Journal of Disability Research

2024 | Volume 3 | Pages: 1–8 | e-location ID: e20230063
DOI: 10.57197/JDR-2023-0063

Using Voice Technologies to Support Disabled People

H. E. Semary[1,2,*], Khamis A. Al-Karawi[3,4] and Mahmoud M. Abdelwahab[1,5]

[1]Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia
[2]Department of Statistics and Insurance, Faculty of Commerce, Zagazig University, Zagazig, Egypt
[3]School of Science, Engineering, and Environment, Salford University, Great Manchester, UK
[4]Department of Computer Science, College of Science, Diyala University, Baqubah, Diyala, Iraq
[5]Department of Basic Sciences, Higher Institute of Administrative Sciences, Osim, Egypt

Correspondence to:
H. E. Semary*, e-mail: hesemary@imamu.edu.sa, Tel.: +201003527613
Khamis A. Al-Karawi, e-mail: k.a.yousif@edu.salford.ac.uk
Mahmoud M. Abdelwahab, e-mail: mmabdelwahab@imamu.edu.sa

Received: August 31 2023; Revised: November 29 2023; Accepted: November 30 2023; Published Online: January 5 2024

ABSTRACT

In recent years, significant strides have been made in speech and speaker recognition systems, owing to the rapid evolution of data processing capabilities. Utilizing a speech recognition system facilitates straightforward and efficient interaction, especially for individuals with disabilities. This article introduces an automatic speech recognition (ASR) system designed for seamless adaptation across diverse platforms. The model is meticulously described, emphasizing clarity and detail to ensure reproducibility for researchers advancing in this field. The model's architecture encompasses four stages: data acquisition, preprocessing, feature extraction, and pattern recognition. Comprehensive insights into the system's functionality are provided in the Experiments and Results section. In this study, an ASR system is introduced as a valuable addition to the advancement of educational platforms, enhancing accessibility for individuals with visual disabilities. While the achieved recognition accuracy levels are promising, they may not match those of certain commercial systems. Nevertheless, the proposed model offers a cost-effective solution with low computational requirements. It seamlessly integrates with various platforms, facilitates straightforward modifications for developers, and can be tailored to the specific needs of individual users. Additionally, the system allows for the effortless inclusion of new words in its database through a single recording process.

KEYWORDS

speech recognition, mel frequency cepstral coefficients, assistive technology, disabilities

INTRODUCTION

Speech technologies have the potential to greatly assist individuals with disabilities, contributing to their equality and inclusion in society and everyday life (Delić et al., 2013). Speech, being a natural and intuitive form of human communication, holds great potential as an ideal medium for human–computer interaction. The ability of machines to accurately mimic and understand human speech would provide a straightforward solution to this challenge. Among the various applications of speech technology, one of the most promising is the development of spoken dialogue systems, which enable users to access information in a simple, direct, and hands-free manner (Aggarwal and Dave, 2012; Zheng and Li, 2017; Ross et al., 2020; Alenizi and Al-Karawi, 2023a). This becomes particularly crucial when users have disabilities that hinder their ability to interact with systems using standard methods. However, achieving reliable automatic speech recognition (ASR) is not straightforward (Rosdi and Ainon, 2008; Al-Karawi, 2015; Al-Karawi and Ahmed, 2021). Factors such as variations in speech patterns among different speakers, background noise levels, variations in pronunciation speed, and user mood can introduce errors that significantly impact the accuracy of speech recognition systems, leading to low success rates. As a result, many current systems are limited to controlled environments or cater to specific user groups, often requiring particular microphone positioning or other constraints (Vieira et al., 2022). This limitation results in interfaces that may feel unnatural and restrictive (Bedoya and Muñoz, 2012; Al-Karawi and Mohammed, 2019). Efforts to overcome these challenges are ongoing, with researchers and developers striving to improve the robustness and adaptability of ASR systems.

By addressing factors such as speaker variability, noise robustness, and user-centric design, the goal is to create more inclusive and user-friendly interfaces that accommodate a more comprehensive range of users, including those with disabilities. Despite the complexity of the task, advancements in speech technology continue to pave the way for more seamless and effective human–computer interaction

(Busatlic et al., 2017). As research progresses, speech recognition systems are expected to become more accurate, versatile, and accessible, enabling individuals with disabilities to benefit from natural and intuitive technological interactions (Delić et al., 2013; Al-Karawi, 2021). Over the years, numerous approaches have been proposed for ASR (Abushariah et al., 2010; Ibarra and Guerrero, 2010; Bedoya and Muñoz, 2012). Among these, the most robust methods are based on Hidden Markov Models (HMM) (Al-Karawi and Ahmed, 2021; Al-Karawi and Mohammed, 2021). While these HMM-based systems have achieved significant accuracy, they continue to face challenges related to high computational costs.

