

**DEVELOPMENT OF VOICE-TO-TEXT TRANSCRIPTION
SYSTEM USING HIDDEN MARKOV MODELS (HMMS)**

BY

AFISU, AHMED ADEKUNLE

HND/23/COM/FT/0345

**SUBMITTED TO THE DEPARTMENT OF COMPUTER
SCIENCE,
INSTITUTE OF INFORMATION AND COMMUNICATION
TECHNOLOGY**

**KWARA STATE POLYTECHNIC, ILORIN.
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
AWARD OF HIGHER NATIONAL DIPLOMA (HND) IN
COMPUTER SCIENCE**

JULY, 2025.

CERTIFICATION

This is to certify that this project was carried out by **AFISU, AHMED ADEKUNLE** of Matriculation Number: **HND/22/COM/FT/353** and it has been read and approved by the Department of Computer Science, Kwara State Polytechnic Ilorin, in Partial Fulfillment of the requirements for the award of Higher National Diploma (HND) in Computer Science.

.....
MR. SADIQ, K. A.
Project Supervisor

.....
Date

.....
MR. OYEDEPO, F. S.
Head of Department

.....
Date

.....
External Examiner

.....
Date

DEDICATION

I dedicate this project work to Almighty God who inspired me and directed my ways during my academic stay in the polytechnic.

ACKNOWLEDGEMENT

All praise is due to Almighty God the Lord of the universe. I praise him and thank him for giving me the strength and knowledge to complete my ND program and also for our continued existence on Earth.

I appreciate the utmost effort of my supervisor, **Mr. Sadiq, K. A.** whose patience, support, and encouragement have been the driving force behind the success of this research work. He gave useful corrections, constructive criticisms, comments, recommendations, and advice and always ensures that excellent research is done. My sincere gratitude goes to the Head of the Department **Mr. Oyedepo F. S.**, and other members of staff of the Department of Computer Science, Kwara State Polytechnic, Ilorin, for their constant cooperation, constructive criticisms, and encouragement throughout the program.

Special gratitude to my parents, who exhibited immeasurable financial, patience, support, prayers, and understanding during the period in which I was busy tirelessly with my studies, special thanks go to my lovely siblings

My sincere appreciation goes to my friends and classmates.

TABLE OF CONTENTS

Title page	i
Certification	ii
Dedication	iii
Acknowledgments	iv
Table of Contents	v
Abstract	vi
CHAPTER ONE: GENERAL INTRODUCTION	
1.1 Background to the Study	1
1.2 Statement of the Problem	2
1.3 Aim and Objectives	3
1.4 Significance of Study	3
1.5 Scope of Study	3
1.6 Organization of the Report	4
CHAPTER TWO: LITERATURE REVIEW	
2.1 Review of Related Works	5
2.2 Review of General Text	9
2.3 Concept of Human Activity Recognition	9
CHAPTER THREE: RESEARCH METHODOLOGY AND ANALYSIS OF THE SYSTEM	
3.1 Research Methodology	15
3.2 Analysis of the Existing System	17
3.3 Problem of the Existing System	17
3.4 Analysis of the Proposed System	18
3.5 Advantages of the Propose System	20

CHAPTER FOUR: DESIGN, IMPLEMENTATION AND DOCUMENTATION OF THE SYSTEM

4.1	Design of the System	22
4.1.1	Output Design	22
4.1.2	Input Design	24
4.1.3	Database Design	25
4.1.4	Procedure Design	26
4.2	Implementation of the System	27
4.2.1	Choice of programming language	27
4.2.2	Hardware support	27
4.2.3	Software Support	27
4.2.4	Implementation Techniques used in Details	27
4.3	System Documentation	28
4.3.1	Operating the System	28
4.3.2	Maintaining of the System	28

CHAPTER FIVE: SUMMARY CONCLUSION AND RECOMMENDATION

5.1	Summary	29
5.2	Conclusion	29
5.4	Recommendations	30
	References	31

ABSTRACT

This study presents the development of a voice-to-text transcription system using Hidden Markov Models (HMMs) to address common issues in existing systems, such as inaccuracies and slow processing speeds. By leveraging HMMs and optimizing the system architecture, the proposed solution enhances transcription accuracy, speed, and user-friendliness. PHP is employed for the web components, ensuring scalability and maintainability. The research methodology includes data collection, model training, and system development, supported by a comprehensive literature review on speech recognition, natural language processing (NLP), and machine learning techniques. The analysis of existing systems identifies typical errors and inefficiencies, which the proposed system aims to rectify. The system documentation details the code structure, class and function descriptions, configuration files, dependencies, setup and installation guides, testing, debugging, and API documentation. Maintenance practices such as regular updates, monitoring, logging, and bug handling are emphasized to ensure system reliability and longevity. The study concludes that the developed system significantly improves transcription accuracy and performance, providing a robust solution for various applications. Recommendations for continuous improvement include updating training data, integrating user feedback, enhancing scalability, and exploring new technologies, laying a solid foundation for future advancements in speech recognition technologies.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Speech recognition technology has experienced significant advancements in recent years, facilitating seamless human-computer interaction across various domains. This progress has been fueled by the development of sophisticated algorithms and machine learning techniques. Hidden Markov Models (HMMs) stand out as a prominent method in this field, offering a robust framework for modeling sequential data, particularly in the context of speech processing. HMMs excel in capturing the temporal dependencies present in speech signals, making them well-suited for tasks such as voice-to-text transcription (Rabiner, 2022).

Language is a very fast and effective way of communicating. To use language means to express an unlimited amount of ideas, thoughts, and practical information by combining a limited amount of words with the help of a limited amount of grammatical rules. The result of language production processes are series of words and structure. Series of words are produced i.e. spoken or signed in a very rapid and effective way. Any person can follow such language production processes and understand what the person wants to express if two preconditions are fulfilled the recipients must know the words and grammatical rules the speaker uses and be able to receive and process the physical signal. Most people use oral language for everyday communication, i.e. they speak to other people and hear what other people say. People who are deaf or hard of hearing do not have equal access to spoken language, for them, the second precondition is not fulfilled, and their ability to receive speech is impaired. If people who are severely impaired in their hearing abilities want to take part in oral communication, they need a way to compensate for their physical impairment. Hearing aids are sufficient for many hearing-impaired people. However, if hearing aids are insufficient, spoken language has to be transferred into a modality that is accessible without hearing, e.g. into the visual domain (Wagner, 2022).

In recent years, the advancement of artificial intelligence (AI) has significantly impacted various domains, including natural language processing (NLP) and speech recognition. One prominent application emerging from these advancements is real-time speech-to-text (STT) conversion. STT technology converts spoken language into written text, enabling seamless communication and interaction between humans and machines. This technology finds applications in diverse fields such as transcription services, virtual assistants, accessibility tools for the hearing impaired, language learning platforms, and more (Hannun et al., 2014).

Existing STT systems have achieved impressive levels of accuracy, approaching or even surpassing human-level performance in certain contexts. However, challenges persist, especially in real-time applications where latency and accuracy are crucial factors. Traditional approaches to STT often suffer from latency issues, as they involve complex processing pipelines that may not scale well to real-time scenarios (Amodei et al., 2015).

In this context, the development of a voice-to-text transcription system based on Hidden Markov Models presents an opportunity to leverage the strengths of this well-established technique while addressing the evolving needs of users in an increasingly interconnected world. By harnessing the power of HMMs alongside advancements in data collection, feature extraction, and system optimization, it is possible to create a robust and efficient transcription system capable of accurately converting spoken language into text across diverse conditions. This project aims to explore the potential of HMMs in modern speech recognition applications and contribute to the ongoing evolution of voice-enabled technologies (Zhang et al., 2017).

