An Investigation of Multi-Dimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification

[1]Ahmed Al-nasheri, [1]Ghulam Muhammad, [1]Mansour Alsulaiman, [1,2]Zulfiqar Ali, [3]Tamer A. Mesallam, [3]Mohamed Farahat, [3]Khalid H. Malki and [1]Mohamed A. Bencherif

[1]Digital Speech Processing Group, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
a.alnashari@yahoo.com, {zuali, ghulam, msuliman}@ksu.edu.sa

[2]Centre for Intelligent Signal and Imaging Research (CISIR), Department of Electrical and Electronic Engineering Universiti Tekhnologi PETRONAS, Tronoh 31750, Perak, Malaysia

[3]ENT Department, College of Medicine, King Saud University, Riyadh, Saudi Arabia
[3]Research Chair of Voice, Swallowing, and Communication Disorders, King Saud University, Riyadh, Saudi Arabia

Summary

Background and Objective: Automatic voice-pathology detection and classification systems may help clinicians to detect the existence of any voice pathologies and the type of pathology from which patients suffer in the early stages. The main aim of this paper is to investigate Multidimensional Voice Program (MDVP) parameters to automatically detect and classify the voice pathologies in multiple databases, and then to find out which parameters performed well in these two processes.

Material and Methods: Samples of the sustained vowel /a/ of normal and pathological voices were extracted from three different databases, which have three voice pathologies in common. The selected databases in this study represent three distinct languages: (1) the Arabic voice pathology database; (2) the Massachusetts Eye and Ear Infirmary database (English database); and (3) the Saarbruecken Voice Database (German database). A computerized speech lab program was used to extract MDVP parameters as features, and an acoustical analysis was performed. The Fisher discrimination ratio was applied to rank the parameters. A t test was performed to highlight any significant differences in the means of the normal and pathological samples.

Results: The experimental results demonstrate a clear difference in the performance of the MDVP parameters using these databases. The highly ranked parameters also differed from one database to another. The best accuracies were obtained by using the three highest ranked MDVP parameters arranged according to the Fisher discrimination ratio: these accuracies were 99.68%, 88.21%, and 72.53% for the Saarbruecken Voice Database, the Massachusetts Eye and Ear Infirmary database, and the Arabic voice pathology database, respectively.

Key Words: MDVP parameters; AVPD; SVD; MEEI; SVM;

Introduction

Voice pathologies affect the vocal folds, producing irregular vibrations due to the malfunctioning of many factors contributing to vocal vibrations. Vocal fold pathologies exhibit variations in the vibratory cycle of the vocal folds due to their incomplete closure. Voice disorders also affect the shape of the vocal tract (supra-glottal) and produce irregularities in spectral properties [1]. In addition, voice pathologies affect vocal-fold vibration differently depending on the type of disorder and the location of the disease in the vocal folds, making them produce different basic tones.

The number of dysphonic patients has increased significantly. In the United States, approximately 7.5 million people have vocal difficulties [2]. It has been found that 15% of all visitors to King Abdul Aziz University Hospital, Riyadh complain of a voice disorder [3]. The impact of voice problems on teaching professionals is significantly greater than for non-teaching professionals. Studies revealed that, in the United States, the prevalence of voice pathologies during a lifetime is 57.7% for teachers and 28.8% for non-teachers [4]. Approximately 33% of male and female teachers in the Riyadh area suffer from voice pathologies [5]. The Communication and Swallowing Disorders Unit, King Abdul Aziz University Hospital, examines a high volume of voice disorder cases (almost 760 cases per annum) in individuals with various professional and etiological backgrounds. The use of computers to detect or identify pathological problems in speech, a non-invasive method, is advancing over time. In the last decade, much research has been done on the automatic detection of vocal-fold pathologies, which continues to require further investigation due to the lack of standard automatic diagnostic approaches/equipment for voice pathologies. Detection of pathology is the first crucial step to correctly diagnose and manage voice pathologies. Objective assessment, including acoustical analysis, is independent of human bias and can assist clinicians in making decisions. We firmly believe that clinicians have the final decision regarding medical diagnosis; objective assessment can only be used as an assistive tool. On the other hand, subjective measurement of voice quality is based on individual experience, which may vary. Automatic voice-pathology detection can be accomplished by various types of long-term and short-term signal analysis. Long-term parameters can be derived from acoustic analysis [6], [7] of speech, and short-term parameters can be calculated using linear predictive coefficients (LPC) [8], [9], linear predictive cepstral coefficients (LPCC) [10], Mel-frequency cepstral coefficients (MFCC) [11], [12], and so on [42]. Different pattern-matching techniques, such as a Gaussian mixture model (GMM) [13], [14], hidden Markov model (HMM) [15], support vector machine (SVM) [16], artificial neural networks (ANN) [17], and so on have been used to differentiate

between disordered and normal samples. Multiple long-term acoustic features, namely, pitch, shimmer, jitter, APQ (amplitude perturbation quotient), PPQ (pitch perturbation quotient), HNR (harmonic-to-noise ratio), NNE (normalized noise energy), VTI (voice-turbulence index), SPI (soft-phonation index), FATR (frequency amplitude tremor), and glottal-to-noise excitation ratio (GNE) are frequently used to diagnose voice pathology (referenced in [14] as [2]-[12]). Furthermore, jitter and shimmer capture vocal-fold vibratory characteristics for both pathological and normal people, and both parameters are widely used for clinical research purposes [18]. Seven acoustic parameters, including shimmer and jitter, are extracted by means of an iterative residual-signal estimator in Rosa et al. [19], and jitter provided 54.8% accuracy of detection for 21 pathologies. Thirty-three different long-term acoustic parameters with their definitions, derived from the Multi-Dimensional Voice Program (MDVP) [31], are listed in Arjmandi et al. [20]. Twenty-two acoustic parameters were selected from the list extracted from voice samples in the Massachusetts Eye and Ear Infirmary (MEEI) database. Fifty dysphonic patients and 50 normal persons were used for detection. The 22 parameters were calculated for each sample and fed to six different classifiers to compare their accuracies. Two feature-reduction techniques were also used before applying classification methods. Binary classifier SVM showed the best results compared to other classifiers, with a recognition rate of 94.26%. In [21], MFCC and six acoustic parameters—jitter, shimmer, NHR, SPI, APQ, and RAP (Relative Average Perturbation)—were extracted, with the results compared to the NN-based voice pathology detection system [22]. Sáenz-Lechón et al. compared their proposed parameters based on wavelet transform with some of the MDVP parameters to discriminate between pathological and normal voices [23]. To ensure the reliability of the acoustic MDVP parameters, some of them were compared to the same parameters extracted using Praat; results showed no significant difference between the two computer software approaches [24]. Recently, MPEG-7 audio descriptors and multi-directional, regression-based features have been used in voice-pathology detection, with good accuracy [26, 27]. Another recent study investigated the most discriminative frequency region for voice-pathology detection [28]. In general, MDVP parameters are well able to discriminate between normal and pathological voices, as are other tools that are used to extract acoustic parameters, such as WPCVox [25].

