation (AME) waveform was used as a weighting function attenuates the samples of open phase region compared to the close phase samples of glottal cycles, which results in the better estimate of vocal tract model. Finally, glottal flow waveform was estimated by inverse filtering the speech signal with vocal tract model.
- **Zero Frequency Filtering** is an epoch extraction technique in which speech signal is passed through a fourth order zero frequency resonator [31]. This process attenuates the higher order harmonics corresponding to vocal-tract system and emphasises the excitation source characteristics. The resonator output grows/decays with polynomial degree of order three [32]. The trend in the resonator response is removed by subtracting the local average of it. The trend removed response is referred to as zero frequency filtered (ZFF) signal. The positive to negative zero crossings in the ZFF signal are referred to as epoch locations. The knowledge of epoch locations was used to estimate the intonation features (discussed in Subsection 3.3.2). Moreover in this study, the ZFF signal is considered as an approximate of glottal waveform, and computed the cepstral features from ZFF signal as in [33].
- **Linear Prediction (LP) analysis** is a source-filter model separates the excitation source and vocal-tract system components from speech. The LP-coefficients and LP-residual obtained from the speech signal are considered as vocal-tract parameters and glottal excitation, respectively. To compute the LP-residual, $12^{th}$ order LP-analysis was performed on speech segments with 20 $ms$ frame size and 10 $ms$ frame shift. In this study, LP-residual is computed to obtain the excitation source features by using the cepstral analysis.

## 3.3. Feature Extraction

### *3.3.1. Glottal features*

The glottal flow waveform which is estimated from the QCP analysis (discussed in Subsection 3.2) is used to compute glottal parameters as in [34]. The glottal parameters include time-domain features, namely open quotients (OQ1, OQ2), close quotient (CQ,CQa), speed quotient (SQ1,SQ2), and amplitude quotients (AQ, NAQ, QOQ). Amplitude difference of 1st and 2nd glottal harmonics (H1-H2), parabolic spectral parameter (PSP), and harmonic richness factor (HRF) are frequency domain parameters calculated from the spectrum of the glottal waveform to obtain voice quality in spectral domain. A 192-dimensional glottal-feature vector (16*12=192) is obtained for each utterance in the database by computing the 16 statistics of

12-dimensional glottal parameters. More details of the glottal features can be observed in [16, 34].

#### *3.3.2. Intonation feature*

Knowledge of epoch locations is important to obtain the perturbation measures corresponding to the vocal fold vibration. In this work, epoch locations are obtained from speech using zero-phase zero frequency filtering technique [32]. This study used the epoch locations to find fundamental frequency (F0) contour, strength of excitation (SoE) contour, and energy of excitation (EoE) contour. The F0, SoE, and EoE contours have been used to obtain 76 dimensional feature vector which is referred to as intonation feature vector (as in [35]) in this work. Intonation feature vector includes 5 statistics of F0, 66 perturbation parameters (22 jitter parameters of F0, 22 shimmer parameters of SoE, and 22 shimmer parameters of EoE), 4 harmonic-to-noise-ratio parameters and a pitch perturbation entropy measure.

#### *3.3.3. Mel frequency cepstral coefficients of LP-residual, and ZFF signal*

The studies in [33], revealed that Mel frequency cepstral coefficients of excitation source components are useful to identify the phonation type. Hence this study explored Mel frequency cepstral coefficients of LP-residual (MFCC-Residual) and ZFF signal (MFCC-ZFF) for the assessment of voice disorders. The MFCC-residual and MFCC-ZFF features were obtained from segments of LP-residual and ZFF signal, respectively with a frame-length of 20 ms and a frame-shift of 5 ms. They are 39 dimensional cepstral coefficients consisting in 13 static coefficients, and their first and second order derivatives. Finally, 4 statistics, namely mean, standard deviation, kurtosis and skewness were calculated, resulting 156 dimensional MFCC-residual and MFCC-ZFF feature vectors.

#### *3.3.4. Vocal-tract system features*

Conventional Mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) coefficients obtained from speech, capture the vocal-tract system information. In this work, the MFCC and PLP features are computed using speech segments of 20 ms frame size with a 5 ms frame shift. First 13 dimensional static features and corresponding delta, and delta-delta features were computed resulting in 39-dimensional MFCC and PLP features. Further, 156 dimensional MFCC and PLP feature vectors obtained by computing the statistics as discussed in Subsection 3.3.3.

