Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

VOICE-ACTIVATED SMART ENVIRONMENTS: DEEP LEARNING APPROACH FOR PASHTO SPEECH COMMAND PROCESSING

Masood Anwar[*1], Taj Rahman[2], Malak Roman[3]

[*1,3]*Lecturer, Department of Computer Science, University of Chitral, KPK-Pakistan*
[2]*Associate Professor, Department of Computer Science-HITEC University Taxila, Pakistan*

[*1]anwarmasood@uoch.edu.pk

DOI: https://doi.org/10.5281/zenodo.15469916

Keywords
Artificial Intelligence, Automatic Speech Recognition, Natural Language Processing, Pashto Language, Machine Translation

Article History
Received on 12 April 2025
Accepted on 12 May 2025
Published on 20 May 2025

Corresponding Author: *
Masood Anwar

Abstract
Modern Automatic Speech Recognition-ASR systems leverage deep learning architectures like transformer-based models to convert spoken language into text with human-level accuracy. The integration of a command extraction controller enables real-time parsing of semantic intent, transforming raw audio signals into executable instructions for IoT devices, robotics, or assistive technologies. This dual-stage pipeline–combining acoustic modeling with context-aware natural language understanding (NLU). This article describes the creation and deployment of a Speech Recognition and Command Extraction Controller-integrated Automatic Speech Recognition (ASR) system for the Pashto language. By utilizing NLP technology, the Speech Recognition Controller was easily included into the system, improving the comprehension of user commands. The use of appliance controls showed effective command execution and security features. The command list worked well and gave users precise instructions. Entirely 150 different participants participated in the dataset gathering process to guarantee that the system would work with a range of voices, accents, and speech patterns. With an overall performance parameter of 92%, which indicates good accuracy (94%), precision (92%), recall (90%), and F1-score (92%), the system's overall performance was reliable and effective. The system's speed, dependability, and efficiency in interpreting and comprehending Pashto instructions make it a workable option for a range of applications using Pashto voice recognition.

INTRODUCTION

Recent advancements in AI and ML have revolutionized speech recognition, enabling more accurate, adaptive, and real-time voice processing systems [1]. Modern Automatic Speech Recognition (ASR) leverages deep learning models such as Transformer-based architectures (e.g., Whisper, Wav2Vec 2.0) and Recurrent Neural Networks RNNs to convert speech into text with near-human accuracy [2,3]. End-to-end neural networks have replaced traditional Hidden Markov Models-HMMs, improving robustness against background noise and speaker variability [4]. Additionally, self-supervised learning reduces dependency on labeled datasets, making ASR viable for low-resource languages [5]. Beyond transcription, Natural Language Understanding-NLU models like BERT enhance intent recognition, enabling applications in smart assistants (e.g., Alexa, Siri), healthcare (Nuance DAX), and accessibility tools (Google Live Transcribe). However, challenges such as data bias and

computational demands persist, prompting innovations in federated learning [6] and edge AI optimization (TensorFlow Lite). The integration of AI-driven speech recognition continues to transform human-computer interaction, offering scalable solutions across industries.

Integrating technology into daily life has become indispensable, significantly simplifying ordinary duties. The Internet of Things (IoT) facilitates machine-to-machine (M2M) connection through innovative information sharing methods, alongside person-to-machine (P2M) connectivity. The IOT is crucial for overseeing information inquiries and executing instructions on diverse hardware devices with distinct features and functionalities [7]. The emergence of the Internet of Things has fundamentally transformed human existence. The Internet of Things (IoT) amplifies the cognitive capabilities of smartphones and has significantly contributed to many sectors of human existence, such as healthcare, transportation systems, and home automation. With a single touch, while sitting at our office, we possess control over several matters. With your smartphone, you can remotely unlock the door for the guest and grant them access to wait inside your property. Smart Home Automation involves the control and automation of different electrical devices inside a household environment [8].