In this paper, the well-known MDVP parameters are investigated in three different databases—(1) AVPD; (2) MEEI [29]; and (3) SVD [30]—to detect and classify voice pathology. MDVP parameters are commonly used by physicians/clinicians to assess voice pathology; however, MDVP is commercial software. The objective of this study is to investigate the capability of MDVP parameters to detect and classify voice pathologies in a cross-database scenario and to find out which of these parameters perform best in each individual database.

Material and Methods

Data

In this study, we used three different databases: (1) MEEI; (2) SVD; and (3) AVPD. We chose only three types of pathological voices—(1) vocal fold cyst; (2) unilateral vocal fold paralysis; and (3) vocal fold polyp—because only these pathologies are common in all three databases. We selected these three databases in our study due to the following reasons:

- MEEI database is one of the most popular databases in the field of voice pathology. It is considered as the basis of many studies with voice pathology assessment; however, it has some limitations as mentioned in subsection of MEEI database description. Therefore, for comparison purpose, we used it, but we did not solely rely on it.
- SVD is a German database that is free downloadable with rich variation of samples. This variation of samples makes possible to carry various type of experiments in different research purposes. Its use in voice pathology is very little.
- AVPD is our Arabic developed database and this is the first time involving it in research.
- Other used databases in many research are private and not available on the net.

Voiced signals can be seen in three types as qualitatively classified by Titze in [37]. Type 1 signals are nearly periodic, Type 2 signals contain strong modulations or bifurcations, while Type 3 signals are irregular and aperiodic. It has been suggested that traditional perturbation methods of voice signal analysis, such as jitter and shimmer, are appropriate only for Type 1 or Type 2 signals. For the MEEI database, some experiments are performed by excluding the Type 3 signals.

The number of samples in each database is shown in Table 1, where the number of male and female speakers are shown, respectively, inside parentheses. The three used databases are each described below.

Table 1: Normal and pathological samples from three different databases.

Database	Normal	Pathological			
		Cysts	Paralysis	Polyp	Total
AVPD	118 (93, 25)	13 (7, 6)	32 (16, 16)	30 (14, 16)	75
MEEI	53 (19, 34)	10 (6, 4)	66 (34, 32)	19 (8, 11)	95
SVD	262 (100, 162)	6 (1, 5)	195 (64, 131)	43 (25, 18)	244

Massachusetts Eye and Ear Infirmary (MEEI) Voice Disorder Database

This database was developed by the MEEI Voice and Speech Lab. It includes more than 1,400 voiced samples of sustained vowel /a/ and the first part of the Rainbow Passage. It is commercialized by Kay Elemetrics [29]. It was recorded in two different environments. The sampling frequency for normal samples was 50kHz, while that of the pathological samples was 25kHz or 50kHz. It is used in most studies of voice-pathology detection and classification even though it has many disadvantages, such as the different environments and sample frequencies used to record normal and pathological voices. In this database, many tools were used to evaluate voice condition, including stroboscopy, acoustic aerodynamic measures, and a physical examination of neck and mouth (this information is provided by Kay Elemetrics). Many voice pathologies can be addressed by observing changes in the muscles of the voice that can activate and improve the efficiency of the voice. In the CD Kay Elemetrics provided, we filtered the filenames according to the three diseases; if there were multiple pathologies for a file or if there were missing MDVP parameters for a file, we ignored that file. For normal speakers, we selected all available 53 samples.

Saarbruecken Voice Database (SVD)

The SVD is a freely downloadable database [30], recorded by the Institute of Phonetics of Saarland University. This database contains sustained vowels /a/, /i/, and /u/ with different intonations: normal, low, high, and low-high-low, along with a spoken sentence in German “Guten Morgen, wie geht es Ihnen?” which means, in English, “Good morning, how are you?” These attributes make it a good database for researchers to conduct experiments. All recorded voices in the SVD database were sampled at 50 kHz with

16-bit resolution. This database is new; very few studies of voice-pathology detection have been done using it. We downloaded the files from the website mentioned in [30] using the criteria of the three diseases. We selected only sustained vowel /a/ samples produced at normal pitch.

Arabic Voice Pathology Database (AVPD)

The voice and speech samples in this database were collected in different sessions at the Communication and Swallowing Disorders Unit [4], King Abdul Aziz University Hospital, Riyadh, Saudi Arabia, by experienced phoniatricians in a sound-treated room using a standardized recording protocol. This database collection was one of the major tasks of the ongoing project funded by the National Plans for Science and Technology (NPST), Saudi Arabia, over the duration of two years. The protocol of the database was designed such that it should avoid various shortcomings of the MEEI database [23]. This database has recordings of sustained vowels as well as the speech of patients who have vocal-fold pathologies, along with the same recordings of persons with normal speech. Normal and pathological vocal folds were determined after clinical assessment using laryngeal stroboscopy. In case of pathology, the perceptual severity of voice disorders was rated on a scale of 1–3, where 3 represents the most severe case. A severity rating was associated with each sample based upon the consensus of a panel of three expert medical doctors. The recording has different types of texts: (1) three sustained vowels with onset and offset information; (2) isolated words including Arabic digits and some other common words; and (3) continuous speech. The selected text was carefully selected to cover all Arabic phonemes. All speakers recorded three utterances of each vowel /a/, /u/, and /i/, while isolated words and continuous speech were recorded once to avoid burdening patients. The sampling frequency in the database was 44 kHz, and the speech was recorded using the computerized speech lab (*CSL*) program. The voice disorders recorded in this database were evaluated and validated by different specialist doctors at King Abdul Aziz University Hospital. Among the recorded samples, only recordings of patients having vocal-fold cyst, vocal-fold polyp, and unilateral vocal-fold paralysis pathologies were included in this study. We selected only sustained vowel /a/ samples. We understand that acoustic analysis is vulnerable and there are many factors that affect its sensitivity as a diagnostic tool. However, in our study we tried our best to control these factor as much as possible by controlling the mic-mouth distance, sound-treated rooms.