The high cost of commercial ASR systems and copyright restrictions limit access for users and researchers. To address these challenges, an in-house ASR system is needed. This solution is simple, computationally efficient, accessible, reliable, and adaptable to any platform, reducing computational burden and allowing for customization (Mohammed et al., 2021; Alenizi and Al-Karawi, 2023a). Our ASR system aims to provide a cost-effective, reliable, and accessible solution for computational resources, promoting openness, collaboration, and innovation in speech recognition, by eliminating the limitations of existing commercial systems (Mohammed et al., 2020; Al-Karawi, 2023). Our ASR development approach offers flexibility in system improvements, customization, and exploration, enhancing speech recognition technology accessibility, affordability, and adaptability for various applications. The remaining sections of this paper are organized as follows. The following sections describe speech recognition systems for disabled people: Speech Recognition Techniques, State of the Art, Proposed Models, Experiments and Results, and Conclusions.

SPEECH RECOGNITION SYSTEMS FOR DISABLED

Technological advancements, particularly speech recognition technology, are revolutionizing the lives of disadvantaged and disabled individuals, enhancing their daily lives and overall well-being (Noyes and Frankish, 1992). Speech recognition technology is the foundation of popular voice assistants like Siri, Amazon Echo, and Google Assistant, enabling computers to understand and process human spoken language (Jiang et al., 2000; Alenizi and Al-Karawi, 2023b). Speech recognition technology enables natural, intuitive communication through voice interaction, providing a transformative solution for individuals with disabilities or limitations that make traditional input methods challenging (Azam and Islam, 2015; Alenizi and Al-karawi, 2022). Speech recognition technology is advancing, enabling individuals to control devices, access information, perform tasks, and engage with digital platforms, enhancing independence and quality of life. This technology holds promise for creating inclusive environments, enabling everyone to fully participate (Noyes et al., 1989). ASR, or voice recognition, helps disabled individuals with limited mobility or visual impairments by converting human speech into machine-readable language. As speech recognition technology advances, it transforms devices into digital assistants, enhancing efficiency and productivity, especially for those with limited upper-limb mobility.

Speech recognition technology can also assist older people and individuals with speech and hearing impairments (Gonzalez et al., 2016; Alenizi and Al-Karawi, 2023a). Considering the estimated 15 million disabled individuals in the United States alone, with millions worldwide, the potential benefits of speech recognition are immense. By leveraging the power of speech recognition, we can empower disabled individuals to navigate and interact with digital devices more effectively, fostering greater independence and inclusivity. The widespread adoption of speech recognition technology promises to transform millions of individuals' lives, enabling them to quickly overcome communication barriers and access the digital world (Isyanto et al., 2020; Al-Karawi and Mohammed, 2023).

SPEECH RECOGNITION TECHNIQUES

As the market continues to witness the proliferation of devices like Siri on iPhone and Alexa from Amazon, speech recognition has gained significant influence in our daily lives. At its core, speech recognition enables machines to hear, comprehend, and respond to the information conveyed through speech. The primary objective of ASR is to assess, extract, analyze, and recognize spoken speech to obtain meaningful information (Gaikwad, 2010). Therefore, it becomes crucial to comprehend and examine the techniques involved in the comprehensive identification and understanding of speech. The speech recognition system comprises three main stages, as depicted in Figure 1: feature extraction, modeling, and performance evaluation.

STATE OF THE ART

A comprehensive analysis of the current state of ASR systems was conducted using the Tree of Science tool developed at Universidad Nacional de Colombia. This systematic review aimed to identify influential articles that have contributed to advancements in the accessibility, accuracy, and efficiency of ASR systems. In the early stages of speech processing research, the short-term spectral amplitude technique, employing the minimum mean square error estimator, was widely used (Ephraim and Malah, 1985). Although complex, this algorithm offered higher accuracy than other methods available at that time. Subsequently, the utilization of more robust approaches based on HMM gained prominence. These techniques incorporated mel frequency cepstral coefficients (MFCCs) for feature extraction (Gales and Young, 2008). HMM continues to be an essential method for continuous speech recognition systems with large vocabularies due to its reliable performance. Another unique method mentioned in the literature is PARADE (Periodic Component to Aperiodic