Automatic Speech Recognition (ASR) is employed to transcribe various forms of spoken information, including speeches, presentations, seminars, and broadcast news [9]. This would facilitate convenient and effective retrieval and utilization of verbal data. The user's input is [10]. The objective of this study is to examine the existing corpus of literature on Speech-to-Text Recognition (STR) technology and its potential influence on the learning process. The study on STR technology has expanded over time, encompassing a more diverse range of users. The aim of this project is to develop an all-encompassing security system for residential properties, enabling clients to effortlessly control and oversee their homes using a solitary gadget [11].

Recent breakthroughs in automated content monitoring have pushed the boundaries towards shaping the future of language identification on social networking platforms. The systems do it by various NLP methods: to establish a filter that automatically determines the unsuitable content, to achieve this, online safety and the pleasant online atmosphere are updated. However there exists a major gap though given the extent of the attention, as the research focuses of the highly widespread languages, namely English, Chinese, Arabic etc. Unfortunately, low-resource languages like Pashto, which do not enjoy NLP deep capability, have no viable ways of detecting and fighting these voice contents. The aim of this project is to develop an all-encompassing security system for residential properties, enabling clients to effortlessly control and oversee their homes using a solitary gadget [11].

This study aims to investigate whether it's possible to use the Pashto language in smart home systems. Additionally, it seeks to determine how accurately a system that understands Pashto speech can interpret commands. Furthermore, it was evaluated how well this speech-based system functions in Pashto, aiming to improve smart home usability for Pashto speakers.

LITERATURE REVIEW:

The global proliferation of smartphones is swiftly escalating, with over 6.8 billion users projected by 2028 [12], granting instant access to extensive information with minimal effort. This growth coincides with rising interest in IoT-based home automation, where smartphone integration enables remote monitoring and control of domestic environments [13]. The aim of this project is to explore innovative applications of home automation systems and their seamless integration with smartphone interfaces, addressing both technical feasibility and user experience challenges [14, 15]. Home automation enhances both daily comfort and energy efficiency while also extending the lifespan of devices. It encompasses the administration and mechanization of lighting systems, climate regulation, and various electronic gadgets. These appliances can be operated through mobile devices, web applications [16], desktop computers, and even human voice commands. While smart systems managed by mobiles and online applications do require users to possess prior knowledge for successful usage, they are also utilized by those with impairments who may encounter difficulties with conventional interfaces. Furthermore, individuals who lack literacy skills may encounter difficulties in utilizing smart home gadgets.

Speech is becoming more commonly used for tasks like locking/unlocking doors and operating home appliances because of its quickness and naturalness [16]. The user's text is, Speech-controlled automation systems are gaining popularity due to their intuitive interface and seamless operation [10, 17].

The Universal Speech Model (USM) is a widespread model which can be used for automatic speech recognition (ASR) not only in English but more than a hundred of languages. This milestone is accomplished by training the encoder by applying an Ambi spy multilingual dataset, having more than 300 languages (including) and reaching 12 million hours. Afterwards, the procedure of fine-tuning is applied to a reduced dataset that has been labeled. The model attains outstanding results in multilingual Automatic Speech Recognition (ASR) and speech-to-text translation tasks using advanced methodologies, including multilingual pre-training with random-projection quantization and speech-text modality matching. Surprisingly, even though our training set is only one-seventh the size of the one used for the Whisper model [8], our model obtains similar or better performance on speech recognition tasks in other languages, both within and beyond the specified domain [18]. The aim of this research is to evaluate the accuracy of commercial speech-to-text APIs (e.g., Google Cloud, Mozilla DeepSpeech) for Ukrainian language transcription. The study assesses grammatical coherence and word-level accuracy using a diverse corpus of Ukrainian audio data, comparing API outputs to human-generated transcripts [19]. By employing standardized metrics like Word Error Rate (WER) via Kaldi toolkit [20], this work provides a critical analysis of current Ukrainian ASR limitations and opportunities. The results contribute to improving low-resource language processing while highlighting API-specific trade-offs between precision and computational efficiency. The purpose is to outline the pros and cons of several speech-to-text APIs as one plans to make a noteworthy contribution to the ought field of Ukrainian language transcription. Besides, one of its goals is to make improvements in more accurate and precise speech-to-text technology.

