Computer-based Blind Diagnostic System for Classification of Healthy and Disordered Voices

Zulfiqar Ali
School of Computer Science and Electronic Engineering
University of Essex
Colchester, United Kingdom.
z.ali@essex.ac.uk

Alba G. Seco De Herrera
School of Computer Science and Electronic Engineering
University of Essex
Colchester, United Kingdom.
alba.garcia@essex.ac.uk

Tamer A. Mesallam
Department of Otolaryngology, Head and Neck Surgery
College of Medicine
King Saud University
Riyadh, Saudi Arabia.
tmesallam@ksu.edu.sa

Ghulam Muhammad
Department of Computer Engineering
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia.
ghulam@ksu.edu.sa

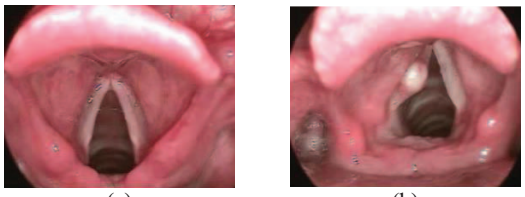
***Abstract*—A large population around the world is suffering from voice-related complications. Computer-based voice disorder detection systems can play a substantial role in the early detection of voice disorders by providing complementary information to early-career otolaryngologists and general practitioners. However, various studies have concluded that the recording environment of voice samples affects disorder detection. This influence of the recording environment is a major obstacle in developing such systems when a local voice disorder database is not available. In addition, sometimes the number of samples is not sufficient for training the system. To overcome these issues, a blind detection system for voice disorders is designed and implemented in this study. Hence, without any prior knowledge of voice disorders, the proposed system has the ability to detect those disorders. The developed system relies only on healthy voice samples which can be recorded locally in the desired environment. The generation of a reference model for healthy subjects and decision criteria to detect voice disorders are two major tasks in the proposed systems. These tasks are implemented with two different types of speech features. Moreover, the unsupervised reference model is created by using DBSCAN and k-means algorithms. The overall performance of the system is 74.9% in terms of the geometric mean of sensitivity and specificity. The results of the proposed system are encouraging and better than the performance of Multidimensional Voice Program (MDVP) parameters which are widely used for disorder assessment by otolaryngologists in clinics.**

***Keywords*—Blind voice disease detection, judgment reference model, vocal fold disorders, DBSCAN, objective analysis, unsupervised learning.**

I. INTRODUCTION

The air pressure generated by the lungs causes the vocal folds to vibrate for producing the voice. Then, this voice travels through the mouth and becomes sound after the application of oral cavities [1]. The voice of personnel is considered to be healthy if they can meet their personal and professional requirements without facing any fatigue and vocal problems [2].

Vocal folds open and close at regular intervals during phonation for generating a healthy voice. However, due to abnormal growth of tissues on their surface or injury to nerves controlling them, they exhibit irregular vibrations. Consequently, the voice becomes strained and harsh due to the tight closure of vocal folds, and sometimes the excessive distance between them makes the voice breathy, weaker, and whispering [3]. The abnormal growths of vocal folds and injury to nerves are known as voice pathologies or vocal folds disorders. Some common types of voice disorders are vocal folds nodules, cysts, polyps, paralysis, and sulcus. Normally, they appeared due to poor hydration, alcohol consumption, smoking, and vocal misuse including screaming, and excessive talking. Healthy and disordered vocal folds (suffering from vocal folds cysts) are shown in Fig. 1.



(a) (b)

Fig. 1. Vocal folds (a) healthy (b) suffering from vocal folds cysts [4].

Subjective evaluation using different rating scales such as Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) is a common practice of clinicians to assess voice disorders [5-7]. However, human error, attention, memory lapses of raters, interpretation of rating scales, experience, and knowledge of the clinicians may affect this way of evaluation [8, 9]. On the other hand, objective evaluation using computer-based diagnostic (CBD) systems is a non-invasive approach and independent of human bias.

