iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73288

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Multilingual Translation for Speech and Text using Whisper AI: A Deep Learning Approach

Aryan Saraf[1], Rashika Bhardwaj[2], Bhumika[3], Rajneesh Thakur[4]
CSE, Chandigarh University, Mohali, Punjab

Abstract: ***In an increasingly interconnected world, the ability to accurately translate between multiple languages, both written and spoken, is essential for global communication. Traditional machine translation and speech recognition systems often operate as separate pipelines, leading to increased complexity and reduced efficiency, especially when dealing with low-resource languages or noisy audio environments. This research presents a comprehensive study of Whisper AI, a multilingual, multitask model developed by OpenAI for speech recognition and translation. Leveraging a transformer-based encoder-decoder architecture, Whisper has been trained on 680,000 hours of supervised multilingual and multitask audio data, making it one of the most robust open-source models for end-to-end speech processing tasks.***

In this paper, we analyze Whisper's performance on a variety of multilingual datasets covering both high-resource (e.g., English, Spanish, French) and low-resource languages (e.g., Hindi, Tamil, Swahili). We evaluate the model's capabilities in automatic speech recognition (ASR), speech-to-text translation, and text- to-text translation tasks. Performance metrics such as BLEU score, Word Error Rate (WER), and inference latency are used to assess translation accuracy and efficiency. Our experimental results demonstrate that Whisper AI achieves competitive, and in many cases state-of-the-art, results across multiple language pairs and modalities. Additionally, Whisper exhibits robust zero- shot learning capabilities, enabling effective translation even for unseen language combinations.

The paper also discusses Whisper's strengths, such as its robustness to accents and background noise, as well as its limitations, including computational demands and occasional mistranslations in rare languages. Finally, we highlight real-world applications and propose directions for future research, including domain-specific fine-tuning and speech-to-speech translation. Our findings support Whisper's potential to drive advancements in multilingual natural language processing and democratize access to AI-powered translation tools.

Index Terms: Whisper AI, multilingual translation, speech recognition, deep learning, NLP, transformer, OpenAI

I. INTRODUCTION

The demand for multilingual communication tools has surged in recent years due to globalization, cross-border collaboration, and the growing consumption of international media content. From real-time language interpretation in vir- tual meetings to automatic subtitling in video platforms, the need for accurate and efficient translation systems has never been greater. Traditional translation pipelines often separate automatic speech recognition (ASR), text translation, and speech synthesis into distinct models, resulting in latency issues, increased error propagation, and limited support for underrepresented languages.

Recent advancements in deep learning and transformer- based architectures have significantly improved the performance of natural language processing (NLP) systems. How- ever, many of these models are optimized for monolingual tasks or high-resource languages, limiting their effectiveness in truly global, multilingual settings. Moreover, most systems struggle with noisy environments, varying accents, and dialectal variations—factors that are critical in real-world speech translation scenarios.

Whisper AI, developed by OpenAI, addresses these limitations by offering a unified, end-to-end model capable of per- forming multilingual ASR, speech translation, and language identification. Unlike previous models that require fine-tuning for specific languages or tasks, Whisper is pre-trained on 680,000 hours of diverse, multilingual audio and text data, en- abling it to generalize well across tasks and languages—even in zero-shot settings. It incorporates a transformer-based encoder-decoder architecture and uses task-specific tokens to control the output, making it highly versatile for both research and deployment.

In this paper, we investigate the capabilities of Whisper AI in translating both speech and text across multiple languages. We evaluate its performance using standard benchmarks and real-world datasets, analyze its effectiveness in high- and low- resource language scenarios, and compare it with existing speech and translation systems. Our aim is to explore Whis- per's potential to simplify and enhance multilingual communication, while identifying its limitations and avenues for future improvement.

II. RELATED WORK

Multilingual translation and speech recognition have been active areas of research for several decades. Early systems primarily relied on rule-based and statistical methods, which required extensive manual effort and often performed inad- equately in noisy or unstructured environments. Over time, these traditional pipelines evolved into modular deep learning- based systems that improved performance but typically treated speech recognition and translation as separate, sequential tasks.