Methods

A subset of the sustained vowel /a/ samples of normal and pathological voices were taken from these three databases. MDVP parameters for the selected samples were extracted using the Kay Pentax CSL Model

4300 program [31]. This software is the most commonly used and cited software for acoustic analysis. It can perform acoustic analysis based on 33 quantitative voice parameters, including fundamental frequency, amplitude, spectral energy, shimmer, jitter, the presence of any sonority gap, and other features. Indeed, this software has been widely used to perform this kind of analysis during the period between 1991 and 1995, as mentioned in [34]. In our study, we used only 22 of the possible parameters, those that have statistical significance, while the others were ignored because they did not reflect voice quality or they were not produced for some voices [20].

Many of the MDVP parameters do not correlate with the other voice assessment measures but not all of them. For example perturbation measures have been shown to correlate with certain voice problems that would affect more frequency-related parameters. That is why it could be a significant point to look for these parameters that may correlate much with the underling voice pathology and can give us more insight about its acoustic correlates. Moreover, some of the stroboscopic abnormalities can predict deviation in the acoustic analysis; for example aperiodicity, and asymmetry in mucosal waves of vocal folds vibration can predict abnormalities in perturbation measures of the acoustic analysis. This is the base on which we build our hypothesis that there might be certain correlation where the results of this study will prove or disprove. In literature, Eskenazi et al. [38] found out that the percent jitter presented a correlation of 0.55 with the traditional perceptual rating of breathiness. In addition, Shrivastav et al. [39] reported an improved correlation of 0.86 for the percent jitter parameter with ratings of breathiness. On the other hand, Marin et al. [40] discovered that there is a very bad correlation between percent jitter and breathiness ratings. Other investigations in different studies that reported the effectiveness of HNR measures as an acoustical correlate of pathological voice quality [38], [40], [41].

In this study, the Fisher discriminative ratio (FDR) was used between the two classes (normal and pathological) for all parameters in each database individually. The purpose of using FDR is to find which features better detect each pathology. FDR can be calculated as shown in (1).

$$FDR_i = \frac{(\mu_N - \mu_P)^2}{\sigma_N^2 + \sigma_P^2} \quad where \quad i = 1,2,3,.... 22 \qquad (1)$$

where μ_N and μ_P represent the mean for classes of normal and pathological samples, respectively, while σ_N and σ_P represent variances for each class, respectively, and *i* represents the feature index number. In the experiments, the features were fed to a support-vector machine classifier to make a decision about whether the subject was normal or pathological. Features were sorted in descending order based on FDR, and the top certain number of features were selected. In addition, we performed a *t*-test for the three highest features, ordered according to FDR. *T*-test is a statistical test that allows the comparison of means from two populations. We performed the *t*-test between two classes of normal and pathological samples on the three

different databases separately with the following null hypothesis: "there is no significant difference between the two classes." The *p*-value probability of the *t*-test contributes in making a decision about the null hypothesis and therefore to make ourselves more confident about our achieved accuracy of both classification and detection processes. If the *p*-value is high (greater than 0.05) then this indicates that the probability of the observed result is high and so we accept the null hypothesis at the 5% (0.05) significance level. On the other hand, if the *p*-value is low (less than 0.05) then this indicates that the probability of the observed result is low and we reject the null hypothesis. Consequently, we infer that there is a significant difference between two classes.

For classification, a 10-fold, cross-validation approach was utilized, in which the data were randomly divided into 10 groups. For each iteration, nine groups were used in training, while the remaining group was used in testing. After 10 iterations of this procedure, all groups were tested.

Results

The results of the performed experiments for pathology detection and classification are expressed in terms of accuracy (ACC: the ratio between correctly detected samples and the total number of samples), sensitivity (SN: the proportion of pathological samples that are positively identified), specificity (SP: the proportion of normal samples that are negatively identified), and the area under the Receiver Operating Characteristic (ROC) curve, called AUC. All of these are shown in Table 2. These terms can be calculated using the following distinct equations:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

$$SN = \frac{TP}{TP + FN} \qquad (3)$$

$$SP = \frac{TN}{FP + TN} \qquad (4)$$

where true negative (TN) means that the system detects a normal subject as a normal subject, true positive (TP) means that the system detects a pathological subject as a pathological subject, false negative (FN) means that the system detects a pathological subject as a normal subject, and false positive (FP) means that the system detects a normal subject as a pathological subject.

To verify the validity of the selected parameters for the detection and classification process of pathological samples that were extracted from the three different databases, various experiments were performed. For

example, in the case of the detection process, for each database, four different experiments were performed with various numbers of parameters. To ensure accuracy, every experiment was repeated ten times (10 folds and ten times, which equal to 100 runs), and then we reported the average. First, we selected the 22 parameters as used and defined in [20]. We performed the experiments with the 22 individual parameters from these databases. After that, we sorted the parameters in descending order by FDR. We chose the top 10 parameters by FDR and performed the experiments with these parameters. Next, we chose the three top parameters from each database and performed the experiments with these. To develop a general system independent of the databases, we chose the four most common parameters from among the top 10 parameters in each individual database and performed the experiment with these. *Table 2* shows the results of these experiments. The achieved accuracies varied from one database to another with the same number of MDVP parameters. The best achieved accuracy was 99.68% when using the top three parameters belonging to the SVD database.

Table 2: Results of using different MDVP parameters from three databases (pathology detection)

Parameters	Database	ACC%	SN%	SP%	AUC
22	AVPD	71.63	52.69	84.23	0.71
	MEEI	76.36	93.22	45.2	0.69
	SVD	72.58	62.17	82.4	0.72
Top -10	AVPD	70.42	51.34	83.34	0.68
	MEEI	89.71	92.17	86.27	0.89
	SVD	68.52	51.77	84.3	0.69
Top - 3	AVPD	72.53	49.89	86.83	0.68
	MEEI	88.21	90.63	84.83	0.88
	SVD	99.68	99.75	99.63	0.99
4 – Common	AVPD	71.16	43.09	88.85	0.68
	MEEI	81.71	80.52	84.97	0.82
	SVD	67.86	46.91	87.96	0.67

The top three features that gave high accuracy with the SVD were vAm (peak amplitude variation period-to-period), APQ (amplitude perturbation quotient), and PFR (phonatory fundamental frequency). vAm has an FDR value of 1.470, which is far greater than the FDR values of 0.538 and 0.510 for PFR and APQ, respectively. *Figure **1*** shows a scatter plot using these three parameters.

