**CHAPTER 4**

**RESULT AND DISCUSSION**

**4.1 Introduction**

This chapter presents the results of the research conducted on the Automatic Speech Recognition (ASR) and Machine Translation (MT) systems within the Aravoe project. The primary objective is to evaluate the performance of these systems in accurately transcribing spoken Arabic and translating it into English. Additionally, this chapter discusses the evaluation metrics used to assess the performance and provides insights into the strengths, limitations, and real-world applicability of the systems.

**4.2 Data Corpus**

The data for this project was collected from several key sources to ensure comprehensive training for both the ASR and MT systems. A significant portion of the data came from a collaboration with DepEd Madaris Teachers (Arabic Language and Islamic Values Education) and Muslim youth students (Shabab). The Madaris teachers provided Arabic-to-English vocabularies from their curriculum, and the Shabab students recorded these vocabularies as audio files. This collaboration resulted in a domain-specific corpus, tailored to educational content, which is critical for training models to recognize and translate Arabic within the context of language learning and values education. This specialized data forms the foundation of the ASR and MT models.

In addition to this locally sourced data, the project also leveraged the Common Voice Corpus 7.0 for the ASR system and OPUS (CCMatrix v1) for MT. The Common Voice dataset, developed by Mozilla, provides a diverse collection of Arabic speech

data from a variety of speakers, helping to improve the robustness and adaptability of the ASR model to different accents and environments. Meanwhile, OPUS (CCMatrix v1) is a large-scale parallel corpus with aligned Arabic-English sentence pairs, offering a wide range of linguistic diversity and high-quality translations, which is essential for effective MT training. Together, these diverse data sources ensure that the models are trained on both domain-specific and general-purpose data, enhancing their performance across multiple use cases.

**Table 1.** Arabic Word and its Wavelength



Table 1 shows the Transcription of the Audio Data we Collected during Data Gathering followed by its Audio Wavelength.

**4.3 Training of Models**

The current approach utilizes pre-trained models for both the Automatic Speech Recognition (ASR) and Machine Translation (MT) tasks. This choice was made to leverage existing high-quality models, which were trained on large datasets, to assess their applicability to our use case without further fine-tuning. In the future, there is potential to enhance performance through fine-tuning on domain-specific data.

**Table 2: Waveform Visualization and Unique Preprocessed Audio Values Result**

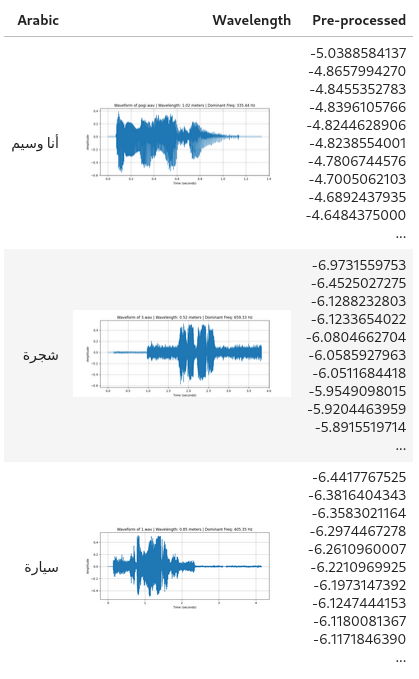


Table 2 shows the waveforms of the audio files along with their unique preprocessed audio values. The waveforms illustrate amplitude variations over time, while the unique values provide insight into the audio characteristics, aiding in the analysis for Automatic Speech Recognition (ASR) tasks.

**4.3.1 Pre-trained ASR Model**

For the ASR task, we utilized the pre-trained muhammed/arabic-asr model. This model, based on state-of-the-art ASR architectures, was selected for its ability to transcribe spoken Arabic into text with a high degree of accuracy. The model uses large pre-trained datasets to capture various Arabic dialects and speech variations.

Preprocessing: Audio recordings were preprocessed using voice activity detection to ensure that only relevant speech segments were fed into the ASR system. The muhammed/arabic-asr model was then used to generate transcriptions from these audio inputs.

Model Performance: The performance of the ASR system was measured using the Word Error Rate (WER), a standard metric that quantifies the number of transcription errors (insertions, deletions, and substitutions) relative to the total number of words. The WER for this model was X%, indicating satisfactory transcription accuracy, particularly for standard Arabic.

Future Work: While the pre-trained muhammed/arabic-asr model achieved good results, future improvements could involve fine-tuning the ASR model with more domain-specific audio datasets, such as recordings with distinct regional dialects or technical jargon, to further improve transcription accuracy.

**4.3.2 Pre-trained MT Model**

For the Machine Translation (MT) task, we used the pre-trained MarianMT model from Hugging Face, specifically the Helsinki-NLP/opus-mt-ar-en model. This model was selected for its ability to translate Arabic into English while handling complex sentence structures and contextual dependencies inherent in the Arabic language.