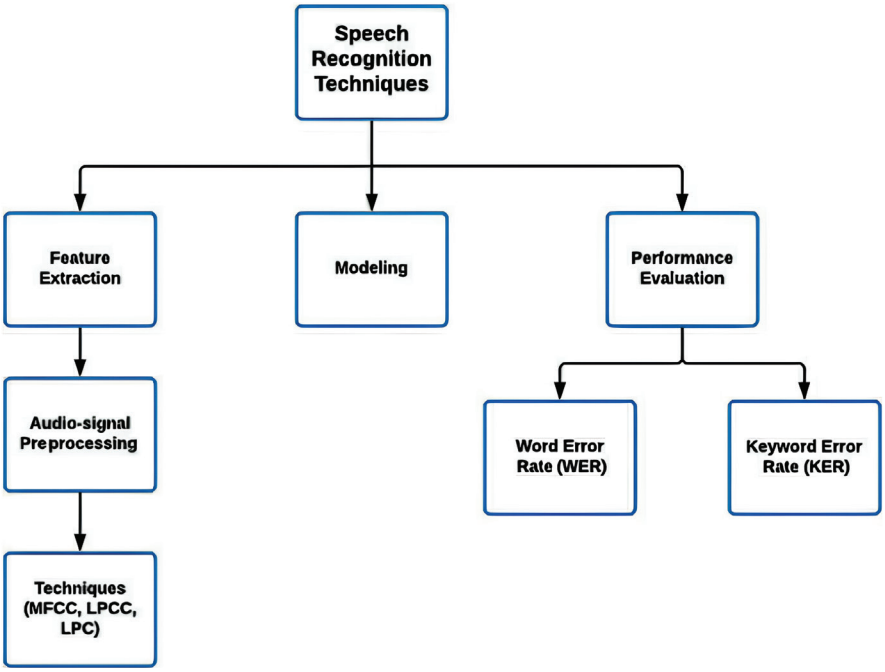

Figure 1: Primary subdivisions in speech recognition systems (Al-Karawi and Mohammed, 2023).

Component Ratio-based Activity Detection), combined with a feature extraction technique known as SPADE (Subband-based Periodicity and Aperiodicity Decomposition) (Ishizuka et al., 2010). This approach has demonstrated significantly improved accuracy in word recognition. Dynamic time warping (DTW), a widely used algorithm known for its low computational cost, has been discussed by Zhang et al. (2014). However, DTW is limited to small vocabularies. Current research efforts focus on achieving accurate speech recognition and developing tools for word segmentation, which involves identifying individual words' start and endpoints. This segmentation aims to reduce the complexity associated with continuous speech recognition (Komatani et al., 2015). By examining the progress made by the scientific community, this systematic review provides valuable insights into the evolution of ASR systems. The adoption of robust methods such as HMM, along with advancements in feature extraction techniques and word segmentation, has contributed to improved accuracy and efficiency. Continued research in this field aims to further enhance the recognition of speech, expand the vocabulary size, and develop innovative tools to simplify the processing complexity of continuous speech recognition systems.

PROPOSED MODEL

We propose a system that facilitates the recognition of a specific set of voice commands and can be seamlessly integrated with various applications, including virtual learning tools. Interacting with applications through voice commands can be a powerful tool for inclusion, especially when it offers user-friendly features and accessibility options for both end users and developers. Our model was explicitly designed to cater to individuals with physical and sensory disabilities, focusing on enabling interaction with digital learning resources. While already existing tools like job access with speech (JAWS) serve similar purposes, we aimed to create an experimental tool that could be easily integrated with other developments to address diverse needs. In particular, our system was incorporated into the global astrometry interferometer for astrophysics (GAIA) tools framework, which is dedicated to constructing accessible learning objects for individuals with visual disabilities. The initial version of GAIA tools includes various authoring tools such as a dictionary, text editor and reader, a game for learning, and assessment through questionnaires.

Furthermore, it guides designers in developing learning objects and enables visually impaired users to interact effectively with them in educational activities (Gonzalez et al., 2016). By proposing this system, we aim to enhance accessibility and inclusivity in the learning environment, ensuring that disabled individuals can engage in educational activities more effectively. Integrating voice commands and interaction capabilities with the GAIA tools' framework provides new possibilities for accessible learning and opens doors for further developments in this field. In the context of developing countries, where socioeconomic limitations are more prominent, tools like the one proposed here hold particular significance. The motivation behind focusing on an audio recognition system as a complement to educational tools stems from the understanding that education plays a fundamental role in overcoming socioeconomic challenges. Currently, the system can recognize 10 isolated words in low-noise environments. A straightforward process can expand the system to cater to the specific needs of individual users.

