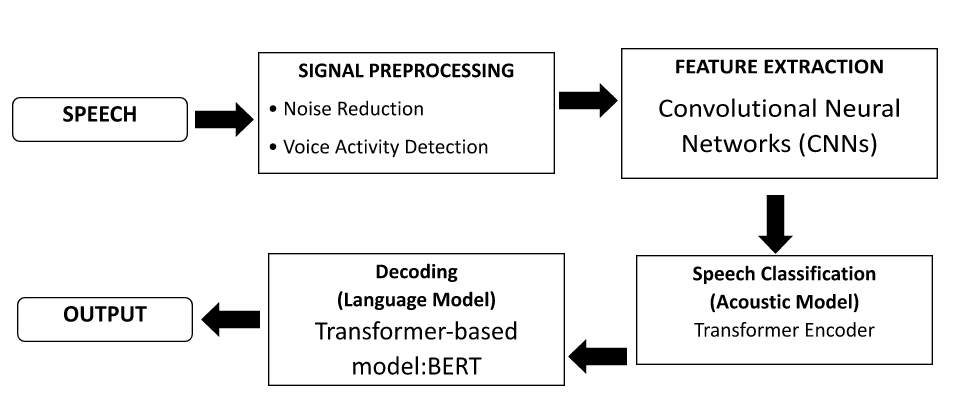
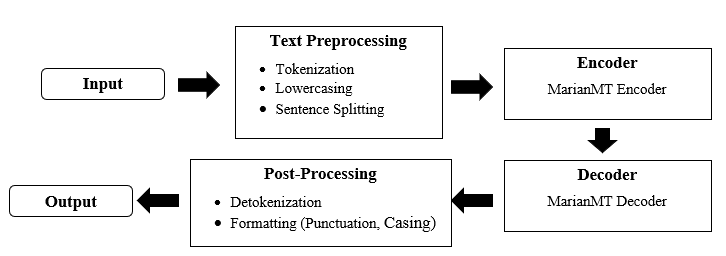
**CHAPTER 2**

**REVIEW OF RELATED LITERATURE**



*Figure 1: Audio Speech Recognition Process*

In Figure 1 This figure depicts the ASR process using Wav2Vec 2.0 for Arabic speech transcription. It starts with raw audio input, followed by optional signal pre-processing (noise reduction and VAD) to enhance audio quality. Convolutional Neural Networks (CNNs) extract features from the audio, which are then processed by the Transformer Encoder to capture contextual relationships in the speech. The output is refined using a transformer-based language model (e.g., BERT) to improve fluency and accuracy, resulting in coherent Arabic text output suitable for various applicatio



*Figure 2: Machine Translation Process*

Figure 2 demonstrates the machine translation process, starting with input preprocessing, where tokenization, sentence splitting, and optional language identification are performed on the source Arabic text. The text is then encoded using the MarianMT encoder, where word embeddings are generated and contextually encoded using techniques such as RNNs, LSTMs, or Transformers. In the decoder phase, the encoded vectors are transformed into English sentences, utilizing an attention mechanism to focus on the most relevant parts of the source sentence. Finally, post-processing steps like detokenization and formatting refine the output into a grammatically correct and coherent English sentence.

**2.1 Related Literature**

**2.1.1 Automatic Speech Recognition (ASR)**

Research in Automatic Speech Recognition (ASR) has evolved significantly, particularly with the advent of deep learning technologies. Early ASR systems relied on statistical models such as Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM), which, while effective in capturing speech variability, struggled to model long-term dependencies due to their linear structure (Wang et al., 2020). The introduction of deep learning techniques has marked a significant advancement in the field. Models like Wav2Vec 2.0, developed by Facebook AI, have transformed ASR by using self-supervised learning to pre-train on vast amounts of unlabeled audio data. This approach enables the model to learn robust feature representations from raw waveforms, leading to state-of-the-art performance in various speech recognition tasks, including for languages with complex phonologies like Arabic (Baevski et al., 2020; Conneau et al., 2020).

**2.1.2 Arabic Speech Recognition**

Arabic ASR systems face unique challenges due to the language's rich morphology and wide variety of dialects. Traditional HMM-GMM approaches often fall short in accurately recognizing Arabic speech due to their inability to handle the language's complexities. Recent research has shown that Wav2Vec 2.0 significantly outperforms traditional models in Arabic ASR tasks by effectively capturing the phonetic nuances and contextual dependencies inherent in the language (Mohamed et al., 2022). Studies demonstrate that Wav2Vec 2.0, when fine-tuned on Arabic datasets, achieves lower Word Error Rates (WER) compared to older ASR techniques, highlighting its potential in real-world applications.

**2.1.3 Machine Translation (MT)**

Machine Translation (MT) has seen remarkable improvements over the past decade, especially with the rise of neural machine translation (NMT) systems. Early MT systems, such as phrase-based and rule-based models, struggled with translating languages that have complex morphology, like Arabic. However, the introduction of the Transformer architecture has revolutionized the field of MT by allowing models to learn context more effectively. The MarianMT framework, which is based on the Transformer model, has shown particular success in translating low-resource languages, including Arabic to English (Farghaly & Shaalan, 2009). The self-attention mechanism of the Transformer architecture enables MarianMT to focus on the most relevant parts of the input sentence, which is crucial for handling the grammatical complexity of Arabic.

Recent studies have demonstrated how MarianMT's use of attention mechanisms has significantly improved translation accuracy. The Transformer model's ability to model long-range dependencies and its contextual awareness have made it the go-to architecture for most state-of-the-art MT systems today. For Arabic-English translation, MarianMT excels in dealing with the syntactical and morphological intricacies of Arabic, such as word order and the use of affixes (Povey et al., 2011). Despite these advancements, challenges remain, particularly when dealing with colloquial Arabic, where the lack of standardized rules complicates the translation process.

