Intra- and Inter-database Study for Arabic, English, and German Databases

Ali, Z., Alsulaiman, M., Muhammad, G., Elamvazuthi, I., Al-nasheri, A., Mesallam, T. A., Farahat, M., & Malki, K. H. (2017). Intra- and Inter-database Study for Arabic, English, and German Databases: Do Conventional Speech Features Detect Voice Pathology? *Journal of Voice*, *31*(3), 386.e1-386.e8. https://doi.org/10.1016/j.jvoice.2016.09.009

Published in:
Journal of Voice

Publication Status:
Published (in print/issue): 01/05/2017

DOI:
10.1016/j.jvoice.2016.09.009

Document Version
Author Accepted version

Intra- and Inter-Database Study for Arabic, English, and German Databases: Do Conventional Speech Features Detect Voice Pathology?

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

[1]Digital Speech Processing Group, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia.

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

[3]ENT Department, College of Medicine, King Saud University, Riyadh, Saudi Arabia

Email: zuali@ksu.edu.sa

ABSTRACT

A large population around the world suffers from voice complications. Various approaches for subjective and objective evaluations have been suggested in the literature. The subjective approach strongly depends on the experience and area of expertise of a clinician, and human error cannot be neglected. On the other hand, the objective/automatic approach is non-invasive. Automatic developed systems can provide complementary information that may be helpful for a clinician in the early screening of a voice disorder. At the same time, automatic systems can be deployed in remote areas where a general practitioner can use them and may refer the patient to a specialist in order to avoid complications that may be life threatening. Many automatic systems for disorder detection have been developed by applying different types of conventional speech features such as the Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), and Mel-frequency Cepstral Coefficients (MFCC). This study aims to ascertain whether conventional speech features detect voice pathology reliably, and whether they can be correlated with voice quality. To investigate this, an automatic detection system based on MFCC was developed and three different voice disorder databases used in this study. The experimental results suggest that the accuracy of the MFCC-based system varies from database to database. The detection rate for the intra-database ranges from 72% to 95%, and that for the inter-database is from 47% to 82%. The results conclude that conventional speech features are not correlated with voice, and hence are not reliable in pathology detection.

Keywords: Disorder detection, Intra-database, Inter-database, English, German, Arabic, MFCC, GMM

1 INTRODUCTION

A well-known speech features extraction algorithm, the Mel-frequency Cepstral Coefficients (MFCC) [1], is implemented in this study to develop an automatic voice disorder detection system. In the developed system, MFCC features are extracted from normal and pathological subjects to differentiate between them. The aim of the study is to determine whether the implemented conventional speech features are capable of detecting voice disorders reliably. Moreover, it explores whether these features can be correlated with voice quality. To answer the underlying questions, the speech features are extracted from three different voice disorder databases: the Massachusetts Eye and Ear Infirmary (MEEI) database [2] (English database), the Saarbrucken Voice Database (SVD) [3] (German database), and the Arabic Voice Pathology Database (AVPD). During the investigation of the features, two approaches are used. In the first approach, the developed system is trained and tested with the same database, and this is referred to as an intra-database approach. In the second approach, the system is trained and tested with different databases, and this is referred to as an inter-database approach.

Around one-third of the global population suffers from voice-related problems [4, 5], and approximately 7.5 million of these affected people are in the USA [6]. Voice disorders may be the result of various pathologies such as benign lesions (growth of abnormal tissues on the vocal folds) [7], paralysis (one of the main reasons is injury to the recurrent laryngeal nerve) [8], or sulcus vocalis (scarring or mucosal cover of the vocal folds) [9-11]. Benign lesions are further classified as vocal fold nodules [12], cysts [9], and polyps [13]. Due to voice disorders, vocal folds exhibit irregular vibrations and make the voice sound strained, hard, weak, whispering, or breathier [14], ultimately affecting the personal and professional life of a person. The most common reasons for the occurrence of voice disorders are excessive talking, poor dehydration, alcohol consumption, and smoking [15, 16].

Voice disorders can be diagnosed by subjective and objective evaluations. The former is the most common method of diagnosis in medical clinics [17-19]. Perceptual evaluation and visual investigation of the vocal folds are used by medical doctors during a subjective evaluation. Three scales are practiced for perceptual evaluation in clinics: Grade, Breathiness, Roughness, Asthenia, and Strain (GBRAS) [17], Roughness, Breathiness, and Hoarseness (RBH) [20], and Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). However, there are some limitations to the scales, including the size of the assessment panel [21], human error, attention, memory lapses of raters [21, 22], professional background of raters [23], and disagreement of judgement between slight and moderate types of voice disorders [21, 22, 24]. In addition, video laryngostroboscopy (VLSC) [25] is used for the visual inspection of vocal folds to diagnose a voice disorder, and the results require a subjective interpretation. Different rating scales are thus introduced [26, 27] to avoid this, but no standard approach is available at present for the interpretation of VLSC results [28, 29]. Subjective evaluation is invasive and strongly depends on the experience and area of expertise of a clinician, and thus runs the risk of human error.

Therefore, many automatic voice detection systems have been developed by researchers for the objective evaluation of vocal fold disorders [30-33]. Automatic disorder detection systems can be used by general practitioners for early screening, and in the case of a voice problem, they can refer the patient to a specialist. The early detection of voice disorders can avoid severe complications such as keratosis [34], which is a pre-cancerous pathology that can be life threatening. Automatic detection systems are non-invasive in nature and easy to use, and they can be deployed in remote areas where specialized clinics are not available. Due to advances in computer technology, constraints such as computational power and storage no longer exist during the development and implementation of the various algorithms. Many complex algorithms have been implemented in the development of healthcare-related medical applications.

