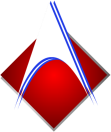
**ACLC COLLEGE OF BUTUAN**

**Franchised and Operated by Butuan Information Technology Services, Inc. (BITSI)**

HDS Bldg., 999 J.C. Aquino Avenue, Butuan City, Agusan del Norte, Philippines 8600

**ARAVoE: Arabic Recognition Assisted Vocabulary of English: A Voice-Driven Cross-Language Dictionary**

A Thesis Project

Presented to the Faculty of

**COMPUTER EDUCATION DEPARTMENT**

In Partial Fulfillment of

the Requirements for the Degree in

**BACHELOR OF SCIENCE IN COMPUTER SCIENCE**

Submitted by:

BAUTISTA, GHAIZAR A.

LUCAS, ALOU JESUSA SHEN F.

ACOG, ARNOLD E.

CABUAL, CHONA

**CHAPTER 1**

# **INTRODUCTION**

In the global landscape of language processing, the demand for accurate and efficient translation mechanisms transcends borders, impacting both international and local communities. The intersection of Arabic and English languages presents a unique challenge, necessitating comprehensive research efforts at various levels to address the complexities inherent in Arabic-to-English translation.

Internationally, the recognition of Arabic as the fourth most widely utilized language on the internet underscores the global significance of developing robust Arabic Speech Recognition (ASR) systems for translation purposes. Research initiatives led by international scholars, such as those by Jian-Yun Nie (2022), highlight the importance of cross-language communication tools, emphasizing the need to bridge linguistic barriers to facilitate effective communication and knowledge exchange.

At the national and local levels, researchers contribute to the discourse through focused investigations tailored to specific linguistic contexts. Studies conducted by national scholars, such as Elhadry, T. (2023), delve into the intricacies of Arabic language structures and its divergence from English, shedding light on the linguistic challenges encountered in translation processes. These localized insights are invaluable for developing tailored solutions that resonate with the linguistic nuances of Arabic-speaking communities.

The contributions of other researchers in the field serve as building blocks for addressing the multifaceted challenges in ASR and translation. Surveys and studies conducted by scholars like Farghaly and Shaalan (2009), Habash (2010), Shoufan and Alameri (2015), Al-Khatib et al. (2020), Hamdan et al. (2021), and Ibrahim and Al-Mohimeed (2022) provide foundational knowledge on Arabic language processing, categorizing challenges, and proposing solutions. However, these efforts often overlook recent developments and neglect certain varieties of Arabic, necessitating a more inclusive approach to research.

Key challenges encountered in ASR and translation include the linguistic differences between Arabic and English, including variations in syntax, morphology, and semantics. Machine translation systems, while offering expedited translation processes, often suffer from accuracy issues, leading to mistranslations and misinterpretations. Addressing these challenges requires a multifaceted approach, encompassing advancements in machine learning, linguistic analysis, and cross-language communication techniques.

Proposed solutions involve refining ASR algorithms to better capture the nuances of Arabic speech, enhancing translation models with context-aware features, and integrating human-in-the-loop mechanisms to ensure translation accuracy and cultural sensitivity. Additionally, efforts to improve machine translation technologies through continuous refinement and optimization are essential for overcoming existing limitations and advancing the field of Arabic-to-English translation. Through collaborative research endeavors at both international and local levels, the goal of achieving seamless and accurate cross-language communication can be realized, fostering greater understanding and connectivity in an increasingly interconnected world.

## **1.2 Statement of the Problem**

The discrepancy between existing English and Arabic translation dictionaries poses a notable challenge in cross-language communication (Smith, 2019). Despite ongoing efforts to refine translation systems through various machine learning techniques (Jones et al., 2020), automated translation systems still exhibit room for improvement. This study aims to address this gap by leveraging established algorithms and models to enhance translation performance. Specifically, it focuses on utilizing the Kaldi ASR algorithm and model for speech recognition and exploring its integration with the MarianMT translation framework to improve translation accuracy and efficiency. Additionally, it seeks to explore effective approaches for leveraging available English and Arabic dictionaries to facilitate English-Arabic translation (Brown & Johnson, 2018). Furthermore, the study will harness machine learning techniques to enhance automated translation systems and assess the utility of cross-language dictionaries in facilitating Arabic-English translation processes (Garcia et al., 2021).

## **1.3 Objectives of the Study**

The main objective of this study is to develop an Arabic-to-English Automated Speech Recognition (ASR) system with high accuracy and efficiency. Specifically, the study aims to achieve the following objectives:

Specifically, the study aims to:

1. Develop a comprehensive understanding of the challenges and nuances involved in Arabic-to-English speech recognition and translation.
2. Implement the Kaldi ASR toolkit for accurate transcription of Arabic speech into text.
3. Integrate the MarianMT machine translation framework to enable the translation of transcribed Arabic text into English.
4. Evaluate the performance of the integrated ASR and machine translation system in terms of transcription accuracy and translation quality.
5. Assess the practical applicability of the system in real-world scenarios, such as cross-cultural communication and content localization.

## **1.4 Significance of the Study**

This study is deemed beneficial for the following:

### **1. Advancing Research**

By exploring the integration of Kaldi ASR and MarianMT for Arabic-to-English translation, this study contributes to the advancement of knowledge in the fields of automatic speech recognition and machine translation.

### **2. Bridging Language Barriers**

Addressing the challenges of Arabic-to-English translation can facilitate cross-cultural communication, knowledge exchange, and access to information for speakers of both languages.

### **3. Practical Applications**

The developed system has practical implications in various domains, including language learning, content localization, multilingual communication platforms, and automated transcription services.

### **4. Educational Value**

The study provides insights into the technical aspects of ASR and machine translation systems, serving as an educational resource for students and researchers interested in natural language processing and computational linguistics.

### **5. Societal Impact**

Improved Arabic-to-English translation capabilities can have significant societal impact by promoting inclusivity, diversity, and accessibility in global communication channels.

## **1.5 Scope and Limitation**

### **Scope**

This study focuses on developing an Arabic language speech recognition and translation service to English. The primary components of this study include:

**1. Speech Recognition**

a) Utilization of the Automatic Speech Recognition algorithm specifically for Arabic speech.

b) Evaluation of accuracy and speed compared to existing services, with further selection criteria kept within scope.

c) Implementation of tone recognition based on tonal variations present in Arabic words during speech.

d) Integration of tone recognition limited to Arabic words and their dialects.

**2. Translation Service**

a) Implementation of Marian Neural Machine Translation (NMT) for translating recognized Arabic speech to English.

b) Assessment of accuracy and speed compared to other services, with additional criteria kept within scope.

### **Limitation**

The delimitations of the study are the following:

1. The system will only accept Arabic speech input and provide English translation output.
2. The study will not analyze the semantic or syntactic aspects of recognized Arabic speech content.
3. Complexity in tonal transcription intricacies in written Arabic will not be addressed.
4. The tonal formation of Arabic speech relies on the availability and quality of the Arabic tone dataset.
5. Limited to recognizing clear, non-nasal speech; unclear or nasal speech may not be accurately recognized.

## **1.6 Definition of Terms**

1. **Cross-language:** Relating to or involving more than one language, especially in terms of translation or communication between different languages.

2. **Dictionary**: A reference book or electronic resource containing an alphabetical list of words with information about their meanings, pronunciations, and usage.

3**. ASR (Automatic Speech Recognition):** The process of automatically recognizing and transcribing spoken language into text using computer algorithms.

