https://doi.org/10.51153/kjcis.v5i2.87

Automatic Speech Recognition on Non-Pathological Dataset of Urdu Language

Anoshia Imtiaz[1] Munaf Rashid[2] Sidra Abid Syed[3]*
Hira Zahid[4] Muzzaffar Iqbal[5] Akhtar Ali Khan[6]

Abstract

Voice is a primary tool for communications and voice disorders bring atypical characteristics in the voice which influence the quality of voice. Voice disorders are abnormal conditions that influence the quality of voice. Several protocols, including acoustic analysis, can detect clinical voice pathology. Based on the computerized acoustic analysis, machine learning algorithms and non-invasive systems may play a vital part in the initial detection, tracking, and growth of proficient pathological speech analysis. The methodology proposes to collect a non-pathological dataset, i.e., a healthy voice dataset, and offers a unique combination of feature extraction techniques combining Mel-Frequency Cepstral Coefficients (MFCC), and Pitch. Support Vector Machine (SVM) was used as a machine learning classifier for the training and testing of the dataset model. The SVM algorithms demonstrated satisfactory training and testing accuracy rate, i.e., 85.886%, which proves to be a milestone on the Urdu language dataset.

Keyword: Voice dataset, Urdu language, SVM, MFCC, Pitch.

1. Introduction

The human voice is the fundamental means of communicating and delivering verbal meaning [1]. It has been reported by the ASHA (American Speech–Language–Hearing Association) [2] that vocal disorders, also known as voice pathology, have a significant influence on people's day-to-day and professional lives. Disordered voices can cause social disadvantages and inferiority complexes, especially those already marginalized. When a person's quality, pitch, or volume differs or is unsuitable for their age, sex, culture, or geographic region, they typically report having a voice problem. Expresses concern over

[1,4,5,6]*Biomedical Engineering Department, Ziauddin University Faculty of Engineering Science Technology and Management, Karachi, Pakistan*
[2]*Electrical Engineering Department & Software Engineering Department, Ziauddin University Faculty of Engineering Science Technology and Management, Karachi, Pakistan*
[3]*Biomedical Engineering Department, Sir Syed University of Engineering & Technology, Karachi, Pakistan*
Corresponding Author email id: sidra.agha@yahoo.com

developing a voice that differs from one's usual one and does not satisfy basic demands, although others do not notice the difference [3]. Generally, sounds may be divided into three categories: voiced sounds (such as vowels and nasals), unvoiced sounds (such as fricatives), and stop-consonants (e.g., plosives). Although speech originates in the lungs, it is created when air travels through the larynx and vocal cords [4]. The sound can be categorized as follows based on the health of the larynx's vocal folds: the time-periodic and harmonic voiced sound; the more noise-like unvoiced sound [5]. Speech processing is a broad and well-studied issue with applicability in telecommunications, audiovisual, and other disciplines. Real-time speech processing offers more problems than offline action. The processing includes various features, such as differentiating between utterances, identifying the speaker, etc.

Waveform and source coding are the two central coding schemes used in speech modeling [6]. When the researchers first started, they tried to copy the sounds exactly, which they dubbed waveform coding. This technique uses quantization and redundancies to try to keep the actual waveform. Instead of separating the sound into different components, it can be divided up and then each one can be modeled independently. Source coding is the term used to describe this approach of employing variables. For the identification of the spoken phrases, gender, identification of the speaker, and further speech aspects might be used. Pitch is one of the essential aspects of speech. The difference in pitch between speech signals is substantial. Vocal fold oscillation frequency influences pitch: for example, a rise of 300 Hz is produced by oscillating the folds 300 times per second. While the air travels through the folds, integer iterations of the fundamental frequency (harmonics) are also made—the pitch changes with the singer's age. In the period leading up to adulthood, the pitch is about 250 Hz. Slope ranges from 60 to 120 Hz for mature males and 120 to 200 Hz for adult women [7]. For generating speech, Rabiner, and Schafer's [8] discrete-time model utilizes linear prediction. An impulsive oscillator simulates voiced speech excitement; a glottal shaping filter then processes the impulses. A random noise generator generates the unvoiced speech. Ideally, any characteristics chosen for a speech model (1) should not be purposefully influenced by the speaker, (2) should not be independent of their physical state, and (3) resistant to any ambient noises. Although a speaker's pitch may be readily adjusted, it can also be a low-pass filter that can be applied as a feature to remove any noise and interference.