The detection of vocal fold disorders is one such application and this has been developed with various types of extraction algorithms. Some features are developed to determine the quality of voice (e.g., shimmer [35] and jitter [36]), whereas others are taken from speech processing to develop automatic systems.

The rest of the paper is organized as follows. Section 2 presents related works on automatic detection systems developed with different speech features. Section 3 describes voice disorder databases and the system developed for the investigation of the speech features. Section 4 provides the experimental results for the intra- and inter-database experiments. Section 5 provides the discussion, and Section 6 concludes the study.

2 RELATED WORKS

Generally, speech features are divided into two categories: the human hearing system and the human speech production system. The MFCC belongs to the first type of speech feature, and simulates the human auditory system where the inner part of the human ear plays a very important role in separating the frequencies. Higher frequencies are localized at the basal turn and lower frequencies towards the apex of the cochlea. Each point on the basilar membrane is a band pass filter, and these are referred to as critical bands. The phenomenon is incorporated into the MFCC by applying Mel-scaled band pass filters. By contrast, linear prediction coefficients (LPC) [1] fall under the second category of speech features. Voice disorders disturb the vocal folds, causing irregular vibrations in the folds due to voice box malfunctioning. Voice pathologies also affect the shape of the vocal folds and produce abnormalities in spectral characteristics. Human vocal tract characteristics can be modelled by using LPC features with the help of the all-pole model. LPC represents the vocal tract resonance characteristics in the acoustic spectrum and highlights the formant structure of a speaker [1, 37]. A number of automatic pathology detection systems have been developed by using both types of features.

LPC and LPC-based cepstral coefficients (LPCC) [38] have been used in many studies [39-42] to develop a voice pathology assessment system. The correct acceptance rate of 73% with LPC and 73% with LPCC was obtained in [39], when edema was detected from normal samples and other pathologies such as cysts, nodules, paralysis, and polyps. The efficiencies for LPC and LPCC were 85% and 80%, respectively. To conduct this study, 120 subjects were considered, including 67 patients and 53 normal persons from the MEEI database, and experiments were performed by using the sustained vowel /ah/. MFCC were also calculated to make a comparison with LPC and LPCC and this achieved an efficiency of 52%, very low compared with LPC and LPCC. The high false acceptance rate of 74% showed that MFCC was unable to detect edema from other pathologies as well as other features. However, when all normal persons were grouped in one class and all pathologies were combined in a second class, the results of MFCC were much better than those of the other features, which shows that MFCC can perform well in the detection of disorders but is not as good at discriminating between types of disorders. In addition, MFCC was used for the development of many pathology detection systems [40, 43-48] and performed better than LPC in pathology detection.

In [40], MFCC and LPC fed a support vector machine (SVM) and k-nearest neighbours (KNN) for the classification of three classes: healthy, diffuse, and nodular. The database used for the study contained sustained vowels only and was recorded at the Department of Medicine, Lithuania. The classification rate obtained for MFCC was 73.08% and that for LPC was 67.31%. In [46], multi-dimensional voice program (MDVP) [49] features and MFCC extracted from all voice samples of the sustained vowel sound /a/ of the MEEI database were used to build a voice disorder detection system. Many experiments were performed by providing extracted features to different modelling techniques. The highest accuracy for MDVP features with the sustained vowel by using the Gaussian mixture model (GMM)-based system was 97.67%. The extracted MFCC with pitch and without was fed to the hidden Markov model (HMM) for disorder detection, and the highest achieved accuracy was 97.75% with MFCC alone. In [45], a database at the ENT department of the Busan National University Hospital, South Korea, was built for pathology assessment. It contained the sustained vowel sound /a/ recorded by disorder patients and normal persons. The extracted MFCC was used with SVM, artificial neural networks, GMM, and HMM for disorder detection. The recorded disorders were nodule, polyp, edema, cyst, glottis cancer, and laryngitis. The highest detection rate (95.2%) was achieved with GMM. In [44], MFCC was extracted with a temporal derivative and showed good results as a pathology detection system. MFCC with a different number of temporal derivatives provided an accuracy of 95%.

[47], an extension of [46], classified five types of disorders carried out using MFCC and fundamental frequency with sustained vowels. All speech samples of ventricular compression, gastric reflux, hyper function, paralysis, and A-P squeezing were considered to develop the disorder. The maximum accuracy was obtained with paralysis, and the minimum was achieved with hyper function. An average classification rate of approximately 70% was attained for the five disorders. Accuracy decreased when MFCC was used in a multi-class problem, and the results support the fact that MFCC is good for detection systems and performs better than other speech features. Therefore, in this study, we develop an MFCC-based detection system and use it for the inter- and intra-database experiments.

3 METHOD

Three different voice disorder databases were used in the experiments to investigate the role of conventional speech features for pathology detection. The MEEI database was recorded at the Massachusetts Eye and Ear Infirmary Lab and openly commercialized by Kay Elemetrics. The database [2] contains samples of the sustained vowel /ah/ recorded by normal and voice-disordered subjects. The SVD [3] was recorded by the Institute of Phonetics of Saarland University, and it is freely available. This also contains the sustained vowel /ah/ recorded by normal and pathological subjects. The AVPD was recorded by our group at King Abdul Aziz University Hospital, Riyadh. Like other databases, it also contains the sustained vowel sound /ah/. All three databases have samples from disordered subjects suffering from different kinds of voice pathologies. In this study, only common voice disorders are considered. Five voice disorders are common among the AVPD and MEEI: cysts, sulcus, polyps, nodules, and paralysis. However, only three of these disorders are also common to the SVD: polyps, cysts, and paralysis. The distribution of the samples for the MEEI, SVD, and AVPD is provided in Table 1.