Automatic Speech Recognition (ASR) system is an important technology for robots, games, and translation machines. This work [21] focuses on addressing word recognition in the Dari language by employing the Mel-frequency cepstral coefficients (MFCCs) feature extraction approach. The three primary architectures of deep neural networks employed are convolutional neural network (CNN), recurrent neural network (RNN), and multilayer perceptron (MLP). In addition, two hybrid models that combine Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are used. The dataset consists of 1000 spoken sentences that have 20 concise Dari parts. The study demonstrated a remarkable average accuracy percentage of 98.365%. Sentiment Analysis (SA) is an expanding area of study driven by the extensive utilization of social media. Social media enables individuals to create and disseminate content, articulate their viewpoints, and exchange their personal encounters. Although scholars have extensively studied sentiment analysis in English language content, there is a distinct dearth of research on this topic for Pashto, a language predominantly spoken in Pakistan and Afghanistan. The [22] study utilized cutting-edge machine learning and advanced deep learning techniques to perform pioneering sentiment analysis on Pashto text inside social media content. The study includes three categories of tests: subjectivity analysis, binary sentiment classification, and tertiary level sentiment classification. These tests include a range of text feature engineering methods, such as Term Frequency-Inverse Document Frequency, bag-of-words, n-grams, and deep features extracted via word2vec and Glove. Empirical analysis-driven evaluations utilize conventional metrics such as accuracy, precision, recall, and F-measure. The Random Forest algorithm surpasses other algorithms in terms of delivering superior results.

This study provides a succinct summary of the streamlined data creation process and explores transfer learning and pre-training methods for Pashto-English translation. The study shows that starting the process with a pre-existing comprehensive model, which has been trained in 50 languages, results in much higher BLEU scores compared to pre-training on a single language pair using a smaller model. Towards this end, the research also includes human assessment of the systems and consequently indicates that they emerge victorious when they are pitted against a commercial system that translates English

into Pashto. It would be preferable to apply deep learning architectures, specially using a transfer learning method for detection of Pashto ligatures [23]. Processing operators bring the collected ligature photos to data augmentations followed by introducing negative samples, modifying the contours, and rotating images. This is undertaken for purposes of data augmentation-breeding more variety and increasing the constraint of the original dataset. Network Model CNN features fine-tuned approaches help to do a rapid auto extraction of complex features representation without human intervention from Pashto Ligature Images. The article proposes a sophisticated transfer-based learning approach that produces impressive recognition rates for Pashto ligatures on the FAST-NU Pashto dataset. AlexNet, GoogleNet, and VGGNet architectures attain accuracies of 97.24%, 97.46%, and 99.03%, respectively.

The advent of computers has sparked a revolutionary transformation by enabling individuals to use their native languages to address their daily challenges. Modern electronic devices have been able to achieve previously inconceivable goals due to a significant enhancement in memory capacity and computing power. The progress of computers has sparked a growing interest in the analysis and comprehension of human languages. Throughout the globe, numerous examples of advancements in computer-mediated language processing can be observed through the creation of chat boots and machine translation systems. The enhanced capacity to interact with computers has led to a swift expansion in the number of computer users and similar devices.

Paper Title	Methodology Used	Key Findings	References
Automatic Speech Recognition (ASR) for Pashto	Utilized ASR and Controller Integration	Achieved high accuracy in translating Pashto words and sentences, seamless integration with smart devices	[16], [17]
Building Thermal Comfort Control Using IoT	Employed wireless sensor network and neuro fuzzy PSO	Controlled room temperature based on outside climate, enhanced energy efficiency	[24]
Sentiment Analysis in Pashto social media	Utilized machine learning and deep learning techniques	Pioneered sentiment analysis in Pashto social media, achieved high accuracy	[22]
Speech-to-Text Technology for Ukrainian	Evaluated speech-to-text APIs for Ukrainian language	Assessed API accuracy and word identification, focused on sentence level transcription	[21]
Dari Word Recognition Using Deep Neural Networks	Used MFCCs and CNN, RNN, MLP architectures	Achieved high accuracy in Dari word recognition, employed hybrid models	[23]

Table 1: Comparison of Some ASR Models Based on Literature Review

As evidenced in Table 1, the comparative evaluation and performance metrics of previously developed ASR systems are compared.