Much study has been conducted on voice processing/recognition in English, Spanish, German, and Arabic, but implementing these time-tested methods to the Urdu language has not been explored much. As a result, we should first lop a non-pathological dataset in the Urdu language to perform the speech recognition and then continue the effort for pathological datasets because voice disorders are highly effective psychological problems [2].

1. Urdu Language

The Urdu language is the official language of Pakistan. Urdu was hugely affected by Persian and Turkish and is written in a modified version of the Arabic script. The Persians adopted the Arabic letter in the 8th century, altering a few characters to represent Persian consonants found in Arabic. Some characters, such as the first two letters of the Urdu Alif and Alif MADD, may go directly beneath letters to alter the sound they make [9]. In table 1, five Urdu letters that have been chosen to form the Urdu speech dataset to conduct this study are shown with their written pronunciation.

Table 1. Pronunciation of selected Urdu letter in this study

Sr.	Letter	Pronunciation
1.	ا	Alif (‘ā)
2.	ب	Bā’y
3.	غ	Ghayn [gh(ġ)]
4.	ک	Kāf
5.	ي	Yā’ [y(ī.ay)]

2. Related Work

Al-Nasheri in Focus on an accurate and efficient method for detecting and classifying vocal disorders based on extracted features by studying the use entropy in various frequency bands of autocorrelation. The autocorrelation was used an objective of to collect the maximal peak and lag values from each spoken signal frame for disease identification and categorization. After normalizing his values as features, in addition, we calculated the entropy of the speech signal at each frame. To evaluate the contributions of each band to the sensing and classification process, these characteristics were examined across a range of frequency ranges. Several samples were extracted from three databases in for the continuation of a vowel, both for ordinary and pathological voices. The classifier was a support vector machine. If the averages of healthy and diseased samples vary considerably, then the U-tests were conducted. The best detection and classification accuracies achieved differ depending on the band, method, and database used. The most

significant bands were for both detection and classification, a frequency range of 1000 Hz to 8000 Hz is recommended. [10]. Further Al-Nasheri investigate the parameters of the Multidimensional Voice Program (MDVP) to detect, classify, and automatically classify the voice pathology in multiple data banks. The experimental results show a clear difference in the performance of these database MDVP parameters. The parameters highly ranked differentiated between databases. Three MDVP parameters adjusted in accordance with the Fisher discrimination rate yielded the highest accuracy and the most accurate parameters were obtained [11]. Lastly Al-Nasheri et, al. work focuses on developing a precise and robust extraction of features for determining, classifying, and investigating voice pathologies using correlation functions in different frequency bands. In the MEEI, they describe a new algorithm for exploring voice pathologies from the sounds of sustained vows, particularly in the light of a possible gap: the classification of the co-existing problems, which are the same as the principal phonic symptom, which implies similar interclass characteristics. Signal energy, null crossing rates, and signal entropy (SE) are used in the proposed technique to classify speech signals using the DPM, which provides an overview of the combined time and frequency data map [12].

Four diseases from the SVD dataset, including laryngitis, cyst, non-fluency syndrome, and dysphonia, were selected for analysis by Sidra et al. [13]. They extracted features from the audio signals and compared the results of four machine learning algorithms, including SVM, Nave Byes, decision tree, and ensemble classifier. They employed a comparison technique and a novel combination of features to identify laryngitis, cysts, non-fluency syndrome, and dysphonia in the SVD dataset using a purposeful sampling technique and the new features. To diagnose voice disorders with greater accuracy, a combination of particular 13 MFCC (Mel-frequency cepstral coefficients) characteristics, pitch, ZCR (zero-crossing rate), spectral flux, spectral entropy, spectral centroid, and short-term energy is used. An audio sample of 10ms has been shown to provide the best outcome when a mixture of characteristics is extracted. In the inter-classifier comparison, four machine learning classifiers, SVM (93.18%), Naive Bayes (99.45%), decision tree (100%), and ensemble classifier (51%), were used. Naive Bayes and the decision tree have the highest detection rates among these precisions. The suggested methodology's chosen collection of characteristics yields the best results using naive Bayes and decision trees. Furthermore, the SVM has been shown to be the most often utilized method for determining voice conditions. For the most part, clinical identification of voice abnormalities using machine learning algorithms has been the focus of most research, according to Sidra et al. As a result, we were able to improve the convolutional neural network's accuracy by 87.11 percent compared to the previously reported accuracy by the use of the suggested technique. The present neural network's accuracy is comparable to CNN's, and the implications were almost identical. Working with the SVD dataset's neural network for the identification of voice disorders will provide improved results in the future [14].