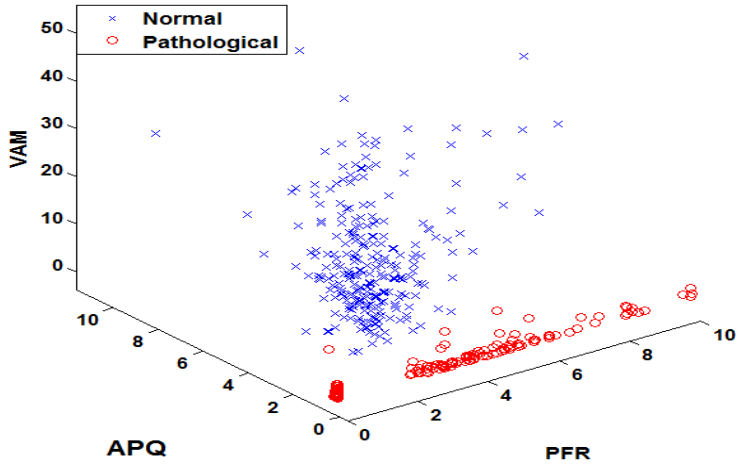

Figure 1: Scatter plot for the three top MDVP parameters of the SVD database

The scatter plot in *Figure 1* shows that the vAm parameter has a better ability to discriminate pathological samples than the other two parameters. The top three parameters extracted from the SVD database have better performance than the top three parameters extracted from the AVPD and MEEI databases. Figure 2 illustrates that the probability density functions (PDF) of normal and pathological samples using the vAm parameter have almost no overlap, further suggesting that it has the most capability to differentiate between normal and pathological voices. It is obvious that when there is more overlap between the probability density functions of normal and pathological samples, the ability to discriminate between normal and pathological sample will be reduced.

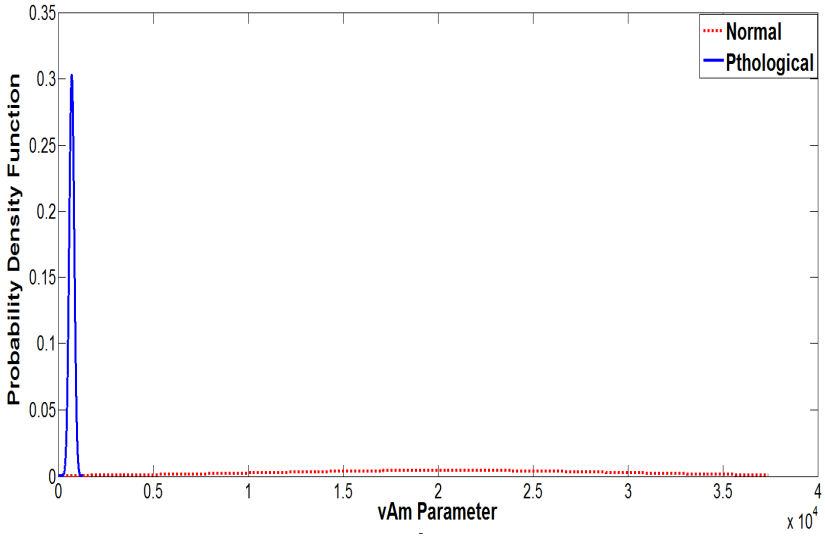

Figure 2: PDF for the vAm Parameter

Figure 3 shows the ROC curves of the top three parameters from each of the three databases. It demonstrates that the best performance is obtained with the features extracted from SVD. One of the reasons for this is that the pathological samples in the SVD database are highly severe, presenting a clear difference between normal and pathological samples. The 95% confidence interval is [0.9449 0.9870], and the one-tail p-value is zero (<0.05), describing the significance of the data in the two classes.

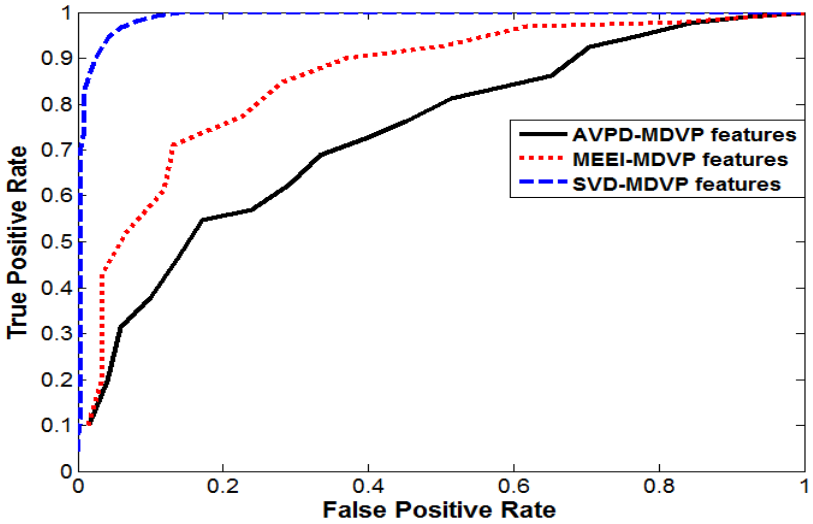

Figure 3: ROC curves for the top three features for the three databases

When using the four MDVP parameters in common between the three databases, accuracy decreased by 7% in the case of MEEI while the accuracies for the other two databases, AVD and AVPD, remained almost unchanged (with respect to the accuracies obtained by using 10 parameters). The reason for that is that

every database is independent of one another, and the four common parameters have different values of FDR for each database. The four common parameters between the three databases are Jitta, Jitt, Rap, and PPQ. Jitta is absolute jitter, evaluating in microseconds the period-to-period variability of pitch within an analyzed voice sample. Jitt evaluates the variability in percent of the pitch period within the analyzed voice sample. Rap, or relative average perturbation, evaluates the variability of the pitch period within the analyzed voice sample using a smoothing factor of three periods. Finally, PPQ, or pitch perturbation quotient, evaluates in percent the long-term variability of the pitch period within the analyzed voice sample using a smoothing factor of three periods.