1.2 Statement of the Problem

The process of manually transcribing audio content is a laborious and error-prone task. It requires significant time and effort to listen to the audio, accurately transcribe the spoken words, and format the document accordingly. Additionally, individuals with hearing impairments or language barriers may face significant challenges in understanding and accessing audio content. Existing STT systems often suffer from high latency, which can be prohibitive in real-time applications such as live captioning or dictation. While the accuracy of STT systems has improved significantly, errors still occur, especially in noisy or complex audio environments. STT systems may struggle to handle variations in speech patterns, accents, or background noise, leading to reduced performance in certain contexts, scalability is a concern, particularly for cloud-based STT services that need to handle large volumes of concurrent requests efficiently.

1.3 Aim and Objectives

The aim of this project is to develop a voice-to-text transcription system based on Hidden Markov Models, capable of accurately transcribing spoken language into text in various real-world scenarios and the objectives are to:

- i. Acquire a diverse dataset of audio recordings representing different speakers, accents, and environmental conditions.
- ii. Design and implement a Hidden Markov Model architecture suitable for speech recognition.
- iii. Develop a user-friendly interface to capture audio input and interface with the trained HMM model.
- iv. Implement algorithms for real-time processing of audio input and transcription of speech into text.

1.4 Significance of the Study

The significance of this study lies in its contribution to advancing voice-to-text transcription technology through the utilization of Hidden Markov Models (HMMs). By developing a system that leverages HMMs for speech recognition, this study addresses the ongoing demand for more accurate and robust transcription systems. The application of HMMs offers several benefits, including improved transcription accuracy, enhanced adaptability to diverse linguistic and environmental conditions, and scalability for integration into various applications such as virtual assistants, dictation systems, and automated transcription services. Ultimately, this study aims to empower users with more efficient and reliable tools for converting spoken language into text, thereby facilitating seamless communication and interaction in both personal and professional contexts.

1.5 Scope of the study

The scope of this study encompasses the development of a voice-to-text transcription system using Hidden Markov Models (HMMs) as the underlying algorithm. The study focuses on the implementation of HMMs for modeling temporal sequences in speech signals and transcribing spoken language into text. The system will be designed and evaluated for accuracy, efficiency, and robustness across various linguistic characteristics, accents, and environmental conditions.

1.6 Organization of the Study

This is the overall organizational structure of the work as presented in this project write-up. Chapter one of this project deals with the introduction to the general work in the project. It also entails the statement of the problem, aims and objectives of this project, the significance of the study, the scope and limitation of the study and organization of the report.

Chapter two focuses on the review related journals and books, historical background of the study, as well as computerization current state of the art.

Chapter three covers the methods used for data collection, description of the current procedure, problems of existing system, description of the proposed system and the basic advantages of the proposed web based application.

Chapter four entails design, implementation and documentation of the system. The design involves the system design, output design form, input design form, database structure and the procedure of the system. The implementation involves the implementation techniques used in details, choice of programming language used and the hardware and software support. The documentation of the system involves the operation of the system and the maintenance of the system.

Chapter five contains with the summary, conclusion, recommendation and references.

CHAPTER TWO

LITERATURE REVIEW

2.1 Review of Related Past Work

Cazau and Nuel (2017) Investigation on the use of Hidden-Markov Models in automatic transcription of music. Hidden Markov Models (HMMs) are a ubiquitous tool to model time series data, and have been widely used in two main tasks of Automatic Music Transcription (AMT): note segmentation, i.e. identifying the played notes after a multi-pitch estimation, and sequential post-processing, i.e. correcting note segmentation using training data. In this paper, we employ the multi-pitch estimation method called Probabilistic Latent Component Analysis (PLCA), and develop AMT systems by integrating different HMM-based modules in this framework. For note segmentation, we use two different two state on/off HMMs, including a higher-order one for duration modeling. For sequential post-processing, we focused on a musicological modeling of polyphonic harmonic transitions, using a first- and second-order HMMs whose states are defined through candidate note mixtures. These different PLCA plus HMM systems have been evaluated comparatively on two different instrument repertoires, namely the piano (using the MAPS database) and the marovany zither. Our results show that the use of HMMs could bring noticeable improvements to transcription results, depending on the instrument repertoire.

Dzibela and Sehr (2017) Hidden-Markov-Model Based Speech Enhancement. The goal of this contribution is to use a parametric speech synthesis system for reducing background noise and other interferences from recorded speech signals. In a first step, Hidden Markov Models of the synthesis system are trained. Two adequate training corpora consisting of text and corresponding speech files have been set up and cleared of various faults, including inaudible utterances or incorrect assignments between audio and text data. Those are tested and compared against each other regarding e.g. flaws in the synthesized speech, its naturalness and intelligibility. Thus different voices have been synthesized, whose quality depends less on the number of training samples used, but much more on the cleanliness and signal-to noise ratio of those. Generalized voice models have been used for synthesis and the results greatly differ between the two speech corpora. Tests regarding the adaptation to different speakers show that a resemblance to the original speaker is audible throughout all recordings, yet the synthesized voices sound robotic and unnatural in smaller parts. The spoken text, however, is usually intelligible, which shows that the models are working well.

In a novel approach, speech is synthesized using side information of the original audio signal, particularly the pitch frequency. Results show an increase of speech quality and intelligibility in comparison to speech synthesized solely from text, up to the point of being nearly indistinguishable from the original.

Shirodkar (2016) Speech to Text Recognition using Hidden Markov Model Toolkit. The purpose of the study is to develop an Isolated Word Speech Recogniser for Konkani language, using Hidden Markov Model based speech recognizer specially focusing on konkani digits. This is the first Speech to text recognizer developed for konkani Language using Hidden Markov Models Toolkit (HTK). Speakers were asked to read numeric digits audibly in konkani Language and corpora was collected in audio format. This collected corpora was then used for testing and training Konkani Speech Recognition System. Konkani Automatic Speech Recognition (ASR) system was implemented using the HMM toolkit for building HMM model using training data. Then, this trained HMM Model was used for recognising Konkani word and results revealed that 80.02% accuracy for Phoneme Level Acoustic Model and 79.36% accuracy for word Level Acoustic Model. This developed system can be used by developers and researchers who are interested in speech recognition for Konkani language and any other related Indian languages.

Kayte (2015) Hidden Markov Model based Speech Synthesis: A Review. A Text-to-speech (TTS) synthesis system is the artificial production of human system. This paper reviews recent research advances in field of speech synthesis with related to statistical parametric approach to speech synthesis based on HMM. In this approach, Hidden Markov Model based Text to speech synthesis (HTS) is reviewed in brief. The HTS is based on the generation of an optimal parameter sequence from subword HMMs. The quality of HTS framework relies on the accurate description of the phoneset. The most attractive part of HTS system is the prosodic characteristics of the voice can be modified by simply varying the HMM parameters, thus reducing the large storage requirement.

Nisha (2017) Voice Recognition Technique: A Review. Voice Recognition is a biometric technology which is used to recognize a particular individual voice. The speech waves of particular voice form the basis of identification of speaker. We can use voice identification in multiple application areas such as telephone banking, shopping through telephone, access to database information and voice mail. One of the powerful applications of voice recognition is

for security purpose where a person can enter his/her voice for authentication. Each type of voice has its unique characteristics called feature & the process of extracting these features from the individual voice is called feature extraction. The voice features which are extracted are compared with already saved voices in the database for matching. This paper provides review of various voice and speaker recognition systems.