3. Dataset

The dataset collected to conduct this study contains non-pathological audio clips, i.e., a healthy voice at the data based on audio files that could serve the purpose of automatic speech recognition. There are a total of 37 letters in the Urdu language, out of which we have chosen five letters that can be seen in table 1. Depicting that there are many five classes in the proposed dataset. Each class contains 41 samples (each participant for every letter), so there are 205 recordings in the dataset. This dataset is still in the initial stages, adding more notes in the corpus. Dataset was prepared using the microphone of the iPhone 7 because iPhone has the best microphone installed that could easily filter the jittering in the recordings. Table 2 further represents the specifications of the proposed dataset.

Table 2. Specifications of the dataset that was developed in the proposed methodology

Specification Table	
Subject	Urdu language non-pathological voice dataset
Specific subject area	Computerized speech recognition with the help of a machine learning classifier
Type of data	Audio files
How data were acquired	Using a microphone of an iPhone
Data format	Raw, Analyzed
No. of classes	5 classes/letters
No. of samples	41 samples of each letter = 41 x 5 = 205
Parameters of data	Data was collected in a noise free environment
No. of participants	41
Demographics of participants	Forty-one participants include both men and women in between the range of 18 till 36 years.

4. Methodology:

In figure 1. SVM (Support Vector Machine) is used as a classifier to test the dataset collected in the first step of this paper. The training and testing dataset is divided into the ratio of 80% and 20%. MATLAB classification app is used to train the model. SVM is a decent tool for designing a speech recognition system. It tries to set up a threshold among classes, allowing labels to be anticipated from more than one vectors. This option, known as the hyper-plane, is chosen so that it is as far away as possible from the closest data points in each class. Support vectors are defined as the points that are closest to each other [15]. SVM classifier creates the most innovative hyperplane in the transformed entrance space, differentiates the exceptional groups, and maximizes the distance to nearby

cleanly separated instances. The variables of the hyperplane approach are a quadratic optimization problem [16]. To label a dataset, do the following:

$$(x1,y1)\ldots\ldots(xn,yn),xi \in Rd \qquad (1)$$

Where xi is a vector representation of a characteristic and yi is a class mark (negative or positive) of a practice formula i. The optimal hyperplane is then described as follows:

$$wxT + b=0 \qquad (2)$$

Where w denotes the weight matrix, x denotes the input vector, and b denotes the bias W and b must satisfy the following in equalities for all components of the training collection.

$$wxT+b \geq +1 \qquad if \quad yi= +1 \qquad (3)$$

$$wxT+B \leq -1 \qquad iF\ yi=-1 \qquad (4)$$

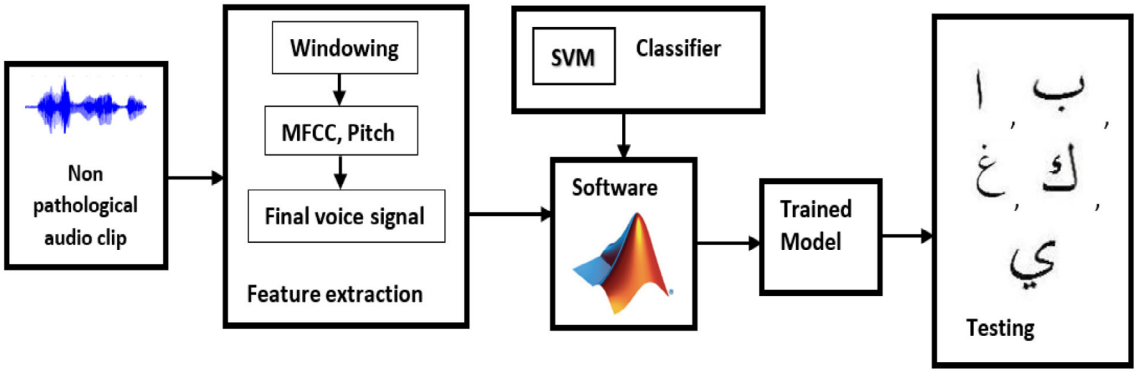

Figure 1. The proposed methodology that used two core features, namely MFCCs and pitch, are extracted from each audio clip and fed to an SVM classifier to predict 5 Urdu letters.