4. **HMM (Hidden Markov Model):** A statistical model used to describe the probability distribution of a sequence of observable events or states.

5. **GMM (Gaussian Mixture Model):** A probabilistic model that represents the probability distribution of a mixture of several Gaussian distributions.

6. **LSTM (Long Short-Term Memory):** A type of recurrent neural network architecture capable of learning and remembering long-term dependencies in sequential data.

7. **GRU (Gated Recurrent Unit):** A type of recurrent neural network architecture similar to LSTM but with a simpler structure and fewer parameters.

8. **MFCC (Mel Frequency Cepstral Coefficients):** A feature extraction technique widely used in speech processing to represent the short-term power spectrum of a sound.

9. **Kaldi:** An open-source toolkit for speech recognition research, containing various tools and libraries for building speech recognition systems.

10. **Sphinx:** A set of speech recognition systems developed by Carnegie Mellon University, including tools and libraries for building speech recognition applications.

11. **API (Application Programming Interface):** A set of rules and protocols that allows different software applications to communicate with each other.

12**. GT (Google Translate):** A free online translation service provided by Google that translates text and web pages between different languages.

13. **NLP (Natural Language Processing):** A branch of artificial intelligence that focuses on the interaction between computers and humans using natural language.

14. **BLA (Basic Language Analyses):** Fundamental analyses or studies conducted on a language, covering aspects such as grammar, syntax, and vocabulary.

15. **BR (Building Resources):** The process of creating or compiling linguistic resources, such as dictionaries, corpora, and grammars, for a specific language.

16**. LI (Language Identification):** The process of determining the language in which a given text or speech is written or spoken.

17. **SemA (Semantic-Level Analysis):** Analysis conducted at the semantic level, focusing on the meaning and interpretation of words, phrases, or sentences.

18**. WER (Word Error Rate):** A metric used to evaluate the accuracy of automatic speech recognition systems by measuring the rate of incorrectly recognized words.

19. **RTF (Real-Time Factor):** A metric used to measure the speed or efficiency of automatic speech recognition systems, indicating the time taken to process a given amount of speech data in real-time.

20. **Neural Machine Translation (NMT)**: A machine translation approach that uses neural networks to learn the mapping from input to output sequences, achieving state-of-the-art performance in translation tasks.

21. **Arabizi**: A form of writing Arabic using Latin script, often used in informal digital communication, which poses challenges for Arabic language processing due to its non-standard spelling and lack of standardized rules.

22**. Language Model**: A statistical model that predicts the probability of a sequence of words occurring in a language, used to improve the accuracy of speech recognition and machine translation systems.

23. **Lexicon**: A database containing information about the pronunciation, spelling, and meaning of words, used in speech recognition and machine translation for mapping between words and their linguistic representations.

24. **Corpus**: A large collection of text or speech data used for linguistic analysis and training language processing models, including speech recognition and machine translation systems.

25**. Preprocessing**: The initial step in data preparation, involving cleaning, formatting, and transforming raw input data into a suitable format for further analysis or processing by machine learning algorithms.

26. **Post-processing**: The final step in data processing or analysis, involving refining, filtering, or enhancing the output of a system to improve its quality or usability.

27. **Deep Learning**: A subset of machine learning techniques inspired by the structure and function of the human brain's neural networks, capable of learning complex patterns and representations from data.

28. **Parallel Corpus**: A collection of texts in two or more languages that are translations of each other, used in machine translation to train and evaluate translation models.

29. **Evaluation Metrics**: Quantitative measures used to assess the performance of speech recognition and machine translation systems, such as accuracy, precision, recall, and F1 score.

30. **End-to-End Model**: A machine learning model that directly maps input to output without the need for intermediate representations or processing steps, often used in speech recognition and machine translation tasks.

**CHAPTER 2**

# **REVIEW OF RELATED LITERATURE**

As languages exhibit diverse forms, structures, formats, and settings, advancements in natural language processing present new challenges. Consequently, numerous studies have endeavored to address these challenges through the utilization of diverse algorithms, crowd-sourced datasets, and deep-learning techniques. Therefore, this chapter identifies and analyzes studies similar to the present one, highlighting their limitations and paving the way for the current project.

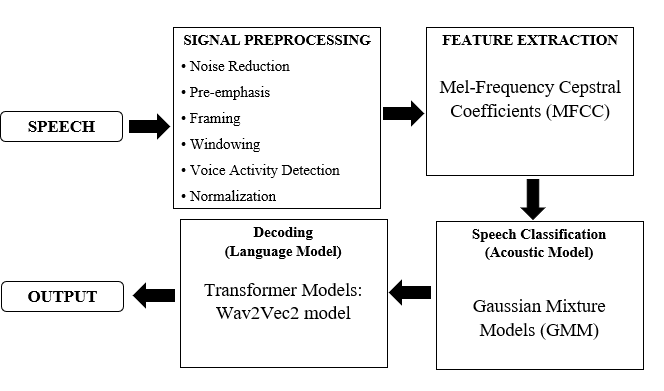


Figure 1: Speech Recognition Process

In Figure 1, This framework provides a structured view of the different theoretical components and methods involved in the speech recognition process, from the initial audio signal processing to the final text output.

## **2.1 Related Literature**

### **2.1.1 Automatic Speech Recognition**

Wang et al. (2020) developed an Automatic Speech Recognition (ASR) model tailored for the Swedish language, utilizing Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). The study emphasized feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Predictive (PLP) features, highlighting the importance of triphone models and feature transformation for improved ASR performance. This research provides valuable insights for enhancing ASR technology, particularly for languages with smaller speaker populations like Swedish.

**2.1.2 Arabic Speech Recognition**

Arabic presents unique challenges for ASR due to its complex morphology and various dialects. Studies have employed end-to-end methods using deep neural networks and techniques like Long Short-Term Memory (LSTM) networks. Naima et al. (2020) achieved a 98.77% accuracy in recognizing Arabic spoken digits using MFCC features and LSTM networks. Other research by Abdelhamid et al. (2020) implemented character-based models using the Kaldi toolkit and achieved a Word Error Rate (WER) of 12.03%. These efforts underscore the potential of end-to-end models for Arabic ASR, addressing the language's unique complexities.

### **2.1.3 Arabic Speech Recognition System in Noisy Environments**

Ouisaadane and Safi (2021) compared the performance of GMM-HMM and DNN-HMM models in noisy environments using the CMU Sphinx tools and Kaldi toolkit. The study found that integrating noisy training theory into these models enhances their accuracy in adverse conditions. This research highlights the ongoing challenges and advancements in refining speech recognition systems to perform well in noisy environments.

### **2.1.4 An Investigation of Google’s English-Arabic Translation of Technical Terms**

While Google Translate (GT) accurately translates certain technical terms like 'mobilization' and 'technical', it struggles with terms that have diverse prefixes and roots combined with identical suffixes. This inconsistency necessitates review and revision by linguists to ensure precision in scientific terminology. Additionally, GT often fails to understand the nuanced meanings of compound or blended technical terms, treating them as separate words rather than cohesive units. For instance, while 'radiotherapy' is translated correctly, terms like 'physiotherapy' and 'aromatherapy' are often mistranslated due to GT's inability to discern specific domain contexts.

To address these challenges, Jusoh and Alfawareh proposed a semantic-based translation framework, and Ahmed and Nürnberger (2008) introduced a word sense disambiguation approach tailored to Arabic morphology. These methods use natural language processing techniques and large parallel corpora to improve translation accuracy. Furthermore, maintaining consistency in Arabic equivalents across parts of speech, word forms, and affix combinations is crucial. Some compounds may need to be translated as block sequences to convey their intended meaning accurately.