Many CBD systems have been developed for the detection of voice disorders [10-14]. Such systems can play a significant role in the reliable detection of voice disorders by providing complementary information to otolaryngologists. In addition, the CBD system can detect voice disorders at an early stage as some cancerous disorders like keratosis become life-threatening if they are not treated on time. However, the reported results of developed voice disorder detection systems vary from one database to another even if the same set of features and machine learning algorithms are implemented. For instance, in the study conducted by Ali et al. [15], Mel-frequency Cepstral Coefficients (MFCC) are extracted from voice samples of three voice disorder databases: Massachusetts Eye and Ear Infirmary (MEEI) database [16], Saarbrücken Voice Database (SVD) [17], and Arabic Voice

Pathology Database (AVPD) [18]. Then, several experiments are conducted for the classification of healthy and disordered subjects. The respective best-obtained accuracies for these databases are 94.6%, 83.65%, and 80.2%. The accuracy of voice disorder detection for MEEI is 11% better than AVPD and 14% better than SVD. One of the potential reasons for MEEI's high accuracy is its different recording environments for healthy and disordered subjects as indicated by Sáenz-Lechón et al. [19] "*Normal and pathological voices were recorded at different locations (Kay Elemetrics and MEEI Voice and Speech Lab., respectively), assumedly under the same acoustic conditions, but there is no guarantee that this fact has no influence in an automatic detection system*." Due to the recording of healthy and disordered subjects at different locations, the corresponding sample of these subjects becomes easily differentiable. This is the reason that MEEI database yields higher accuracies for disorder detection as compared to other databases.

A similar trend is found by Al-nasheri et al. [20] where the best-obtained accuracy for MEEI database is 89% and that for AVPD and SVD is 70% and 68.5%, respectively. These accuracies are obtained with top-10 Multidimensional Voice Program (MDVP) parameters (out of 22). These 10 parameters are selected based on their Fisher discriminant ratio [21]. The list of all twenty-two MDVP parameters is provided in [22].

In [15] and [20], when cross-database experiments are conducted with MEEI, SVD, and AVPD, the results of voice disorder detection become worst, i.e., 47% to 82% using MFCC [15] and 38.89% to 70.27% using MDVP parameters [20]. Similarly, the best F1- score for the detection of voice disorders obtained by Harar et al. [14] is 0.733 (or 73.3%) using MFCC and MDVP parameters. The following different databases are used for the experiments MEEI, SVD, AVPD, and Príncipe de Asturias Database (PDA) [23]. One of the major factors for such varying accuracies is the recording environment of these databases [14].

The cross-database results signify that the generated models for healthy and disordered subjects using one database do not make good references for classification when testing is done with another database. Therefore, for good results, the training and testing samples should be recorded in the same environment. However, a CBD system for voice disorder cannot be developed for a community if a local voice disorder database is not available. Because if they use a database that is recorded somewhere else to train the system, the testing samples recorded in the local environment will not have a similar environment. Eventually, the diagnosis of voice disorder will be affected.

To overcome the unavailability of a voice disorder dataset, a first attempt is made to design and implement a blind detection system for voice disorders in this study. The proposed system will detect voice disorders without having any prior knowledge about them. The developed system only needs to be trained with samples of healthy people which can easily be recorded locally in the desired environment. In addition, this study will fill the gaps of unsupervised techniques in this area as no work has been reported for such systems [24]. The unsupervised reference model for healthy people is generated by using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and *k*-means algorithms [25, 26]. Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) features [27, 28] are extracted from the voice sample and given to the DBSCAN algorithm to identify the dense region of features. The other important task in the proposed blind detection system is the decision criteria to determine the class of test samples, and the fractal dimension (FD) of the voice samples is used as one of the measures in it [29].

The rest of the paper is organised as follows: Section II describes the pre-processing of speech signals, the extraction of two types of speech features, and the generation of an unsupervised reference model for healthy people. Section III explains the voice disorder dataset, the creation of decision criteria, and experimental results. Section IV analyses the proposed system for decision-making and compares it with other studies. Finally, Section V draws some conclusions.

II. Generation of Reference Model For Proposed System

The first objective in developing the proposed blind detection is the generation of a reference model using healthy subjects. The block diagram of the proposed system is depicted in Fig. 2 and each of its components is described in the following sections.

A. Feature Extraction: RASTA-PLP and Fractal Dimension

Speech rapidly changes over time, and therefore, its analysis becomes difficult. To make the speech fairly stationary, each signal S is divided into short frames f as expressed in Eq. (1).

$$S = \left[f_{i,n} \right] \tag{1}$$

where i represents the total number of frames, and it varies from one speech signal to another. In this study, a frame of length $n = 1024$ (~40 milliseconds) is used with an overlap of 50% with the previous.