The total number of normal subjects in the MEEI is 53, which makes the overall sample equal to 166. In the SVD and AVPD, normal subjects are considered in such a way that the overall number of samples remains close to each other.

Table 1: Distribution of samples for the MEEI, AVPD, and SVD

Subjects	**Type**	**Distribution of samples**		
		MEEI	**AVPD**	**SVD**
Disordered	Cysts	7	9	6
	Sulcus	3	20	--
	Polyps	20	19	45
	Nodules	19	10	--
	Paralysis	67	20	50
Total disordered samples:		113*	78	101
Normal		53	87	60
Overall samples:		166	165	161

*Two files are common in cysts and nodules and one file is common in polyps and nodules. Therefore, the total files are 113 instead of 116.

To observe the role of conventional features in disorder detection, the setup for the experiments needs to be the same. This is the reason that the same type of text is used for feature extraction in this study, and the only common text in all databases is the sustained vowel /ah/. The setup for the intra- and inter-database experiments for disorder detection and the MFCC-based developed systems are described in the next section.

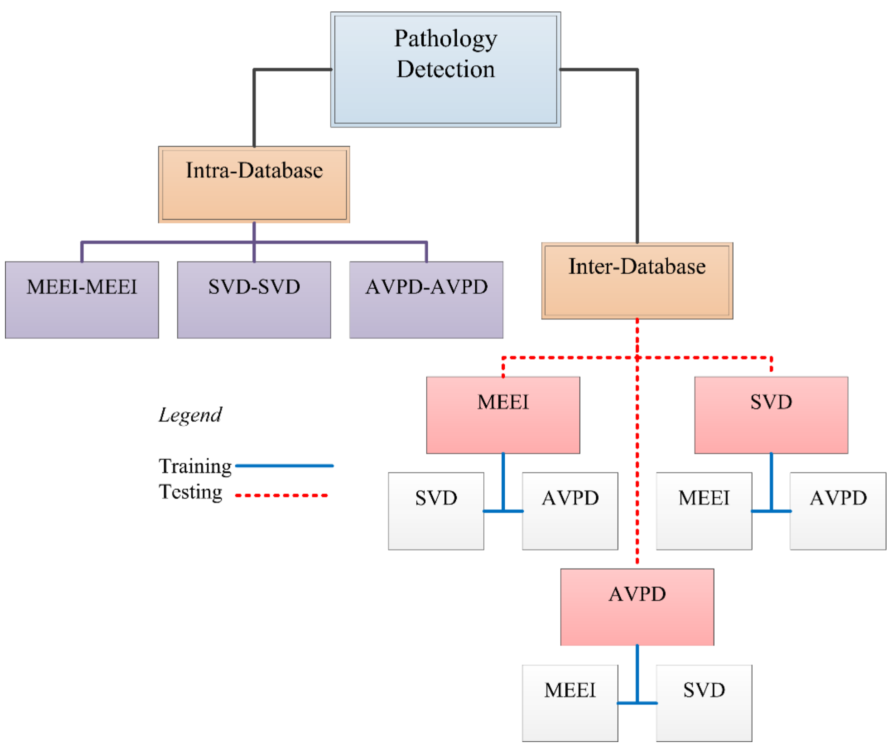


Figure 1: Inter- and intra- database approaches for pathology detection

The complete setup for automatic pathology detection with the three databases, the MEEI, SVD, and AVPD, is presented in Figure 1. For the inter-database experiment, the solid line connected to the MEEI indicates that the testing database is the MEEI, and the dotted lines connected to the SVD and AVPD show that they are the testing databases.

To perform the intra- and inter-database experiments for pathology detection, an automatic detection system based on well-known conventional speech features, MFCC, was developed. The MFCC-based pathology detection system (MPDS) has two main components: feature extraction and pattern matching techniques. The features are extracted by applying the MFCC algorithm and pattern matching is done by the GMM [50]. The MPDS has two important phases: training and testing. In the training phase, the MPDS extracts MFCC speech features from normal and disordered subjects by following the computational steps mentioned in [51], and generates an acoustic model for each subject. Once models are generated, the MFCC of an unknown speech sample is compared with the generated models in the testing phase. The likelihood of each model is then computed with the unknown testing sample. The model that has the maximum likelihood is the unknown testing sample. The block diagram of the MPDS is depicted in Figure 2.