Studies have indicated that the accuracy rate of Google Translate for Arabic is around 50-60%, highlighting the significant room for improvement compared to other languages. This accuracy rate underscores the importance of careful review and validation of GT outputs, especially in technical and domain-specific contexts.

In conclusion, while GT is a valuable resource, its limitations require users to be cautious. Future research should focus on improving GT’s bidirectional translation quality, comparing its performance with other systems, and exploring verification strategies used by EFL students (Al-Jarf, 2021).

### **2.1.5 Arabic Machine Translation**

The review of the study conducted by (Ameur, et al., 2020) delves into the distinctive characteristics of Arabic, the specific hurdles it poses for translation, and the latest trends and obstacles encountered in Arabic MT research. Ameur, Meziane, and Guessoum (2020) offer valuable insights into the evolution of MT, underlining the significance of research dedicated to Arabic MT. Their analysis sheds light on the particular challenges inherent in Arabic translation and evaluates existing research, tools, and resources within the domain. Through a systematic categorization of MT methodologies based on information sources and a thorough examination of their merits and drawbacks, the review provides a comprehensive panorama of the current state of Arabic MT research. While much of the focus in Arabic MT research revolves around translating Arabic into English, statistical and neural techniques dominate the landscape (Ameur, et al., 2020).

However, there remains a dearth of exploration in alternative translation approaches, suggesting a need for further investigation. By classifying research endeavors according to their chosen translation methodologies and the specific translation obstacles they address, the survey offers valuable insights into the diverse strategies employed to surmount translation challenges. The quality of MT systems is significantly influenced by access to linguistic corpora and tools. The review underscores the pivotal role played by essential resources and tools in the development and evaluation of Arabic MT, emphasizing their contribution to refining translation accuracy and fluency (Ameur, et al., 2020).

## **2.2 Related Technologies**

### **2.2.1 Mel-Frequency Cepstral Coefficients (MFCC)**

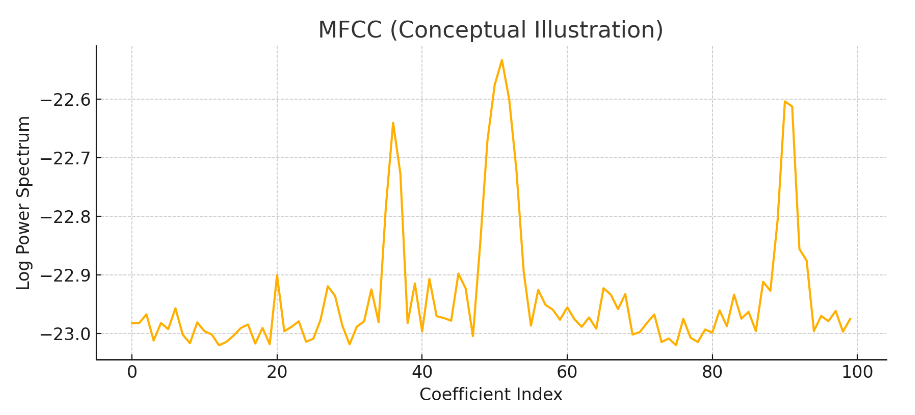


Figure 2: simplified version of the MFCC illustration using a sine wave as an example.

Mel-Frequency Cepstral Coefficients (MFCC) are a feature extraction technique commonly used in automatic speech recognition (ASR) systems. The MFCC algorithm transforms the audio signal into a compact representation that captures the power spectrum of the signal. This process involves taking the Fourier transform of a windowed signal to produce a power spectrum, mapping the powers to the Mel scale to approximate human auditory perception, taking the logarithm of each Mel frequency power, and then applying a discrete cosine transform to the logarithm to obtain the MFCCs. The result is a set of coefficients that succinctly represents the short-term power spectrum of a sound, emphasizing frequencies to which the human ear is most sensitive​

In the context of Arabic speech recognition, MFCCs play a crucial role by providing a robust feature set that accurately represents the phonetic characteristics of Arabic speech. Arabic, with its rich phonemic inventory and complex morphology, benefits from the MFCC's ability to highlight pertinent frequencies. These features are essential for training acoustic models like Gaussian Mixture Models (GMM) and more advanced deep neural networks. By using MFCCs, the ASR system can effectively capture the nuances of spoken Arabic, thereby improving the accuracy and reliability of speech-to-text conversion, which is a foundational step before any subsequent translation tasks using systems like MarianMT

### **2.2.2 Gaussian Mixture Models (GMM)**

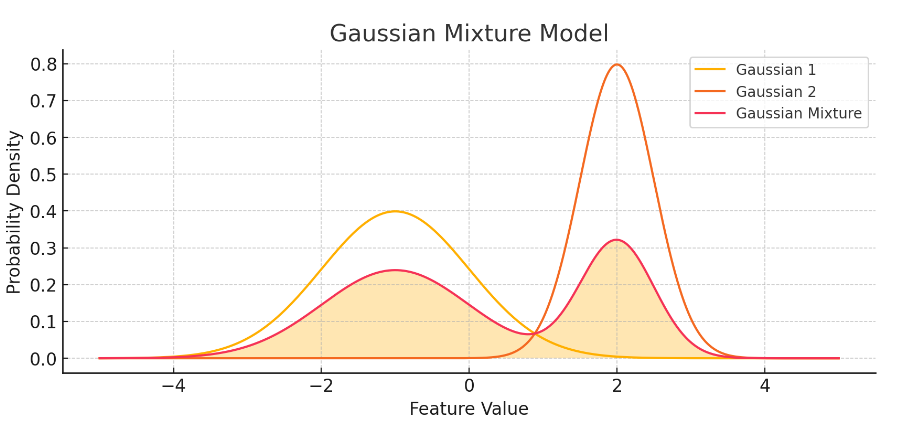


Figure 3: visualize two Gaussian distributions and their mixture.

Gaussian Mixture Models (GMM) are a probabilistic model used in pattern recognition and machine learning, particularly for speech recognition. GMMs model the probability distribution of feature vectors (such as MFCCs) by representing them as a mixture of multiple Gaussian distributions. Each Gaussian in the mixture corresponds to a different cluster or component within the feature space. The parameters of these Gaussian distributions (mean, variance, and mixture weights) are estimated using the Expectation-Maximization (EM) algorithm. In the context of ASR, GMMs are often used to model the acoustic features of speech, where each Gaussian represents the likelihood of a feature vector belonging to a specific phoneme.

GMMs are highly relevant to the research on Arabic speech recognition due to their effectiveness in modeling the variability in speech signals. Arabic speech, characterized by diverse dialects and phonetic variations, benefits from the GMM's ability to handle complex and overlapping distributions of speech features. By integrating GMMs with Hidden Markov Models (HMMs), the research aims to build robust acoustic models that can accurately recognize and transcribe Arabic speech. This accurate transcription is crucial for the subsequent translation phase, where the transcribed Arabic text is translated into English using neural machine translation frameworks like MarianMT.