Both ends of the frames are tapered closer to zero by applying hamming window. This process does not only exhibit the periodicity in the successive frames but also avoids spectral leakage after the application of Fast Fourier Transformation (FFT). The other important steps in the computation of RASTA-PLP are critical band analysis and inverse filtering to get the source signal. To estimate critical bands, the Bark scale is implemented which is linear up to 500 Hz and increased by 20% of the center frequency beyond it. This analysis simulates the human auditory system. Whereas the linear prediction (LP) analysis determines the formant structure and cancels its effect from the speech to get the source signal [10]. The LP analysis of the R^{th} order divides the vocal tract into R linear tubes. It means the current sample is estimated by R previous samples. This analysis mimics the human speech production system. Therefore, the extracted features simulate both human auditory and speech production systems.

In this study, 1024 points hamming window and FFT are applied on every frame. Twenty-four filters are used for critical band analysis and 11^{th}-order LP analysis is applied to extract twelve RASTA-PLP features from each frame f. The obtained set of features for a signal S is represented by $F_{i,j}$, where j represents the dimension of features which is equal to 12.

Another type of feature, FD, is extracted from each signal. It measures the complexity of a signal. Due to the presence of

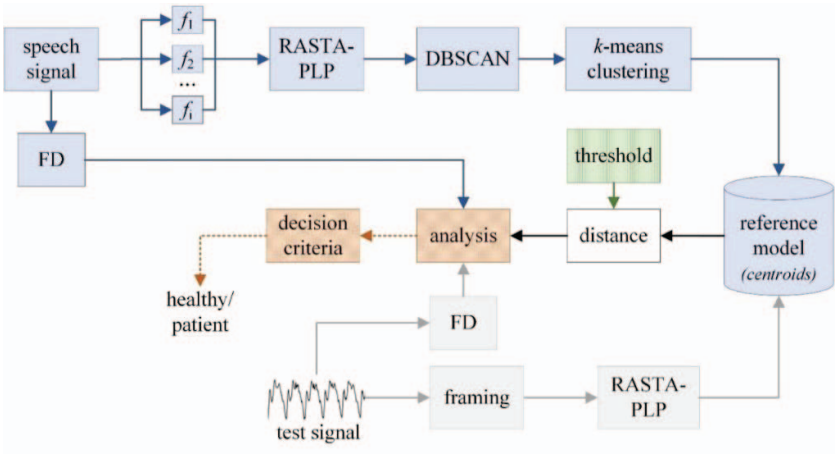


Fig. 2. Block diagram of the proposed system for blind detection of voice disorders, where RASTA-PLP and fractal dimension (FD) are two different types of extracted speech features. For the reference model, two clustering algorithms are implemented, i.e., DBSCAN and *k*-means. The proposed system detects voice disorders without having any prior knowledge about them.

voice pathology on vocal folds, they exhibit irregular vibrations during phonation. It makes the speech signal of a patient more complex/transient as compared to a healthy person. Katz's and Higuchi's algorithms are widely used to estimate FD [30, 31]. However, the Higuchi algorithm is not sensitive to amplitude which makes the KATZ algorithm a strong choice in this study. For instance, two synthetic signals with different amplitudes are shown in Fig. 3(a) and 3(b). The maximum amplitude in 3(a) is 10 and that in Fig. 3(b) is 5. FDs with the Higuchi algorithm are the same for both signals, whereas, they are different when computed using the KATZ algorithm.

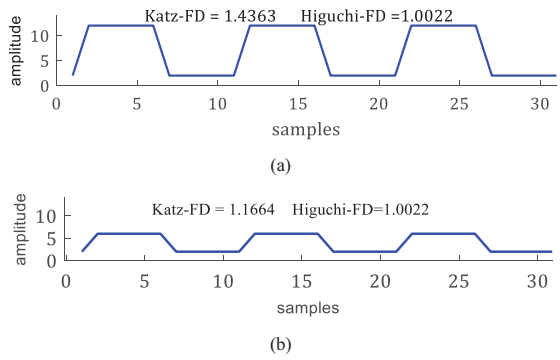


Fig. 3. Fractal dimension using two algorithms, KATZ and Higuchi, for two synthetic signals of different amplitude.