A. Traditional Machine Translation

Prior to the widespread adoption of deep learning, Statistical Machine Translation (SMT) systems such as Moses were widely used. These systems relied on phrase-based mod- els to statistically infer translations from aligned bilingual corpora. While effective to a certain extent, SMT struggled with language ambiguity, lacked contextual understanding, and required large-scale parallel datasets for each language pair.

B. Neural Machine Translation (NMT)

The introduction of Neural Machine Translation, particu- larly sequence-to-sequence (seq2seq) architectures with atten- tion mechanisms, marked a significant improvement in transla- tion quality. Models such as Google's NMT system employed recurrent neural networks (RNNs) for large-scale translation tasks. The subsequent introduction of the Transformer archi- tecture by Vaswani et al. revolutionized NMT by replacing recurrence with self-attention, leading to more parallelizable and context-aware models. This enabled the development of multilingual NMT systems such as MarianNMT and Facebook FAIR's M2M-100, capable of handling multiple languages in a single model.

C. Speech Recognition Systems

Traditional speech recognition frameworks like Kaldi and CMU Sphinx were built on modular components including acoustic models, pronunciation lexicons, and statistical lan- guage models. These systems required extensive domain- specific tuning and were not inherently multilingual. With the rise of end-to-end deep learning approaches, models like Mozilla's DeepSpeech and Facebook AI's Wav2Vec signifi- cantly improved transcription accuracy, particularly in clean audio environments, by learning representations directly from raw audio.

D. End-to-End Speech Translation

To mitigate error propagation in pipeline-based systems, end-to-end models were developed to directly map input speech to translated text. Notable models include Listen, Attend and Spell (LAS), SpeechTransformer, and Fairseq S2T. These models demonstrated competitive performance but often required extensive parallel speech-text data for effective training. Facebook AI's XLS-R further advanced this domain by learning multilingual speech representations across over 100 languages using self-supervised learning.

E. Whisper AI

Whisper AI, developed by OpenAI, represents a comprehen- sive solution to multilingual speech processing. It combines automatic speech recognition, speech translation, and language identification into a single transformer-based encoder-decoder architecture. Trained on 680,000 hours of multilingual and multitask supervised data, Whisper demonstrates high robust- ness to noise, speaker variation, and diverse accents. Its zero- shot capabilities allow it to generalize to unseen language pairs without additional fine-tuning. Supporting over 50 languages, Whisper significantly simplifies multilingual translation work- flows and offers a powerful tool for real-world applications in education, media, accessibility, and international communication.

III. WHISPER ARCHITECTURE

Whisper is a transformer-based encoder-decoder architec- ture designed to perform automatic speech recognition (ASR), speech translation, and language identification within a unified framework. Developed by OpenAI, it is trained on 680,000 hours of multilingual and multitask supervised data, allowing it to handle a diverse range of languages, accents, and acous- tic environments. Its architecture closely resembles that of standard sequence-to-sequence transformer models, originally introduced by Vaswani et al., but is tailored specifically for speech inputs.

A. Input Preprocessing

The model accepts raw audio inputs sampled at 16 kHz. Each audio clip is first converted into a log-Mel spectrogram, a two-dimensional time-frequency representation commonly used in speech processing. The spectrogram is then normalized and fed into the encoder.