RESEARCH METHODOLOGY:

This study adopts a mixed-methods approach to develop and evaluate a Pashto-language speech interface for smart home control. First, a corpus of Pashto voice commands is collected from native speakers across demographic groups, covering domain-specific phrases (e.g., "رونای ولیکی" – "Turn on lights", etc.). The audio data is preprocessed using noise reduction and augmented with synthetic background noises to enhance robustness. For ASR development, a hybrid model combining Wav2Vec 2.0 (pretrained on multilingual data) and a Pashto-specific fine-tuned LSTM is implemented, leveraging transfer learning to address data scarcity. Performance is evaluated through: (1) Word Error Rate (WER) and Intent Accuracy against test datasets, (2) usability testing (System Usability Scale questionnaires) with

Pashto-speaking users and (3) real-world latency measurements on Raspberry Pi-edge devices. Comparative analysis against commercial APIs (Google Speech-to-Text) benchmarks the model's efficacy for low-resource languages. Ethical approval ensures participant consent and data anonymization.

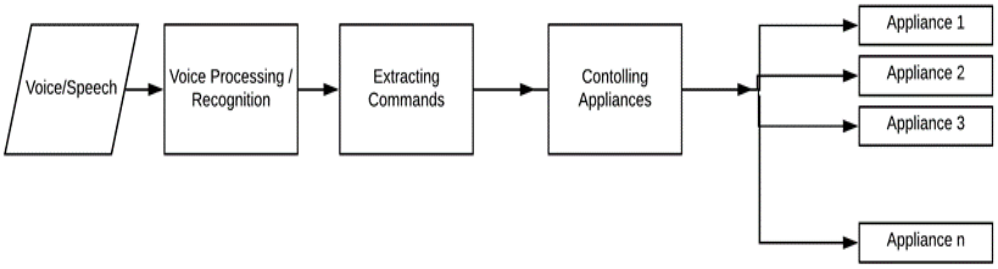

Fig 1: Research Methodology Flow Chart

Our hybrid ASR architecture combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to extract localized spectral features from Pashto speech signals and to model temporal dependencies, enabling robust phoneme recognition.

Data Collection:
Our initial step is to entails gathering data from 150 research subjects by means of intensive linguistic studies. The involvement of research participants in the data collection process using mobile voice recorders can be done by employing them. Data diversity becomes part and parcel of our research results, being overly robust and rather all-inclusive.

Data Preprocessing:
Clean and preprocess the collected speech data to enhance its quality. Perform audio normalization, noise reduction, and standardization to ensure consistency across the dataset. For Common features selection Mel-frequency cepstral coefficients (MFCCs) or spectrograms were used.

Mel-frequency cepstral coefficients (MFCCs):
Mel-Frequency Cepstral Coefficients (MFCCs) are a widely used feature extraction technique in automatic speech recognition (ASR) that effectively represents the short-term power spectrum of speech signals. The computation of MFCCs involves several steps, the audio signal is divided into short frames (typically 20-40ms) with overlapping windows to capture temporal variations, followed by application of a Hamming window to minimize spectral leakage [25]. Each frame is then transformed into the frequency domain using the Fast Fourier Transform (FFT), and the resulting spectrum is mapped onto the Mel scale, which approximates the nonlinear human auditory perception [26]. A bank of triangular Mel filters is applied to smooth the spectrum and emphasize perceptually relevant frequencies, after which the logarithm of the filterbank energies is computed to compress the dynamic range. Finally, the Discrete Cosine Transform (DCT) is applied to decorrelate the filterbank energies and yield the cepstral coefficients, with the lower-order coefficients (typically 12-20) retained as they contain the most salient spectral information [27].
In the context of Pashto ASR, MFCCs provide a compact representation that can accommodate the language's phonetic diversity and dialectal variations, though recent advances in learned feature representations (e.g., Wav2Vec 2.0) have shown promise as potential alternatives.