Users are required to record new words to increase the system's database, allowing for customization. It is worth noting that while the system was initially designed to recognize words in Spanish, there is no limitation if someone

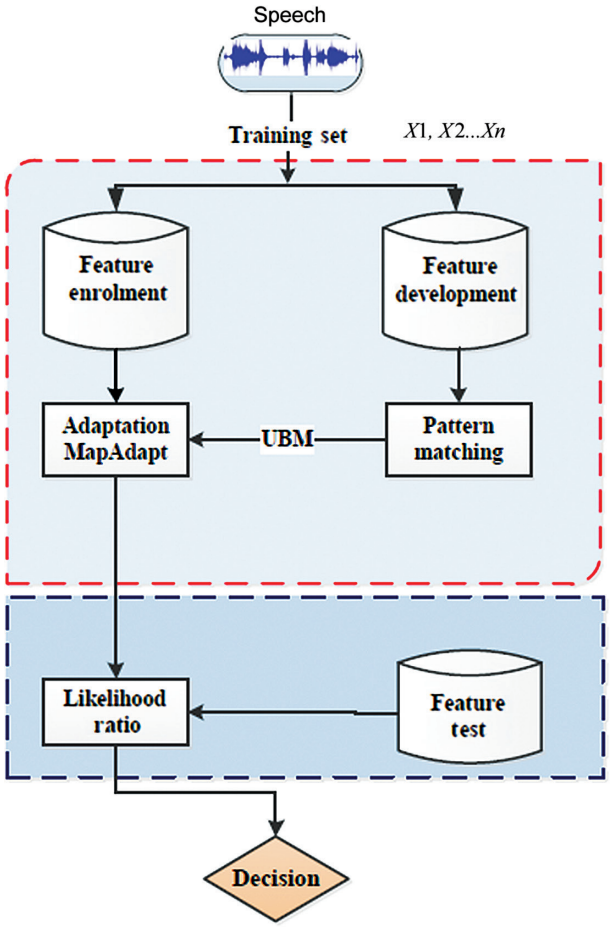

Figure 2: The architecture of the proposed model (Al-Karawi and Ahmed, 2021).

needs to add new commands to the database in another language. The proposed system is divided into four stages, as illustrated in Figure 1. The first stage involves acquiring the audio signal that contains the information to be recognized. Subsequently, a preprocessing stage is implemented to filter out noise and unwanted segments, such as silence at the beginning and end of the recording. The next stage is feature extraction, where a matrix containing MFCCs is calculated. Finally, the system employs an algorithm to compute the Euclidean distance, comparing the feature matrices with the corresponding patterns stored in the database. This decision-making process completes the recognition process. By offering a customizable and efficient audio recognition system, we aim to address the specific needs of users in educational settings, particularly in developing countries. This tool can potentially empower individuals by enhancing access to educational resources, thereby contributing to the overall socioeconomic development of these regions.

Speech samples

In this stage, the user speaks the word to be recognized. The system records an audio vector consisting of $t * Fs$ samples, where t represents the duration of the recording, and Fs denotes the sampling frequency. In this case, the recording duration is 2 seconds, and the sampling frequency is 44,100 Hz. The audio recording is in the mono format, meaning only one audio channel is recorded.

Preprocessing

The preprocessing stage comprises three steps: filtering, normalization, and silence suppression. To perform filtering, a spectral analysis process was conducted on multiple audio recordings to identify expected noise frequencies in various environments (office, outdoor, home, etc.). A general frequency range where helpful information is expected was established. The best-performing filter was a Hamming windowing finite impulse response (FIR) filter, specifically a band-pass filter with a sampling frequency of 44,100 Hz and cut-off frequencies of 200 Hz and 8000 Hz. Once the signal was filtered, it underwent normalization to restrict its values to a standardized range between 0 and 1. This normalization step is particularly important when comparing multiple audio samples with different amplitude ranges due to varying speaker intensity or noise levels. Normalizing the signal facilitates the subsequent steps in the process. The next step involves silence detection and suppression. The signal is segmented, and the energy of each segment is calculated. Segments with energy values below a defined threshold are considered nonuseful and are discarded. The energy threshold was determined by comparing the noise energy values with those obtained from randomly pronounced words. This approach helps reduce the computational cost of the algorithm by excluding signal segments that do not provide relevant information and may hinder the recognition process.