To investigate the effect of the MDVP parameters using Type 1 and Type 2 signals, we performed several additional experiments on the MEEI database. We analyzed the samples of the MEEI database in our study, and we found that most of the samples are of signal Type 1 and Type 2, and the others are of signal type 3. We excluded all the samples that belong to signal type 3 as listed in Table 3, and some of them are plotted in Figure 4 using two periods. It can be observed from the Figure 4, the signals are strongly aperiodic. We repeated the experiments by using 22-parameters, top 10 parameters and the three parameters to see if there are any significant differences of performances between the previous experiments and these new experiments. We found that the detection accuracy improved by 12% compared with the achieved accuracy in the previous experiment using 22 parameters. However, using the top 10 parameters and the top three parameters, the accuracies remained almost the similar, because these parameters do not include jitter and its variants. So we conclude that the acoustic analysis of signals of type 3 with the MDVP measurements have negative effect on the accuracy of the detection process, if we include 22 parameters, and we will analyze the rest of the samples that belong to the SVD and the AVPD databases in the future work. *Table 4* shows the achieved accuracies after we excluded the mentioned samples in Table 3.

Table 3: Type 3 signal not included in the MEEI acoustic analysis

Files Name	
CAR10AN.NSP	JPP27AN.NSP
CTY09AN.NSP	JTG18AN.NSP
DVD19AN.NSP	JXS14AN.NSP
EDG19AN.NSP	MPB23AN.NSP
EJH24AN.NSP	PDO11AN.NSP
HWR04AN.NSP	RPJ15AN.NSP
JDO14AN.NSP	

Table 4: Performance measures of the MEEI database without Type 3 samples

Number of Parameters	SN%	SP%	ACC%
22-Parameters	88.909	89.214	88.692
10-Parameters	89.138	89.497	89.538
3 - Parameters	91.492	82.688	87.385

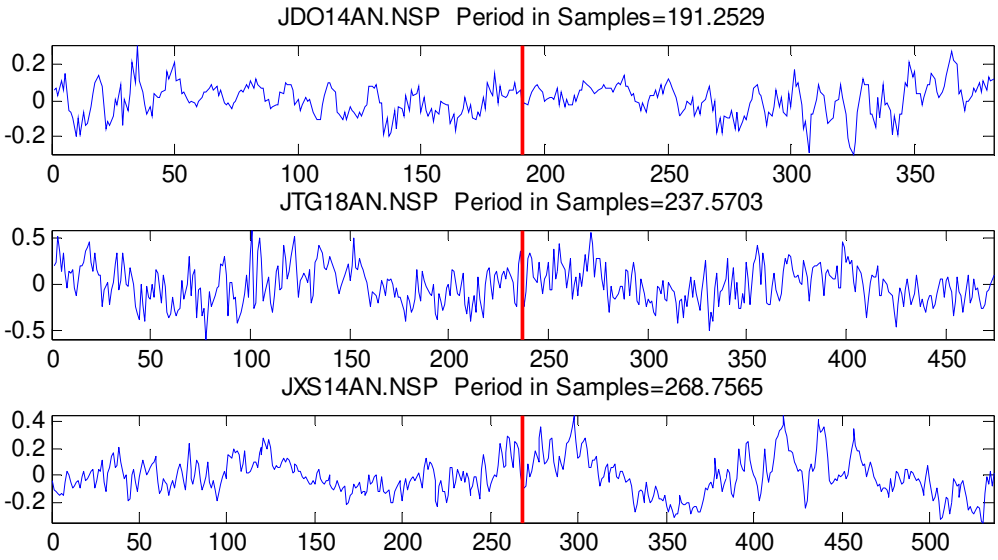

Figure 4: Examples of Type 3 signals in the MEEI database.

Some additional cross-database experiments were performed to make sure that the extracted MDVP parameters yield the same detection ability and to avoid unfairly over-fitting as a result of error estimation. For this, we used only the four parameters in common between the databases. First, we use one database to train and one to test. Then, we combined two database as trainers and used the other one as a test. The results of these experiments are expressed in terms of accuracy. *Table 5* shows the achieved accuracies, which varied from one database to another. The achieved accuracies were 70.27% when using the SVD database for training and MEEI for testing and 70.94% when using SVD for training and AVPD for testing. When we combined two databases for training and used the third one for testing, accuracies were reduced.

Table 5: Accuracies (%) of cross-database experiments with the four common parameters for pathological detection

Databases		Testing		
		MEEI	**SVD**	**AVPD**
Training	MEEI	-	48.22	38.89
	SVD	70.27	-	70.94
	AVPD	47.97	52.37	-
	MEEI + SVD	-	-	66.32
	SVD + AVPD	64.19	-	-
	MEEI + AVPD	-	65.81	-

We performed three distinct experiments on each database with 22 features to find the validity of the classification using MDVP parameters for each individual database. *Table 6* shows the three different classification experiments performed for each database. Three distinct voice pathologies were used. Two were used as a training set, while the third was used as a testing set. The best achieved accuracy was 97.5%, with the SVD database.

Table 6: Classification for three different databases

Classification type	Database accuracy (%)		
	AVPD	SVD	MEEI
Cyst vs (Paralysis & Polyp)	82.86	97.5	88.89
Paralysis vs (Cyst & Polyp)	57.14	79.17	65.56
Polyp vs (Cyst & Paralysis)	60	82.08	30

Moreover, cross-database experiments were performed on the voice pathology samples. *Table 7* shows the results when two databases were used as a training set while the other was used as a testing set. These results show that different pathologies had distinct accuracies in various combinations of training and testing sets.

Table 7: Cross-database experimental results (pathology classification)

Classification Type	Training DB	Testing DB	% Accuracy
Cyst vs (Paralysis & Polyp)	SVD+MEEI	AVPD	82.67% (62/75)
Paralysis vs (Cyst & Polyp)	SVD+MEEI	AVPD	57.33% (43/75)
Polyp vs (Cyst & Paralysis)	SVD+MEEI	AVPD	60.00% (45/75)
Cyst vs (Paralysis & Polyp)	AVPD+MEEI	SVD	97.54% (238/244)
Paralysis vs (Cyst & Polyp)	AVPD+MEEI	SVD	79.92% (195/244)
Polyp vs (Cyst & Paralysis)	AVPD+MEEI	SVD	82.38% (201/244)
Cyst vs (Paralysis & Polyp)	AVPD+SVD	MEEI	89.47% (85/95)
Paralysis vs (Cyst & Polyp)	AVPD+SVD	MEEI	30.53% (29/95)
Polyp vs (Cyst & Paralysis)	AVPD+SVD	MEEI	80.00% (76/95)

The highest achieved accuracy was 97.54% when we used the MEEI and AVPD databases as a training set and the SVD database as a testing set.

t-tests were performed for the three highest-ordered features in each database to determine whether the differences in the means between the two classes (normal and pathological samples) were significant for each feature. *Table 8* shows the results of this *t*-test for the highest-ordered three features, along with their p-value probability.