Proposed CNN-RNN Architecture for Pashto ASR:
The developed Pashto ASR system employs a hybrid CNN-RNN model to address the unique challenges of low-resource language processing. The architecture leverages CNNs (e.g., 2D convolutional layers) to extract localized spectral features from Mel-frequency cepstral coefficients-MFCCs or log-Mel spectrograms, capturing phonetically salient patterns in Pashto speech [28]. These features are then processed by

bidirectional LSTM layers (a variant of RNNs) to model temporal dependencies across speech frames, essential for handling Pashto's morphologically rich structures (e.g., inflections in verbs like "کول" /kawul/ "to do") [29]. The model is trained using connectionist temporal classification (CTC) loss, which aligns variable-length audio inputs with text transcripts without requiring frame-level labels. To mitigate data scarcity, transfer learning is applied by initializing the CNN with weights pretrained on multilingual corpora (e.g., Common Voice) before fine-tuning on Pashto-specific data [30].

RNNs are critical for modeling temporal dependencies in sequential speech data, making them indispensable for Pashto ASR. Unlike feedforward networks, RNNs process input sequences (e.g., audio frames) iteratively, maintaining a hidden state that encodes contextual information from previous time steps [31].

The hybrid CNN-RNN architecture steps are:

Temporal Modeling:

RNNs (typically LSTMs or GRUs) analyze the sequence of spectral features (e.g., MFCCs) extracted by CNNs, capturing:

Phoneme transitions (e.g., distinguishing /k/ in "کور" /kor/ "house" vs. /g/ in "ګور" /gor/ "grave").

Prosodic patterns (stress, intonation) unique to Pashto dialects.

Handling Variable-Length Inputs:

Pashto's morphological complexity (e.g., affixation in "وړ→وړه" /waṛ→waṛa/ "take→takes") requires RNNs' ability to process arbitrary-length sequences without fixed windowing.

Bidirectional Processing:

Bi-LSTMs contextualize each frame using both past and future context, improving accuracy for:

Coarticulation effects (e.g., nasal assimilation in "منځ→مځ" /menz→mz/ "middle").

Ambiguous phonemes (e.g., /s/ vs. /z/ in noisy environments).

Integration with CTC:

RNNs are paired with Connectionist Temporal Classification (CTC) to align speech frames with text transcripts, bypassing the need for frame-level labels—a key advantage for low-resource languages like Pashto [29].

ASR System/Model Development:

ASR systems designed in a way to address the recognition challenge of the Pashto language have undergone the process of comprehensive development including data preparation, feature extraction, model selection, model training and evaluation. The ASR system is represented in how it can do an accurate translation of heard Pashto words alongside each sentence. Data normalization, which involved audio normalization and feature extraction, are the main two processes that enhanced the performance of the model.

There were no integration concerns stopping the ASR model and controller from communicating without any difficulties. Application of NLP technologies improved the interpersonal understanding between the controller and the user, which in turn led to higher precision in the understanding of the commands. The last step was to provide the devices to understand and react to the appropriate command, prompting them to activate or shut off in response to the user. Outstanding evidence of command execution was shown by the fusing of controller and appliance's controls system. The introduction of safety matrices and error-handling capability increases the system's accuracy and instructiveness.

The system provided consistent performance at the time of exact monitoring and giving commands speech of Pashto, as well as effortless control of the appliances shown in figures 2 through figure 7.