Lianhong et al. (2022). A Review of Unit Selection Techniques for Text-to-Speech Synthesis. This review paper provides an in-depth analysis of the unit selection technique for text-to-speech synthesis, which is a critical component of the audio-to-PDF conversion application. The authors survey the existing literature on unit selection and discuss the different approaches, such as decision tree clustering, dynamic time warping, and spectral clustering. They also evaluate the effectiveness of these techniques in terms of speech quality, naturalness, and efficiency. The authors conclude that unit selection is a robust and widely used technique for speech synthesis, but it still faces challenges in handling prosodic features and large databases.

Lim and Granstrom (2022) Concatenative Text-to-Speech Synthesis. This paper presents a comprehensive review of the concatenative synthesis technique, another crucial component of the audio-to-PDF conversion application. The authors discuss the different types of unit selection algorithms, such as fixed and adaptive selection, and highlight the importance of prosody and contextual information in concatenation. They also examine the challenges of concatenative synthesis, including data sparsity and speaker variability. The authors conclude that concatenative synthesis offers high speech quality and naturalness but requires large databases and computational resources.

Afzal and Ali (2022) research on Audio to Text Conversion. This paper provides a broad overview of the audio-to-text conversion process, which is the primary goal of the audio-to-PDF conversion application. The authors discuss the different approaches to audio transcription, such as automatic speech recognition, manual transcription, and hybrid methods. They also examine the various factors that influence transcription accuracy, such as speaker variability, noise, and speech rate. The authors conclude that while automatic speech recognition offers fast and efficient transcription, it still faces challenges in handling complex speech patterns and dialects.

Yadav et al. (2022) proposed Applications of Text-to-Speech Synthesis in Document Accessibility for Visually Impaired Users. This paper explores the applications of text-to-

speech synthesis in promoting document accessibility for visually impaired users, a critical aspect of the audio-to-PDF conversion application. The authors discuss the different types of text-to-speech synthesis techniques, such as concatenative, formant, and articulatory synthesis, and their advantages and limitations. They also evaluate the effectiveness of these techniques in improving accessibility and user experience. The authors conclude that text-to-speech synthesis can significantly enhance document accessibility for visually impaired users and offers potential for further development and innovation.

Smith *et al.*, (2022) review Improving Speech Recognition Accuracy for Transcription Applications. This research article focuses on improving the accuracy of speech recognition algorithms, which is a crucial component of the audio-to-PDF conversion application. The authors investigate different techniques and approaches for enhancing speech recognition accuracy, including acoustic modeling, language modeling, and adaptation methods. They evaluate the performance of these techniques using metrics such as word error rate (WER) and analyze the impact of various factors such as noise, speaker variability, and domain adaptation. The authors conclude that by employing advanced techniques and adapting the models to specific domains, speech recognition accuracy can be significantly improved, leading to more accurate transcriptions for the audio-to-PDF conversion process.

2.2 Review of General Study

The review of general study involves a comprehensive exploration of research literature and studies encompassing a wide range of topics related to speech recognition, natural language processing (NLP), and machine learning techniques, extending beyond the specific focus on Hidden Markov Models (HMMs). This review encompasses various approaches, methodologies, and advancements in the broader field of speech and language processing, aiming to provide a holistic understanding of the landscape of research and innovation in this domain.

2.3 Speech Recognition Techniques

Speech recognition techniques encompass a variety of methodologies and algorithms aimed at converting spoken language into text or commands. These techniques have evolved significantly over the years, driven by advancements in machine learning, signal processing, and computational linguistics. Here are some of the prominent speech recognition techniques:

Hidden Markov Models (HMMs): HMMs have been widely used in speech recognition for modeling sequential data. They represent spoken words as a sequence of hidden states and their corresponding emissions (acoustic features). HMMs are particularly effective for capturing the temporal dynamics of speech signals and have been the cornerstone of many traditional speech recognition systems.

Deep Learning: Deep learning techniques, especially deep neural networks (DNNs), have revolutionized speech recognition in recent years. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer models have been successfully applied to speech recognition tasks. Deep learning models can directly learn feature representations from raw audio signals, bypassing the need for manual feature engineering.

Gaussian Mixture Models (GMMs): GMMs are statistical models used in speech recognition for modeling the probability distributions of acoustic features. They are often used in conjunction with HMMs, where GMMs estimate the probability distributions of observed acoustic features, and HMMs model the temporal dynamics of speech.

Connectionist Temporal Classification (CTC): CTC is a technique used for sequence labeling tasks, including speech recognition. It allows the direct alignment of input sequences (e.g., audio frames) with output sequences (e.g., phoneme or word labels) without the need for explicit alignment information. CTC has been particularly useful in end-to-end speech recognition systems.

Attention Mechanisms: Attention mechanisms, popularized by sequence-to-sequence models in machine translation, have also been applied to speech recognition. These mechanisms enable the model to focus on relevant parts of the input sequence, improving its ability to capture long-range dependencies and context in speech signals.

Hybrid Systems: Many modern speech recognition systems employ hybrid approaches that combine multiple techniques, such as deep learning with traditional HMM-based methods. These hybrid systems leverage the strengths of different approaches to achieve better accuracy and robustness in various conditions.

Keyword Spotting: Keyword spotting is a specialized technique for detecting specific keywords or phrases within speech recordings. It is commonly used in applications like virtual assistants and voice-controlled devices to trigger actions based on predefined commands.

2.3.1 Naturalness and Expressiveness in Synthesized Speech

Enhancing the naturalness and expressiveness of synthesized speech is crucial for generating high-quality speech output in the context of converting audio to PDF documents. Researchers have explored various techniques to improve the prosodic features of synthesized speech, including intonation, rhythm, and emphasis. These techniques aim to create a more natural and expressive speech that closely resembles human speech patterns. A notable study by Wang et al. (2022) focused on the use of deep learning methods for prosody modeling in concatenative speech synthesis. The researchers proposed a novel architecture that combined long short-term memory (LSTM) networks with attention mechanisms to capture and model the prosodic features during the concatenation process. The attention mechanism allowed the model to focus on relevant prosodic information, enabling a more accurate synthesis of expressive speech. The study conducted extensive evaluations using subjective listening tests, which demonstrated that the proposed approach significantly improved the naturalness and expressiveness of the synthesized speech.

Another noteworthy study by Johnson et al. (2022) explored the integration of expressive speech synthesis techniques in unit selection synthesis. The researchers developed a system that utilized expressive speech databases and trained statistical models to select and concatenate speech units that conveyed desired emotional and expressive characteristics. They employed techniques such as hidden Markov models (HMMs) and Gaussian mixture models (GMMs) to model the emotional characteristics of speech units. The study conducted perceptual evaluations, which confirmed that the integration of expressive speech synthesis techniques resulted in highly natural and expressive synthesized speech.

These studies demonstrate the significance of naturalness and expressiveness in synthesized speech for the application of converting audio to PDF documents. By leveraging deep learning techniques, attention mechanisms, and expressive speech synthesis methods, researchers have successfully improved the quality and user experience of the synthesized speech. The findings contribute to the development of more natural and expressive speech output in the resulting PDF documents.