#### *3.3.5. OpenSMILE features*

The OpenSMILE is publicly available toolkit for audio and music application designed for extracting acoustic features [36]. In our experiment, two feature sets of this toolkit are used, namely ComParE feature set [37] and eGeMAPS feature set [38]. The 2013 Interspeech Computational Paralinguistics Challenge features set (ComParE) is a brute-forced acoustic feature set contains 6373 features, whereas, extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) is a 88-dimensional features vector designed to extract paralinguistic information from speech.

Details about the feature extraction and corresponding MATLAB implementations are provided in the following link: `https://github.com/gurugubelllik/IS20-Supporting-material.git`.

### 3.4. Classifier

Support vector machine (SVM) classifier is the most widely used classifier in voice disorder detection as it gives consistence performance even on small dataset [39]. The present study used the SVM classifier for the detection and assessment of voice disorders. This study performed classification by using different other classifiers like decision tree, logistic regression, k-nearest neighbour, and ensemble classifier. Among all these classifiers, SVM classifier outperforms for most of the tasks. Moreover, different kernel functions like linear, polynomial and radial basis functions were also explored. Best performance was observed with polynomial kernel of order 2. Further, grid search approach is explored to select best parameters for quadratic kernel. In this regard, kernel parameter (box constraint level) is changed from 0.1 to 1000 with multiples of 10 and the kernel parameters for which the classifier has the best classification accuracy are considered for further analysis. The experiments were conducted with five fold cross validation and the average classification accuracy of all folds is referred to as the performance of the system.

## 4. Results and discussion

The main objective of this work is to assess the voice disorders in clinical approach. This study explored the excitation source features (MFCC-Residual, MFCC-ZFF, Glottal, and Intonation features) for the assessment of voice disorders and compared their performance with baseline features, namely vocal-tract system features (MFCC and PLP) and OpenSMILE features (ComParE and eGeMAPS) discussed in Subsection 3.3. In this regard, classification systems for the detection and assessment of voice disorders are developed by using support vector machine classifier (discussed in Subsection 3.4) with individual excitation source feature sets and baseline feature sets. Further, to investigate the complementary nature of excitation source features and baseline feature sets, experiments have been performed using combinations of feature sets.

Table 2: *Performance of voice disorder detection and assessment systems in terms of classification accuracy (in %) for individual feature set on SVD database. Here, Exp. 1: classification of healthy and voice disorder, Exp. 2: classification of organic and non-organic voice disorders, Exp. 3: classification of structural and neurogenic voice disorders, and Exp. 4: classification of functional and psychogenic voice disorders.*

| Feature type | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 |
|---|---|---|---|---|
| ComParE | **82.8** | 71.7 | **74.3** | **65.3** |
| eGeMAPS | 76.0 | 70.1 | 67.3 | 57.5 |
| MFCC | 74.4 | 72.4 | 67.8 | 63.4 |
| PLP | 74.2 | **72.7** | 70.5 | 64.1 |
| Glottal | 67.4 | 64.8 | 59.9 | 58.3 |
| Intonation | 69.3 | 66.0 | 60.2 | 52.8 |
| MFCC-Residual | 67.4 | 70.8 | 64.3 | 61.0 |
| MFCC-ZFF | 68.5 | 69.2 | 66.4 | 64.2 |

In this study, five fold cross validation is used so that the recordings correspond to 80% and 20% of total speakers were used as training and testing data, respectively. A total of four experiments were conducted in speaker independent approach using SVD database (discussed in Subsection 3.1). Experiment 1 is carried out to classify voice disorders from healthy class. Experiment 2 classifies organic voice disorder from non-organic

Table 3: *Performance of voice disorder detection and assessment systems in terms of classification accuracy (in %) for combination of feature sets on SVD database. Here, Exp. 1: classification of healthy and voice disorder, Exp. 2: classification of organic and non-organic voice disorders, Exp. 3: classification of structural and neurogenic voice disorders, and Exp. 4: classification of functional and psychogenic voice disorders.*