Feature extraction

The time-domain waveform of a speech signal, representing the amplitude variations over time, contains essential auditory information. However, to extract meaningful information from the waveform, it is necessary to condense the data of each segment into a limited number of parameters or characteristics while retaining the signal's discriminatory power (Delić et al., 2013; Delić et al., 2014; Ajibola Alim and Rashid, 2018; Alenizi and Al-Karawi, 2023b). As the number of input voice samples increases, the accuracy of the speech recognition systems tends to decrease (Gaikwad, 2010; Virkar et al., 2020). This highlights the significance of feature extraction in achieving accurate speech processing. Feature extraction plays a critical role in representing a speech signal using a predetermined set of signal components that are more distinctive and reliable. These features should effectively capture the characteristics of each segment, enabling the grouping of similar segments based on their shared characteristics (Shrawankar and Thakare, 2013). Feature extraction is an initial step in ASR preprocessing or front-end signal processing. Over the years, several approaches have been developed for extracting features from audio signals, drawing from extensive research in mathematics, acoustics, and speech technology (Ajibola

Alim and Rashid, 2018). Feature extraction is closely intertwined with model variable selection, which is another crucial aspect that can significantly impact the performance of a speech processing system. Proper selection and inclusion of relevant model variables are essential for achieving accurate and reliable results. The MFCCs are employed as the feature representation for each command in the ASR system. These coefficients have been recognized in the state-of-the-art review as effective for achieving accurate results while maintaining low computational costs. A series of steps are involved to compute the MFCCs.

First, a perceptually spaced triangular filter bank is applied to the discrete Fourier-transformed signal. This filter bank captures the important frequency components of the audio signal. The resulting filter outputs undergo logarithmic compression to compress the energy values. Next, the discrete cosine transform is applied to the logarithmically compressed filter-output energies. This transformation decorates the coefficients, making them more suitable for speech recognition tasks. The MFCCs are derived from this process, resulting in decor-related parameters (Hossan et al., 2010; Delić et al., 2013; Alenizi and Al-Karawi, 2023a). First, the audio signal is segmented into intervals of 1024 sample length, with a 410-sample overlap between segments to avoid information loss during windowing using a Hamming window in the time-domain. This windowing process attenuates the beginning and end of each segment. A total of 14 MFCCs (excluding the 0th coefficient) are calculated for each segment. This is achieved by applying a logarithmically compressed filter bank, derived from a perceptually spaced triangular filter bank, to the discrete Fourier-transformed signal. In this case, 30 filters are used in the filter bank, with a low-end frequency of 0 and a high-end frequency of 0.1815. The calculated MFCCs are stored in a matrix of size $N \times C$, where N represents the number of segments into which the audio signal was divided, and C represents the number of MFCCs calculated (which is 14 in this case). The number of rows in the feature matrices may vary due to the silence suppression process applied to the recorded signals, which can result in different lengths. The parameters for calculating the MFCCs were selected based on the results obtained through various tests, including those described in the next section. For each command that the system can recognize, a "pattern matrix" of features is calculated and stored, representing the class of the command. Additionally, whenever a new recording is entered for recognition, a new matrix of features, known as the "new matrix," is calculated.

$$d(x,y) = \sqrt{\sum_j (x_i - y_j)^2} \quad (1)$$

Decision stage

This stage aims to associate the newly calculated matrix with a specific command to determine the corresponding recording. This is achieved by comparing the new matrix with the pattern matrices and identifying the best match. To compare the matrices, it is essential to note that each row represents a segment of the audio signal in terms of the MFCCs. Therefore, comparing each row of the new matrix with the rows of the pattern matrix is equivalent to comparing each segment of the new recording with the segments of the pattern recordings. However, there is a challenge regarding time alignment, as the two recordings may not match in terms of pronunciation duration, resulting in the comparison of different audible segments of the same word. To address this issue, an individual error is calculated using the Euclidean distance between one row of the new matrix and each row of the pattern matrix. The minimum individual error is then determined, and this value is considered as the contribution of the analyzed row to the total error (Al-Karawi, 2015; Zheng and Li, 2017; Ross et al., 2020; Al-Karawi and Ahmed, 2021; Alenizi and Al-Karawi, 2023a). This process is repeated for each row of the new matrix. Ultimately, the total errors are calculated for each class, and by identifying the minimum error, it becomes possible to determine which command was pronounced in the entered recording. Although this process may appear complex, it is performed through algebraic arrangements rather than iterative calculations, significantly reducing computational costs. Finally, it has been determined that the minimum total error value must be below a certain threshold to ensure reliability. If the minimum error exceeds this threshold, the user is prompted to repeat the command more clearly to achieve an acceptable level of reliability.