Some frequency bands are more discriminant in the classification of healthy and disordered signals. Especially, lower frequencies from 1-1562 Hz have shown good performance for disorder detection [10, 32]. The lower frequencies of speech are heavily source dependent due to the low-frequency glottal formant, while the higher frequencies are less dependent on the source signal. Therefore, before computing FD using KATZ algorithm, this frequency band is achieved by applying Discrete Wavelet Transformation [33]. To compute FD, only the first second of all voice samples is considered. These FDs of the 1-1562 Hz band of signals are used in the decision criteria.

Now, to generate the reference model, the extracted RASTA-PLP features $F_{i,j}$ are given to DBSCAN.

B. Identification of a Dense Region in Feature Space F

A dense region (DR) in feature space $F_{i,j}$ is determined using the DBSCAN algorithm to make sure that the generated reference model is a good representative of healthy subjects. The region is obtained by tunning two parameters of the algorithm: the number of minimum points clustered together for the region (*mPts*) and a threshold to locate the neighborhood points (ε).

In this study, a large value for the minimum points and a low value for the threshold, *mPts* = 1500 and ε = 0.005 are used to get the dense region. This region is represented by yellow 'x' in Fig. 4 and contains 46.5% of the total feature $F_{i,j}$. For visualisation, only two features (features 3 and 4) of all frames are depicted in Fig. 4.

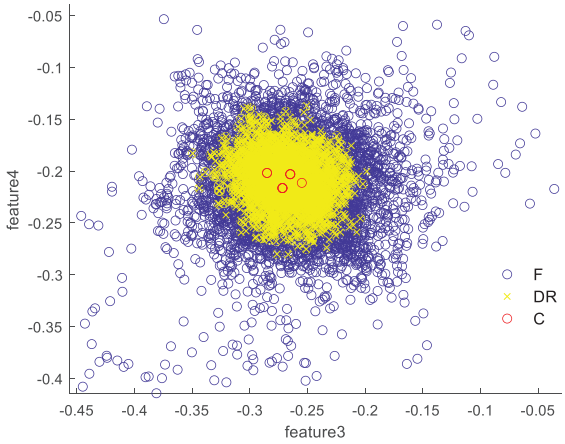


Fig. 4. Feature space *F* and the dense region DR which is further divided into four clusters having centroids C_k.

Now, a signal should be compared with the mean of the dense region (*mDR*) to determine the class, i.e., healthy or disordered. However, to use the entire dense region, DR is divided into four sub-regions using the *k*-means clustering algorithm where each resulting region is indicated by its mean

C1, C2, C3, and *C4*. These means are highlighted by red 'o' in Fig. 4. Eventually, the decision with four regions (C_k, k=1, 2, 3, 4) will be more reliable than using the single *mDR*. The centroids C_k is the desired reference model for the healthy voice samples.

III. Decision Criteria and Experimental Results

To develop the decision criteria, the AVPD database is divided into three partitions: training, validation, and testing. The database and its partitions are described in the following section.

A. AVPD database

The Computerized Speech Lab model 4500 (CSL 4500) was used to record both healthy and disordered subjects in the AVPD database. All subjects were recorded in a sound-treated room at the Communication and Swallowing Disorders Unit of King Abdulaziz University Hospital by expert clinicians. The samples were recorded at a bit rate of 16 bits with a sampling frequency of 48 kHz. The distance between the mouth and the microphone was kept constant at 15 cm for all recordings, which were then saved in two different audio formats. Five voice disorders: vocal fold cysts, nodules, paralysis, polyps, and sulcus were recorded in AVPD. These disorders fall under the category of organic voice disorders because they appeared due to abnormal growth of tissues on the vocal folds or injury to the nerves controlling the vocal folds.

In addition, all healthy subjects are recorded following clinical evaluation to confirm that they are healthy and do not suffer from any disorder in the past. Each subject signed a consent form to indicate their consent and to state that they had no problems to utilise their samples in research. Information about the individual's gender, age, and smoking habits was also obtained. Moreover, the perceptual severity of voice quality disorders was graded on a scale of 1 to 3, with 1 denoting mild, 2 denoting moderate, and 3 denoting severe disorders.