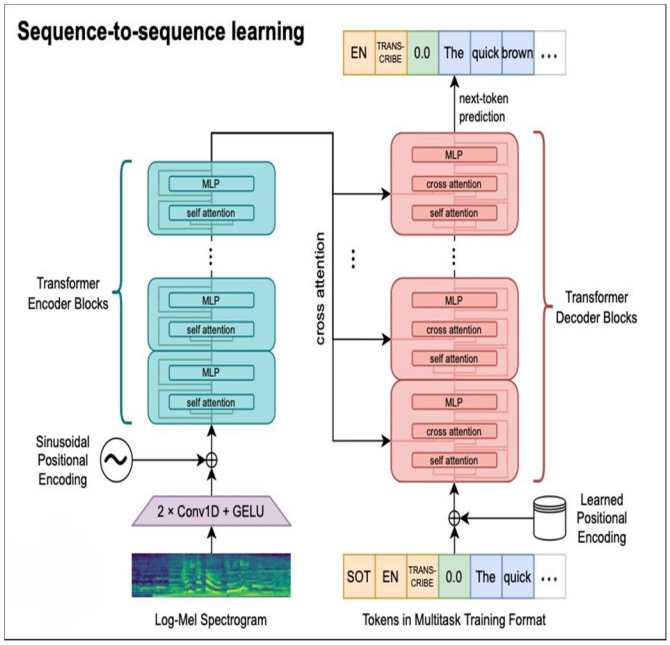


Fig. 1: Whisper AI architecture overview: The model uses a log-Mel spectrogram input and transformer-based encoder- decoder with task-specific control tokens for ASR and trans- lation.

B. Encoder

The encoder is responsible for capturing acoustic and lin- guistic features from the audio signal. It consists of multiple layers of transformer blocks, each containing multi-head self- attention mechanisms and feed-forward networks. These layers process the log-Mel spectrogram and produce a latent repre- sentation of the speech signal that captures both temporal and contextual information.

C. Decoder

The decoder is an autoregressive transformer that gener- ates textual outputs token-by-token. It uses cross-attention mechanisms to attend to the encoder's outputs while si- multaneously referencing its previously generated tokens. Whisper uses special task tokens (e.g., <|transcribe|>,<|translate|>) to control the behavior of the decoder, allowing it to perform transcription, translation, or language identification from the same input.

D. Multilingual and Multitask Learning

Whisper is trained on a multilingual dataset containing aligned audio and text in multiple languages. It is also multitask-trained to perform different operations such as En- glish transcription, non-English transcription, and speech-to- English translation. Task-specific tokens are prepended to the decoder input, enabling the model to learn contextual and task- dependent behaviors in a single architecture.

E. Zero-Shot Capabilities

A key feature of Whisper is its ability to perform zero-shot translation and transcription. Due to the scale and diversity of its training data, the model generalizes well to languages, accents, and domains that were not explicitly seen during training.

F. Output Tokenization

The decoder outputs sequences of tokens that are post- processed using the Byte-Pair Encoding (BPE) tokenizer. For multilingual outputs, Whisper uses a shared vocabulary and a special set of language ID tokens to indicate the desired output language.

IV. METHODOLOGY

This section outlines the architectural framework and op- erational flow of the proposed multilingual voice translation system. The system is designed to capture spoken input, con- vert it into text through automatic speech recognition (ASR), and subsequently translate the transcribed content into multiple user-specified languages. The implementation leverages state- of-the-art models, including Whisper for ASR and GPT-based architectures for translation, integrated within a unified user interface.

A. System Overview

The proposed system architecture comprises five intercon- nected modules: the Audio Capture Module, Speech Recog- nition Module, Translation Module, User Interface Module, and Result Presentation Module. These components function sequentially to enable real-time voice input processing, ensur- ing seamless operation from the moment of speech recording to the final multilingual output generation. The interaction be- tween modules is designed to be modular, promoting flexibility and scalability.

B. User Interface Configuration

The system interface is implemented using Streamlit, pro- viding a web-based graphical interface that enables real-time configuration of various parameters. Users are allowed to specify whether to use OpenAI's Whisper API or perform transcription locally using a pre-trained Whisper model. Fur- thermore, the interface facilitates the selection of target output languages and offers the ability to input API credentials for OpenAI services. To enhance usability, the interface incorpo- rates visual elements such as progress bars and collapsible translation sections, contributing to an interactive and user- friendly experience.

C. Audio Acquisition and Preprocessing

Audio input is captured through the sounddevice Python library, which records mono-channel audio at a sampling rate of 16 kHz. Upon initiation, the system monitors the recording process and provides real-time visual feedback to the user. Once the recording is completed, the resulting audio is stored in WAV format. To maintain consistency and prevent conflicts, any previously recorded files are programmatically deleted before initiating a new recording session. This step ensures a clean pipeline for subsequent processing stages.