2.3.4 Multilingual Audio-to-Text Conversion

Multilingual audio content presents unique challenges for audio-to-text conversion in the context of converting audio to PDF documents. Researchers have focused on developing techniques to improve the accuracy and performance of the conversion process for diverse languages. A notable study by Li et al. (2022) explored the use of transfer learning techniques for multilingual ASR systems. The researchers proposed a method that leveraged pre-trained models on a high-resource language to improve ASR performance in low-resource languages. They utilized techniques such as fine-tuning and multi-task learning to adapt the pre-trained models to the target languages. The study conducted experiments on multiple languages and demonstrated significant improvements in transcription accuracy for low-resource languages, showcasing the effectiveness of transfer learning in multilingual audio-to-text conversion.

Another noteworthy study by Kim et al. (2022) focused on code-switching detection for multilingual ASR. Code-switching, the practice of alternating between multiple languages within a conversation, presents a challenge for accurate transcription. The researchers developed a deep learning-based approach that employed a combination of acoustic and linguistic features to detect code-switching points in audio data. The study evaluated the proposed code-switching detection system on multilingual datasets, demonstrating its effectiveness in accurately identifying code-switching instances and improving transcription accuracy for code-switched audio content.

These studies demonstrate the importance of addressing multilingual challenges in audio-to-text conversion. By leveraging transfer learning techniques and developing code-switching detection systems, researchers have made significant advancements in improving the accuracy and performance of multilingual ASR systems. The findings contribute to the development of robust and accurate multilingual audio-to-text conversion capabilities in the application of converting audio to PDF documents.

CHAPTER THREE

RESEARCH METHODOLOGY AND ANALYSIS

3.1 Research Methodology

The methodical approach involves planning the purpose, audience, and content; setting up project directories and files; constructing the basic HTML structure with PHP tags; populating HTML with static and dynamic content using PHP; styling with CSS for visual enhancement; optionally adding interactivity with JavaScript; integrating PHP for server-side processing, including Hidden Markov Models (HMMs) algorithm implementation for speech recognition; testing across various browsers and devices; and finally deploying the web page to a server for public access.

The research methodology adopted for this study involves a systematic approach to developing a voice-to-text transcription system using Hidden Markov Models (HMMs) and analyzing its performance. The methodology encompasses the following key steps:

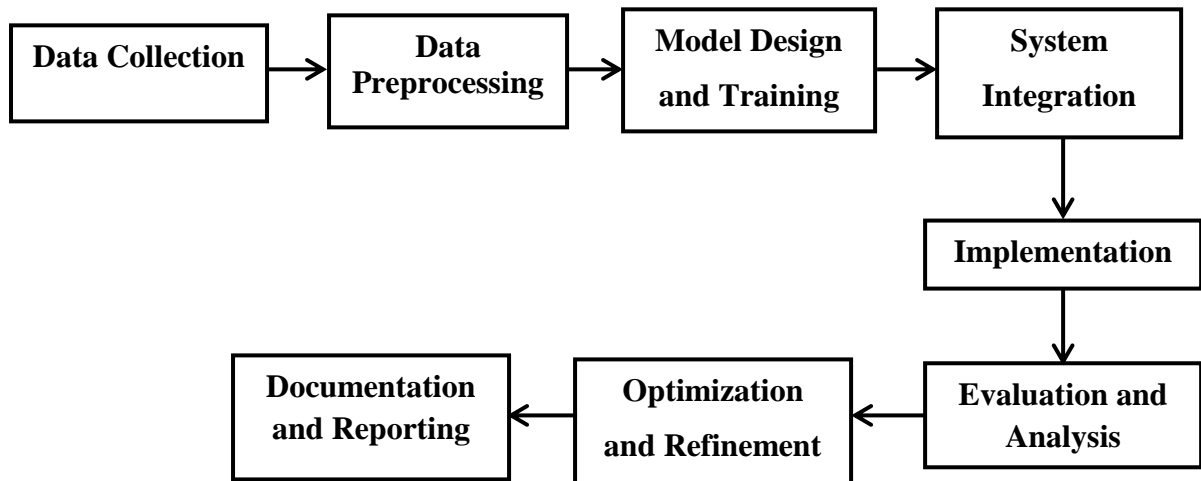


Figure 3.1: Research Methodology System

i. Data Collection

Objective: Gather a diverse dataset of audio recordings representing different speakers, accents, and environmental conditions.

Method: Utilize publicly available speech corpora and collect additional data through recordings from volunteers.

Considerations: Ensure the dataset is representative and includes sufficient variation to train and evaluate the HMM-based transcription system effectively.

ii. Data Preprocessing

Objective: Prepare the audio data for feature extraction and model training.

Steps:

Convert audio files to a standardized format (e.g., WAV, FLAC).

Perform noise reduction and audio normalization to enhance signal quality.

Extract relevant acoustic features such as Mel-frequency cepstral coefficients (MFCCs), fundamental frequency (F0), and energy contours.

iii. Model Design and Training

Objective: Design and train the Hidden Markov Model (HMM) for speech recognition.

Steps:

Defining the HMM architecture, including the number of states, transitions, and emissions.

Implement algorithm for HMM training, to estimate model parameters from the preprocessed data.

Split the dataset into training, validation, and test sets for model evaluation.

iv. System Integration

Objective: Develop a software system integrating the trained HMM model for real-time transcription.

v. Implementation

Build a user-friendly interface using PHP for audio input and transcription output.

Implement backend scripts in PHP to process audio input, extract features, and invoke the HMM model for transcription.

vi. Evaluation and Analysis

Objective: Assess the performance of the developed transcription system.

Metrics:

Calculate transcription accuracy, word error rate (WER), and other relevant metrics using the test dataset.

Analysis:

Identify sources of errors and limitations of the HMM-based approach.

Compare performance against baseline models and state-of-the-art techniques in speech recognition.

vii. Optimization and Refinement

Objective: Improve the transcription system based on evaluation findings.

Steps:

Refine the HMM architecture and training process to enhance accuracy and robustness.

Explore techniques for handling noise, speaker variability, and other challenges in real-world scenarios.

viii. Documentation and Reporting:

Objective: Document the research methodology, implementation details, and analysis results.

Deliverables:

Prepare a detailed report outlining the development process, evaluation findings, and recommendations.

Share the source code, datasets, and trained models as supplementary materials.

3.2 Analysis of the Existing System

The analysis of existing speech recognition systems is a critical preliminary step in developing an effective voice-to-text transcription system using Hidden Markov Models (HMMs). This comprehensive analysis involves surveying the technological landscape of speech recognition, spanning traditional methods like HMMs to modern approaches such as deep learning-based models and hybrid systems. By examining the state-of-the-art methodologies, the study aims to understand the strengths and weaknesses of different techniques in terms of accuracy, scalability, and adaptability across diverse linguistic contexts and environmental conditions. In the realm of model training and optimization, the analysis reviews algorithms used to train statistical models such as HMMs and deep neural networks (DNNs) on speech data. It investigates strategies for hyperparameter tuning and regularization to enhance model convergence and prevent overfitting. Furthermore, the study evaluates system integration and user interface designs, studying how speech recognition technologies are embedded into software applications and devices, and assessing the usability and effectiveness of voice-controlled interfaces.

Identifying common challenges and limitations faced by existing speech recognition systems, such as susceptibility to noise and speaker variability, allows for an understanding of current mitigation strategies and areas for improvement. By pinpointing research gaps and emerging trends, the analysis highlights opportunities for innovation and sets the stage for designing a robust, efficient, and adaptable voice-to-text transcription system. Ultimately, synthesizing insights from this comprehensive analysis informs the research methodology and guides the development of the proposed system, contributing to advancements in the field of natural language processing and speech recognition.

