Al-Farouk journal of sciences
Volume (2), Issue (3) (JUNE, 2026)
ISSN: 3135-2359

مجلة الفاروق للعلوم
المجلد (2) العدد 3 يونيو 2026
ردمد: 3135-2359

Design and Implementation of a Real-Time Speech-to-Text and Translation System Using Speech Recognition and Natural Language Processing Techniques in Python.

AHED ALI ABOALKASEM
Department of Computer Science
University of Sabratha
ahedkr3@gmail.com
GHADA ABDULMUMEM MOHAMMED
Department of Computer Science
University of Sabratha
ghadashowsh@gmail.com
HUDI AMHIMMID BASHEER ATEEYAH
Department of Computer Science
University of Sabratha
huda.eatia@sabu.edu.ly
JAZIYAH RAMAD HAN ALI
Department of Computer Science
University of Sabratha
jazia.zreg@sabu.edu.ly

تاريخ الاستلام: 2026/04/20 تاريخ المراجعة 05/20/ 2026 تاريخ القبول:2026/06/07- تاريخ النشر: 22/06/ 2026

Abstract

This study aims to design and implement a real-time speech-to-text and translation application using speech recognition and natural language processing techniques in Python. The system is developed to provide an integrated solution that converts spoken language into text and optionally translates the recognized content between Arabic and English through a simple and interactive graphical user interface. The application combines speech recognition, machine translation, audio processing, and multithreading techniques to ensure efficient and responsive performance.

The system was implemented using Python and several supporting libraries, including SpeechRecognition for speech-to-text conversion, Deep Translator for machine translation, SoundDevice for audio acquisition, and Tkinter for graphical user interface development. The application allows users to select the input language, record speech, display recognized text, perform real-time translation, and automatically save recognized content with timestamps for future reference.

The results demonstrate that the proposed application can successfully recognize spoken input and generate translated output in real time while maintaining an accessible and user-friendly environment. The integration of speech recognition and translation technologies provides a practical solution for bilingual communication, language learning, and assistive applications.

The findings indicate that combining modern speech recognition and neural machine translation technologies within a single lightweight application can improve usability and accessibility while demonstrating the practical benefits of recent advances in natural language processing and artificial intelligence.

Keywords: Speech Recognition, Natural Language Processing, Machine Translation, Speech-to-Text, Real-Time Translation, Python, Neural Machine Translation.

Introduction

Natural Language Processing (NLP) and speech technologies have experienced significant advancements in recent years due to the rapid development of artificial intelligence and deep learning techniques. These technologies have become widely used in various applications such as machine translation, speech recognition, virtual assistants, and multilingual communication systems. Their ability to process human language more accurately and efficiently has contributed to improving interaction between humans and computers and expanding the practical use of intelligent language systems in everyday life (Bahdanau et al., 2015; Vaswani et al., 2017).

Speech recognition has become one of the most important areas of NLP, as it enables computers to convert spoken language into written text automatically. This technology is increasingly used in education, communication, accessibility services, and productivity applications. However, achieving accurate speech recognition remains challenging due to factors such as background noise, variations in pronunciation, speaking speed, and differences in accents. These challenges become more significant when speech recognition is integrated with machine translation systems that require accurate textual input to produce reliable translations.

Machine translation has also undergone major transformations over the past decade. Traditional translation systems relied on statistical methods that often struggled to preserve contextual meaning and sentence fluency. The introduction of Neural Machine Translation (NMT) provided a more effective approach by employing deep neural networks to learn translation patterns directly from large datasets. Encoder–decoder architectures became the foundation of these systems, allowing models to generate more coherent and context-aware translations than previous statistical approaches (Bahdanau et al., 2015).

Despite the success of early neural translation models, researchers identified limitations related to representing long and complex sentences using fixed-length vectors. To address this issue, Bahdanau et al. (2015) introduced the attention mechanism, which allows translation models to focus on the most relevant parts of the input sequence during translation. Later, Vaswani et al. (2017) further advanced this concept through the Transformer architecture, which relies entirely on attention mechanisms and has become the foundation of many modern language processing systems.

Although speech recognition and machine translation technologies have achieved remarkable progress independently, integrating them into a single real-time application remains a practical challenge. Such systems must efficiently capture audio input, convert speech into text, process the recognized content, generate translations, and present results to users without noticeable delays. In addition, user-friendly interfaces and efficient processing techniques are required to ensure smooth interaction and usability.

This study aims to design and implement a real-time speech-to-text and translation application using Python and modern NLP technologies. The proposed system integrates speech recognition, machine translation, audio processing, and graphical user interface components within a single platform. The objective is to provide a practical and accessible solution that supports bilingual communication between Arabic and English while demonstrating the application of recent advances in speech recognition and neural machine translation technologies.

Literature Review

Several previous studies have contributed to the development of speech recognition, machine translation, and natural language processing technologies. These studies have provided the theoretical and practical foundations for modern language systems that can process spoken and written language with increasing levels of accuracy and efficiency.