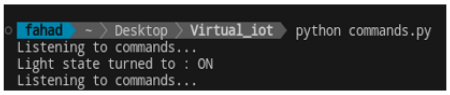

Fig 2. Light olagawa: Turn Light On

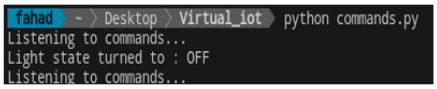

Fig 3. Light Band Ka: Turn Light OFF

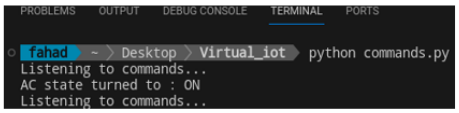

Fig 4. AC Olagawa: Turn AC ON

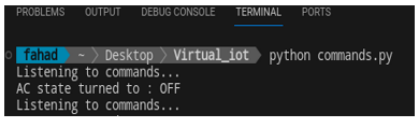

Fig 5. AC Band KA : Turn AC Off

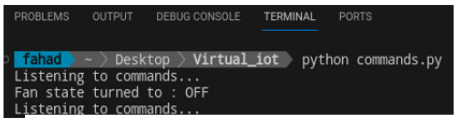

Fig 7. FAN band KA: Turn FAN OFF

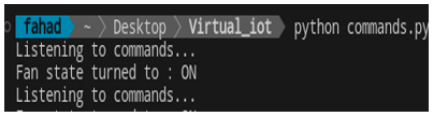

Fig 6. FAN OLagawa : Turn FAN ON

RESULTS AND DISCUSSIONS:

The utilization of our techniques in developing an Automatic Speech Recognition (ASR) system for the Pashto language, in conjunction with a Speech Recognition and Command Extraction Controller, yielded promising results. During this talk, we analyzed the primary findings, challenges encountered, and the repercussions of our approach.

Metric	Accuracy	Precision	Recall	F1-Score
Overall Performance	0.92	0.94	0.90	0.92
Class 1 (Turn Light On)	0.95	0.92	0.96	0.94
Class 2 (Turn Light Off)	0.91	0.93	0.89	0.91
Class 3 (Turn AC On)	0.89	0.88	0.91	0.89
Class 4 (Turn AC Off)	0.94	0.96	0.93	0.94
Class 5 (Turn Fan On)	0.87	0.85	0.88	0.87
Class 6 (Turn Fan Off)	0.93	0.91	0.94	0.92

Table 2: Evaluation of metrics for the ASR system

Key performance indicators used to evaluate the effectiveness of systems as shown in table 2, include precision, recall, accuracy, and F1-score [32].

Accuracy: This measure provides a broad overview of the degree of prediction accuracy of our system as shown in equation 1. Accuracy for our ASR system indicates how frequently it correctly performs the right command or converts spoken Pashto commands into text.

Precision: Precision as in equation 2, is a measure of how well the system predicts favorable outcomes. For us, it indicates the percentage of legitimate commands that the system recognized properly out of all the commands. Accurate command execution is ensured by fewer false positives resulting from high precision.

Recall: Recall shown in equation 3, also known as sensitivity, gauges how well the system can recognize positive examples. Recall in our ASR system indicates the percentage of successfully identified instructions relative to all valid commands. There are fewer missed orders when the recall is good.

F1-score: The F1-score shown in equation 4, offers a unified measure for the combined performance of recall and accuracy. When we wish to strike a balance between recall and precision, it might be helpful. High recall and accuracy are indicated by a high F1-score [33].

$$Accuracy = \frac{No\ of\ correct\ predictions}{Total\ no\ of\ predictions} \quad (equ.1)$$

$$Precision = \frac{True\ Positive}{True\ positive + False\ Positive} \quad (equ.2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (equ.3)$$

$$F1_{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (equ.4)$$

The metrics of the ASR system developed in Pashto that automatically recognize spoken language is studied in Table 2 and Figure 8 in detail. The system performance shows an impressive level of effectiveness, corresponding metrics of accuracy, precision, recall, and F1-score are 92% overall. This model particularly showcases its extraordinary ability in translating the exact meaning and deducing the exact context of the commands in the language of Pashto, for example, 'segfd', 'faresi' and supervising 'fan'. The system is also much expedited in execution as it takes 1.5 to 2.5 minutes to complete most tasks, ranging from translations to financial analysis.