Each subject in the AVPD recorded a variety of texts. In this study, the sustained vowel /ah/ is used. Healthy samples are split into the 70% as training and 30% as test set and are denoted by P_{tr} and P_{ts}. P_{tr} does not contain any disordered samples. All disordered samples are in P_{ts}. For the tuning of the thresholds in the decision criteria, the training subset P_{tr} is further divided into two parts, P_T and P_V, where P_T contains 70% of P_{tr}'s samples and P_V consists of the remaining 30%. The distribution of the samples in these subsets is provided in Table I.

TABLE I. Distribution of Healthy and Disordered Samples in Three Partitions of the AVPD Database

Partitions	Samples		Total
	Healthy	***Disordered***	
Training subset P_T	59	-	59
Validation subset P_V	25	-	25
Testing Subset P_{ts}	36	97	133

B. Decision Criteria

The criteria to differentiate between normal and pathological signals are of prime importance in the proposed blind detection system. To develop the decision criteria, two questions need to be addressed. The first question is what distance from the reference model will declare a frame healthy. The second is how to declare a signal as healthy or disordered.

To answer the first question, the distortion of each frame of every healthy sample in the training partition P_T is computed with the generated model (C_k). The distortion (d_i) of the i^{th} frame of signal X is computed using Eq. (2).

$$d_i = \frac{1}{k}\sum_{k=1}^{4}\sqrt{\sum_j (fx_i - C_k)^2} \qquad (2)$$

where fx_i is RASTA-PLP of i^{th} frame of X, C_k are the centroids in the reference model, and the dimension of features and centroids is the same which is j.

Now, a threshold (*thresh*) on the distortion needs to be adjusted for deciding its class. To set *thresh*, the computed distortions of all signals in P_T are averaged and the resulting value is 0.1250. So, it implies that any frame having a distance less than *thresh*=0.1250 will declare as a member of the healthy class. A healthy voice sample with 402 frames is shown in Fig. 5. 61.6% of its frames are less than the *thresh* which indicates that they are closer to the generated reference model of the healthy subjects. This percentage of frames for a signal S is represented by pFrame[S].

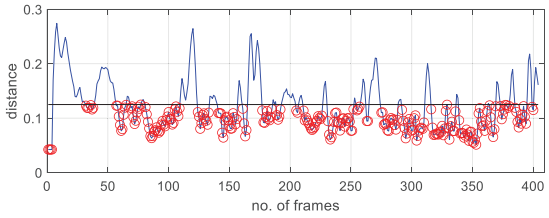


Fig. 5. Distortion of all frames of a healthy signal. Frames less than the *thresh* = 0.1250 are indicated by red 'o' and the horizontal black line is representing the *thresh*.

To answer the second question, the overall class of the signal will be determined using the percentage of the frame below the *thresh*. The threshold on pFrame is denoted by *thFrame* and initially set to 50%. Another measure, that is the FD of signals (FD[S]), is also used to find the class of the signal. The threshold on FD (*thDF*) is set to the average FD of all healthy signals in P_T which is 1.0015. The FD of the healthy signal shown in Fig. 5 is 1.0011. Finally, the standard deviation of distortions (i.e., STD[d_i]) of a signal is also considered. A threshold (*thSTD*) on it is set to the average of the standard deviation (STD) of all signals in P_T, i.e., *thSTD*=0.042. The STD of the distortion for the healthy signal in Fig. 5 is 0.451. The purpose of using these three measures is to make the decision reliable.

The next important task is tunning of the adjusted thresholds. To do it, the experiments are performed using the validation subset. This subset consists of healthy subjects as the generated model in the proposed system is also for healthy subjects only. In addition, there are no available criteria to differentiate between healthy and disordered samples at this stage. Healthy subjects are labelled as a negative class, therefore, specificity is used to measure the performance of the generated model. To get the optimal specificity, grid search is performed in the ranges 50%±5, 1.0014±0.002, and 0.042±0.04 for *thFrame*, *thFD*, and *thSTD*, respectively. The best specificity for the validated subset P_V is 80% and is obtained with *thFrame*=47%, *thFD*=1.0013, and

thSTD=0.0395. Therefore, these values of thresholds will be used in the decision criteria to differentiate between healthy and disordered samples.