Table 8: t-test for the highest-ordered three features (pathology detection case)

Database	*Features*	*Mean (N-P)*	*sd (N-P)*	*t-value*	*df*	*p-value*
AVPD	STD	4.19 - 11.22	2.8-12	-3.76	79.17	0.00003
	Jitta	2.16 - 175.16	2.4-133.24	-9.22	74.03	0.00001
	vF0	2.24 - 6.16	1.7-6.7	-5.58	79.97	0.00001
SVD	PFR	2.60 - 5.53	1.23 - 4.40	-9.38	278.22	0.00001
	APQ	2.62 - 5.20	1.32 - 4.09	-9.35	289.52	0.00001
	Vam	17.74 - 21.81	9.67 - 11.14	-4.23	482.62	0.00003
MEEI	shim	2.21 - 8.72	0.92 - 5.29	-11.68	103.99	0.00001
	APQ	1.63 - 6.34	0.72 - 3.98	-11.23	325.72	0.00001
	sAPQ	2.64 - 7.58	1.16 - 4.61	-9.89	113.91	0.00001

Finally, we compared the results cross-database and performed a *t*-test with the 22 features to determine whether there are significant differences for all of the used features between the two classes (normal and

pathology). *Table 9* shows the *p*-values (at 95% confidence) of all 22 features taken from the three databases. From the listed *p*-values in this table, we can infer that not all MDVP parameters had significant differences in all type of the classification process. In addition, we can notice from this table that the performance of classification on SVD is better than the other two databases in all type of classification.

Table 9: p-values for the three databases with 22 features (pathology classification case)

MDVP	Databases								
	AVPD			SVD			MEEI		
	Classification Type			Classification Type			Classification Type		
	PRL vs ALL	PLP vs ALL	CYST vs ALL	PRL vs ALL	PLP vs ALL	CYST vs ALL	PRL vs ALL	PLP vs ALL	CYST vs ALL
Fo	0.50	0.98	0.31	0.00	0.02	0.03	0.03	0.02	0.80
Fhi	0.52	0.88	0.19	0.00	0.03	0.01	0.00	0.00	0.33
Flo	0.54	0.82	0.20	0.01	0.05	0.07	0.36	0.15	0.64
STD	0.67	0.90	0.60	0.11	0.25	0.00	0.09	0.25	0.04
PFR	0.96	0.94	0.97	0.26	0.59	0.00	0.01	0.11	0.03
Fftr	0.85	0.58	0.67	0.61	0.99	0.38	0.86	0.61	0.68
Fatr	0.70	0.49	0.68	0.87	0.97	0.82	0.65	0.90	0.51
Jita	0.36	0.93	0.11	0.20	0.44	0.00	0.19	0.47	0.14
Jitt	0.27	0.94	0.02	0.11	0.32	0.00	0.04	0.11	0.27
RAP	0.23	0.87	0.02	0.12	0.35	0.00	0.05	0.15	0.22
PPQ	0.30	0.89	0.04	0.19	0.48	0.00	0.05	0.14	0.23
sPPQ	0.60	0.66	0.86	0.90	0.70	0.01	0.35	0.44	0.75
vFo	0.54	0.92	0.60	0.43	0.71	0.00	0.16	0.54	0.03
Shim	0.99	0.64	0.68	0.14	0.49	0.00	0.08	0.10	0.59
APQ	0.81	0.94	0.77	0.30	0.79	0.00	0.11	0.11	0.65
sAPQ	1.00	0.71	0.72	0.74	0.90	0.07	0.26	0.19	0.88
vAm	0.26	0.19	0.63	0.29	0.59	0.16	0.05	0.13	0.30
NHR	0.98	0.93	0.93	0.96	0.77	0.00	0.15	0.54	0.10
VTI	0.32	0.90	0.02	0.93	0.55	0.00	0.18	0.89	0.01
SPI	0.62	0.19	0.38	0.78	0.86	0.75	0.13	0.98	0.04
FTRI	0.62	0.95	0.39	0.54	0.31	0.00	0.59	0.22	0.74
ATRI	0.46	0.53	0.92	0.92	1.00	0.80	0.34	0.25	0.98

To compare with other methods using the MEEI and SVD databases, we provide in *Table 10* some of the best reported accuracies found using these databases in different research studies.

Table 10: Comparison of accuracies between methods (pathology detection)

Methods	MEEI	SVD
This paper MDVP(22)	76.36%	72.58%
This paper MDVP(3)	88.21%	99.68%
Method [35]	-	81%
Method [36]	74.10%	-
Method [14]	94.07%	-

Discussion

We investigated the use of the MDVP parameters in three databases for voice pathology detection and classification. Based on the results, mentioned above, we can infer that the variation in the achieved accuracies from one database to another may be caused by different reasons, namely: (1) the severity of voice pathologies, which are not the same between the three databases, as shown, for instance, in Table 2, where sensitivity (to pathological samples) varies from one database to another; (2) the recording environment and the regulation of the recording are not the same between the three databases; (3) in the case of the MEEI database, the recording environments for pathological and normal samples were not the same; and (4) the numbers of samples taken from each database in this study are not the same. Indeed, our results are comparable to many previous studies that used the same parameters and one database. For example, with 22 parameters, the highest accuracy presented here is 76.36% with the MEEI database, which is comparable to Arjmandi et al. [32], who used the same database and the same classifier but different pathological samples. Moreover, the highest achieved accuracy with AVPD was 71.63%, which is comparable to the work performed using SVD in [33]. The accuracies with the two other databases, SVD and AVPD, are comparable (slightly less) to the accuracy obtained by MEEI. To the best of our knowledge, this is the first instance evaluating the MDVP parameters in these two databases, and they need more investigation. Using the top 10 parameters, accuracy increased by 11% with the MEEI samples, while remaining almost constant with SVD and AVPD. The reason for this may be that the top 10 parameters are not the same in all databases, and their FDR values are also different. Moreover, the selected 10 parameters contribute more than the rest in differentiating between normal and pathological voices. Using the top three MDVP parameters dramatically increased the accuracies, especially with the samples taken from the SVD database. The best accuracies were 99.68%, 88.21%, and 72.53% for SVD, MEEI, and AVPD, respectively. In each type of experiment mentioned above, we compared how well MDVP parameters detected voice

pathology in these databases and how much their samples contributed to discrimination between pathological and normal voices. Additional, cross-database experiments were performed to make sure that the extracted MDVP parameters from these databases produce the same ability (or not) to detect pathological voices and to avoid unfair over-fitting that results from error estimation. By performing cross-database experiments, it becomes easy to see how well a system trained on one database can classify samples from another database. As shown in Table *5:* 5 (first part), the cross-database experiments used one database for training and the other for testing; that table depicts increased accuracy with SVD used for training. In addition, three more experiments were performed, in which two databases were used for training and the other used for testing, as depicted in the second part of Table *5:* 5. The accuracies are comparable among the three experiments. The diverse accuracies across the three databases indicate that we need to choose appropriate features for a particular setup. May be a little training of the system by voices uttered by a particular language will help. All MDVP parameters cannot be directly applied to all types of setups to assess voice pathologies.