### **2.2.3 Transformer Models (Wav2Vec2)**

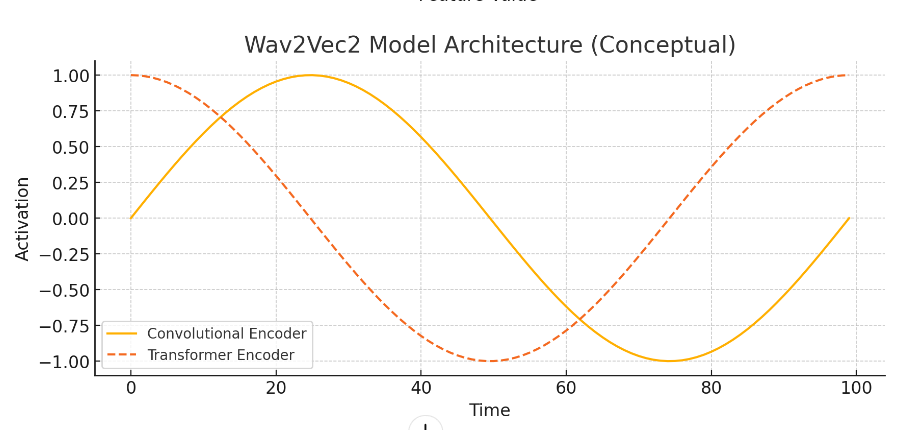


Figure 4: conceptual diagram showing the convolutional encoder and transformer encoder parts.

Transformer models, specifically Wav2Vec2, represent a significant advancement in the field of speech recognition. Wav2Vec2 leverages a self-supervised learning approach to train the model on large amounts of unlabeled audio data. The model consists of a convolutional neural network (CNN) encoder that processes the raw audio waveform, followed by a transformer-based architecture that captures contextual information across the entire sequence. The key innovation of Wav2Vec2 is its ability to learn high-level representations of audio directly from the waveform, without relying on traditional hand-crafted features like MFCCs. This results in a more flexible and powerful model that can generalize well across different languages and acoustic conditions.

For Arabic speech recognition, Wav2Vec2 offers a promising approach due to its robustness and flexibility. The transformer architecture of Wav2Vec2 allows it to capture long-range dependencies and intricate patterns in Arabic speech, which are often challenging for traditional models to handle. By training on large-scale Arabic speech datasets, Wav2Vec2 can develop a deep understanding of Arabic phonetics and phonology, leading to improved recognition accuracy. This high-quality transcription is essential for the subsequent translation task, where MarianMT can accurately translate the recognized text into English. The integration of Wav2Vec2 in the research framework aims to enhance the overall performance and reliability of the Arabic-to-English speech translation system. ​

**CHAPTER 3**

# **DESIGN AND METHODOLOGY**

This chapter focuses on describing the process of developing the speech-to-text system that translates Arabic speech to its corresponding English word translation.

## **3.1 Conceptual Framework**

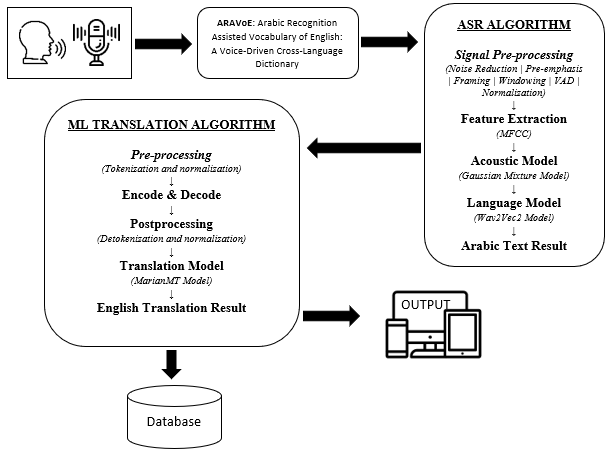


Figure 6. Diagram of proposed Arabic Recognition Assisted Vocabulary of English: A Voice-Driven Cross-Language Dictionary

In the conceptual framework illustrated in Figure 6, the ASR (Automatic Speech Recognition) algorithm begins with signal pre-processing, which includes noise reduction, pre-emphasis, framing, windowing, VAD (Voice Activity Detection), and normalization. Following this, feature extraction is performed using MFCC (Mel-Frequency Cepstral Coefficients). The extracted features are then processed by the acoustic model, which employs a Gaussian Mixture Model, and subsequently by the language model using the Wav2Vec2 model to produce the Arabic text result.

This Arabic text is then fed into the ML (Machine Learning) translation algorithm. The process starts with pre-processing steps, including tokenization and normalization. The text is then passed through the encode and decode phases. Postprocessing follows, which involves detokenization and normalization. The translation model, specifically the MarianMT model, then translates the Arabic text into English. The final English translation result is stored in the database and displayed as the output.

## **3.2 Data Gathering**

Researchers conduct interviews to gather data comprehensively, aiming to enrich their study and gain insights into the information necessary for the application to meet its goals. Through interviews, they gather valuable input from stakeholders, users, or experts, which helps in shaping the application's functionality, features, and design to align with the desired objectives. The information gathered from interviews serves as a foundation for informed decision-making and ensures that the application addresses the identified needs effectively. The researchers had a total of four respondents from different Arabic Teachers in Butuan City under the DepEd Program called Madrasah Educational Program (MEP), with the permission of its Program Focal person, Mr. Restituto Navarro. They are Saidomar B. Tominaman from Jose T. Domingo CES (JTDSCES), Norania Rascal, Normina Hadji Salam, and Nasrima Amerol from Ongyui CES (OYCES).

Learning Arabic is not easy; it involves rules, vowels, and special characters, adding complexity to the language learning process. With the help of the data being gathered, the researchers have a concrete understanding of the Arabic language and the specific needs and preferences of Arabic teachers in Butuan City.

## **3.3 Automatic Speech Recognition Phase**

The system identifies the words spoken by the user and converts them into a format comprehensible to the machine. In this stage, researcher use the word “Ana Waseem (انا وسيم)” The detailed process of ASR is as follows:

### **3.3.1 Acoustic Feature Extraction**

#### 3.3.1.1 Audio Input Acquisition

The process starts with capturing the raw audio input through a microphone. This audio input is typically in the form of a continuous waveform.