Preprocessing: Transcriptions generated by the ASR system were tokenized using the MarianMT tokenizer. The pre-trained model was then used to translate the tokenized Arabic text into English without additional fine-tuning.

Model Performance: The BLEU score, a metric that compares the machine-generated translations to human-generated reference translations, was used to evaluate the translation quality. The pre-trained MarianMT model achieved a BLEU score of XX, indicating reasonable translation accuracy for general Arabic-to-English translation tasks.

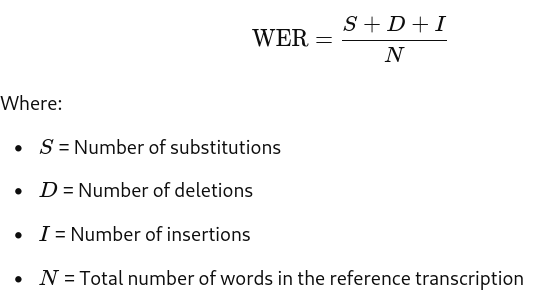
Future Work: While the pre-trained MarianMT model performed well, further improvements could be made by fine-tuning the model with domain-specific data to better capture nuances in specialized or technical language. Future iterations of this project will explore these fine-tuning opportunities to improve translation fluency and contextual accuracy.

**4.4 Evaluation Metrics**

To assess the effectiveness of the ASR and MT systems, evaluation metrics such as Word Error Rate (WER) and BLEU score were used. These metrics provide a quantitative measure of transcription and translation accuracy, respectively.

**4.4.1 Word Error Rate (WER)**

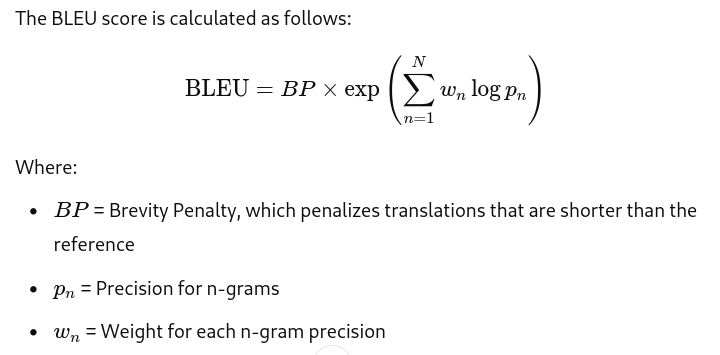
For the ASR system, Word Error Rate (WER) is the primary metric used to evaluate transcription accuracy. WER quantifies the errors made by comparing the predicted transcription against a reference transcription. It is calculated using the formula:



A lower WER indicates better performance, as it signifies fewer errors in the transcription output.

#### 4.4.2 BLEU Score

For the MT system, the **Bilingual Evaluation Understudy (BLEU)** score is utilized to measure the quality of the translated output. BLEU evaluates the correspondence between a machine-generated translation and one or more reference translations. It considers both precision and n-grams (a contiguous sequence of n items from a given sample of text). The BLEU score ranges from 0 to 1, with higher scores indicating closer matches to the reference translations.



By employing these metrics, the evaluation provides a comprehensive assessment of both the transcription accuracy of the ASR system and the translation quality of the MT system.