In 2015, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio introduced a neural machine translation model based on a novel attention mechanism. The study aimed to overcome the limitations of traditional encoder–decoder architectures that relied on compressing an entire input sentence into a single fixed-length vector. The researchers proposed a method that allows the model to focus on relevant parts of the source sentence while generating each target word. The results demonstrated significant improvements in translation quality, particularly for long and complex sentences. The study concluded that the attention mechanism enables more effective context representation and substantially improves neural machine translation performance (Bahdanau et al., 2015).

Similarly, Ashish Vaswani and his colleagues presented the Transformer architecture in 2017. The study introduced a new neural network architecture based entirely on self-attention mechanisms without relying on recurrent neural networks. The researchers demonstrated that the Transformer achieved superior performance in machine translation tasks while reducing training time and improving scalability. The findings showed that self-attention mechanisms are highly effective in capturing long-range dependencies within text sequences, making the Transformer a foundation for many modern natural language processing systems (Vaswani et al., 2017).

In addition, recent advances in deep learning have played a significant role in improving both speech recognition and machine translation systems. Deep learning models have enabled computers to learn complex language patterns directly from large datasets, resulting in more accurate recognition and translation capabilities. These developments have contributed to the widespread adoption of intelligent language technologies in communication, education, and multilingual applications.

Although previous studies have made substantial contributions to neural machine translation and language processing, most of them focused on developing theoretical models and improving translation accuracy. Limited attention has been given to the practical integration of speech recognition and machine translation technologies within a single real-time application that can be easily used by end users.

Accordingly, this study seeks to bridge this gap by designing and implementing a real-time speech-to-text and translation application that combines speech recognition, machine translation, audio processing, and user interface components within a unified system. The proposed application demonstrates how modern natural language processing technologies can be integrated into a practical and accessible solution for bilingual communication.

Significance of the Study

The primary objective of this study is to demonstrate the practical application of modern speech recognition and machine translation technologies within a single integrated system. The study contributes to the field of natural language processing by combining multiple technologies into a real-time application that supports bilingual communication between Arabic and English. It also provides a simple and user-friendly platform that can be utilized in educational, assistive, and communication environments. Furthermore, the study highlights how recent advances in neural machine translation and speech processing can be transformed into practical software solutions that improve accessibility and facilitate real-time language interaction.

Objectives of the Study

This study aims to design and implement a real-time application that combines speech recognition and machine translation technologies within a single interactive system. The primary goal is to facilitate communication by converting spoken language into written text and

providing optional translation between Arabic and English. The study seeks to demonstrate how recent advances in Natural Language Processing (NLP) can be applied in a practical software solution that is simple, efficient, and accessible to users.

In addition, the study aims to develop a graphical user interface that enables users to interact with the system easily while maintaining real-time responsiveness during speech processing and translation tasks. The proposed application also incorporates multithreading techniques to ensure smooth operation without interrupting the user interface during audio recording and recognition processes.

Furthermore, the study seeks to integrate speech recognition, machine translation, audio processing, and data logging functionalities into a unified environment. By doing so, it provides a practical example of how existing NLP technologies can be combined to create useful tools for communication, language learning, and accessibility support.

Ultimately, the study aims to bridge the gap between theoretical research in speech recognition and machine translation and their practical implementation in real-world applications, highlighting the potential of these technologies to support multilingual communication in an efficient and user-friendly manner.

Theoretical Framework

This study is based on a set of theoretical concepts related to Natural Language Processing (NLP), Automatic Speech Recognition (ASR), and Machine Translation (MT), which together form the foundation of modern language technologies. These fields aim to enable computers to process human language, convert spoken speech into text, and translate information between different languages automatically. Advances in deep learning have significantly improved the performance of these technologies, making them increasingly suitable for real-time applications (Jurafsky & Martin, 2023; Goodfellow et al., 2016).

Within the field of machine translation, Neural Machine Translation (NMT) has emerged as a powerful approach that replaces traditional rule-based and statistical methods with end-to-end neural networks. Early encoder–decoder architectures demonstrated promising results; however, their performance was limited when processing long and complex sentences. To overcome this limitation, Bahdanau et al. (2015) introduced the attention mechanism, which enables the model to focus on relevant parts of the input sequence during translation, leading to substantial improvements in translation quality and contextual understanding.

Further progress was achieved with the introduction of the Transformer architecture by Vaswani et al. (2017). Unlike previous sequential models, the Transformer relies entirely on self-attention mechanisms to process linguistic information, allowing more efficient training and better handling of long-range dependencies. This architecture has become the foundation of many modern language processing systems and has significantly influenced the development of contemporary translation and language models.

In parallel, Automatic Speech Recognition technologies have evolved to convert spoken language into text with increasing accuracy. These systems combine acoustic processing and language modeling techniques to interpret human speech and generate textual representations. When integrated with machine translation systems, speech recognition technologies enable real-time multilingual communication by transforming speech into translated text within a single workflow (Jurafsky & Martin, 2023).

Based on these theoretical foundations, the proposed system integrates speech recognition and machine translation technologies into a unified application. By combining audio processing, speech-to-text conversion, translation services, and a graphical user interface, the system demonstrates how modern NLP techniques can be applied to create practical tools that support communication, language learning, and accessibility in bilingual environments.