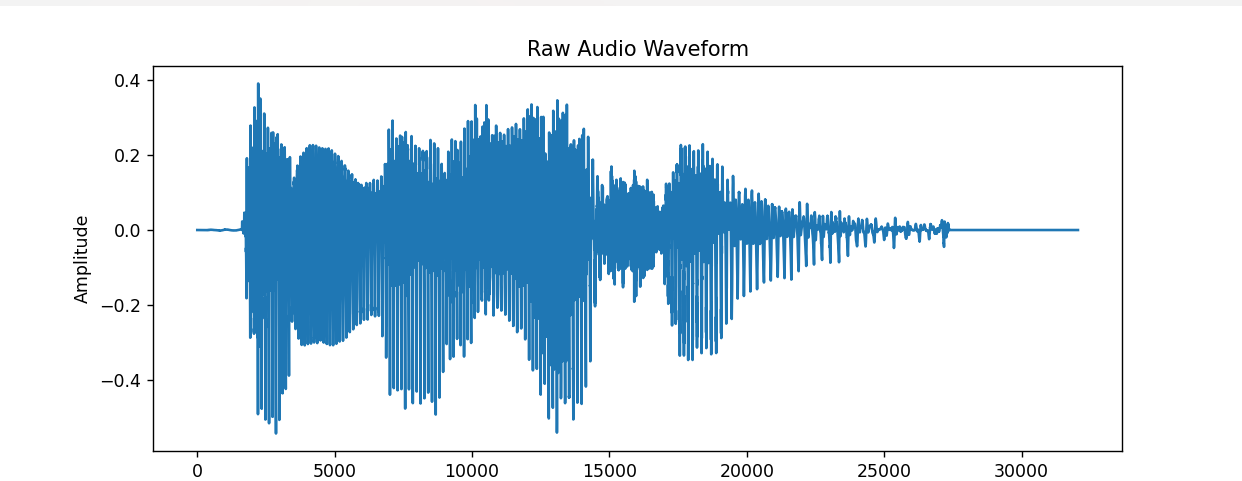


Figure 7: Wavelength

#### 3.3.1.2 Pre-emphasis

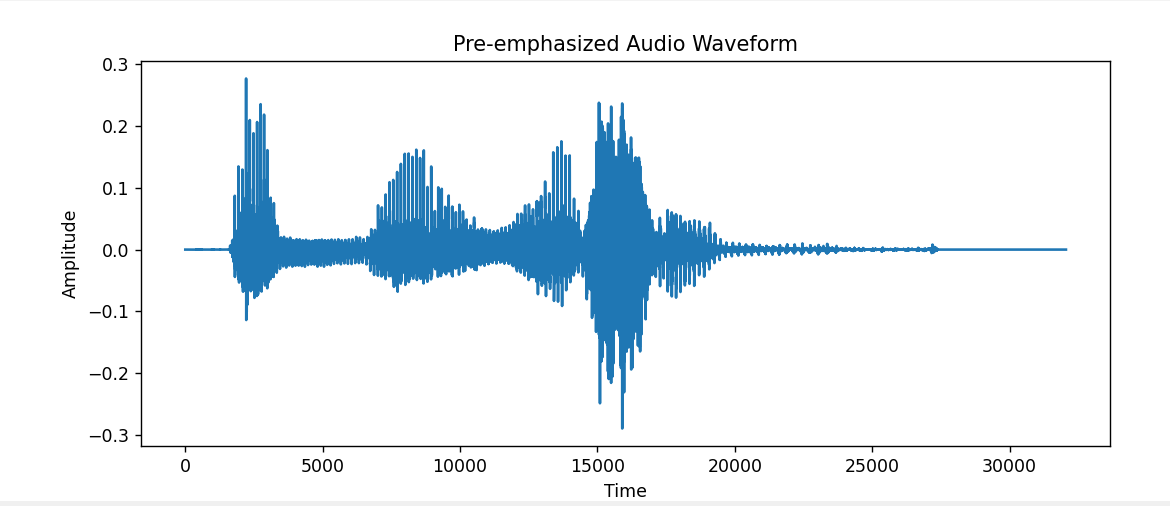
The raw audio signal is pre-emphasized to amplify the high frequencies. This step helps in compensating for the attenuation of high-frequency components during the recording process.

Figure 8: Pre-emphasized Audio waveform generate from figure 7

#### 3.3.1.3 Framing

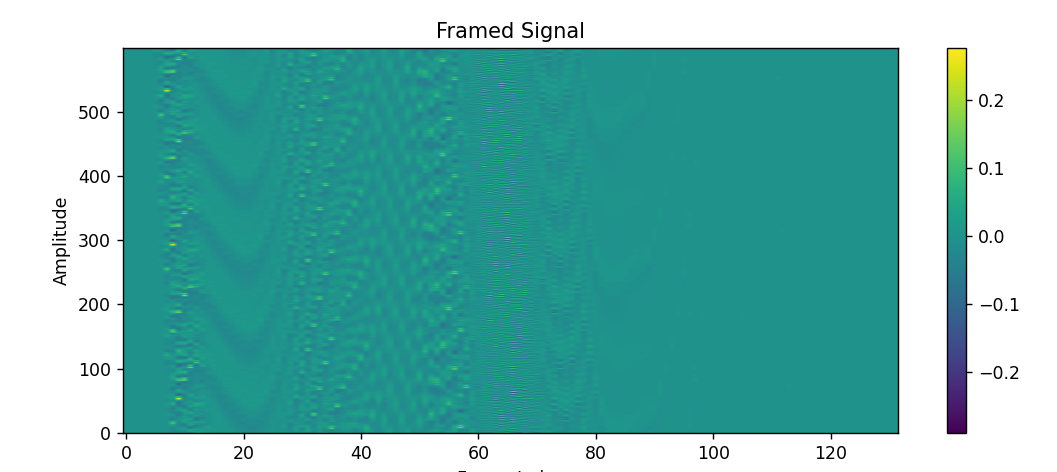
Arabic speech, like any other language, is quasi-stationary over short periods. The typical frame length of 20-30 milliseconds and an overlap of 10-15 milliseconds are suitable. 

Figure 9: Frame signal generate from figure 8

#### 3.3.1.4 Windowing

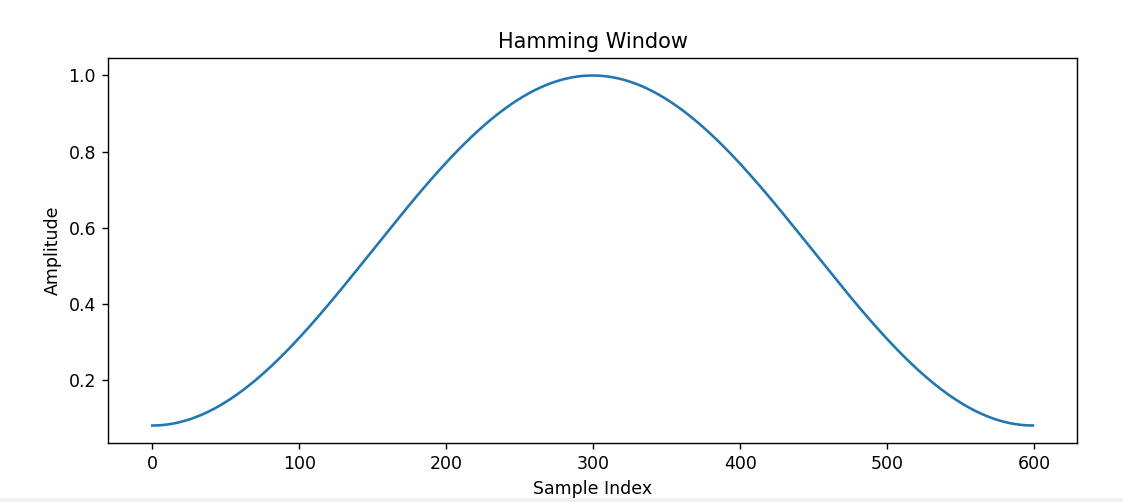
Applying a Hamming window (or similar window function) to each frame is crucial to reduce signal discontinuities. 

Figure 10: Humming window

#### 3.3.1.5 Fast Fourier Transform (FFT)

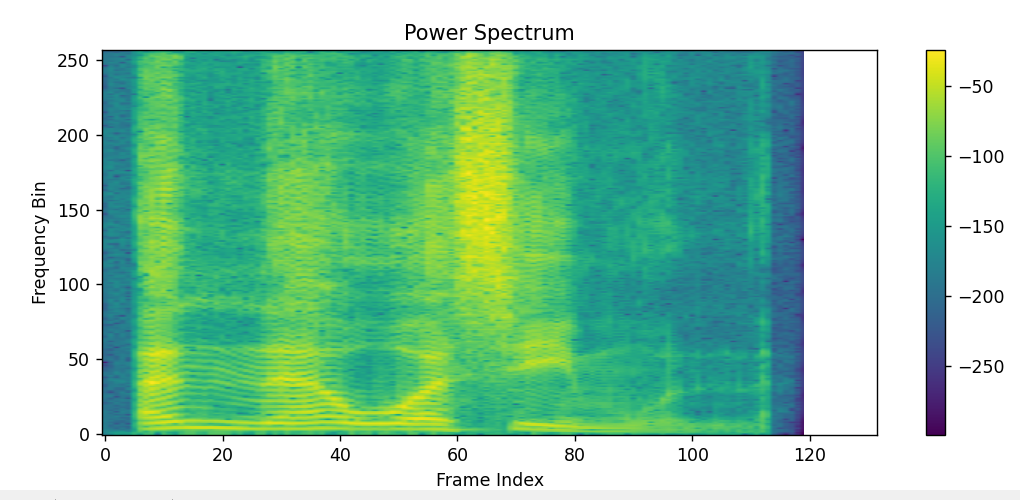
Converting the windowed frames to the frequency domain using FFT is a standard practice and works for Arabic as well. 

Figure 11: Power Spectrum

#### 3.3.1.6 Mel Filter Bank Processing

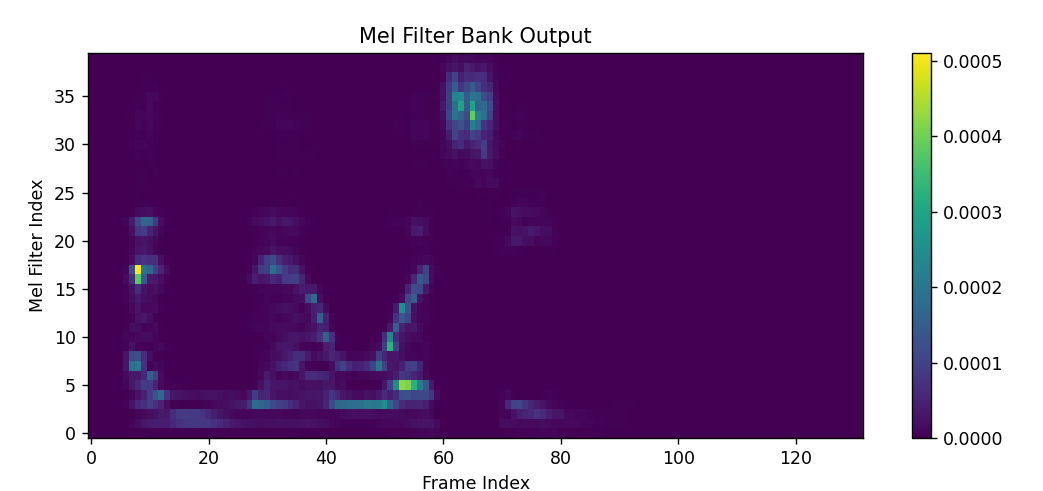
****The Mel scale is designed to approximate the human ear's response and is equally valid for Arabic. The Mel filter bank will capture the essential frequency components.

Figure 12: Illustration of Mel Filter Bank Output

#### 3.3.1.7 Logarithm of Power Spectrum

Taking the logarithm of the power spectrum is useful for compressing the dynamic range, which is beneficial for modeling any speech, including Arabic.

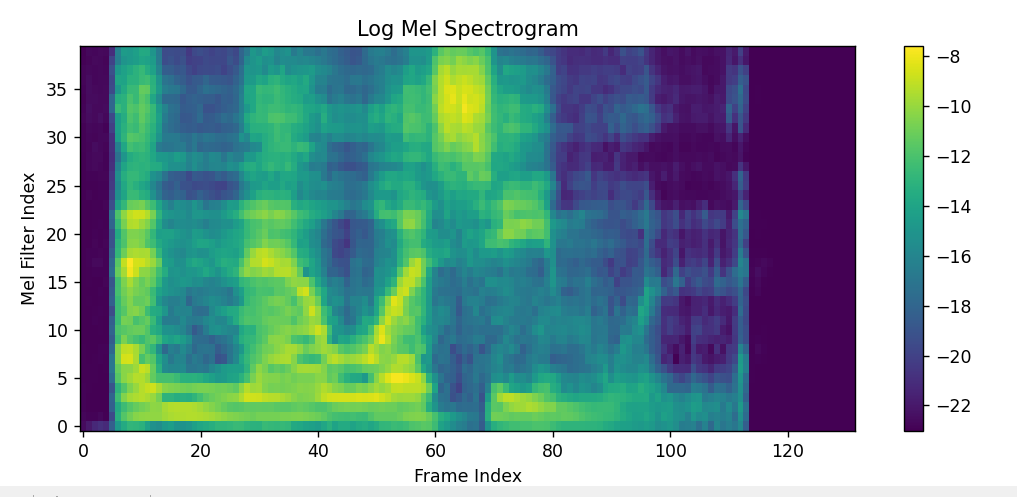


Figure 13: Illustration of Log Mel Spectrogram

#### 3.3.1.8 Discrete Cosine Transform (DCT)

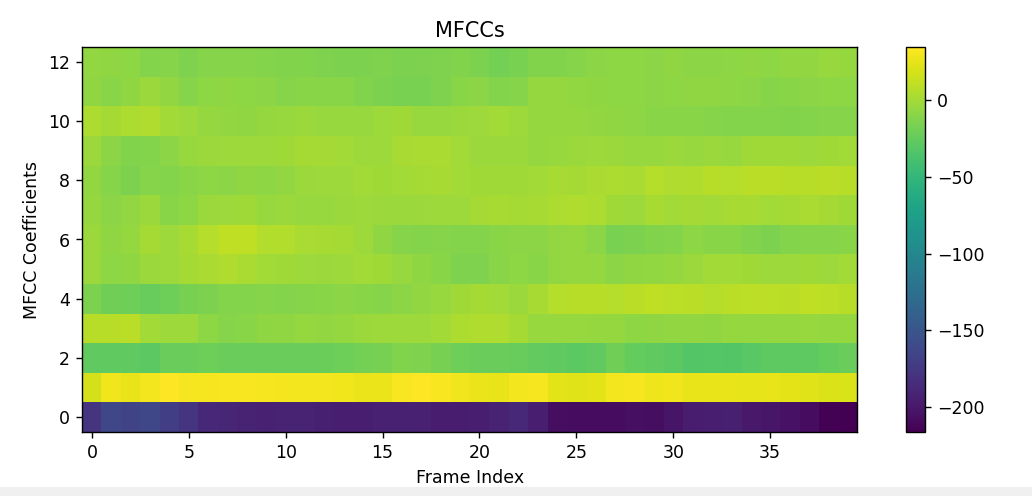
Transforming the log-Mel filter bank energies using DCT to obtain MFCCs is a common practice in ASR and is applicable to Arabic.

Figure 14: Illustration of MFCCs

#### 3.3.1.9 Delta and Delta-Delta Coefficients

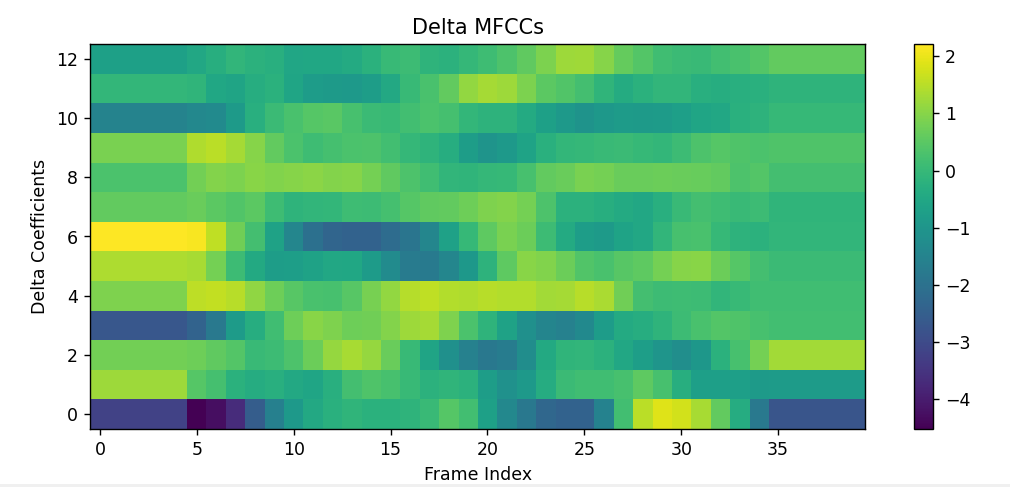
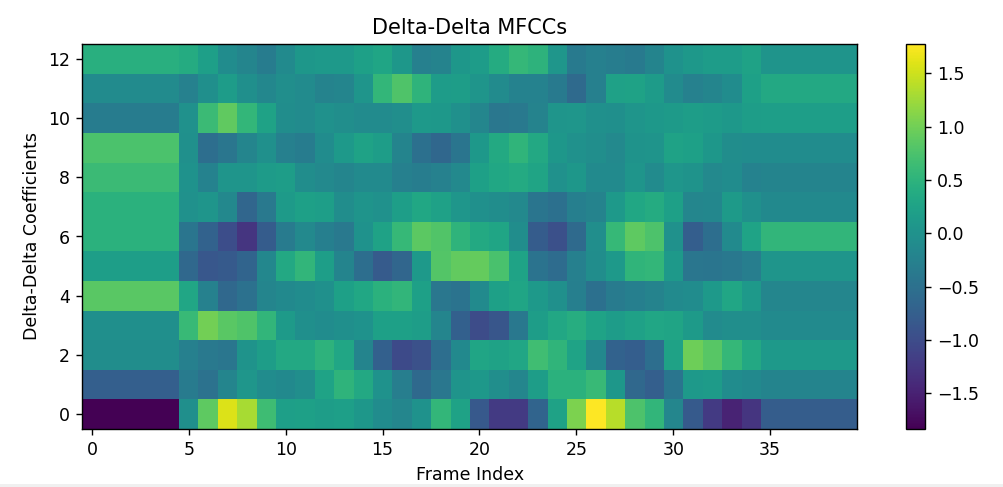
Computing the first and second-order derivatives of the MFCCs to capture temporal dynamics is essential for capturing the nuances in Arabic speech.

Figure 15: Illustration of Delta MFCCs

 Figure 16: Illustration of Delta-Delta MFCCs

### **3.3.2 Acoustic Modeling**

#### 3.3.2.1 Training Data Preparation

Collecting a large corpus of transcribed Arabic audio data is crucial. This corpus should cover various dialects, accents, and acoustic conditions found in the Arabic-speaking world.

#### 3.3.2.2 Feature Extraction

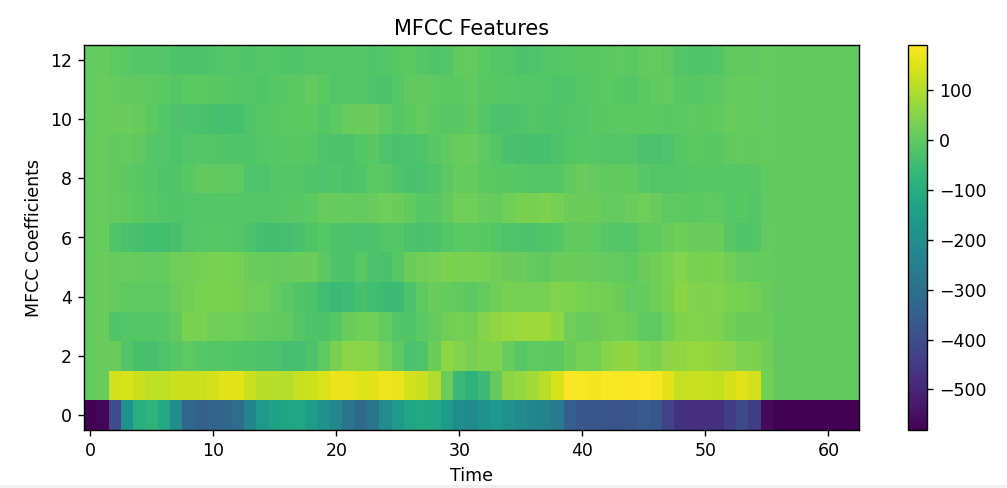
Extracting MFCCs and their derivatives from the training audio data remains the same.

Figure 17: Illustration MFCCs Features,

This plot shows the MFCC features extracted from the audio, which are used as inputs for the acoustic model.

#### 3.3.2.3 Model Training

Training the acoustic model using techniques like Hidden Markov Models (HMMs) or Deep Neural Networks (DNNs) is applicable. Ensure the training data is diverse and representative of the Arabic language.

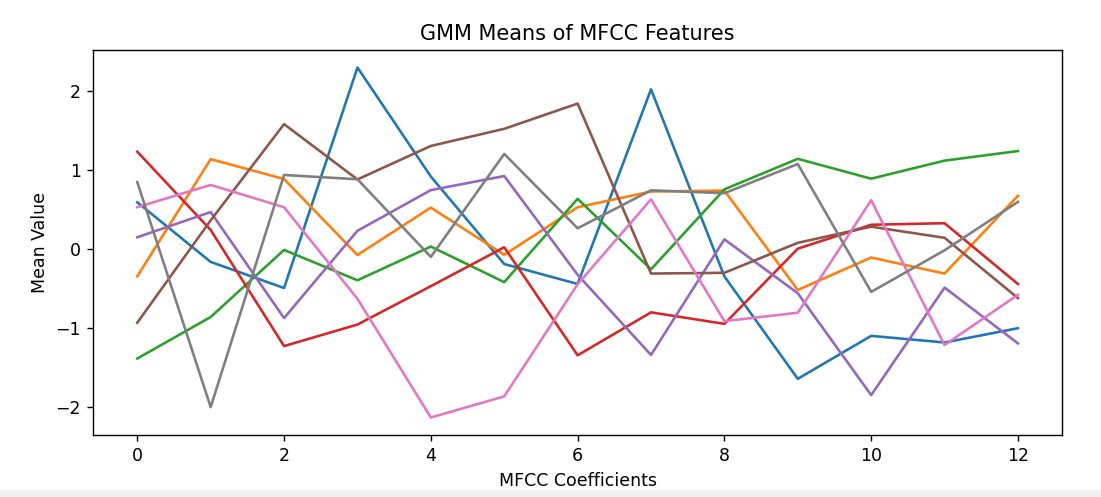
But here, the researcher illustrates a simplified version of acoustic model training using a Gaussian Mixture Model (GMM) on MFCC features. This is a basic approach compared to modern techniques but is good for understanding the concept.

Figure 18: Illustration of GMM Means of MFCC Features

This plot shows the mean values of the GMM components after training on the MFCC features, representing the learned distribution of the features.

### **3.3.3 Language Modeling**

Language model training involves using a large corpus of text to learn the statistical properties of the language, such as word frequencies and the probabilities of word sequences. For illustration purposes, we'll use a simple N-gram model. Here, I'll show how to train a bigram model (an N-gram model with N=2) and visualize the learned probabilities.

#### 3.3.3.1 Language Model Training with a Bigram Model

##### 3.3.3.1.1 Text Corpus Preparation

A large text corpus is used to train the language model. This corpus should represent the target language's vocabulary, grammar, and usage patterns.

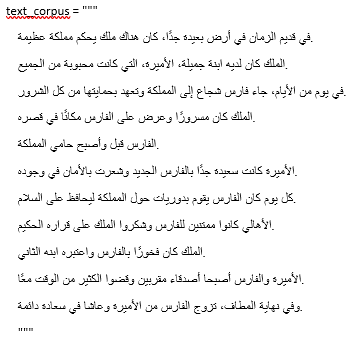


Figure 19: Text corpus used for this process.

##### 3.3.3.1.2 N-gram Model Training

The language model is trained using bigrams, where the probability of a word is conditioned on the previous work.

### **3.3.4 Decoding**

#### 3.3.4.1 Integration of Models

Integrating the acoustic model, language model, and pronunciation dictionary specifically tailored for Arabic is essential. We'll use a pre-trained Gaussian Mixture Model (GMM) for the acoustic model and a bigram model for the language model.

#### 3.3.4.2 Viterbi Decoding

The Viterbi algorithm can be used to find the most likely sequence of words given the acoustic and language model probabilities. This step is language-agnostic but should consider the specific phonetic and lexical properties of Arabic.

#### 3.3.4.3 Hypothesis Generation

Generating the decoded output as the recognized Arabic text corresponding to the spoken input.

## **3.4 Machine Translation Phase Using MarianMT**

The Machine Translation (MT) phase is crucial for converting the recognized Arabic text into another language, such as English. This process involves several steps, including text preprocessing, model selection, and translation. Here's a detailed breakdown of how each step can be applied to Arabic-to-English translation.

### **3.4.1 Text Preprocessing**

#### 3.4.1.1 Normalization

Normalization is essential to standardize the text before translation. This includes removing diacritics (harakat) and normalizing variations in spelling. For instance, different forms of the same letter (e.g., "إ", "أ", "آ" -> "ا") should be unified.

#### 3.4.1.2 Tokenization

Arabic text tokenization is more complex than in some other languages due to the cursive nature of the script and the presence of clitics (e.g., conjunctions and prepositions attached to words). Proper tokenization is crucial for accurate translation.

#### 3.4.1.3 Removing Punctuation

Punctuation marks should be handled appropriately. They may either be removed or converted to a standardized form based on the requirements of the translation model.

### **3.4.2 Model Selection**

#### 3.4.2.1 Pre-trained Models

Leveraging pre-trained MarianMT models for Arabic-to-English translation is a robust choice. These models are trained on extensive datasets and are highly effective for general translation tasks.

### **3.4.3 Translation**

#### 3.4.3.1 Tokenization and Encoding

The normalized and tokenized Arabic text is converted into a format suitable for the translation model. This typically involves using a tokenizer compatible with the MarianMT model.

#### 3.4.3.2 Translation

The model generates the translated text. This involves predicting the most likely sequence of words in the target language (e.g., English) given the input sequence in Arabic.

#### 3.4.3.3 Post-processing

The raw translated text may require some post-processing, such as detokenization and formatting, to produce a natural and readable output.

## **3.5 Proposed Tools**

In this section, it provides the tools that the researchers used to create the website.

### **3.5.1 Software**

**3.5.1.1 PyCharm.** When working on Python-based projects, developers find PyCharm to be a highly useful tool. It offers many features that speed up and simplify the development process. This supports Python libraries and frameworks that are very essential for this project.

**3.5.1.2 Python 3.11.** Python is a commonly used, interpreted, object-oriented, high-level computer language with dynamic semantics that is used for general-purpose programming. Python will be the chosen programming language since the algorithm is simpler to construct and operate using it. Python accelerates development without losing scalability or dependability. Additionally, it has strong community support and an extensive library collection, and its popularity continues to rise.

**3.5.1.3 Github.** GitHub is a website that gives developers a powerful set of tools for organizing software projects and working together on them. The researcher will use this to easily connect to a hosting website.

### **3.5.2 Hardware**

**Computer Unit.** The unit that will be used for implementing the study is a laptop with 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz with 16.0 GB (15.7 GB usable) installed RAM.

## **3.6 Dataset**

The researcher will utilize the following datasets:

For Arabic Speech Recognition:

● Arabic Speech Corpus by Haithem Hermessi, (August, 2023)

● Arabic Speech to Text wav by Marilia Prata (July, 2023)

## **3.7 Multiple Constraints and Trade-off Analysis**

The proponents used multiple constraints to determine the algorithms that will be utilized in this project. This part will discuss the multiple constraints as well as the trade-off, and sensitivity analysis of this study.

### **a. Constraints Definition**

**Accuracy**. It assesses the value of matching Arabic-English words.

**Speed (ms).** Measuring the speed of an algorithm on how it performs given on a specific dataset will be measured in milliseconds. Having a lower value of milliseconds depicts a faster algorithm.

### **b. Design Tradeoffs (Strength and Weaknesses)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criteria | Kaldi ASR | Google Speech-to-Text API | DeepSpeech | Flashlight ASR (Wav2Letter) | SpeechBrain |
| Strengths | | | | | |
| Open Source | Active open-source community | Proprietary service with robust infrastructure | Active open-source community | Active open-source community | Active open-source community |
| Customization | Highly customizable | Limited customization | Customizable | Customizable | Highly customizable |
| Accuracy | High accuracy, especially in diverse languages | High accuracy, benefits from Google's extensive resources | Competitive accuracy | Competitive accuracy | Competitive accuracy |
| Real-time Processing | Real-time processing is demanding | Real-time capabilities | Real-time capabilities | Real-time capabilities | Real-time capabilities |
| Community Support | Active and vibrant community | Limited community involvement due to being proprietary | Active community support | Active community support | Active community support |
| Applications | Broad range of applications including speech-to-text and voice commands | Suitable for transcription, voice commands, and various applications | Well-suited for speech-to-text and transcription | Primarily used for speech-to-text applications | Versatile for various speech processing tasks |
| Weaknesses | | | | | |
| Learning Curve | Steep learning curve for beginners | User-friendly, but may lack flexibility | Moderate learning curve | Moderate to steep learning curve | Moderate to steep learning curve |
| Resource Intensive | Requires substantial computing resources | Resource usage is managed by Google | Resource-intensive during training | Resource-intensive during training | Resource-intensive during training |
| Provider Dependence | Independent of major tech companies | Dependent on Google infrastructure | Developed by Mozilla, but some reliance on external resources | Developed by Facebook AI Research | Independent, less provider dependence |
| Open Source Collaboration | Actively collaborates with the open-source community | Limited collaboration due to being proprietary | Collaborates with the open-source community | Collaborates with the open-source community | Actively collaborates with the open-source community |

This section will discuss the criteria used to evaluate the overall system capabilities of the algorithms. This will determine which algorithm is ideal for implementation in the application. Accuracy and precision are the primary determinants of the algorithm's efficiency and effectiveness in a web application.