4.1. MFCC

The Mel frequency cepstral coefficient is influenced by the human auditory cortex. As per the perception research, the human auditory system does not work on a linearly inbound sound with an initial frequency 'f' recorded in hertz (Hz) and a pitch defined on the Mel scale [17]. MFCC is defined as coefficients deduced from audio signals. The voice input is the audio signal input that goes through the framing process. Before framing, the audio signal goes through a pre-emphasis process, which helps achieve accuracy and efficiency. This process compensates for the higher frequency suppressed in the human auditory

system throughout sound production.

$$C2\ (n) = c(n) - d * c(n-1) \quad (1)$$

Here, c2(n) represents the output signal, and the recommended values for d are 0.9 and 1. The z transform of the filter is as follows:

$$H\ (Z) = 1 - D * Z^{-1}) \quad (2)$$

Following the pre-emphasis process, the goal is to divide the entire audio signal into several frames so that each frame signal can be easily analyzed and interpreted. The audio signal is divided into 10 ms frames, whereas the standard framing size is 25 ms [18]. It demonstrates that the frame length for a 50 kHz audio signal is 50 k * 0.01 = 500 samples. The framing step allows the frames to overlap. There is a frame step of 10 ms for the first 500 samples. It starts at sample 0 and continues till the final audio signal has been heard, at which point the 500-sample structure ends. Since vowel recordings are included in the SVD dataset, the dataset specifies a recording duration of 10 ms, hence the 10 ms signal is used. When utilizing the Hamming window function, spectral artefacts may be minimized after framing. Convolution in the frequency domain results from the combination of short-term spectrum and window transfer function (hamming). Each frame must be multiplied with hamming window [19] to maintain continuity between the first and last marks in the frame. The hammering window function is described below.

$$K(n) = 0.54 = 0.64 \cos(2nnN - 1) \quad (3)$$

K(n) denotes the window, and Q (n) means the output, whereas X (n) represents the input frame signal.

$$Q\ (n) = Xn) * w(n) \quad (4)$$

$$Q(w) = FFT[k(t) * X(t)] \quad (5)$$

$$Q(w) = k(w) * X(w) \quad (6)$$

The signal's original strength is now transformed to the Mel frequency using the Mel filter bank. Neither filter has a constant distance between them, nor is the number of filters in the higher frequency range less than those found in the lower frequency range. Filter banks are the only ones that can be used on signals in both the time domain and frequency domain. When processing Mel frequency cepstral coefficients in the frequency domain, it is important (MFCC). Figure 2 shows the filter applied in the lower and higher frequency regions to demonstrate the frequency change. Pitch and frequency may be linked by using the Mel scale. Low-frequency pitch variations may be distinguished from those occurring at higher frequencies by the human hearing system. [20] The Mel scale

is used to extract elements that are specific to human hearing. Using this mathematical technique, the frequency response may be converted into the Mel Scale:

$$M(f) = 1125 * \ln(1 = f700) \qquad (7)$$

Where f is the frequency of the audio signal.

5. Results:

In graph 1, the classifier's accuracy is calculated from the below formula. Whereas after finding the individual accuracies, the average accuracy is 85.866%.

$$\text{AUC} = \frac{\text{true positive+true negative}}{\text{true positive+true negative+false positive+false negative}}$$