Methodology

This study adopts a design and implementation methodology to develop a real-time speech-to-text and translation application based on speech recognition and natural language processing technologies. The methodology focuses on analyzing system requirements, integrating software components, implementing the application, and evaluating its functionality in a practical environment. The objective is to demonstrate how speech recognition and machine translation technologies can be combined into a single interactive system that supports real-time communication.

The development process began with identifying the core functionalities required by the application, including audio acquisition, speech recognition, language translation, text display, and data storage. Appropriate software libraries and tools were then selected to support these functions, including SpeechRecognition for speech-to-text conversion, Deep Translator for machine translation, SoundDevice for audio recording, and Tkinter for graphical user interface development.

The proposed system was implemented using the Python programming language due to its extensive support for natural language processing and rapid application development. The application was designed to support both Arabic and English speech input, allowing users to convert spoken language into text and optionally translate the recognized content between the two languages. Multithreading techniques were incorporated to ensure that audio processing operations run independently from the graphical user interface, thereby maintaining system responsiveness during real-time operation.

Data Collection

The data used in this study were obtained through practical testing of the developed application. Various speech samples in both Arabic and English were used to evaluate the system's ability to recognize spoken input and generate translated output. The testing process focused on verifying the functionality of the speech recognition module, the translation component, and the interaction between different system modules within a real-time environment.

Data Analysis

The system was evaluated using descriptive analysis based on the observed performance of the implemented application. The analysis focused on the accuracy of speech recognition, the effectiveness of translation output, the responsiveness of the graphical user interface, and the overall integration of the system components. In addition, the behavior of the application during real-time processing was examined to assess its suitability for communication and language-support tasks.

The findings were analyzed by comparing the expected functionality of the system with the actual outputs generated during testing. This evaluation provided insights into the strengths of the proposed implementation as well as the limitations associated with speech recognition accuracy, translation quality, and dependence on external online services.

Ethical Considerations

This study was conducted in accordance with ethical principles related to scientific research and software development. The proposed application was designed to process speech data solely for recognition and translation purposes without collecting sensitive personal information from users. During the implementation and testing phases, all generated data were used exclusively for evaluating system functionality and performance.

Particular attention was given to data privacy and security because speech input may contain personal information depending on the content provided by users. To address this concern, the application stores recognized text locally on the user's device rather than in external databases. This approach helps reduce privacy risks and provides users with greater control over their information.

The study also ensured transparency in the operation of the system by clearly defining the functions of speech recognition, machine translation, and data storage components. Users are

informed that recognized speech may be converted into text, translated, and optionally saved for future reference. No hidden data collection or background monitoring mechanisms were incorporated into the application.

Furthermore, the development process relied on publicly available software libraries and services, including speech recognition and translation tools. These technologies were used exclusively for educational and research purposes while respecting their licensing conditions and intended usage policies. The implementation did not involve modifying third-party services or using them in a manner that could compromise their integrity or violate their terms of use.

Since the study focused on the design and implementation of a software application rather than human-subject experimentation, no personal demographic information was collected and no participants were exposed to physical, psychological, or social risks. Consequently, the ethical considerations of this work primarily focused on privacy protection, responsible technology use, transparency, and secure handling of speech data throughout the development and evaluation stages.

Results

The results obtained from the implementation and testing of the proposed system demonstrated the successful integration of speech recognition and machine translation technologies within a single application. The system was able to capture audio input, convert spoken language into written text, and provide translation between Arabic and English in real time.

The findings showed that the speech recognition component performed effectively in recognizing spoken input under normal operating conditions. The application successfully processed speech in both supported languages and displayed the recognized text through the graphical user interface. Recognition performance was generally improved when speech was clear and environmental noise was minimal.

The results also indicated that the translation module was capable of generating translated text within a short processing time, enabling users to receive translated output immediately after speech recognition. This functionality enhanced the system's ability to support bilingual communication and language accessibility.

Furthermore, the integration of different system components, including audio recording, speech recognition, machine translation, text display, and data storage, was achieved successfully. The interaction between these components was smooth and allowed the application to perform multiple tasks within a unified environment.

The findings additionally revealed that the implementation of multithreading contributed to maintaining application responsiveness during audio processing operations. The graphical user interface remained functional while speech recognition and translation tasks were executed in the background, resulting in a more efficient user experience.

Finally, the system successfully stored recognized text together with timestamps in an external file, providing a simple mechanism for preserving speech records and supporting future review and analysis. Overall, the results demonstrate the feasibility of combining speech recognition and machine translation technologies into a lightweight and practical real-time application.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. Proceedings of the International Conference on Learning Representations (ICLR 2015). arXiv:1409.0473.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Cambridge, MA: MIT Press.

Latif, M. B., & Saeid, S. A. M. (2026). The Impact of Teaching English Literature as an Elective Course to Fourth-Year ESL Students in a University Context. *Al-Farooq Journal of Sciences*, *2*(3), 1166-1179.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, 29(6), 82–97.

Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Stanford University.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI Technical Report.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30, 5998–6008.