EXPERIMENTS AND RESULTS

To evaluate the performance of the implemented model, a test was conducted involving 30 participants with diverse characteristics, including different genders and ages. The test was conducted in moderate noise environments such as study rooms and bedrooms. The experiment consisted of two parts: the first involved creating a database with the participants' voices, and the second focused on validating the system. During the database-creation phase, participants were instructed to speak clearly and naturally while repeating 10 specific words in Spanish: open, back, center, right, left, save, home, help, view, and internet. Each participant repeated these words four times, resulting in 40 recordings per user. The system then determined whether it recognized the spoken word successfully in each attempt. The detailed results of these tests are presented in Tables 1 and 2.

Performance evaluation

Evaluating the performance of a classification model is crucial to assess its effectiveness in achieving a desired outcome. Performance evaluation metrics quantitatively assess the model's performance on a test dataset. Selecting appropriate metrics to evaluate the model's performance accurately is

Table 1: Elements of a confusion matrix.

	Predictive values	
Actual values	True positive (TP)	False positive (FP)
	False negative (FN)	True negative (TN)

Table 2: Results of the tests.

User	Number of hits	Percentage (%)
1	36	90
2	39	97.5
3	34	85
4	36	90
5	34	86
6	36	91
7	36	91
8	40	100
9	34	86
10	32	80
11	32	80
12	34	86
13	39	97.6
14	39	97.5
15	34	86
16	39	97.6
17	38	95
18	32	81
19	32	81
20	32	81
21	27	68.6
22	36	66
23	36	91
24	28	71
25	36	91
26	34	86
27	39	97.6
28	34	86
29	33	83.5
30	38	95

essential. Several metrics can be utilized, including the confusion matrix, accuracy, specificity, sensitivity, and more. The following formulas are commonly employed to calculate these performance metrics:

$$Specificity = \frac{TN}{TN + TP} \quad (2)$$

$$True\ Positive\ Rate\ or\ Sensitivity = \frac{TN}{TN + FN} \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TN + TP + FP + FN} \quad (4)$$

In the above mentioned formulas (i.e. accuracy, precision, sensitivity, specificity, and F1) and the confusion matrix, TPs denote true positives, TNs refer to true negatives, FPs signify false positives, and FNs represent false negatives. The confusion matrix provides insights into the percentages of accurate and inaccurate classifications for each class, pinpointing precisely which classes the algorithms encounter the most challenges in classification for the trained models. TP and TN signify the number of data points belonging to the positive and negative classes, respectively, where the model correctly identifies them. Conversely, FP indicates the number of negatives erroneously classified as positives, and FN represents the count of positives mistakenly classified as negatives by the machine.

CONCLUSIONS

The proposed ASR system serves as a tool to complement the development of educational platforms, aiming to enhance accessibility for individuals with visual disabilities, among others. While the system's recognition accuracy may not match that of commercial systems, it offers several advantages, such as low computational cost, seamless integration with other platforms, ease of customization, and the ability to adapt to the specific needs of individual users. Additionally, the system allows for the inclusion of new words in its database with a single recording, providing flexibility and ease of updates. As part of their future work, the authors have outlined several areas for improvement in the ASR system. They plan to explore other characterization methods, such as autoregressive

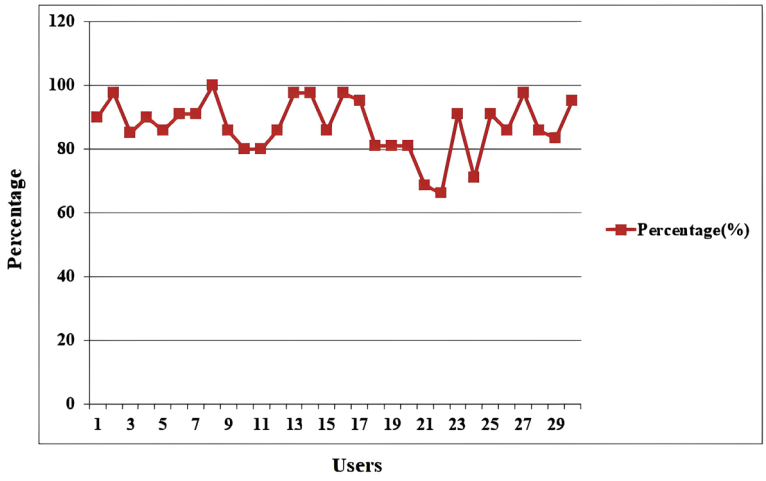

Figure 3: The results of the tests.

coefficients or strategies of classification like HMM, to enable continuous speech recognition and enhance the system's interaction with applications.