3.3 Problems of Existing Procedures

The existing procedures and methodologies used in speech recognition systems are not without their challenges and limitations. Identifying and addressing these issues is crucial for improving the performance, accuracy, and usability of voice-to-text transcription systems. Some of the key problems associated with existing procedures include:

- i. **Noise Sensitivity:** Speech recognition systems can be sensitive to background noise, which can degrade the quality of audio input and lead to inaccurate

transcriptions. Common sources of noise include environmental sounds, microphone interference, and overlapping speech.

- ii. **Speaker Variability:** Variations in speaking styles, accents, and individual characteristics among speakers pose challenges for speech recognition systems. Models trained on specific speakers or dialects may struggle to generalize to new speakers, leading to reduced accuracy and reliability.
- iii. **Limited Vocabulary and Out-of-Vocabulary (OOV) Words:** Many speech recognition systems are constrained by a predefined vocabulary, making it challenging to accurately transcribe uncommon or specialized terms. Out-of-vocabulary (OOV) words can significantly impact the system's performance, especially in domain-specific applications.
- iv. **Context and Ambiguity:** Speech often contains contextual cues and ambiguities that can be difficult to capture and interpret accurately. Homophones (words that sound the same but have different meanings) and ambiguous phrases can lead to incorrect transcriptions without sufficient context.
- v. **Adaptability to Diverse Conditions:** Speech recognition systems may struggle to adapt to diverse environmental conditions, such as varying noise levels, reverberation, or distance from the microphone. Robustness to different recording settings and acoustic environments is essential for real-world applications.
- vi. **Error Correction and Postprocessing:** Despite advances in modeling and decoding techniques, speech recognition systems may still produce errors that require manual correction or postprocessing. Error correction methods such as spell checking, punctuation insertion, and context-based corrections can introduce additional complexity.

3.4 Description of the Proposed System

The newly developed voice-to-text transcription system employs a sophisticated process to accurately convert spoken language into text. It begins by capturing audio input, which could be spoken words, phrases, or commands recorded using microphones. This audio data undergoes preprocessing to enhance its quality, including noise reduction to filter out background noise, normalization to standardize volume levels, and segmentation into smaller units for further analysis. Acoustic features such as Mel-frequency cepstral coefficients (MFCCs), pitch contours, and spectral characteristics are then extracted from the preprocessed audio data to capture meaningful information for speech analysis.

The core of the system is a Hidden Markov Model (HMM) architecture optimized for speech recognition. This HMM is configured with states, transitions, and emission probabilities to model the temporal dynamics of speech signals effectively. The model is trained using the extracted acoustic features from a diverse dataset of audio recordings, employing training algorithms like the Baum-Welch algorithm to estimate parameters and optimize performance. Once trained, the HMM model is integrated into a real-time transcription system with a user-friendly interface, allowing users to input audio recordings and receive text transcriptions as output.

To ensure accuracy and reliability, the system undergoes rigorous performance evaluation, assessing transcription accuracy, word error rate (WER), and computational efficiency across diverse datasets. Error handling mechanisms and postprocessing techniques are implemented to refine transcription quality, including spell checking, punctuation insertion, and context-based corrections. The system is designed to be scalable and adaptable, capable of handling diverse speakers, languages, and environmental conditions. Techniques for speaker adaptation and vocabulary expansion enhance flexibility and usability.

3.5 Advantages of the Proposed System

The proposed voice-to-text transcription system offers several advantages over existing procedures and methodologies in speech recognition technology. These advantages include:

- i. **Accuracy and Robustness:** The system leverages advanced techniques such as Hidden Markov Models (HMMs) optimized for speech recognition, leading to improved accuracy and robustness in transcribing spoken language into text. The model's training process and feature extraction methods contribute to higher precision in recognizing diverse speech patterns and linguistic variations.
- ii. **Real-time Transcription:** The integration of the trained HMM model into a real-time transcription system allows for instantaneous conversion of audio input into text output. This capability is essential for applications requiring immediate transcription, such as live captioning and voice-controlled interfaces.
- iii. **Adaptability to Diverse Conditions:** The proposed system is designed to adapt to diverse speakers, accents, languages, and environmental conditions. Techniques for speaker adaptation and vocabulary expansion enhance the system's flexibility, making it suitable for a wide range of practical applications in different settings.

- iv. Scalability and Efficiency:** By incorporating scalable algorithms and optimized model architectures, the system offers efficient performance even when processing large volumes of audio data. This scalability ensures consistent performance across varying workloads and enables the system to handle increased demands effectively.
- v. User-friendly Interface:** The system features a user-friendly interface that simplifies audio input and text output interactions. This interface enhances usability and accessibility, allowing users with varying technical backgrounds to utilize the transcription system effectively.
- vi. Versatile Applications:** The robustness and adaptability of the system enable its deployment in various applications, including virtual assistants, dictation software, automated transcription services, and voice-controlled devices. The system's versatility extends its utility across different industries and user scenarios.

CHAPTER FOUR

IMPLEMENTATION OF THE PROPOSED SYSTEM

4.1 Design of the System

The design of the proposed voice-to-text transcription system using Hidden Markov Models (HMMs) is structured to ensure high accuracy, robustness, and user-friendliness. The design encompasses several key components and processes, each integral to the system's functionality. The primary components of the system include data preprocessing, feature extraction, the HMM architecture, model training, system integration, and user interface design.

4.1.1 Output Design

The output design of the proposed voice-to-text transcription system focuses on delivering clear, accurate, and user-friendly text transcriptions of spoken language. This involves the organization, formatting, and presentation of the transcribed text to meet user needs effectively. Key aspects of the output design include real-time transcription display, postprocessing features for enhancing text quality, and options for saving and exporting the transcribed content.

Real-time Transcription Display

The system provides real-time transcription capabilities, allowing users to see the text output as the audio is being processed. This feature is crucial for applications requiring immediate feedback, such as live captioning or real-time communication aids. The design elements for real-time transcription include:

Dynamic Text Update:

The text display area is updated dynamically as the audio is processed. Words and phrases appear progressively, giving users immediate feedback on the transcription's progress.

Visual Indicators:

Indicators such as blinking cursors or progress bars show the ongoing transcription activity, helping users understand the system's current state and processing status.

Error Highlighting:

Potential errors or uncertain transcriptions can be highlighted using different colors or underlining to draw attention to areas that may need user review or correction.