| Feature type | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 |
|---|---|---|---|---|
| Glottal + ComParE | **85.2** | 72.7 | 73.1 | 59.2 |
| Glottal + eGeMAPS | 79.0 | 70.8 | 65.5 | 60.1 |
| Glottal + MFCC | 74.4 | 71.2 | 66.7 | 64.1 |
| Glottal + PLP | 78.0 | 71.5 | 67.8 | 63.0 |
| Intonation + ComParE | 84.9 | 72.8 | 74.9 | 60.3 |
| Intonation + eGeMAPS | 81.5 | 68.5 | 68.1 | 60.1 |
| Intonation + MFCC | 77.5 | **75.0** | 65.2 | 64.4 |
| Intonation + PLP | 77.6 | 72.7 | 69.3 | 62.4 |
| MFCC-Residual + ComParE | 84.1 | 73.0 | **76.0** | 65.0 |
| MFCC-Residual + eGeMAPS | 84.3 | 70.9 | 62.6 | 63.3 |
| MFCC-Residual + MFCC | 73.1 | 74.6 | 69.6 | 66.2 |
| MFCC-residual + PLP | 74.2 | 73.0 | 68.4 | 65.3 |
| MFCC-ZFF + ComParE | 84.5 | 72.3 | 74.0 | 67.3 |
| MFCC-ZFF + eGeMAPS | 84.3 | 71.8 | 67.5 | 62.1 |
| MFCC-ZFF + MFCC | 71.7 | 72.3 | 68.7 | 63.6 |
| MFCC-ZFF + PLP | 74.4 | 70.1 | 70.5 | 65.9 |
| Glottal + Intonation + MFCC-Residual + MFCC-ZFF | 75.6 | 72.4 | 67.0 | **70.0** |

voice disorder. In our 3rd experiment structural disorder is classified from neurogenic disorder. Finally in experiment 4, psychogenic disorder is classified from functional disorder. In all the experiments, binary classification systems are trained with different feature sets and corresponding results are tabulated in Table 2 and Table 3.

From Table 2, it is observed that among all individual features sets ComParE feature set shows best performance in experiments 1, 3 and 4. PLP feature produced best performance in experiment 2 than all other individual features. The source feature sets, namely MFCC-ZFF (69.2% and 64.2%) and MFCC-Residual (70.8% and 61%) shown comparable results in experiments 2 and 4. However, in experiment 1 performance of excitation source features was shown to be lower than the baseline features. Further, the performance of voice disorder detection and assessment systems with the combination of features sets can be observed from Table 3. In voice disorder detection, ComParE with glottal feature combination produced best classification accuracy of 85.2%. Intonation features with MFCC, MFCC-Residual with ComParE, and combination of all excitation source feature sets produced best classification accuracies 75%, 76% and 70% in experiments 2, 3 and 4, respectively.

Among all the source features intonation features gave best classification accuracy of 69.3% for voice disorder detection. From this it is anticipated that perturbation parameters captures voice disorder information in better way. On the other hand, MFCC-Residual features shown better performance than other source related features in the classification of organic and non-organic voice disorders. For experiment 4, when all source features were combined, functional and psychogenic voice disorder classification system outperforms with a classification accuracy of 70%. Among the individual features, ComParE feature set shown best performance in most of the experiments. However it is brute-forced acoustic feature set which has very high dimension (6373) compared to the other feature sets. In most of the experiments, combination of baseline features (ComParE, eGeMAPS, PLP and MFCC feature sets) with excitation source feature sets shown significant improvement in performance of assessment systems trained with individual baseline feature sets. It indicates that excitation source features capture the complementary information about voice disorders compare to baseline features. Results of the present study reveals that the detection of voice disorder has a higher classification accuracy compared with the assessment of voice disorders. Moreover, the classification of functional and psychogenic voice disorders is more challenging compared to classification of structural and neurogenic voice disorders.

## 5. Summary and conclusion

This study proposed a multi-level classification approach using excitation source features for automatic detection and assessment of voice disorders from a clinical perspective. More detailed analysis of voice disorders was performed to know whether the disorder is structural, neurogenic, functional or psychogenic. All the experiments were performed on Saarbruecken voice disorder database using support vector machine classifier. Excitation source feature used in this experiments are intonation features, glottal features, MFCC-Residual and MFCC-ZFf. Excitation source features were compared with state-of-art MFCC, PLP, ComParE and eGeMAPs features. Results of experiments showed that OpenSMILE ComParE features outperformed for voice disorder detection system, and the combination of all excitation source features showed comparable performance in terms of classification accuracy. Experiments showed that excitation source features when combined with baseline features improved the performance of detection and assessment systems, indicates the complementary nature of source features. It was also observed that classification of functional from psychogenic class was more challenging.

In our future work, we intended to analyze the features which can improve the performance and efficiently classify voice disorder with more specific details, so it will become easy from clinical point of view. Moreover, we want to explore different neural network approaches for detection and assessment of voice disorders.