Additionally, they aim to enhance the preprocessing stage by implementing more efficient filters, improving the system's robustness, and mitigating issues related to tone, pronunciation, and noise variations. Furthermore, a key objective is to achieve system generalization, allowing it to recognize any user without the need for prior registration of their voices. This expansion will enhance the system's usability and make it more accessible to a broader range of users.

ACKNOWLEDGEMENTS

The authors extend their appreciation to the King Salman Centre for Disability Research for funding this work through Research Group no KSRG-2023-240.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest in association with the present study.

REFERENCES

Abushariah M.A., Ainon R.N., Zainuddin R., Elshafei M. and Khalifa O.O. (2010). Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools. In: *International Conference on Computer and Communication Engineering (ICCCE'10)*, IEEE.

Aggarwal R. and Dave M. (2012). Recent trends in speech recognition systems. In: *Speech, Image, and Language Processing for Human Computer Interaction: Multi-modal Advancements*, IGI Global; pp. 101-127.

Ajibola Alim S. and Rashid K.A. (2018). Some commonly used speech feature, *extraction algorithms.* In: *From Natural to Artificial Intelligence—Algorithms and Applications* (Lopez-Ruiz R., ed.), IntechOpen, London, UK.

Alenizi A.S. and Al-karawi K.A. (2022). Cloud computing adoption-based digital open government services: challenges and barriers. In: *Proceedings of Sixth International Congress on Information and Communication Technology*, Springer, London.

Alenizi A.S. and Al-Karawi K.A. (2023a). Effective biometric technology used with Big data. In: *Proceedings of Seventh International Congress on Information and Communication Technology*, Springer, London.

Alenizi A.S. and Al-Karawi K.A. (2023b). Internet of things (IoT) adoption: challenges and barriers. In: *Proceedings of Seventh International Congress on Information and Communication Technology*, Springer, London.

Al-Karawi K.A. (2015). Automatic speaker recognition system in adverse conditions—implication of noise and reverberation on system performance. *Int. J. Inform. Electron. Eng.*, 5(6), 423.

Al-Karawi K.A. and Mohammed D.Y. (2019). Early reflection detection using autocorrelation to improve robustness of speaker verification in reverberant conditions. *Int. J. Speech Technol.*, 22, 1077-1084.

Al-Karawi K.A. (2021). Mitigate the reverberation effect on the speaker verification performance using different methods. *Int. J. Speech Technol.*, 24(1), 143-153.

Al-Karawi K.A. and Ahmed S.T. (2021). Model selection toward robustness speaker verification in reverberant conditions. *Multimed. Tools. Appl.*, 80, 36549-36566.

Al-Karawi K.A. and Mohammed D.Y. (2021). Improving short utterance speaker verification by combining MFCC and entrocy in noisy conditions. *Multimed. Tools Appl.*, 80(14), 22231-22249.

Al-Karawi K.A. (2023). Face mask effects on speaker verification performance in the presence of noise. *Multimed. Tools Appl.*, 82, 1-14.

Al-Karawi K.A. and Mohammed D.Y. (2023). Using combined features to improve speaker verification in the face of limited reverberant data. *Int. J. Speech Technol.*, 26, 789-799.

Azam G. and Islam M. (2015). Design and fabrication of a voice controlled wheelchair for physically disabled people. In: *International Conference on Physics Sustainable Development & Technology (ICPSDT-2015)*, CUET, Bangladesh.

Bedoya W.A. and Muñoz L.D. (2012). Methodology for voice commands recognition using stochastic classifiers. In: *2012 XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*, IEEE, Medellin, Colombia.

Busatlic B., Dogru N., Lera I. and Sukic E. (2017). Smart homes with voice activated systems for disabled people. *TEM J.*, 6(1), 103-107.

Delić V., Sečujski M., Bojanić M., Knežević D., Vujnović Sedlar N. and Mak R. (2013). Aids for the disabled based on speech technologies-case study for the Serbian language. In: *11th International Conference on ETAI*, Ohrid, Macedonia; pp. E2-1.1-4.