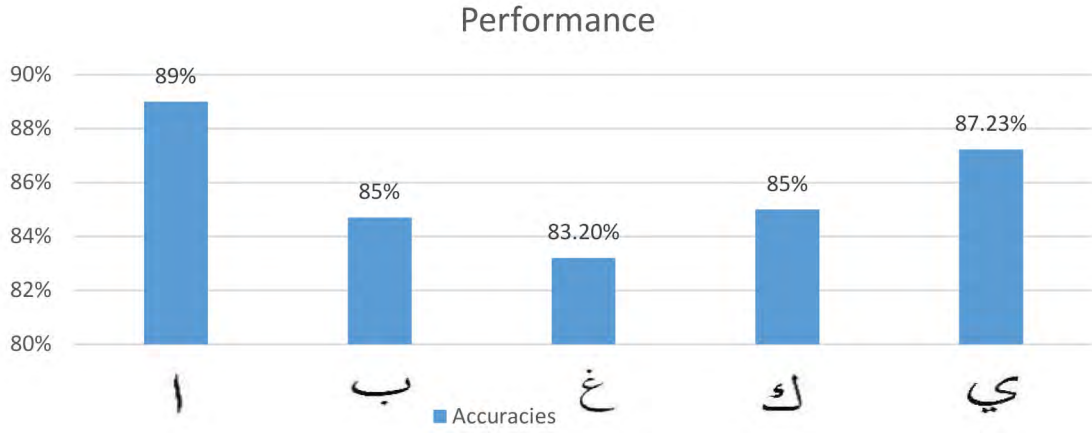

Figure 2. Accuracies of the voice signals of the Urdu letters

6. Conclusion:

Several studies have been conducted to detect voice pathologies. Because all cords are paralyzed, as a result of vocal cord paralysis, individuals often have a limited ability to speak or breathe. Invasive for patients, the screening test used to categorize these diseases of speech is intrusive in nature, so machine learning exploration and development have increased in recent years. The lack of analysis and automatic recognition of speech disorders of Urdu language encouraged authors to collect a non-pathological dataset. The average accuracy of the collected non-pathological dataset when trained and tested through the SVM classifier is 85.886%. In the future, we have planned to create a pathological dataset and perform automatic speech recognition by following the same method.

References

[1] Graham Williamson. Human Communication: A Linguistic Introduction (2nd Edition) 2006.

[2] ASHA Clinical Topics. Voice disorders. Website, 2019. https://www.asha.org/PracticePortal/Clinical-Topics/Voice-Disorders

[3] Michael J. Clark James Hillenbrand, Laura A. Getty, and Kimberlee Wheeler. Acoustic characteristics of American English vowels. Journal of the Acoustical Society of America, 97(1):3099–3111, 1995.

[4] T. Parsons. Voice and Speech Processing. McGraw-Hill College Div., Inc, 1986.

[5] G. C. M. Fant. Acoustic Theory of Speech Production. Mouton, Gravenhage, 1960.

[6] J. R. Deller, J. G. Proakis, J. H. L. Hansen. Discrete-Time Processing of Speech Signals. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1993.

[7] D. O'Shaughnessy. Speech Communication: Human and Machine. Addison Wesley Publishing Co., 1987.

[8] L. R. Rabiner, R. W. Schafer. Digital Processing of Speech Signals. Prentice-Hall, Inc., Englewood Cliffs, 1978.

[9] "ATLAS - Urdu: Urdu Language", Ucl.ac.uk, 2021. [Online]. Available: https://www.ucl.ac.uk/atlas/urdu/language.html. [Accessed: 17- Sep- 2021].

[10] Z. Ali et al., "Intra- and inter-database study for Arabic, English, and German databases: Do conventional speech features detect voice pathology?," J. Voice, vol. 31, no. 3, pp. 386.e1-386.e8, 2017.

[11] A. Al-Nasheri et al., "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," IEEE Access, vol. 6, pp. 6961–6974, 2018.

[12] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of voice pathology detection and classification on different frequency regions using correlation functions," J. Voice, vol. 31, no. 1, pp. 3–15, 2017.

[13] S. A. Syed, M. Rashid, S. Hussain, A. Imtiaz, H. Abid, and H. Zahid, "Inter classifier comparison to detect voice pathologies," Math. Biosci. Eng., vol. 18, no. 3, pp. 2258–2273, 2021.

[14] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative analysis of CNN and RNN for voice pathology detection," Biomed Res. Int., vol. 2021, p. 6635964, 2021.

[15] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of support vector machine (SVM) learning in cancer genomics, Cancer Genomics-Proteomics, 15 (2018), 41–51.

[16] A. Shmilovici, Support vector machines, in Data Mining and Knowledge Discovery

Handbook, Springer, Boston, MA, (2009), 231–247.

[17] S. Memon, M. Lech, L. He, Using information theoretic vector quantization for inverted MFCC based speaker verification, in 2009 2nd International Conference on Computer, Control and Communication, IEEE, (2009), 1–5.

[18] M. Sahidullah, G. Saha, On the use of distributed data in speaker identification, in 2009 Annual IEEE India Conference, IEEE, (2009), 1–4.

[19] Ö. Eskidere, A. Gürhanlı, Voice disorder classification based on multitaper mel frequency cepstral coefficients features, Comput. Math. Methods Med., 2015 (2015), 956249.

[20] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd edition, Wiley-Interscience, USA, 2000.