# 6. References


[1] A. E. Aronson, "Clinical voice disorders," *An interdisciplinary approach*, 1985.

[2] R. D. Kent, "Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders," *American Journal of Speech-Language Pathology*, vol. 5, no. 3, pp. 7–23, 1996.

[3] J.-W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, "An investigation of vocal tract characteristics for acoustic discrimination of pathological voices," *BioMed Research International*, vol. 2013, 2013.

[4] M. Frohlich, D. Michaelis, and H. W. Strube, "Acoustic" breathiness measures" in the description of pathologic voices," in *Proc. ICASSP*, vol. 2. IEEE, 1998, pp. 937–940.

[5] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[6] L. Eskenazi, D. G. Childers, and D. M. Hicks, "Acoustic correlates of vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 298–306, 1990.

[7] J. Laver, S. Hiller, and J. M. Beck, "Acoustic waveform perturbations and voice disorders," *Journal of Voice*, vol. 6, no. 2, pp. 115–126, 1992.

[8] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–9, 2009.

[9] M. Vasilakis and Y. Stylianou, "Spectral jitter modeling and estimation," *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 183–193, 2009.

[10] G. d. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[11] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 537–543, 1997.

[12] J.-W. Lee, S. Kim, and H.-G. Kang, "Detecting pathological speech using contour modeling of harmonic-to-noise ratio," in *Proc. ICASSP*. IEEE, 2014, pp. 5969–5973.

[13] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2218–2224, 1990.

[14] Y. Qi, R. E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *The Journal of the Acoustical Society of America*, vol. 105, no. 4, pp. 2532–2535, 1999.

[15] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of speech, language, and hearing research*, vol. 43, no. 2, pp. 469–485, 2000.

[16] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2020.

[17] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2010.

[18] J. I. Godino-Llorente, S. Aguilera-Navarro, and P. Gómez-Vilda, "Lpc, lpcc and mfcc parameterisation applied to the detection of voice impairments," in *Sixth International Conference on Spoken Language Processing*, 2000, pp. 965–968.

[19] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. F. Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2017.

[20] F. Rubén, G.-L. Juan, Ignacio, S.-L. Nicolás, O.-R. Víctor, and M. G.-A. Juana, "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: sustained vowels and running speech," *Journal of Voice*, vol. 27, no. 1, pp. 11–23, 2013.

[21] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634–641, 2019.

[22] L. Verde, G. De Pietro, and G. Sannino, "Voice disorder identification by using machine learning techniques," *IEEE Access*, vol. 6, pp. 16 246–16 255, 2018.

[23] M. Jičínskỳ and J. Mareš, "Measurable changes of voice after voice disorder treatment," in *Proceedings of the Computational Methods in Systems and Software*. Springer, 2019, pp. 295–305.

[24] B. Woldert-Jokisz, "Saarbruecken voice database," 2007.

[25] American Speech-Language-Hearing Association and others, "Definitions of communication disorders and variations," 1993.

[26] E. Seifert and J. Kollbrunner, "An update in thinking about nonorganic voice disorders," *Archives of Otolaryngology–Head & Neck Surgery*, vol. 132, no. 10, pp. 1128–1132, 2006.

[27] C. A. Rosen and T. Murry, "Nomenclature of voice disorders and vocal pathology." *Otolaryngologic Clinics of North America*, vol. 33, no. 5, pp. 1035–1046, 2000.

[28] M. D. Morrison, H. Nichol, and L. A. Rammage, "Diagnostic criteria in functional dysphonia," *The Laryngoscope*, vol. 96, no. 1, pp. 1–8, 1986.

[29] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 596–607, 2013.

[30] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[31] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[32] K. Gurugubelli and A. K. Vuppala, "Stable implementation of zero frequency filtering of speech signals for efficient epoch extraction," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1310–1314, 2019.

[33] S. R. Kadiri, P. Alku *et al.*, "Mel-frequency cepstral coefficients of voice source waveforms for classification of phonation types in speech," *Proc. INTERSPEECH*, pp. 2508–2512, 2019.

[34] P. Alku, "Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[35] M. H. Javid, K. Gurugubelli, and A. K. Vuppala, "Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech," in *Proc. ICASSP*, 2020, pp. 6754–6758.

[36] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[37] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, 2013, pp. 148–152.

[38] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[39] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *Journal of Voice*, vol. 33, no. 6, pp. 947.e11–947.e33, 2018.