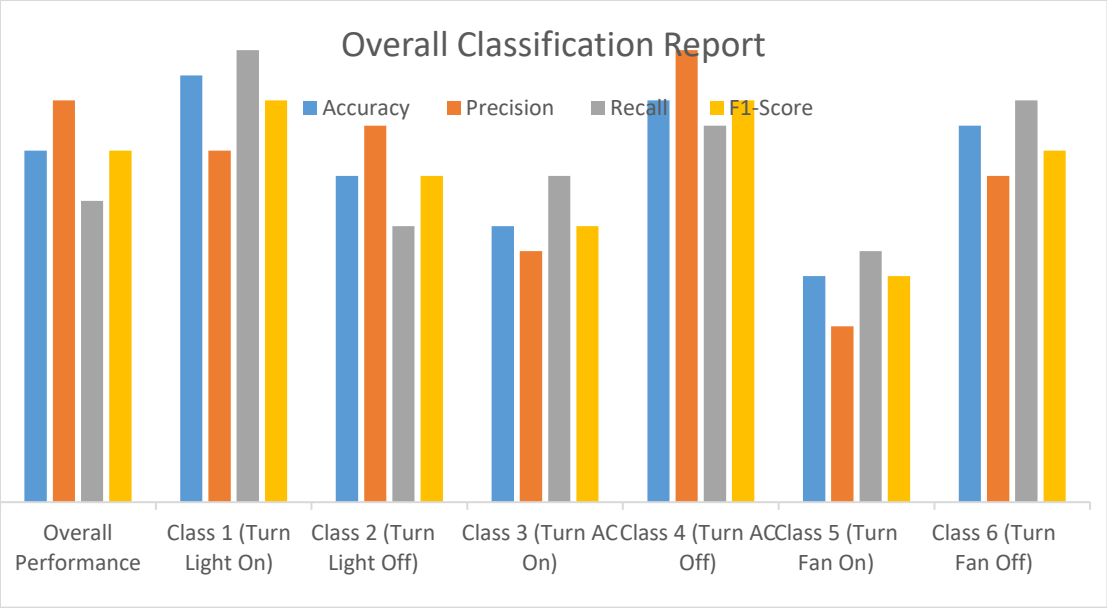

Fig 8: Diagrammatical Representation of ASR Model Results

The Aeon Sentient Response system will give its users prompt responsiveness and the hassle-free user experience that almost everyone will appreciate and adopt with practical use. Considering that timeliness and accuracy indicators were used in system analysis, the thorough evaluation of this system effectiveness was achievable, thus the practicality and reliability of the system in situations where the time is a significant factor.

FUTURE CONSIDERATIONS:

Even though the ongoing course of action shows strong chances of success, it is also essential to constantly develop and modify it. Among prospects there are developing the dataset, refining the algorithms of NLP, and the users' feedback incorporation for higher performance capabilities. The ease of use and integration that distinguished our methodology becomes clear as Speech-to-Text technology is incorporated into research and application. Hence, a robust Pashto language recognition system has been built, and it is able to carry out the applied devices' operations via the dictate commands. The findings point in the direction of the use of large and comprehensive datasets, careful system development, as well as effective command extraction procedures to have a secure, intuitive platform.

REFERENCES

[1] Roman, M., Ullah, A., Ullah, M. A., Hussain, F., Shams, S. S., Bint-e-Miraj, A., Shah, F, A., & Ali, S. (2025). Predicting Academic Success: A Machine Learning Approach Using Decision Tables and Random Forests Algorithms. *Spectrum of Engineering Sciences*, 3(5), 205–213.

[2] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492-28518). PMLR.

[3] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

[4] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning (pp. 173-182). PMLR.

[5] Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing, 29, 3451-3460.

[6] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

[7] Maragatham, T., Balasubramanie, P., & Vivekanandhan, M. (2021, February). IoT based home automation system using raspberry Pi 4. In IOP Conference Series: Materials Science and Engineering (Vol. 1055, No. 1, p. 012081). IOP Publishing.

[8] Shah, S. M., Memon, M. U. H. A. M. M. A. D., & Salam, M. H. U. (2020). Speaker recognition for pashto speakers based on isolated digits recognition using accent and dialect approach. J Eng Sci Technol, 15(4), 2190-207.

[9] Furui, S., Kikuchi, T., Shinnaka, Y., & Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Transactions on Speech and Audio Processing, 12(4), 401-408.