The criteria to declare a sample 'healthy' is given by Eq. (3).

Condition 1: pFrame[S] > *thFrame* **AND** FD[S] < *thFD*

OR (3)

Condition 2: pFrame [S] > *thFrame* **AND** STD[d_i] > *thSTD*

All requirements of the proposed system have been accomplished for the blind detection of disorders such as the generation of the reference model, measures for decision criteria, and their optimal values.

C. Detection of Voice Disorders

To report the results for the test subset P_{tr}, three metrics are used. Disordered and healthy samples are labeled as positive and negative classes, respectively. Consequently, sensitivity (*sen*) is defined as the ratio between correctly identified disordered samples and the total number of disordered samples. Similarly, specificity (*spe*) is defined as the ratio between correctly classified healthy samples and the total number of healthy samples. For the overall performance of the system, accuracy (percentage of the total number of truly detected samples) is not used as imbalanced data affect this measure. As P_{tr} also contains different numbers of healthy and disordered samples. Therefore, geometric means (*GM*) of *sen* and *spe* is used for the overall performance of the system and is defined in Eq. (4).

$$GM = \sqrt{sen \times spe} \quad (4)$$

The performance of the proposed blind detection system is depicted in Fig. 6. The percentage values of *sen*, *spe*, *GM* are computed and listed.

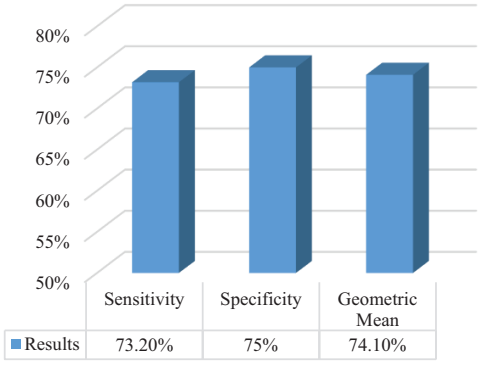


Fig. 6. The results of the proposed system for voice disorder detection over the test subset P_{ts}.

IV. Analysis and Comparisons

As shown in Fig.5, pFrame[S] is 61.6% which means that distortion d_i for the most of frames is less than the adjusted *thresh* = 0.1250. For a signal to be healthy, pFrame[S] should be more than 47% (the tunned threshold *thFrame*) which is satisfied. In addition, the FD of the signals FD[S] is 1.0011 which is less than 1.0013 (the tunned threshold *thFD*). It means that both requirements of condition 1 in the proposed criteria are fulfilled. Therefore, the signal belongs to the healthy class. This signal also satisfied condition 2 because STD(d_i) is 0.0451 which is higher that the *thSTD*. Similarly, in Fig. 6, pFrame[S] is 68.9% and FD[S] is 1.0011. According to condition 1 of the criteria, the corresponding signal also belongs to the healthy class.

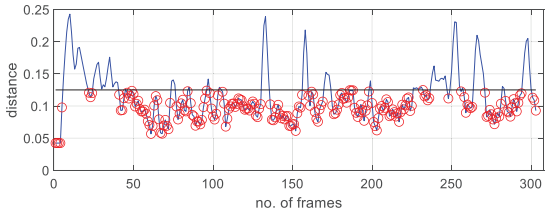


Fig. 6. Distortion of frames below the *thresh* (pFranme = 68.9%) for a healthy signal of the test subset P_{ts}. The fractal dimension of the signal is 1.0011.

For the signal in Fig. 7(a), pFrame[S] is 38.5% which means that most of the frames of the signal are away from the reference model of the healthy subjects and less than the *thframe*. In addition, the FD is 1.0019. Ultimately, both conditions in the criteria become false. Hence, the signal belongs to a disordered patient. Similarly, in Fig. 7(b) and (c), pFrame=27.2% and FD=1.0024, and pFrame=18.1% and FD=1.0020, respectively. Therefore, the corresponding signals are classified as disordered.