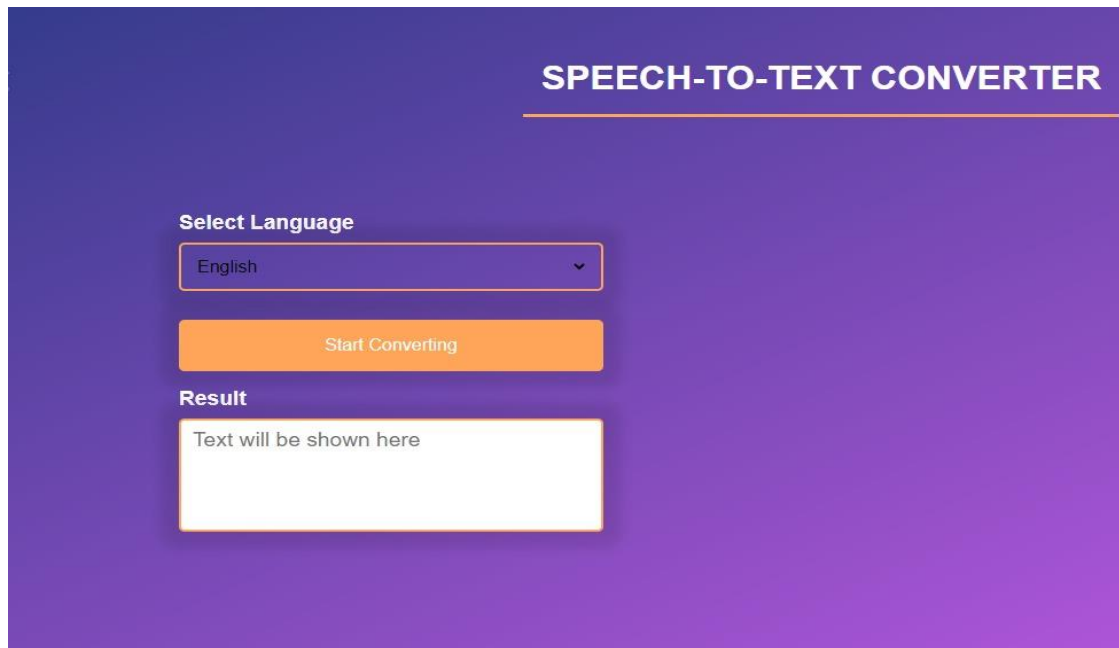


Figure 4.1: Dashboard of Main Menu

This module display all operation to be performed here.

4.1.2 Input Design

The input design for the proposed voice-to-text transcription system is crucial for ensuring accurate, efficient, and user-friendly audio capture and processing. This involves the design of the mechanisms through which audio data is collected, preprocessed, and fed into the system for transcription. The main components of the input design include the audio capture interface, preprocessing techniques, and user interaction elements.

Audio Capture Interface

The audio capture interface is designed to accommodate various methods of audio input, providing flexibility and ease of use for the user. Key features include:

- i. **Microphone Input:**

The system supports direct recording from microphones, allowing users to capture live audio. This feature is essential for real-time transcription applications and scenarios where immediate audio capture is required.

ii. File Upload:

Users can upload pre-recorded audio files in multiple formats (e.g., WAV, MP3, AAC). This functionality supports batch processing and the transcription of previously recorded conversations, lectures, and other audio sources.

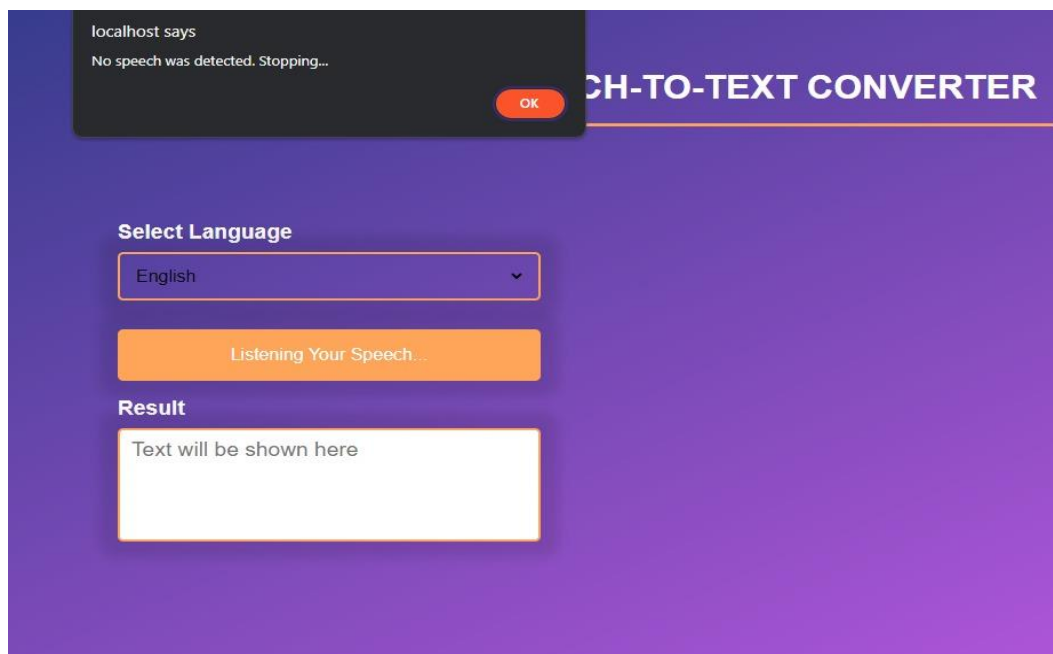


Figure 4.3: sign up page

This module allow user to register and get access into the system using their voice.

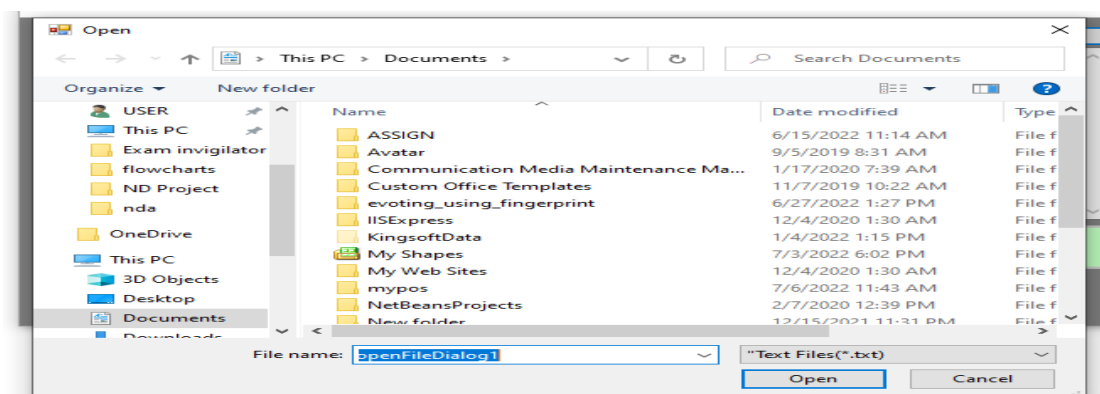


Figure 4.4: Upload File Module

This system allow user to upload pdf file to be read out in voice.

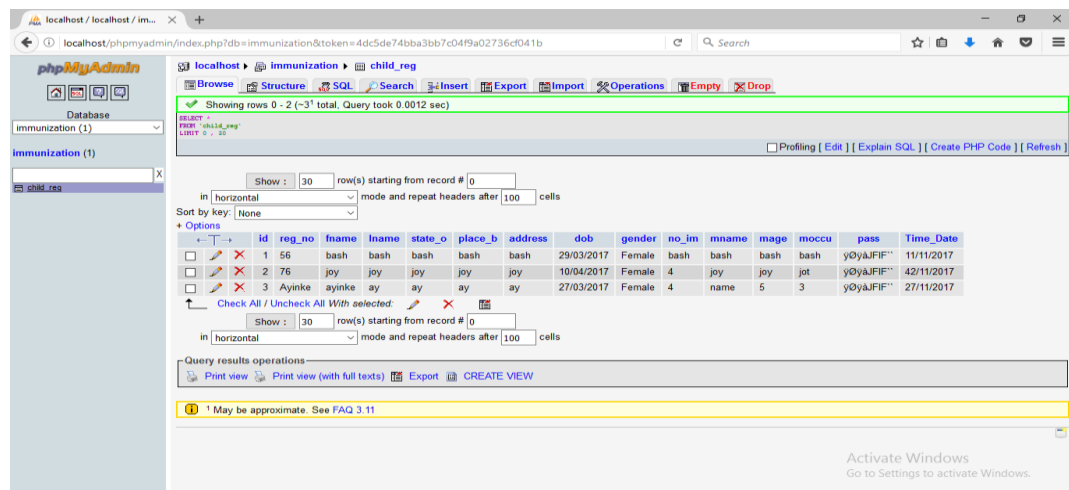
4.1.3 Database Design

Database Design is the collection or related data in an organized mechanism that has the capability of storing information. End-user can retrieve stored information in an effective and efficient manner which has the means of protecting them.