In general, the obtained result can be combined with auditory-perceptual evaluation techniques and laryngoscopic techniques to help the clinician making an accurate diagnosis, and accurate evaluation for voice quality. This type of analysis will help doctors to objectively tracking the progression of their patient and assessing their conditions. Furthermore, acoustic analysis cannot replace the perceptual ratings but both perceptual and acoustic measures can be considered complementary to each other. From the obtained results, we can infer that not all MDVP parameters performed the same in detection and classification processes. For instance, the top three features mentioned before (vAm, APQ, and PFR) have the highest accuracy with SVD and they reflect best performance than the other features. Clinically, these three features could be useful for characterizing and quantifying voice properties, and possibly for differentiating between pathological and healthy voices. As a result, the clinician can depend on these three features in their diagnosis more than the other features.

Finally, it is well known that, acoustic vocal assessment consists of a noninvasive process of obtaining objective measures from signal. This type of acoustic analysis can be used in clinical practice as a tool for monitoring surgical procedures or in speech therapy. These parameters of acoustic analysis can be assessed the vocal quality of patient before and after specific period of endolaryngeal phonosurgery.

From the results of all the experiments, we find the following.

- Training with MEEI does not make the system robust; the system is confused about whether it is classifying the environment or classifying normal versus pathology.
- Training with SVD makes the system more robust than systems trained with other database, because SVD samples are clearly distinguishable as either normal or pathology.

- Training with two databases offsets some of the shortcomings of training with one database.
- One of the major current limitations is that there is no an exact correlation between the numerical parameters of the acoustic analysis with the auditory-perceptual aspects of voice.
- Most of acoustic analysis studied are restricted to use only sustain vowel /a/.
- We did not consider some factors in choosing the used samples from the three databases such as falsetto, or voice abuse signal type.
- Severity of voice pathologies is not addressed in this study as well as the variability of voice pathologies.
- Not all MDVP parameters show high ability to detect and classify voice pathologies.
- Variability of the obtained accuracies for detection and classification in the three databases refers to the use of different recording protocols on each database and also due to the variability of voice disorders as well as to the severity of voice disorders.
- CSL program provides many variation of measurements for the same features which indicate these variation is for commercial purposes not for acoustic analysis that can be used as useful tool to help clinician in their diagnosis.
- There is a need to develop more robust features that can successfully differentiate between normal and pathological samples regardless of the database used. Moreover, the features should be able to classify pathological samples of low severity, as in the case of AVPD.

From *Table 6 and 7,* we can conclude that the accuracies for all types of classification were almost the same for the SVD and AVPD in both same-database and cross-database classification, while, in the case of the MEEI database, the classifications greatly differed between the two types of classification. It is obvious from Table 8 that the *p*-values for the highest-ordered three features were less than the significance level, so we infer that there were significant differences between the two classes of normal and pathology, meaning that these features well differentiated between normal and pathological samples. Not all MDVP parameters showed significant differences between the normal and pathological samples, as shown in Table 9.

Conclusion

In this work, we evaluated MDVP parameters by using three different databases—AVPD, MEEI, and SVD—and four different types of experiments. The detection accuracies varied from one database to another with the same number of MDVP parameters. The best accuracies we obtained were 99.68%, 88.21%, and 72.53% for samples taken from SVD, MEEI, and AVPD, respectively. The obtained

accuracies together with sensitivities in this study are very helpful to the clinicians for primary scanning to detect and classify the voice pathologies. For instance, some of the MDVP parameters have an excellent indication that they have the ability to contribute in detecting and classifying voice pathologies such as vAm, APQ, and PFR. We reiterate that by no means the acoustic analysis can solely be reliable to detect and classify voice pathologies; however, it can greatly assist the clinician to take his or her final decision. It would act as we stated as an adjuvant tool for the clinical assessment battery.

In future work, we will investigate the usage of MDVP parameters to detect the severity of pathology. Voice-disordered patients frequently report worsening of vocal function when they are under stress and when they are suffering from physical fatigued. In addition, there are many factors that can lead to variability in voice disorders such as the location, the size, and the severity of voce disorder. As a result, this variability in voice disorder lead to different results in diagnosing and we will investigate this variability in future work.

Acknowledgment

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-MED-2474-02).

References

[1] G. Muhammad, Z. Ali, M. Alsulaiman, and K. Almutib. "Vocal Fold Disorder Detection by applying LBP Operator on Dysphonic Speech Signal", RAICMS, 222-228, 2014.

[2] National Institute on Deafness and Other Communication Disorders: Voice, Speech, and Language: Quick Statistics, 2014.Availableat http://www.nidcd.nih.gov/health/statistics/vsl/Pages/stats.aspx. Accessed on March, 2015.

[3] Research Chair of Voicing and Swallowing Disorders. Available at http://c.ksu.edu.sa/vas/en/vsb. Accessed on March, 2015.

[4] N. Roy, R.M. Merrill, S. Thibeault, R.A. Parsa, S.D. Gray, and E.M. Smith, "Prevalence of voice disorders in teachers and the general population," J Speech Lang Hear Res., vol.47, no. 2, pp. 281-93, Apr 2004.

[5] K.H. Malki, "Voice Disorders Among Saudi Teachers in Riyadh City", Saudi Journal of Oto-Rhinolaryngology Head and Neck Surgery, 2010.

[6] B. Boyanov, and S. Hadjitodorov, “Acoustic analysis of pathological voices. a voice analysis system for the screening of laryngeal diseases”, Proceedings of IEEE International Conference on Engineering in Medicine and Biology Society, vol.16, pp.74-82, 1997.