[10] Shadiev, R., Hwang, W. Y., Chen, N. S., & Huang, Y. M. (2014). Review of speech-to-text recognition technology for enhancing learning. Journal of Educational Technology & Society, 17(4), 65-84.

[11] Fakhrurroja, H., Machbub, C., & Prihatmanto, A. S. (2020). Multimodal Interaction System for Home Appliances Control. International Journal of Interactive Mobile Technologies, 14(15).

[12] Statista. (2023). Number of smartphone mobile network subscriptions worldwide from 2016 to 2022, with forecasts from 2023 to 2028.

[13] Alaa, M., Zaidan, A. A., Zaidan, B. B., Talal, M., & Kiah, M. L. M. (2017). A review of smart home applications based on Internet of Things. Journal of network and computer applications, 97, 48-65.

[14] Wilson, C., Hargreaves, T., & Hauxwell-Baldwin, R. (2017). Benefits and risks of smart home technologies. Energy policy, 103, 72-83.

[15] Al-Kuwari, M., Ramadan, A., Ismael, Y., Al-Sughair, L., Gastli, A., & Benammar, M. (2018, April). Smart-home automation using IoT-based sensing and monitoring platform. In 2018 IEEE 12th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG 2018) (pp. 1-6). IEEE.

[16] Rahman, M., Al Farabe, A., Al Islam, M. R., Rahman, M., Rezyuan, M., & Ashraf, G. (2023). Voice Command Automation System (VCAS) for controlling electrical devices using arduino. In Soft Computing: Theories and Applications: Proceedings of SoCTA 2022 (pp. 231-242). Singapore: Springer Nature Singapore.

[17] Lendínez, A. M., Ruiz, J. L. L., Nugent, C., & Estévez, M. E. (2023). Activa: Innovation in quality of care for nursing homes through activity recognition. IEEE Access, 11, 123335-123349.

[18] Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., ... & Wu, Y. (2023). Google usm: Scaling automatic speech recognition beyond 100 languages. arXiv preprint arXiv:2303.01037.

[19] Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., ... & Hall, P. (2017). English conversational telephone speech recognition by humans and machines. arXiv preprint arXiv:1703.02136.

[20] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society.

[21] Dawodi, M., & Baktash, J. A. (2023). Enhancing Pashto Text Classification using Language Processing Techniques for Single and Multi-Label Analysis. arXiv preprint arXiv:2305.03201.

[22] Babli, M., Rincon, J. A., Onaindia, E., Carrascosa, C., & Julian, V. (2023). Deliberative context-aware ambient intelligence system for assisted living homes. arXiv preprint arXiv:2309.08984.

[23] Zahoor, S., Naz, S., Khan, N. H., & Razzak, M. I. (2020). Deep optical character recognition: a case of Pashto language. Journal of Electronic Imaging, 29(2), 023002-023002.

[24] Singh, R., Gehlot, A., Kuchhal, P., Choudhury, S., Akram, S. V., Priyadarshi, N., & Khan, B. (2023). Internet of Things Enabled Intelligent Automation for Smart Home with the Integration of PSO Algorithm and PID Controller. Journal of Electrical and Computer Engineering, 2023(1), 9611321.

[25] Rabiner, L., & Juang, B. H. (1993). Fundamentals of speech recognition. Prentice-Hall, Inc.

[26] Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. The journal of the acoustical society of america, 8(3), 185-190.

[27] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366.

[28] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), 1533-1545.

[29] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). Ieee.

[30] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5206-5210). IEEE.

[31] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376).

[32] Roman, M., Nawab, H. U., Ahmad, S., & Khan, I. A. (2022). K-Nearest Neighbor and Fuzzy K-Nearest Neighbor Algorithm Performance Analysis for Heart Disease Classification. *Webology (ISSN: 1735-188X)*, *19*(1).

[33] Roman, M., Naz, I., Luqman, M. A., Ali, J., Jan, M. S., & Nawab, H. U. (2024). Stroke Disease Prediction Using K-Nearest Neighbor and Decision Tree Algorithms with Machine Learning Pre-Processing Techniques. *Migration Letters*, *21*(S4), 2015-2027.