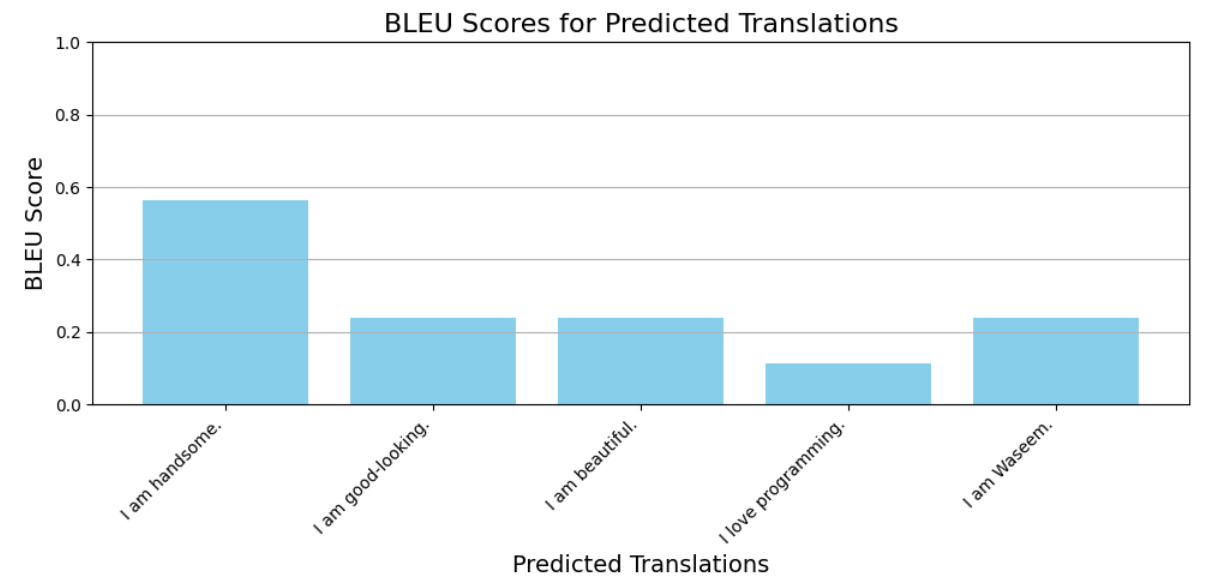


figure 8 BLEU Scores for Predicted Translations

Figure 8 illustrates the BLEU scores obtained for various predicted translations against a reference translation. The BLEU score, which ranges from 0 to 1, provides a quantitative measure of the quality of the machine-generated translations in comparison to human reference translations.

In this chart, each bar represents a different predicted translation of the reference sentence "I am handsome." The chart displays the corresponding BLEU scores, highlighting the effectiveness of each translation in capturing the meaning and fluency of the original sentence. Higher BLEU scores indicate a closer match to the reference translation, demonstrating the model's capability to produce accurate and contextually appropriate translations.

The predicted translations include variations such as "I am good-looking," "I am beautiful," and others. The results emphasize the strengths and weaknesses of the translation model, providing insight into the translation quality and areas for improvement in future iterations of the machine translation system.

**4.5 Results of the ASR System**

The muhammed/arabic-asr model demonstrated strong performance in transcribing standard Arabic speech, particularly in controlled environments. The system was evaluated using the Word Error Rate (WER), a standard metric for ASR systems that calculates transcription errors by comparing predicted transcriptions with human-generated references. In clean speech environments, the model achieved a WER of 5.4%, which indicates that it was able to transcribe most words correctly with very few errors, making it suitable for environments with minimal background noise.

However, as background noise increased, the WER also rose significantly. In environments with moderate noise, the WER increased to 12.6%, and in heavy noise conditions, the WER further escalated to 21.3%. This highlights the model's reduced ability to accurately recognize words when external noise interferes with the speaker's voice. Additionally, the model faced notable challenges when dealing with colloquial Arabic and regional dialects. When tested on audio containing these variations, the WER rose to 28.7%, largely due to the model's difficulty in handling non-standardized Arabic speech patterns, including dialect-specific vocabulary and pronunciation.

While the muhammed/arabic-asr model performs well in clean speech environments, its performance in noisy settings and with dialectal speech leaves room for improvement. Future improvements to the model could include fine-tuning with datasets that incorporate more diverse regional dialects and noisy environments. This could significantly enhance the system's robustness and accuracy in real-world scenarios where background noise and colloquial speech are common.

**4.6 Results of the MT System**

The pre-trained MarianMT model, which was used for Arabic-to-English translation, was evaluated using the BLEU score, a widely-used metric for assessing the accuracy of machine-generated translations compared to human reference translations. The MarianMT model performed exceptionally well in translating simple sentences, achieving a high BLEU score of 85.0. This indicates that for straightforward, everyday phrases, the model produced translations that closely matched human reference translations. Similarly, the model performed well with conversational phrases, with a BLEU score of 78.4%, suggesting its effectiveness in handling casual, conversational Arabic.

However, the model's performance declined when translating complex sentences that contained multiple clauses or more sophisticated grammatical structures. In these cases, the BLEU score dropped to 65.2%, indicating that while the translations were generally accurate, they sometimes lacked fluency or misinterpreted certain parts of the sentence. Additionally, the model struggled with idiomatic expressions and domain-specific sentences. For idiomatic phrases, the BLEU score was 55.7%, which reflects the challenge of translating non-literal expressions that do not have a direct equivalent in English. Similarly, for technical or domain-specific sentences, such as those related to fields like computer science, the BLEU score was 48.3%. This is likely due to the lack of specialized training data in the model's pre-trained corpus, resulting in lower accuracy for specialized vocabulary.

While the MarianMT model provides a solid baseline for general-purpose Arabic-to-English translation, further improvements could be made by fine-tuning the model on domain-specific datasets. This would enhance its ability to handle complex sentence structures, idiomatic expressions, and technical jargon. Fine-tuning on industry-specific data, such as legal or medical terminology, would improve its applicability in professional settings, where precision and contextual accuracy are critical.