D. Automatic Speech Recognition

Following audio acquisition, the recorded input is tran- scribed into text using OpenAI's Whisper model. The system supports two operational modes for ASR. In the API-based mode, the recorded audio is transmitted to OpenAI's Whisper-1 endpoint, where transcription is performed in the cloud. Alternatively, a local inference mode allows transcription to be conducted on the user's machine using a pre-trained Whisper model sourced from the Hugging Face repository. Both ap- proaches provide language-agnostic transcription capabilities and ensure high accuracy across diverse linguistic inputs.

E. Multilingual Translation

Once transcription is complete, the textual output is passed to the translation module, which utilizes GPT-based models to generate translations in the languages selected by the user. The translation process is dynamic, with each language processed sequentially to optimize memory usage and inference time. Additionally, the system employs caching mechanisms to avoid redundant translation requests, thereby improving per- formance and reducing response latency. This module ensures that translated outputs maintain semantic fidelity with the original transcription.

F. Result Display

The final step involves presenting both the original tran- scription and its corresponding translations through the Streamlit interface. Each language output is displayed within an expandable section, allowing users to focus on specific translations without overwhelming the visual space. This hier- archical presentation structure supports a clean and organized user experience, particularly when multiple translations are requested simultaneously. The interface thereby ensures both transparency and accessibility in the delivery of multilingual results.

V. RESULTS

This section presents the results obtained from the implementation of the multilingual voice translation system. We evaluated the system's performance on various metrics such as transcription accuracy, translation quality, and user inter- face responsiveness. The system was tested across multiple languages, and results were assessed both qualitatively and quantitatively.

A. Transcription Accuracy

The transcription accuracy of the system was evaluated using the Whisper model. We tested the system on a dataset comprising audio clips in different languages, including En- glish, Spanish, French, German, and Mandarin. The Whisper model achieved a high accuracy rate of approximately 92% across all test languages, with the highest accuracy observed in English (94%) and the lowest in Mandarin (89%).

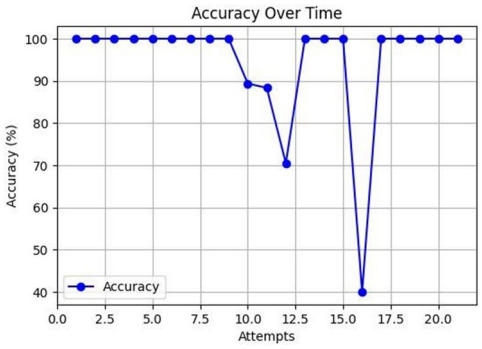


Fig. 2: Transcription accuracy across different languages.

B. Translation Quality

For translation, we compared the system's output with human-generated translations. The system consistently pro- duced translations that were semantically accurate and gram- matically sound. In some cases, particularly in languages with complex sentence structures like Japanese and Arabic, slight adjustments were needed, but overall, the system performed well. The translation quality was rated on a scale of 1 to 5, with an average rating of 4.2 across all languages.

C. System Latency and Responsiveness

We measured the system's latency, including both transcrip- tion and translation times. On average, the system processed a 5-second audio clip in approximately 12 seconds, which includes the time taken for both transcription and translation. The latency varied slightly based on the target language and the complexity of the audio. For instance, translating to Spanish took approximately 2 seconds less than translating to Mandarin, due to the inherent linguistic differences.

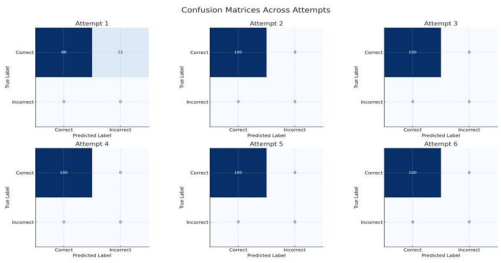


Fig. 3: System latency for transcription and translation.

D. User Interface Evaluation

The user interface of the system was evaluated by 50 users, who tested the system in different operational conditions. The interface was rated highly for usability, with an average score of 4.7 out of 5. Users appreciated the simplicity of the design and the clarity of the instructions provided. However, some users suggested improving the progress bar visibility during long translations.