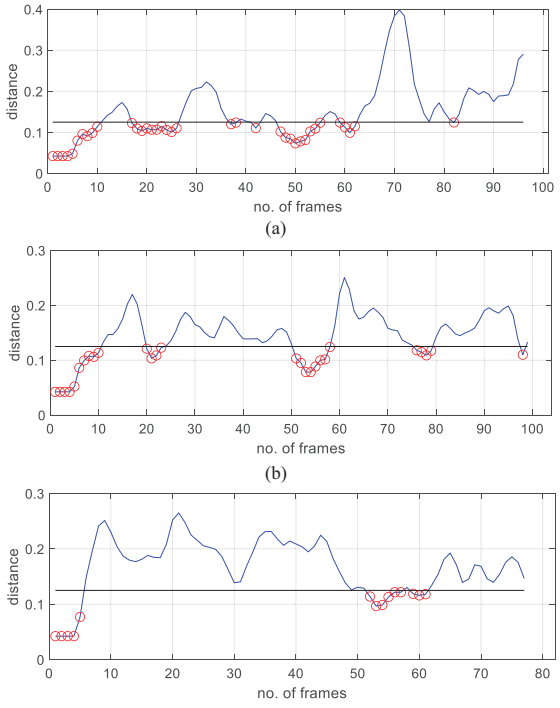


Fig. 7. Measures (percentage of frames below *thresh* and fractal dimension) for disordered signals on the test subset (a) pFrame=38.5% and FD=1.0019 (b) pFrame=27.7% and FD=1.0024 (c) pFrame=18.1% and FD=1.0020.

No CBD system exists for the blind detection of voice disorders. Unlike the proposed system, the existing systems need supervised training to generate reference models for healthy and disordered subjects. Although supervised models are used in [20, 34], the performance of the proposed system is still better than the MDVP parameters-based system where the best-obtained accuracy with the AVPD database is 72.5% for disorder detection [20, 34]. MDVP is a component of Computerized Speech Lab model 4500 (CSL 4500) which is widely used in the clinical evaluation of voice disorders. The

importance of CSL is evident by the fact that all three databases (MEEI, AVPD, and SVD) are recorded by using it.

V. Conclusion

The proposed blind detection system determines the presence of disorders by using the reference models of healthy subjects only. The computed distortions of the frames from the reference model play a significant role in the decision. The other positive aspect of the proposed system is the interpretation of its decision to differentiate between healthy and disordered samples. The developed system can be deployed for voice disorder detection in circumstances when disordered samples are not available or not sufficient to train a system. The proposed system can be enhanced by improving the decision criteria. For instance, adaptive thresholds for the measures can be used. In addition, it is also good to observe how many consecutive frames are below and above the threshold before making the final decision.

References

[1] L. Lopes, V. Vieira, and M. Behlau, "Performance of Different Acoustic Measures to Discriminate Individuals With and Without Voice Disorders," *Journal of Voice,* vol. 36, pp. 487-498, 2022.

[2] R. Jardim, S. M. Barreto, and A. Á. Assunção, "Voice Disorder: case definition and prevalence in teachers," *Revista Brasileira de Epidemiologia,* vol. 10, pp. 625-636, 2007.

[3] P. L. Dhingra and S. Dhingra, *Diseases of ear, nose and throat*, 6 ed.: Elsevier, India, 2014.

[4] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, M. Alsulaiman, and M. Bukhari, "Formant analysis in dysphonic patients and automatic Arabic digit speech recognition," *Biomed Eng Online,* vol. 10, p. 41, 2011.

[5] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich*, et al.*, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," *European Archives of Oto-Rhino-Laryngology,* vol. 258, pp. 77-82, 2001.

[6] P. Carding, E. Carlson, R. Epstein, L. Mathieson, and C. Shewell, "Formal perceptual evaluation of voice quality in the United Kingdom," *Logoped Phoniatr Vocol,* vol. 25, pp. 133-138, 2000.

[7] R. I. Zraick, G. B. Kempster, N. P. Connor, S. Thibeault, B. K. Klaben, Z. Bursac*, et al.*, "Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)," *Am J Speech Lang Pathol,* vol. 20, pp. 14-22, Feb 2011.

[8] I. V. Bele, "Reliability in Perceptual Analysis of Voice Quality," *Journal of Voice,* vol. 19, pp. 555-573, 2005.

[9] J. L. Sofranko and R. A. Prosek, "The Effect of Experience on Classification of Voice Quality," *Journal of Voice,* vol. 26, pp. 299-303, 2012.