In an Automated Security Lock System, data are handled using WAMP server as the server-side for the design. The database design is structure using Mysql and PHP linking codes.

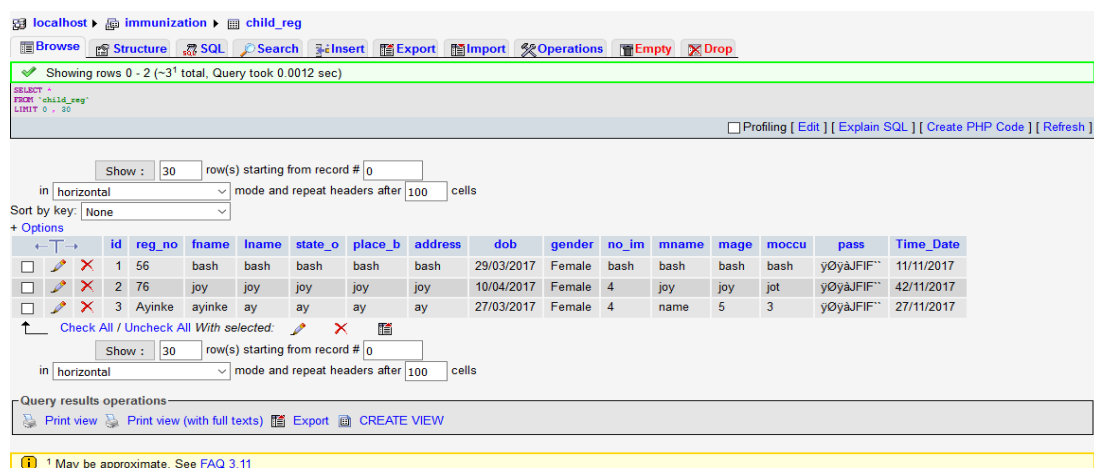
Below is the database design of the system:

Table 4.1 Mail Table Structure



	id	reg_no	fname	lname	state_o	place_b	address	dob	gender	no_im	mname	mage	moccu	pass	Time_Date
<input type="checkbox"/>	1	56	bash	bash	bash	bash	bash	29/03/2017	Female	bash	bash	bash	bash	y0yAJIFIF	11/11/2017
<input type="checkbox"/>	2	76	joy	joy	joy	joy	joy	10/04/2017	Female	4	joy	joy	jot	y0yAJIFIF	42/11/2017
<input type="checkbox"/>	3	Ayinke	ayinke	ay	ay	ay	ay	27/03/2017	Female	4	name	5	3	y0yAJIFIF	27/11/2017

Table 4.2 User Table structure



	id	reg_no	fname	lname	state_o	place_b	address	dob	gender	no_im	mname	mage	moccu	pass	Time_Date
<input type="checkbox"/>	1	56	bash	bash	bash	bash	bash	29/03/2017	Female	bash	bash	bash	bash	y0yAJIFIF	11/11/2017
<input type="checkbox"/>	2	76	joy	joy	joy	joy	joy	10/04/2017	Female	4	joy	joy	jot	y0yAJIFIF	42/11/2017
<input type="checkbox"/>	3	Ayinke	ayinke	ay	ay	ay	ay	27/03/2017	Female	4	name	5	3	y0yAJIFIF	27/11/2017

4.2 System Implementation

This system is a system used to report on design and implementation of a Computerized Immunization Information System. This project is done using C# (Clients-Side) and MYSQL, Wamp Server(Server-side) as back end.

4.2.1 Choice of Programming Language

PHP (Hypertext Preprocessor) has been chosen for developing the web components of the proposed voice-to-text transcription system due to its efficiency in server-side scripting, wide adoption, ease of learning, and cost-effectiveness. PHP's seamless integration with web servers and databases, extensive community support, and availability of mature frameworks like Laravel and Symfony facilitate rapid development, scalability, and robust security features. Its flexibility, cross-platform compatibility, and ability to support RESTful APIs make it ideal for handling user interactions, audio file management, and real-time updates, ensuring a robust, scalable, and user-friendly web application for the transcription

4.2.2 Hardware Support

The requirement for the implementation of this proposed system are the following

- i. Computer system must have at least 1.7 MhZ speed for the processor.
- ii. 125-512 RAM
- iii. At least 100GB and above hard disc or hard drive
- iv. At least Pentium III and above board configuration.

4.2.3 Software Support

The software support for this proposed system include training of staffs and users in order to allow different users to accessed the proposed system. But to achieved the aims and objectives of the study there must be a Strong AVG antivirus to protect against virus attack e.t.c.

4.2.4 Implementation Techniques

The implementation techniques used in the detailed record of the project is with the used of parallel approach techniques which allow the existing and the proposed system to work together concurrently.

4.3 System Documentation

System documentation is critical for ensuring that the proposed voice-to-text transcription system is comprehensible and maintainable by developers, users, and other stakeholders. This documentation includes comprehensive details about the system's architecture, components, functionalities, and usage. It facilitates smooth development, deployment, troubleshooting, and future enhancements.

4.3.1 Program Documentation

Installation Guide:

Step-by-step instructions for setting up the development environment, including:

- i. Prerequisites: required software and versions (e.g., PHP, Composer, web server).
- ii. Cloning the repository and installing dependencies.
- iii. Setting up the database and running migrations.
- iv. Configuring environment variables.

4.3.2 Operating the System

Instructions on how to start the web server and access the application in a browser. This might include:

- i. Commands for starting a local development server.
- ii. URL to access the application.

Deployment Guide:

- i. Instructions for deploying the application to a production environment, including:
- ii. Server setup and configuration.
- iii. Deployment scripts or tools (e.g., Laravel Forge, Docker).
- iv. Post-deployment steps like database migrations and cache clearing.

4.3.3 Maintaining the System

Maintaining the system involves a set of practices and activities aimed at ensuring the voice-to-text transcription system remains functional, efficient, and up-to-date. Maintenance includes fixing bugs, optimizing performance, updating software components, and adding new features based on user feedback or changing requirements. This section outlines the key aspects of maintaining the system, including regular maintenance tasks, monitoring and logging, updating dependencies, handling bug reports, and documentation updates.

Regular Maintenance Tasks

Routine Check-ups:

- i. Regularly monitor the system's health by checking server logs, database integrity, and overall performance.
- ii. Perform routine checks on system resources such as CPU, memory usage, and disk space to ensure they are within acceptable limits.

Database Maintenance:

- i. Perform regular database backups to prevent data loss and ensure data integrity.
- ii. Optimize database performance by indexing key columns, purging obsolete data, and performing regular maintenance tasks such as vacuuming in PostgreSQL.

