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

***,†Zulfiqar Ali, *Mansour Alsulaiman, *Ghulam Muhammad, †Irraivan Elamvazuthi, *Ahmed Al-nasheri, ‡,§Tamer A. Mesallam, ‡Mohamed Farahat, and ‡Khalid H. Malki,** **‡Riyadh, Saudi Arabia, †Perak, Malaysia, and §Shebin Alkoum, Egypt*

Summary: A large population around the world has 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 or automatic approach is noninvasive. 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 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, linear prediction cepstral coefficients, and Mel-frequency cepstral coefficients (MFCCs). 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 were 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.

Key Words: Voice disorder detection–Intra-database–Inter-database–MFCC–GMM.

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 pathologic 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 Saarbrücken 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 has voice-related problems,[4,5] and approximately 17.9 million of these affected people are in the United States.[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] Because of 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[17]; Roughness, Breathiness, and Hoarseness[20]; and Consensus Auditory-Perceptual Evaluation of Voice. 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 judgment between slight and moderate types of voice disorders.[21,22,24] In addition, video laryngostroboscopy[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 video

Accepted for publication September 8, 2016.

From the *Digital Speech Processing Group, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; †Centre for Intelligent Signal and Imaging Research (CISIR), Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, Perak, Malaysia; ‡ENT Department, College of Medicine, King Saud University, Riyadh, Saudi Arabia; and the §ENT Department, College of Medicine, Al-Menoufiya University, Shebin Alkoum, Egypt.

Address correspondence and reprint requests to Zulfiqar Ali, Digital Speech Processing Group, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia. E-mail: zuali@ksu.edu.sa, zulfiqarbutt2000@gmail.com

Journal of Voice, Vol. 31, No. 3, pp. 386.e1–386.e8
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2016.09.009

laryngostroboscopy 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 precancerous pathology that can be life threatening. Automatic detection systems are noninvasive in nature and easy to use, and they can be deployed in remote areas where specialized clinics are not available. Because of 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 (eg, 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. The next section presents related works on automatic detection systems developed with different speech features. The Method section describes voice disorder databases and the system developed for the investigation of the speech features. The Experimental Setup and Results section provides the experimental results for the intra- and inter-database experiments, followed by the Discussion section and the Conclusion section.

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 are localized toward the apex of the cochlea. Each point on the basilar membrane is a bandpass filter, and these are referred to as critical bands. The phenomenon is incorporated into the MFCC by applying Mel-scaled bandpass 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 modeled 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 linear prediction cepstral coefficients (LPCC)[38] have been used in many studies[39–42] to develop a voice pathology assessment system. The correct acceptance rates of 73% with LPC and 73% with LPCC were obtained in Ref. 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/. MFCCs 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 Ref. 40, MFCC and LPC fed a support vector machine and k-nearest neighbors 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 Ref. 46, multidimensional voice program[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 modeling techniques. The highest accuracy for multidimensional voice program features with the sustained vowel by using the Gaussian mixture model (GMM)-based system was 97.67%. The extracted MFCC with and without pitch was fed to the hidden Markov model for disorder detection, and the highest achieved accuracy was 97.75% with MFCC alone. In Ref. 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 support vector machine, artificial neural networks, GMM, and hidden Markov model, 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 Ref. 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%.

Ref. 47, an extension of Ref. 46, classified five types of disorders carried out using MFCC and fundamental frequency with sustained vowels. All speech samples of ventricular compression, gastric reflux, hyperfunction, paralysis, and A-P squeezing were considered to develop the disorder. The maximum accuracy was obtained with paralysis, and the minimum was achieved with hyperfunction. 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