Delić V., Sečujski M., Vujnovic Sedlar N., Miskovic D., Mak R. and Bojanic M. (2014). How speech technologies can help people with disabilities. In: *Speech and Computer: 16th International Conference, SPECOM 2014, Novi Sad, Serbia, October 5-9, 2014. Proceedings 16*, Springer.

Ephraim Y. and Malah D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans Acoust. Speech Signal Process.*, 33(2), 443-445.

Gaikwad V.J. (2010). Application of chemoinformatics for innovative drug discovery. *Int. J. Chem. Sci. Appl.*, 1(1), 16-24.

Gales M. and Young S. (2008). The application of hidden Markov models in speech recognition. *Found. Trends® Signal Process.*, 1(3), 195-304.

Gonzalez R., Muñoz J., Salazar J. and Duque N. (2016). Voice recognition system to support learning platforms oriented to people with visual disabilities. In: *Universal Access in Human-Computer Interaction. Users and Context Diversity: 10th International Conference, UAHCI 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part III 10*, Springer.

Hossan M.A., Memon S. and Gregory M.A. (2010). A novel approach for MFCC feature extraction. In: *2010 4th International Conference on Signal Processing and Communication Systems*, IEEE, Gold Coast, QLD.

Ibarra J.P. and Guerrero H.B. (2010). Identificación de comandos de voz utilizando LPC y algoritmos genéticos en Matlab. *Revista CINTEX*, 15, 36-48.

Ishizuka K., Nakatani T., Fujimoto M. and Miyazaki N. (2010). Noise robust voice activity detection based on periodic to aperiodic component ratio. *Speech Commun.*, 52(1), 41-60.

Isyanto H., Arifin A.S. and Suryanegara M. (2020). Design and implementation of IoT-based smart home voice commands for disabled people using Google Assistant. In: *2020 International Conference on Smart Technology and Applications (ICoSTA)*, IEEE, Surabaya, Indonesia.

Jiang H., Han Z., Scucces P., Robidoux S. and Sun Y. (2000). Voice-activated environmental control system for persons with disabilities. In: *Proceedings of the IEEE 26th Annual Northeast Bioengineering Conference (Cat. No. 00CH37114)*, IEEE, Storrs, CT, USA.

Komatani K., Hotta N., Sato S. and Nakano M. (2015). Posteriori restoration of turn-taking and ASR results for incorrectly segmented utterances. *IEICE Trans. Inf Syst.*, 98(11), 1923-1931.

Mohammed D.Y., Al-Karawi K.A., Husien I. and Ghulam M.A. (2020). Mitigate the reverberant effects on speaker recognition via multi-training. In: *Applied Computing to Support Industry: Innovation and Technology: First International Conference, ACRIT 2019, Ramadi, Iraq, September 15-16, 2019, Revised Selected Papers 1*, Springer.

Mohammed D.Y., Al-Karawi K. and Aljuboori A. (2021). Robust speaker verification by combining MFCC and entrocy in noisy conditions. *Bull. Electr. Eng. Inform.*, 10(4), 2310-2319.

Noyes J.M., Haigh R. and Starr A. (1989). Automatic speech recognition for disabled people. *Appl. Ergon.*, 20(4), 293-298.

Noyes J. and Frankish C. (1992). Speech recognition technology for individuals with disabilities. *Augment. Altern. Commun.*, 8(4), 297-303.

Rosdi F. and Ainon R.N. (2008). Isolated malay speech recognition using Hidden Markov Models. In: *2008 International Conference on Computer and Communication Engineering*, IEEE, Kuala Lumpur, Malaysia.

Ross A., Banerjee S. and Chowdhury A. (2020). Security in smart cities: a brief review of digital forensic schemes for biometric data. *Pattern Recognit. Lett.*, 138, 346-354.

Shrawankar U. and Thakare V.M. (2013). Techniques for feature extraction in speech recognition system: a comparative study. arXiv preprint arXiv:1305.1145.

Vieira A.D., Leite H. and Volochtchuk A.V.L. (2022). The impact of voice assistant home devices on people with disabilities: a longitudinal study. *Technol. Forecast. Soc. Change*, 184, 121961.

Virkar S., Kadam A., Raut N. and Mallick STS. (2020). Proposed model of speech recognition using MFCC and DNN. *Int. J. Eng. Res.*, 9, 5.

Zhang X., Sun J. and Luo Z. (2014). One-against-all weighted dynamic time warping for language-independent and speaker-dependent speech recognition in adverse conditions. *PLoS One*, 9(2), e85458.

Zheng T.F. and Li L. (2017). *Robustness-Related Issues in Speaker Recognition*. Vol. 2, Springer.