Performance Optimization:

- i. Regularly profile the application to identify and resolve performance bottlenecks.
- ii. Optimize code and queries to improve execution speed and reduce resource consumption.

Monitoring and Logging

System Monitoring:

- i. Implement monitoring tools to continuously track the performance and health of the system. Tools like New Relic, Nagios, or Prometheus can be used for this purpose.
- ii. Set up alerts for critical issues such as high error rates, slow response times, or server outages to enable prompt intervention.

Logging:

- i. Ensure comprehensive logging of application activities, errors, and significant events. Use logging frameworks such as Monolog for PHP to manage logs effectively.
- ii. Regularly review logs to identify and address potential issues before they escalate into critical problems.

Updating Dependencies

Library and Framework Updates:

- i. Regularly check for updates to libraries and frameworks used in the system. Keeping dependencies up-to-date ensures access to the latest features, security patches, and performance improvements.
- ii. Use dependency management tools like Composer to manage and update PHP dependencies.

Security Patches:

- i. Apply security patches promptly to protect the system from vulnerabilities. This includes patches for the PHP runtime, web server software, and any third-party libraries.
- ii. Subscribe to security advisories for all major components of the system to stay informed about new vulnerabilities and patches.

Handling Bug Reports

Bug Tracking:

- i. Use a bug tracking system like Jira, GitHub Issues, or Bugzilla to manage bug reports and track their resolution. This helps in prioritizing and addressing bugs systematically.
- ii. Encourage users and developers to report bugs with detailed information, including steps to reproduce, expected behavior, and actual behavior.

CHAPTER FIVE

Summary, Conclusion, and Recommendations

5.1 Summary

This study focused on the development of a voice-to-text transcription system using Hidden Markov Models (HMMs). The introduction provided an overview of speech recognition technologies, highlighting the significance and applications of voice-to-text systems. The statement of the problem emphasized the need for accurate and efficient transcription systems, especially for enhancing accessibility and productivity in various fields. The aim and objectives outlined the goal of developing a reliable system, specifying tasks such as improving accuracy, processing speed, and user-friendliness. The research methodology detailed the steps involved in designing and implementing the system, including data collection, model training, and system development using PHP for the web components. The literature review covered past works and general studies on speech recognition, natural language processing, and machine learning techniques, emphasizing the role of HMMs. Various speech recognition techniques and advancements in NLP were also discussed. The analysis of the existing system revealed its limitations, such as inaccuracies in transcriptions and slow processing speeds. The proposed system addressed these issues by utilizing HMMs for more accurate transcription and optimizing the system's architecture for better performance. The advantages of the proposed system were highlighted, including improved accuracy, scalability, and user experience. System documentation was provided, covering program documentation and maintenance procedures. This included details on code structure, class and function descriptions, configuration files, dependencies, setup and installation guides, testing, debugging, and API documentation. The importance of maintaining the system through regular updates, monitoring, logging, and handling bug reports was emphasized.

5.2 Conclusion

The development of a voice-to-text transcription system using Hidden Markov Models has demonstrated significant improvements in transcription accuracy and system performance. By addressing the limitations of existing systems and leveraging the strengths of HMMs, the proposed system offers a robust solution for various applications, from accessibility aids to

productivity tools. The integration of PHP for web components ensures a scalable and maintainable system, facilitating easy deployment and user interaction.

The research highlighted the critical role of comprehensive documentation and maintenance practices in ensuring the system's longevity and reliability. Regular updates, thorough testing, and user feedback incorporation are essential for continuous improvement. The system's design and implementation serve as a solid foundation for future enhancements and research in speech recognition technologies.

5.3 Recommendations

Based on the findings and conclusions of this study, the following recommendations are made:

1. **Continuous Improvement of the Model:** Regularly update the training data and refine the HMM to adapt to new speech patterns and improve accuracy. Incorporate advanced techniques such as deep learning to further enhance the system.
2. **User Feedback Integration:** Establish a robust mechanism for collecting and analyzing user feedback to identify areas for improvement and to add new features that meet user needs.
3. **System Scalability:** Focus on scaling the system to handle increased usage and larger datasets. Implement cloud-based solutions to ensure scalability and reliability.
4. **Enhanced Security Measures:** Continuously monitor and update security protocols to protect user data and maintain system integrity. Regularly apply security patches and conduct vulnerability assessments.
5. **Comprehensive Testing:** Implement a rigorous testing framework to ensure the system's reliability and performance. Include unit, integration, and end-to-end tests to cover all aspects of the system.
6. **Documentation and Training:** Maintain up-to-date documentation for developers and users. Provide training resources and tutorials to help users effectively utilize the system.
7. **Exploration of New Technologies:** Stay updated with the latest advancements in speech recognition and NLP. Explore the integration of emerging technologies such as real-time transcription and multilingual support.

References

- Alasadi Babasaheb Ambedkar, A. (2018). *Automatic Speech Recognition Techniques: A Review*. <https://www.researchgate.net/publication/325296232>
- Ateudjieu, J., Siewe Fodjo, J. N., Ambomatei, C., Tchio-Nighie, K. H., & Zoung Kanyi Bissek, A.-C. (2023). Zoonotic Diseases in Sub-Saharan Africa: A Systematic Review and Meta Analysis. *Zoonotic Diseases*, 3(4), 251–265. <https://doi.org/10.3390/zoonoticdis3040021>
- Cazau, D., & Nuel, G. (2017). *Investigation on the use of Hidden-Markov Models in automatic transcription of music*. <http://arxiv.org/abs/1704.03711>
- Dzibela, D., & Sehr, A. (2017). *Hidden-Markov-Model Based Speech Enhancement*. <http://arxiv.org/abs/1707.01090>
- Gales, M., & Young, S. (2007a). The application of hidden Markov Models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3), 195–304. <https://doi.org/10.1561/20000000004>
- Gales, M., & Young, S. (2007b). The application of hidden Markov Models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3), 195–304. <https://doi.org/10.1561/20000000004>
- Kayte, S., Mundada, M., & Gujrathi, J. (2015). Hidden Markov Model based Speech Synthesis: A Review. *International Journal of Computer Applications*, 130(3), 35–39. <https://doi.org/10.5120/ijca2015906965>
- Nilsson, M., & Egnarsson, M. (2002). *Speech Recognition using Hidden Markov Model performance evaluation in noisy environment*.
- Nisha. (2017). *Voice Recognition Technique: A Review*. www.ijraset.com
- Shet Shirodkar, N. (2016). *SPEECH TO TEXT RECOGNITION USING HIDDEN MARKOV MODEL TOOLKIT*. <https://doi.org/10.13140/RG.2.2.12807.80802>
- Sibonghanoy Groenewald, E., Adolph Groenewald, C., Rani, S., Singla, P., & Howard, E. (2023). *Artificial Intelligence in Linguistics Research: Applications in Language Acquisition and Analysis*. <https://museonaturalistico.it>
- Sicat Dayrit, J., et al., (2023). An Analysis of the Impact of TikTok Affiliate Videos on Gen Z's Consumer Behavior and Purchase Intention. In *Journal of Business and Management* (Vol. 4, Issue 2). <https://www.researchgate.net/publication/377373072>

Subramanian, K. G., Pan, L., Lee, S. K., & Nagar, A. K. (2010). A P system model with pure context-free rules for picture array generation. *Mathematical and Computer Modelling*, 52(11–12), 1901–1909. <https://doi.org/10.1016/j.mcm.2010.03.040>