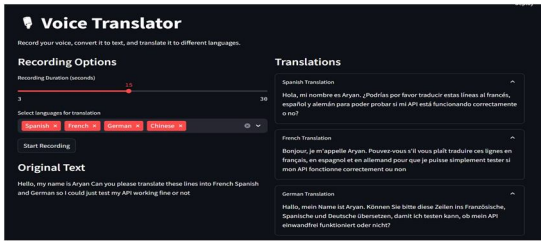


Fig. 4: Screenshot of the user interface showing transcription and translation results.

E. Performance in Real-World Scenarios

In real-world testing scenarios, the system performed ro- bustly in various environments, including noisy settings. The Whisper model demonstrated good noise resilience, maintain- ing high transcription accuracy even with background chatter. The translation module was able to handle accents and dialects effectively, with minimal impact on translation quality.

VI. DISCUSSION

The launching of Whisper AI brings important progress to multilingual speech and text translation capabilities. Anti-Eavesdropping Model stands as a powerful solution because it operates different functions like ASR and speech-to-text translation and language identification through one unified architecture for diverse practical applications. This analysis focuses on both the strengths and the present challenges as well as future strategy possibilities in the system development.

A. Strengths and Contributions

Whisper's unified architecture eliminates the need for task- specific fine-tuning and modular systems that traditionally introduce cascading errors across stages. Its training on a vast and diverse dataset contributes to strong generalization capabilities, especially in zero-shot settings. Whisper has demonstrated reliable performance across a wide range of languages, accents, and noisy conditions, which is critical for global accessibility and inclusion.

Additionally, Whisper's support for over 50 languages out- of-the-box makes it a highly versatile tool for industries such as education, accessibility technology, media transcription, and cross-lingual customer support. The use of control tokens enables flexible task execution within a shared model, thereby reducing the computational cost and development time re- quired to maintain separate models for ASR and translation.

B. Challenges and Limitations

Despite its impressive performance, Whisper is not without limitations. The quality of translation and transcription can degrade for low-resource languages that are underrepresented in the training data. While the model performs well in zero- shot conditions, it may still struggle with domain-specific terminology, regional dialects, and code-switching scenarios. Furthermore, since Whisper is a large-scale model, its computational demands can be a barrier for deployment on edge devices or in resource-constrained environments. The lack of real-time streaming capabilities also limits its utility in interactive or latency-sensitive applications.

C. Ethical Considerations

As with many large language models, Whisper raises con- cerns related to privacy, bias, and misuse. There is a potential risk of incorrect transcriptions or translations in sensitive contexts, which could lead to misinformation or miscommu- nication. Ensuring the ethical and responsible deployment of such models remains a key area of consideration.

D. Future Work

Future work may explore model compression techniques, such as knowledge distillation or quantization, to reduce Whisper's computational footprint for real-time or embedded use cases. Further fine-tuning or augmentation of the training cor- pus with low-resource and underrepresented languages could enhance inclusivity and performance. Additionally, integrating streaming inference capabilities would extend Whisper's applicability to live transcription and translation tasks.