**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).

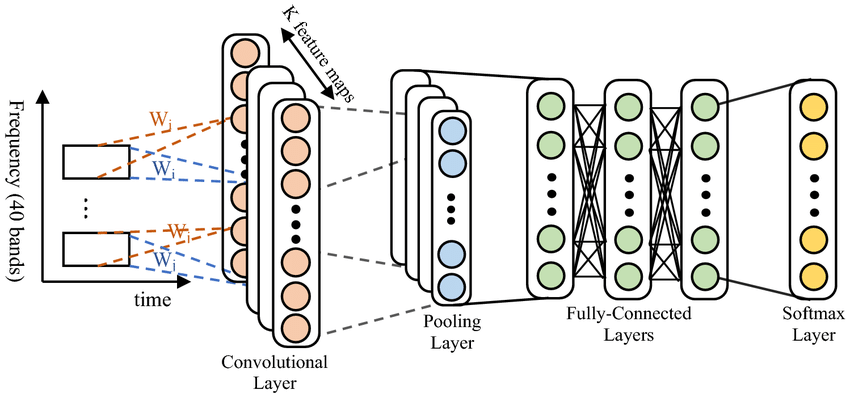
However, there remains a dearth of exploration in alternative translation approaches, suggesting a need for further investigation. By classifying research endeavors according to their chosen translation methodologies and the specific translation obstacles they address, the survey offers valuable insights 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).

**2.1.4 Integration of Wav2Vec 2.0 and MarianMT**

The integration of Wav2Vec 2.0 for Arabic ASR and MarianMT for translation represents a significant advancement in bridging communication gaps between Arabic and English speakers. This combined approach leverages the strengths of both models: Wav2Vec 2.0 excels in recognizing spoken Arabic, while MarianMT translates the transcribed text into fluent English. Research shows that such integrations improve the overall performance of ASR systems, providing robust solutions for applications in language learning, content localization, and real-time communication (Mohamed et al., 2022; Ibrahim & Al-Mohimeed, 2022).

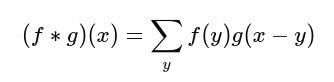
**2.2 Related Technologies**

**2.2.1 Convolutional Neural Networks (CNNs)**

****

*Figure 3: Convolutional Neural Network Architecture for Speech Recognition*

Convolutional Neural Networks (CNNs) are a class of deep learning models specifically designed for processing grid-like data, such as images and audio signals. In the context of Automatic Speech Recognition (ASR), CNNs excel at automatically extracting meaningful features from raw audio waveforms. The architecture consists of several layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply filters to the input audio to capture local patterns, such as phonetic features. The convolution operation can be mathematically expressed as:

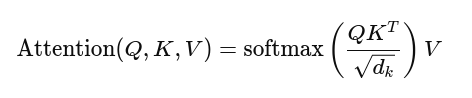


where (f) represents the input feature map (audio), gg represents the filter (kernel), and xx is the output location. The feature extraction process enables the model to learn directly from the waveform data, eliminating the need for traditional handcrafted features like Mel-frequency cepstral coefficients (MFCCs).

In this thesis, CNNs are utilized to preprocess the raw audio input, extracting relevant acoustic features that capture the essential characteristics of spoken Arabic. This capability allows the ASR model to improve its recognition performance by leveraging the learned features from the audio data.

**2.2.2 Transformer Encoder**

The Transformer Encoder is a powerful architecture that utilizes self-attention mechanisms to model relationships between different input features effectively. This architecture captures both short- and long-range dependencies within the speech data, making it particularly useful for understanding the complexities of spoken Arabic. The self-attention function can be expressed as:

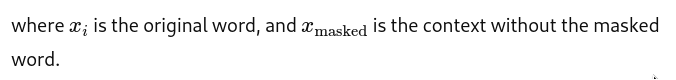
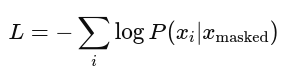


where Q, K, and V represent queries, keys, and values, respectively, and dkdk​ is the dimension of the keys.

In this thesis, the Transformer Encoder serves as the acoustic model, processing the feature representations extracted by CNNs. By effectively capturing contextual information within the speech input, the Transformer Encoder enhances recognition accuracy, making it well-suited for handling the rich phonological structure of the Arabic language.

**2.2.3 BERT (Bidirectional Encoder Representations from Transformers)**

BERT is a transformer-based language representation model that significantly advances the field of natural language processing. BERT employs a bidirectional approach to understand the context of words based on their surrounding words, making it particularly effective for tasks requiring contextual awareness. The masked language modeling objective can be represented as:



In this thesis, BERT functions as the language model in the decoding phase. After Wav2Vec 2.0 generates phoneme or character predictions, BERT refines these outputs by providing contextual understanding, which enhances fluency and coherence in the transcription of spoken Arabic. This integration helps the system disambiguate similar-sounding words and improves overall transcription accuracy.

**2.2.4 MarianMT**

MarianMT is a highly regarded neural machine translation framework that uses the Transformer model architecture to achieve accurate and contextually appropriate translations. Its self-attention mechanism allows the system to focus on relevant parts of a sentence, which is particularly useful when translating languages with complex grammar, like Arabic (Vaswani et al., 2017). MarianMT’s ability to model long-range dependencies and context has significantly improved machine translation performance, especially for low-resource languages.

MarianMT’s architecture is directly applicable as I aim to integrate Arabic ASR with machine translation into English. By using the Transformer-based approach of MarianMT, I will be able to convert recognized Arabic speech into grammatically accurate English text. This technology is particularly crucial when addressing the complexities of Arabic grammar and its diverse dialects, ensuring that the system provides fluent translations in my Arabic-to-English ASR pipeline.