[10] Z. Ali, G. Muhammad, and M. F. Alhamid, "An Automatic Health Monitoring System for Patients Suffering From Voice Complications in Smart Cities," *IEEE Access,* vol. 5, pp. 3900-3908, 2017.

[11] Z. Ali, M. Imran, and M. Shoaib, "An IoT-based smart healthcare system to detect dysphonia," *Neural Computing and Applications,* vol. 34, pp. 11255-11265, 2022.

[12] L.-C. Keung, K. Richardson, D. Sharp Matheron, and V. Martel-Sauvageau, "A Comparison of Healthy and Disordered Voices Using Multi-Dimensional Voice Program, Praat, and TF32," *Journal of Voice,* 2022.

[13] Z. Ali, M. Talha, and M. Alsulaiman, "A Practical Approach: Design and Implementation of a Healthcare Software for Screening of Dysphonic Patients," *IEEE Access* vol. 5, pp. 5844 - 5857, 2017.

[14] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal, "Towards robust voice pathology detection," *Neural Computing and Applications,* vol. 32, pp. 15747-15757, 2020.

[15] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, A. Al-nasheri, T. A. Mesallam*, et al.*, "Intra- and Inter-database Study for Arabic, English, and German Databases: Do Conventional Speech Features Detect Voice Pathology?," *Journal of Voice,* vol. 31, pp. 386.e1-386.e8, 2017.

[16] "Massachusetts Eye and Ear Infirmary Voice and Speech Lab Disordered Voice Database Model 4337 (Ver. 1.03)," ed. Boston, MA: Kay Elemetrics Corp. , 1994.

[17] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit," in *Advances in Speech and Language Technologies for Iberian Languages*. vol. 328, D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández*, et al.*, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 99-109.

[18] T. A. Mesallam, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-nasheri*, et al.*, "Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms," *Journal of Healthcare Engineering,* vol. 2017, p. 8783751, 2017.

[19] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control,* vol. 1, pp. 120-128, 2006.

[20] A. Al-nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat*, et al.*, "An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification," *Journal of Voice,* vol. 31, pp. 113.e9-113.e18, 2017.

[21] Kay Elemetric Corp., "Muti-Dimensional Voice Program (MDVP) Ver. 3.3," ed. Lincoln Park, NJ, 1993.

[22] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, and A. Moqarehzadeh, "Identification of voice disorders using long-time features and support vector machine with different feature reduction methods," *J Voice,* vol. 25, pp. e275-89, Nov 2011.

[23] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology,* vol. 36, pp. 60-69, 2011.

[24] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative Analysis of CNN and RNN for Voice Pathology Detection," *Biomed Res Int,* vol. 2021, p. 6635964, 2021.

[25] J. B. MacQueen, "Some Methods for Classification and Analysis of MultiVariate Observations," presented at the Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," presented at the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996.

[27] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Automatic Voice Pathology Detection With Running Speech by Using Estimation of Auditory Spectrum and Cepstral Coefficients Based on the All-Pole Model," *Journal of Voice,* vol. 30, pp. 757.e7-757.e19, 2016.

[28] M. Alsulaiman, G. Muhammad, and Z. Ali, "Classification of Vocal Fold Diseases Using RASTA-PLP," *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP),* p. 1, 2013.

[29] R. Lopes and N. Betrouni, "Fractal and multifractal analysis: A review," *Medical Image Analysis,* vol. 13, pp. 634-649, 2009.

[30] M. J. Katz, "Fractals and the analysis of waveforms," *Computers in Biology and Medicine,* vol. 18, pp. 145-156, 1988.

[31] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena,* vol. 31, pp. 277-283, 1988.

[32] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Detection of Voice Pathology using Fractal Dimension in a Multiresolution Analysis of Normal and Disordered Speech Signals," *J Med Syst,* vol. 40, p. 20, Jan 2016.

[33] M. H. Farouk, *Application of Wavelets in Speech Processing*: Springer, 2014.

[34] A. Al-nasheri, Z. Ali, G. Muhammad, and M. Alsulaiman, "An Investigation of MDVP Parameters for Voice Pathology Detection on Three Different Databases," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 2952-2956.