[7] C. E. Martinez, and L H. Rufiner, “Acoustic analysis of speech for detection of laryngeal pathologies”, Proceedings of 22nd Annual IEEE International Conference on Engineering in Medicine and Biology Society, vol. 3, pp.2369-2372, 2000.

[8] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and recognition" J. Acoustic. Soc. Amer., vol. 54, no. 6, pp. 1304-1312, 1974.

[9] L. Xugang, and D. Jianwu, “An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification”, Speech Communication’ 07, vol. 50, no. 4, pp. 312-322, Oct 2007.

[10] M. A. Anusuya, S. K. Katti, “Front end analysis of speech recognition: a review”, International Journal of Speech Technology, vol. 14, pp. 99-145, Dec. 2010.

[11] L. Rabiner and B.H. Juang, Fundamentals of speech recognition. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[12] Z. Ali, M. Aslam., and M.E. Ana María, “A speaker identification system using MFCC features with VQ technique”, Proceedings of 3rd IEEE International Symposium on Intelligent Information Technology Application, pp. 115-119, 2009.

[13] W.J.J. Roberts, and J.P. Willmore, "Automatic speaker recognition using Gaussian mixture models", proceedings of Information, Decision and Control, IDC’99, pp. 465 – 470, 1999.

[14] J.I. Godino-Llorente, P. Gomes-Vilda and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters", IEEE Transactions on Biomedical Engineering, vol. 53, no. 10, pp. 1943-1953. Oct. 2006.

[15] L.E. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite state Markov Chains”, Ann. Math. Stat., vol. 37, pp. 1554-1563, 1966.

[16] S. Abe, Support Vector Machines for Pattern Classification. Springer-Verlag, Berlin Heidelberg New York, 2005

[17] T. Ritchings, M. McGillion, and C. Moore, “Pathological voice quality assessment using artificial neural networks,” Med. Eng. Phys., vol. 24, no. 8, pp. 561–564, Sept 2002.

[18] M. Brockmann, M.J. Drinnan, C. Storck, and P.N. Carding, "Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task, Journal of voice, vol. 25, no. 1, pp. 44-53, 2011.

[19] M. Rosa, J.C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," IEEE Trans. Biomed. Eng., vol. 47, no. 1, pp. 96–104, Jan 2000.

[20] 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", Journal of Voice, vol. 25, no. 6, pp. 275-289, Nov 2011.

[21] J. Wang and C. Jo, "Vocal folds disorder detection using pattern recognition method", Proceedings of 29th Annual International Conference of the IEEE EMBS, pp. 3253-3256, Lyon, France, 2007.

[22] T. Li, C. Jo, and S. Wang, "Classification of pathological voice including severely noisy cases", Proceedings of 8th International Conference on Spoken Language Processing, I, Jeju, Korea, pp. 77-80, 2004.

[23] N. Sáenz-Lechón, J.I. Godino-Llorente, Ví. 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, no. 2, pp. 120-128, April 2006.

[24] H. Oğuz, M. A. Kiliç, and M. A. Şafak, "Comparison of results in two acoustic analysis programs: PRAAT and MDVP." Turkish Journal of Medical Sciences 41.5, pp. 835-841, 2011.

[25] J.I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, C. Ramírez-Calvo "Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program," European Archives of Oto-Rhino-Laryngology 265.4, 465-476, 2008.

[26] G. Muhammad and M. Melhem, "Pathological Voice Detection and Binary Classification Using MPEG-7 Audio Features," Biomedical Signal Processing and Controls, 11, pp. 1 – 9, 2014.

[27] G. Muhammad, T. Mesallam, K. Almalki, M. Farahat, A. Mahmood, and M. Alsulaiman, "Multi Directional Regression (MDR) Based Features for Automatic Voice Disorder Detection," Journal of Voice, Vol. 26, No. 6, pp. 817.e19-817.e27, 2012.

[28] A. A-Nasheri, Z. Ali, G. Muhammad, and M. Alsulaiman, "Voice Pathology Detection Using Auto-Correlation of Different Filters Bank," 11th ACS/IEEE International Conference on Computer Systems and Applications, Doha, Qatar, November 10-13, 2014.

[29] Kay Elemetrics Corp., Disordered Voice Database, Version 1.03 (CD-ROM), MEEI, Voice and Speech Lab, Boston, MA (October 1994).

[30] W.J. Barry, P¨utzer, M., Saarbrucken Voice Database, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/

[31] Kay Elemetrics, Multi-Dimensional Voice Program (MDVP) [Computer Program], 2012.

[32] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, A. Moqarehzadeh, "Identification of voice disorders using long-time features and support vector machine with different feature reduction methods," Journal of Voice, 25 (6), pp. e275-e289, 2011.

[33] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba "Voice Pathology Detection on the Saarbruecken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit," Advances in Speech and Language Technologies for Iberian Languages. Springer Berlin Heidelberg, 99-109, 2012.

[34] I. Smits, P. Ceuppens, and M. S. De Bodt, "A comparative study of acoustic voice measurements by means of Dr. Speech and Computerized Speech Lab," Journal of Voice, vol. 19, pp. 187-196, 2005.

[35] D. Martinez, E. Lleida, A. Ortega and A. Miguel, “Score Level versus Audio Level Fusion for Voice Pathology Detection on the Saarbrucken Voice Database, “ Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science Volume 328, pp 110-120, 2012.

[36] Markaki, M. and stylianou, Y., “Voice pathology detection and discrimination based on modulation spectral features", IEEE Trans. Audio, Speech, and Language processing, 19(7): 1938-1948, 2011.

[37] Titze, I. R. "Summary statement: Workshop on acoustic voice analysis." USA: Denver, CO: National Center for Voice and Speech (1995).

[38] L. Eskenazi, D. G. Childers, and D. M. Hicks, “Acoustics correlates of vocal quality,” J. Speech Hear. Res., vol. 33, pp. 298–306, 1990.

[39]] R. Shrivastav, “The use of an auditory model in predicting perceptual ratings of breathy voice quality,” J. Voice, vol. 17, no. 4, pp. 502–512,Dec. 2003.

[40] D. Martin, J. Fitch, and V.Wolfe, “Pathologic voice type and the acoustic prediction of severity,” J. Speech Hear. Res., vol. 38, pp. 765–771, 1995.

[41] V. Parsa and D. G. Jamieson, “Identification of pathological voices based on glottal noise measures,” J. Speech Hear. Res., vol. 43, pp. 469–485, 2000.

[42] G. Muhammad, “Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system,” Cluster Computing, vol. 18, No. 2, pp. 795-802, June 2015.