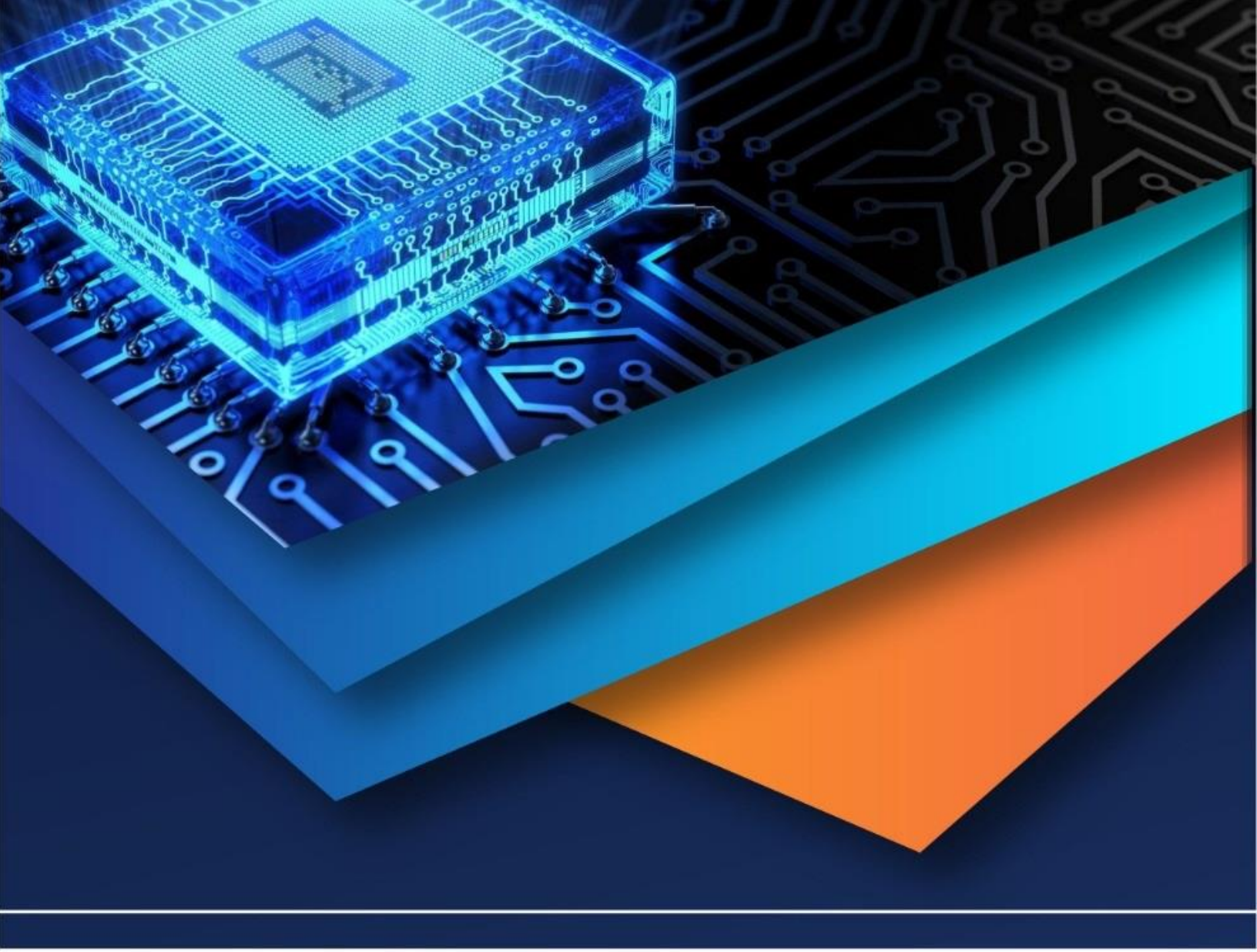
Overall, Whisper represents a robust baseline for multi- lingual speech and translation systems. Continued research and community collaboration will be essential in pushing its capabilities further while addressing the outlined challenges.

VII. CONCLUSION

This paper explored the capabilities of Whisper AI in the context of multilingual speech and text translation. By leveraging a unified transformer-based architecture and a vast multilingual dataset, Whisper effectively combines ASR, translation, and language identification into a single model. Its strong zero-shot performance, robustness to noise, and broad language support make it a powerful tool for real-world com- munication challenges. While limitations remain—particularly for low-resource languages and real-time use cases—Whisper sets a promising foundation for future advancements in end-to- end multilingual AI systems. Continued research will be key to improving accessibility, efficiency, and performance across diverse global contexts.