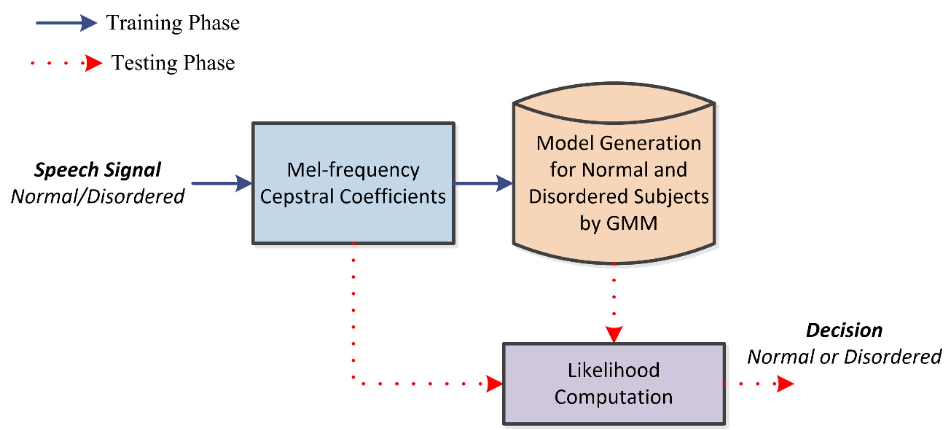


Figure 2: Block diagram of the MPDS

GMM is a state-of-the-art modelling technique that copes with the space of the features, rather than the time sequence of their appearance. In this study, to determine the optimized parameters of GMM, mean, covariance matrix, and prior probability, the Expectation-Maximization (EM) algorithm [52] is implemented. Meanwhile, these parameters are initialized by using the k-means algorithm [53].

4 EXPERIMENTAL SETUP AND RESULTS

All samples of the MEEI, AVPD, and SVD used in this study are down-sampled to 16 kHz. To make the system robust against the training set, all results for disorder detection are obtained by using the five-fold approach, in which the database is divided into five distinct subsets. Each time one of the subsets is used for testing, the remaining four are used to train the system. The performance of the MPDS is provided by measuring sensitivity (SEN), specificity (SPE), and accuracy [54]. During feature extraction, 12 MFCCs are extracted. In addition, the first and second derivatives of the 12 MFCC features, referred to as the delta and delta-delta coefficients, respectively, are computed by using the relation given in Equation (1):

$$\Delta_t = \frac{\sum_{i=1}^{B} i\left(c_{t-1,m} - c_{t+1,m}\right)}{2\sum_{i=1}^{B} i^2} \tag{1}$$

where Δ_t corresponds to a velocity component in the t^{th} frame, $c_{t,m}$ stands for the m^{th} MFCC in the t^{th} frame, and B is the length of the regression window. In all experiments, 12, 24, and 36 MFCCs are used. Twenty-four MFCC features contain 12 static and 12 delta coefficients, while 36 MFCC features contain 12 static, 12 delta, and 12 delta-delta coefficients. To generate the acoustic model for normal and disordered subjects, 4, 8, 16, 32, and 64 Gaussian mixtures are used.

4.1 Intra-Database Pathology Detection Results

In this section, different experiments are performed for all three databases using the intra-database approach. The training and testing of the MPDS are performed by using the samples of the same databases. MEEI-MEEI denotes that the training and testing samples are taken from the MEEI database. Similarly, AVPD-AVPD and SVD-SVD indicate that the training and testing samples are taken from the AVPD and SVD, respectively. All experiments are performed with 12, 24, and 36 MFCC features. However, only the best results of each database are presented in Table 2. The best obtained results for the MEEI are achieved with 36 MFCCs. Similarly, the highest detection rates for the AVPD and SVD are obtained with 36 MFCCs and 12 MFCCs, respectively.

Table 2: Intra-database pathology detection results

Experiments	Features	GMM	Specificity	Sensitivity	Accuracy
MEEI-MEEI	36 MFCC	4	94.55	94.70	**94.60**
		8	86.91	95.53	92.76
		16	83.45	95.61	91.57
		32	60.73	99.13	86.76
		48	50.91	99.13	83.78
		64	58.91	98.22	85.58
AVPD-AVPD	36 MFCC	4	75.88	73.33	74.59
		8	83.92	73.25	78.80
		16	81.70	67.92	75.03
		32	83.86	75.83	80.03
		48	88.50	78.42	**83.65**
		64	89.74	73.17	81.79
SVD-SVD	12 MFCC	4	73.33	71.09	72.13
		8	76.67	84.73	**80.20**
		16	76.67	81.27	78.50
		32	76.67	80.73	78.38
		48	70.00	80.91	74.82
		64	70.00	79.27	73.91

Bold value represents the highest obtained accuracy.

The best accuracy for the MEEI database is 94.60% and the corresponding SPE and SEN are 94.55% and 94.70%, respectively. Accuracy is also good, which shows that the MFCC features perform well in the detection of both types of subjects. For the AVPD, the best obtained accuracy is 83.65%. The corresponding SEN of 78.42% shows that the MPDS detects 78% of pathological subjects correctly, while the SPE of 88.50% suggests that the system recognizes 88% of normal subjects accurately. In the case of the SVD, the maximum obtained accuracy is 80.20%, whereas the SPE and SEN are 76.67% and 84.73%, respectively. The performance of the MFCC features for the AVPD and SVD is not as good as that for the MEEI database.

4.2 Inter-Database Pathology Detection Results

In this section, the inter-database experiments for pathology detection are performed and the results are provided in Table 3. In the inter-database experiments, the MPDS is trained and tested with the samples of different databases. For instance, AVPD-MEEI means that the training database is the AVPD, while the testing database is the MEEI.

Table 3: Inter-database pathology detection results

Testing Database	GMM	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
MEEI		*Training Database AVPD*			*Training Database SVD*		
	4	24.36	75.22	59.02	79.64	77.04	77.72
	8	74.36	71.74	72.34	85.27	80.55	**81.96**
	16	77.45	74.35	**75.31**	85.09	76.01	78.89
	32	69.82	77.91	75.31	85.45	74.31	77.70
	48	88.55	69.01	75.31	83.64	71.78	75.33
	64	86.91	69.21	74.69	80.00	79.72	79.50
AVPD		*Training Database MEEI*			*Training Database SVD*		
	4	0.00	100.00	47.27	27.32	87.42	55.65
	8	1.11	100.00	47.86	22.68	89.58	54.38
	16	0.00	100.00	47.27	28.24	86.08	55.67
	32	0.00	100.00	47.27	44.58	82.17	62.26
	48	0.00	100.00	47.27	57.25	82.08	**69.06**
	64	1.18	100.00	**47.90**	44.71	80.92	61.77
SVD		*Training Database MEEI*			*Training Database AVPD*		
	4	36.67	96.18	63.91	33.33	88.18	58.66
	8	48.33	98.00	**71.07**	65.00	82.18	72.92
	16	31.67	98.18	62.13	71.67	74.55	72.96
	32	26.67	94.36	57.71	73.33	74.00	73.75
	48	16.67	100.00	54.98	78.33	74.18	**76.52**
	64	26.67	100.00	60.43	71.67	70.55	71.15

Bold values represent the highest obtained accuracy.

Table 3 shows that MFCC does not provide robust results for inter-database pathology detection. The best accuracy for AVPD-MEEI is 75.31%, whilst for MEEI-AVPD it is 47.90%, which is very low. The corresponding SPE of the accuracy 47.90% is almost zero (1.18%), which suggests that MFCC is unable to detect normal subjects. Therefore, the false acceptance rate is very high because MFCC detected all samples as disordered subjects.

Furthermore, the best accuracy for SVD-MEEI is 81.96% with SPE and SEN equal to 85.27% and 80.55%, respectively. On the other hand, MEEI-SVD obtains an accuracy of 71.07% but SPE is 48.33%. This indicates that MFCC again fails to detect normal subjects reliably and the false acceptance rate is very high.

All experiments for cross-database pathology detection are performed using 12, 24, and 36 MFCC features. In Table 3, only the results with maximum accuracy are reported. For SVD-AVPD, the maximum accuracy is obtained with 24 MFCCs, while for all the other combinations of the databases the highest accuracy is achieved with 36 MFCCs. The accuracy of the databases with the number of MFCCs is depicted in Figure 3.

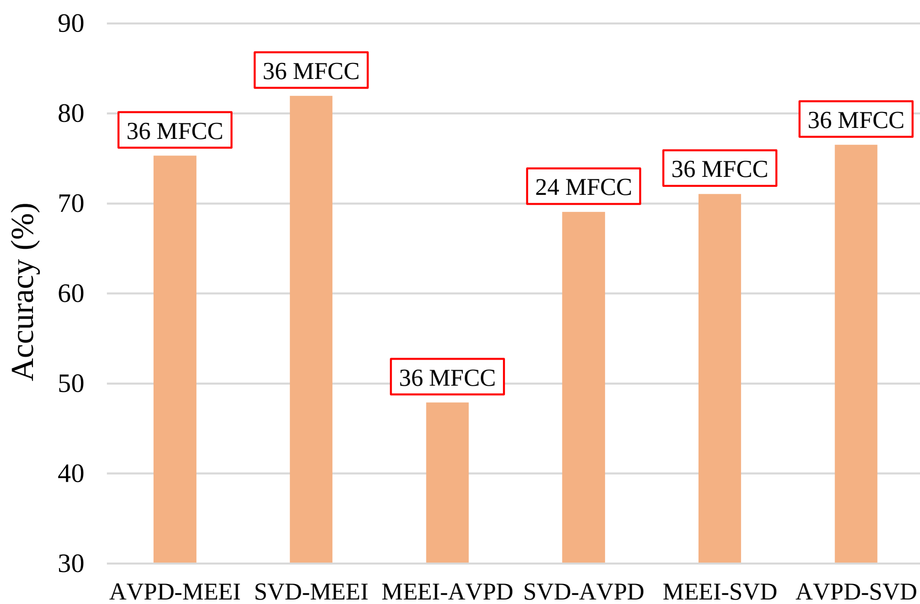


Figure 3. The best obtained accuracy for the inter-database experiments with the number of MFCCs

5 DISCUSSION

To investigate whether speech features can be used to detect voice pathology reliably, intra- and inter-database approaches used the MPDS. The trend of accuracy in the intra-database approach for the MEEI, AVPD, and SVD with 4, 8, 16, 32, 48, and 64 Gaussian mixtures is depicted in Figure 4, which shows that the best accuracy for each database is achieved with a different number of Gaussian mixtures. The best accuracy for the MEEI is obtained with four mixtures, while the maximum accuracy for the AVPD and SVD is obtained with 48 and 8 mixtures, respectively. It can also be observed that the accuracy of the MEEI and AVPD is the same (approx. 83%), with 48 Gaussian mixtures, but SPE and SEN are different as mentioned in Table 2.

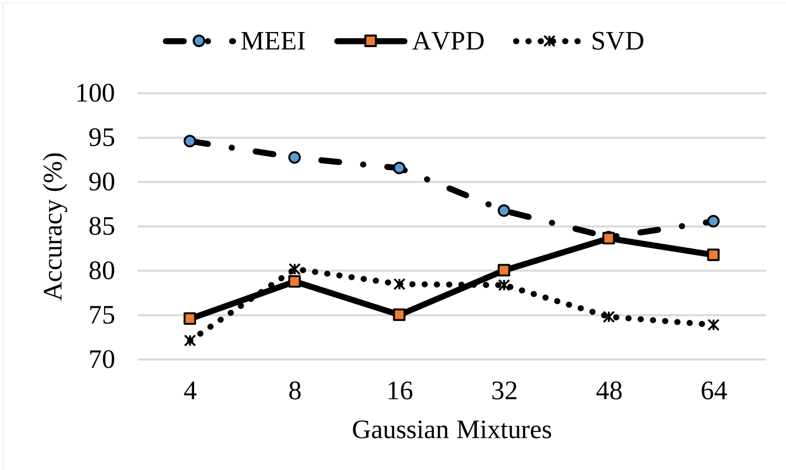
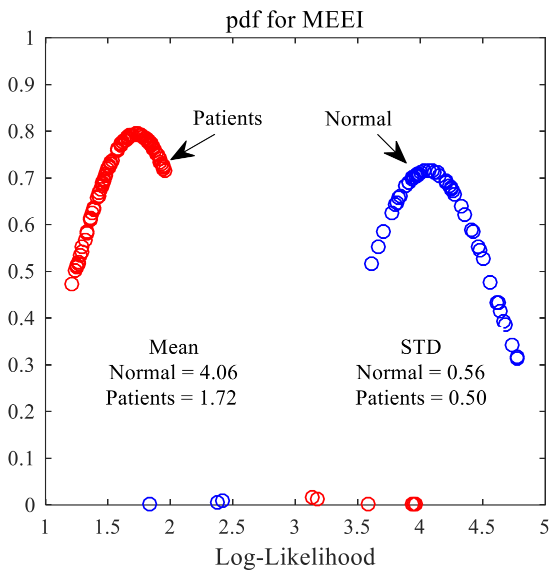


Figure 4. Trend of accuracy for the MEEI, AVPD, and SVD

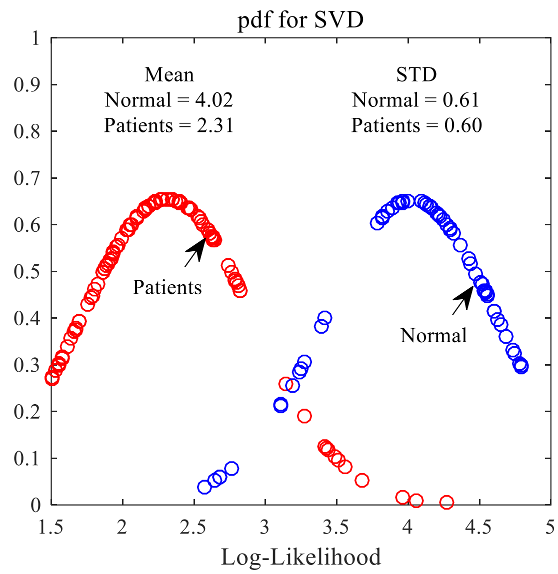
Moreover, the MEEI and AVPD have almost the same number of samples, 166 and 165, respectively, and their types of pathology are also same, including cysts, sulcus, polyps, nodules, and paralysis. Nevertheless, both databases obtain significantly different best accuracies, with 94.60% for the MEEI and 83.65% for the AVPD. A possible reason for this might be MFCC, because these features simulate the human auditory system [55] and act as an clinician, who can discriminate between normal and disordered subjects by hearing. At the same time, a clinician might be confused during the evaluation of normal and mildly disordered voices. Therefore, it can be inferred that MFCC works like a human ear and cannot be correlated with the quality of the voice. This explains the varied accuracies of the different databases, which range from 80% to 95%.

The dimensions of the calculated speech features are 12, 24, and 36, and their interpretation is not possible by human minds. Therefore, to differentiate between normal and disordered subjects, log-likelihood values are computed by using GMM and these are used as a discriminative power to determine the class of a test utterance. During the training phase of the MPDS, a model of a normal subject represented by λ_n is generated. Similarly, a model of a disordered subject is also generated and represented by λ_p. Once the models are generated during the training phases, the next step is the evaluation of the test samples to determine whether they are normal or disordered. The evaluation is carried out by computing the log-likelihood of the test utterance *X*. The log-likelihood of *X* with the model of normal subjects is given by $\log p\left(X \mid \lambda_n\right)$ and the likelihood for disordered subjects is given by $\log p\left(X \mid \lambda_p\right)$. If *X* has maximum likelihood with the normal model, then *X* is the utterance of a normal person. On the other hand, if the likelihood is higher for a disordered model, then *X* is a disordered subject. The probability distribution functions (pdf) for the log-likelihood values are presented in Figures 5(a), 5(b), and 5(c) for the MEEI, SVD, and AVPD, respectively.

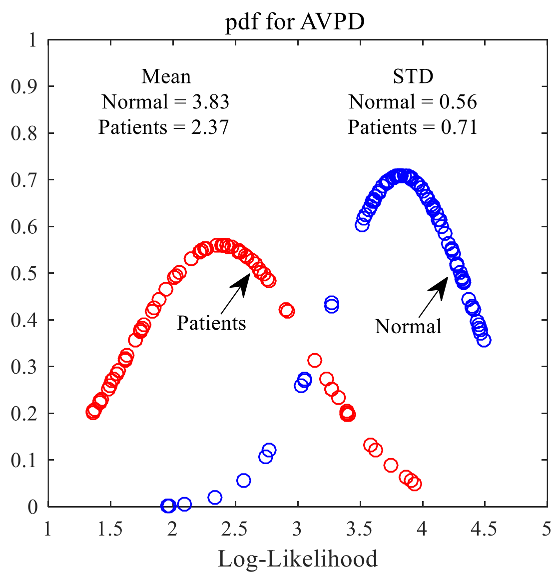
Figure 5 shows that the distributions of the log-likelihood values differ, whereas the same recording material of the sustained vowel /ah/ is used for all databases to train and test the MPDS. For a normal subject, the mean and standard deviation (STD) for the MEEI are (4.06, 0.56), for the SVD are (4.02, 0.61) and for the AVPD are (3.83, 0.56), where the first value stands for the mean and the second for the STD. The mean and STD for pathological subjects are (1.72, 0.50), (2.31, 0.60), and (2.37, 0.71) for the MEEI, SVD, and AVPD, respectively. The different means, STD values, and accuracies of the databases suggest that a common threshold to make the decision for the presence of a disorder cannot be determined. Moreover, for the cross-database results, the accuracy of the system for the MEEI, AVPD, and SVD varies from 47.90% to 81.96%, which leads to the unreliability of the speech features in pathology detection.



(a)



(b)



(c)

Figure 5: Distribution of the log-likelihood values for all databases: (a) MEEI, (b) SVD, and (c) AVPD

6 CONCLUSION

As an automatic pathology detection system, the MPDS is based on well-known speech features MFCC and is implemented to inspect the role of conventional speech features in pathology detection. Three different databases of three different languages (MEEI (English), AVPD (Arabic), and SVD (German)) were used in this investigation. The intra-database results showed that the results vary from database to database even though they have the same numbers and types of speech samples. The results suggested that the speech features only simulate the human auditory system and cannot be correlated with voice quality. In addition, the detection results for the inter-database ranged from 47% to 82%, very different from those obtained for the intra-database (72% to 95%), which strengthens the fact that conventional speech features are not reliable for voice disorder detection. Furthermore, these features should be used carefully for pathology detection, especially when subjects are recorded in different recording environments.

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] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey: Prentice Hall Press 1993.

[2] Massachusette Eye & Ear Infirmry Voice & Speech LAB, "Disordered Voice Database Model 4337 (Ver. 1.03) ", ed. Lincoln Park, NJ: Kay Elemetrics Corp., 1994.

[3] 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: IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández, *et al.*, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 99-109.

[4] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith, "Voice Disorders in the General Population: Prevalence, Risk Factors, and Occupational Impact," *The Laryngoscope*, vol. 115, pp. 1988-1995, 2005.

[5] 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, pp. 281-93, Apr 2004.

[6] (1 Dec 2015). *Quick Statistics: Voice, Speech, and Language*. Available: http://www.nidcd.nih.gov/health/statistics/vsl/Pages/stats.aspx

[7] J. Bohlender, "Diagnostic and therapeutic pitfalls in benign vocal fold diseases," *GMS Curr Top Otorhinolaryngol Head Neck Surg*, vol. 12, pp. 1-19, 2013.

[8] L. H. Rosenthal, M. S. Benninger, and R. H. Deeb, "Vocal fold immobility: a longitudinal analysis of etiology over 20 years," *Laryngoscope*, vol. 117, pp. 1864-70, 2007.

[9] M. Bouchayer, G. Cornut, R. Loire, J. B. Roch, E. Witzig, and R. W. Bastian, "Epidermoid cysts, sulci, and mucosal bridges of the true vocal cord: A report of 157 cases," *The Laryngoscope,* vol. 95, pp. 1087-1094, 1985.

[10] M. Hirano, T. Yoshida, S. Tanaka, and S. Hibi, "Sulcus vocalis: functional aspects," *Ann Otol Rhinol Laryngol,* vol. 99, pp. 679-83, 1990.

[11] P. A. Lindestad and S. Hertegard, "Spindle-shaped glottal insufficiency with and without sulcus vocalis: a retrospective study," *Ann Otol Rhinol Laryngol,* vol. 103, pp. 547-553, 1994.

[12] R. Leonard, "Voice therapy and vocal nodules in adults," *Curr Opin Otolaryngol Head Neck Surg,* vol. 17, pp. 453-457, 2009.

[13] J. Jiang, H.-J. Chen, J. Stern, and N. P. Solomon, "Vocal Efficiency Measurements in Subjects with Vocal Polyps and Nodules: A Preliminary Report," *Annals of Otology, Rhinology & Laryngology,* vol. 113, pp. 277-282, 2004.

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

[15] I. R. Titze, J. Lemke, and D. Montequin, "Populations in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report," *Journal of Voice,* vol. 11, pp. 254-259, 1997.

[16] R. H. Martins, J. Defaveri, M. A. Domingues, and R. de Albuquerque e Silva, "Vocal polyps: clinical, morphological, and immunohistochemical aspects," *J Voice,* vol. 25, pp. 98-106, 2011.

[17] H. P. 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.

[18] 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.

[19] 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.

[20] G. Friedrich and P. H. Dejonckere, "Das Stimmdiagnostik-Protokoll der European Laryngological Society (ELS): erste Erfahrungen im Rahmen einer Multizenterstudie," *Laryngorhinootologie,* vol. 84, pp. 744-752, 2005.

[21] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual Evaluation of Voice QualityReview, Tutorial, and a Framework for Future Research," *Journal of Speech, Language, and Hearing Research,* vol. 36, pp. 21-40, 1993.

[22] B. R. Gerratt, J. Kreiman, N. Antonanzas-Barroso, and G. S. Berke, "Comparing Internal and External Standards in Voice Quality Judgments," *Journal of Speech, Language, and Hearing Research,* vol. 36, pp. 14-20, 1993.

[23] 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.

[24] J. Kreiman and B. R. Gerratt, "Sources of listener disagreement in voice quality assessment," *The Journal of the Acoustical Society of America,* vol. 108, pp. 1867-1876, 2000.

[25] V. Uloza, A. Vegiene, and V. Saferis, "Correlation between the quantitative video laryngostroboscopic measurements and parameters of multidimensional voice assessment," *Biomedical Signal Processing and Control,* vol. 17, pp. 3-10, 2015.

[26] B. J. Poburka, "A new stroboscopy rating form," *Journal of Voice,* vol. 13, pp. 403-413, 1999.

[27] C. A. Rosen, "Stroboscopy as a Research Instrument: Development of a Perceptual Evaluation Tool," *The Laryngoscope,* vol. 115, pp. 423-428, 2005.

[28] S. Deguchi, Y. Ishimaru, and S. Washio, "Preliminary Evaluation of Stroboscopy System Using Multiple Light Sources for Observation of Pathological Vocal Fold Oscillatory Pattern," *Annals of Otology, Rhinology & Laryngology*, vol. 116, pp. 687-694, 2007.

[29] R. Speyer, G. H. Wieneke, W. Kersing, and P. H. Dejonckere, "Accuracy of Measurements on Digital Videostroboscopic Images of the Vocal Folds," *Annals of Otology, Rhinology & Laryngology*, vol. 114, pp. 443-450, 2005.

[30] J. W. Lee, H. G. Kang, J. Y. Choi, and Y. I. Son, "An investigation of vocal tract characteris-tics for acoustic discrimination of pathological voices," *BioMed Research International*, vol. 2013, pp. 1-11, 2013.

[31] A. I. R. Fontes, P. T. V. Souza, A. D. D. Neto, A. d. M. Martins, and L. F. Q. Silveira, "Classification System of Pathological Voices Using Correntropy," *Mathematical Problems in Engineering*, vol. 2014, pp. 1-7, 2014.

[32] L. Jung-Won, S. Kim, and K. Hong-Goo, "Detecting pathological speech using contour modeling of harmonic-to-noise ratio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 5969-5973.

[33] D. Panek, A. Skalski, and J. Gajda, "Quantification of Linear and Non-linear Acoustic Analysis Applied to Voice Pathology Detection," in *Information Technologies in Biomedicine, Volume 4. Advances in Intelligent Systems and Computing*. vol. 284, ed: Springer International Publishing, 2014, pp. 355-364.

[34] T. Mau, "Diagnostic evaluation and management of hoarseness," *Med Clin North Am*, vol. 94, pp. 945-60, 2010.

[35] P. Lieberman, "Perturbations in Vocal Pitch," *The Journal of the Acoustical Society of America* vol. 33, pp. 597-603, 1961.

[36] Y. Horii, "Vocal shimmer in sustained phonation," *J Speech Hear Res*, vol. 23, pp. 202-9, 1980.

[37] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT press, 1999.

[38] M. Anusuya and S. Katti, "Front end analysis of speech recognition: a review," *International Journal of Speech Technology*, vol. 14, pp. 99-145, 2011.

[39] B. Aguiar Neto, J. M. Fechine, S. C. Costa, and M. Muppa, "Feature estimation for vocal fold edema detection using short-term cepstral analysis," in *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, 2007, pp. 1158-1162.

[40] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization," *Comput Methods Programs Biomed*, vol. 91, pp. 36-47, 2008.

[41] D. G. Childers and K. S. Bae, "Detection of laryngeal function using speech and electroglottographic data," *Biomedical Engineering, IEEE Transactions on*, vol. 39, pp. 19-25, 1992.

[42] M. Marinaki, C. Kotropoulos, I. Pitas, and N. Maglaveras, "Automatic detection of vocal fold paralysis and edema," in *8th International Conference on Spoken Language Processing*, 2004, pp. 1-4.

[43] S. C. Costa, B. Aguiar Neto, and J. M. Fechine, "Pathological voice discrimination using cepstral analysis, vector quantization and hidden Markov models," in *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*, 2008, pp. 1-5.

[44] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans Biomed Eng*, vol. 51, pp. 380-384, 2004.

[45] W. Jianglin and J. Cheolwoo, "Vocal Folds Disorder Detection using Pattern Recognition Methods," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 2007, pp. 3253-3256.

[46] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall*

Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint, 2002, pp. 182-183.

[47] A. A. Dibazar, T. W. Berger, and S. S. Narayanan, "Pathological Voice Assessment," in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, 2006, pp. 1669-1673.

[48] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "An improved method for voice pathology detection by means of a HMM-based feature space transformation," *Pattern Recognition,* vol. 43, pp. 3100-3112, 2010.

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

[50] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, pp. 72-83, 1995.

[51] Z. Ali, G. Muhammad, M. Alsulaiman, I. Elamvazuthi, and K. Al-Mutib, "Oriented and Interpolated Local Features for Speech Recognition of Vocal Fold Disordered Patients," *International Journal of Computers and their Applications,* vol. 22, pp. 3-11, 2015.

[52] R. A. Redner and H. F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review,* vol. 26, pp. 195-239, 1984.

[53] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. NJ, USA: Prentice-Hall, 1988.

[54] 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*, 2016.

[55] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing* vol. 18: Prentice Hall Englewood Cliffs, 2001.