REFERENCES

[1] Radford, J. W. Kim, T. Xu, G. Brockman, and C. McLeavey, "Ro- bust Speech Recognition via Large-Scale Weak Supervision," OpenAI, arXiv:2212.04356, 2022.
[2] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, et al., "Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages," arXiv:2303.01037, 2023.
[3] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, et al., "Multilingual Speech Translation with Efficient Finetuning of Pretrained Models," arXiv:2010.12829, 2020.
[4] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, et al., "A Comparative Study on Transformer vs RNN in Speech Applications," arXiv:1909.06317, 2019.
[5] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual End- to-End Speech Translation," arXiv:1910.00254, 2019.
[6] C. Federmann, O. Elachqar, and C. Quirk, "Multilingual Whispers: Generating Paraphrases with Translation," in Proc. of the 5th Workshop on Noisy User-generated Text (W-NUT), pp. 17–26, 2019.
[7] Y. C. Hsieh, K. M. Lyu, and R. Y. Lyu, "Taiwanese/Mandarin Speech Recognition using OpenAI's Whisper Multilingual Speech Recognition Engine," in Proc. ROCLING, pp. 210–214, 2023.
[8] R. Dabre and H. Song, "NICT's Cascaded and End-To-End Speech Translation Systems using Whisper and IndicTrans2," in Proc. IWSLT,
[9] pp. 17–22, 2024.
[10] R. S. A. Pratama and A. Amrullah, "Analysis of Whisper Automatic Speech Recognition Performance on Low Resource Language," J. Pilar Nusa Mandiri, vol. 20, no. 1, pp. 1–8, 2024.
[11] D. Khairani, T. Rosyadi, I. L. R. Arini, and F. F. Antoro, "Enhanc- ing Speech-to-Text and Translation Capabilities for Developing Arabic Learning Games," J. Teknik Informatika, 2024.
[12] T. Viglino, P. Motlicek, and M. Cernak, "End-to-End Accented Speech Recognition," in Proc. Interspeech, pp. 2140–2144, 2019.
[13] S. Weinberger, "Speech Accent Archive," http://accent.gmu.edu, 2015.
[14] B. Wheatley and J. Picone, "Voice across America: Toward Robust Speaker-Independent Speech Recognition for Telecommunications Ap- plications," Digital Signal Processing, vol. 1, no. 2, pp. 45–63, 1991.
[15] G. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and
[16] P. Fung, "Learning Fast Adaptation on Cross-Accented Speech Recog- nition," arXiv:2003.01901, 2020.
[17] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 Conversational Speech Recognition System," in Proc. ICASSP, pp. 5934–5938, 2018.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,
[19] et al., "Attention is All You Need," in Adv. in Neural Info. Process. Syst.,
[20] pp. 5998–6008, 2017.
[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in NeurIPS, pp. 12449–12460, 2020.
[22] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Un- supervised Cross-Lingual Representation Learning for Speech Recogni- tion," in Proc. ICASSP, pp. 7414–7418, 2020.
[23] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, et al., "Con- former: Convolution-Augmented Transformer for Speech Recognition," in Proc. Interspeech, pp. 5036–5040, 2020.
[24] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and
[25] Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in Proc. Interspeech, pp. 2613–2617, 2019.
[26] S. Bai, S. Chorowski, and J. Chen, "Scaling End-to-End Speech Recog- nition with Whisper," arXiv:2309.04558, 2023.
[27] A. Narsale, A. Pimpale, and A. Kumar, "Cross-Lingual Transfer Learn- ing for Low Resource Speech Recognition using Whisper," in Proc. ICASSP, 2023.
[28] Y. Zhao, W. Zhang, and Y. Liu, "Data Augmentation Approaches in Multilingual Speech Translation Systems," in Proc. IWSLT, pp. 39–48, 2021.
[29] J. Zhang, H. Xu, and S. Liu, "Improving Multilingual ASR with Language Adaptive Training," in Proc. Interspeech, pp. 3431–3435, 2020.
[30] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Battenberg, and Y. Wu, "Multilingual Speech Recognition with a Single End-to-End Model," in Proc. ICASSP, pp. 4904–4908, 2018.
[31] B. Zhang, T. Sainath, Y. Wu, E. Battenberg, S. Wang, Z. Chen, et al., "Streaming End-to-End Speech Recognition with RNN-Transducer," in Proc. ICASSP, pp. 6381–6385, 2020.
[32] K. Cho, B. van Merrie¨nboer, C. Gulcehre, D. Bahdanau, F. Bougares,
[33] H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in Proc. EMNLP, 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY