iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VIII **Month of publication:** August 2024

DOI: https://doi.org/10.22214/ijraset.2024.63880

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 12 Issue VIII Aug 2024- Available at www.ijraset.com

Conversion of Speech to Text: Transcribe and Translate Using Whisper AI on iOS

Dr. Shilpa V[1], Chandan Kumar[2], Girija V[3]
Assistant Professor, Dept of CSE, Cambridge Institute of Technology
Dwivedi, CSE, Cambridge Institute of Technology
Assistane Professor, Dept of CSE, Cambridge Institute of Technology

Abstract: ***This thesis investigates the implementation and efficiency of Whisper AI for transcribing and translating speech to text on iOS devices. Leveraging a large-scale weakly supervised dataset, Whisper AI demonstrates robust performance across multiple languages and tasks. The study explores its architecture, implementation on iOS, and performance comparisons with existing models. Findings indicate significant potential for real-world applications, despite some computational and accuracy challenges.***
Keywords*: Speech recognition, Whisper AI, iOS, transcribe, translate, weak supervision, large-scale datasets.*

I. INTRODUCTION

The rapid advancement in speech recognition technology has significantly impacted various fields, including accessibility, communication, and automation. This thesis focuses on the application of Whisper AI for speech-to-text conversion and translation on iOS devices. By leveraging a robust dataset and state-of-the-art machine learning techniques, Whisper AI aims to provide high accuracy and efficiency in real-world scenarios.

A. Background

Speech recognition technology has become progressively necessarily to different applications, extending from virtual associates to translation administrations. The development of robust speech-to-text systems is crucial for enhancing accessibility, improving human-computer interaction, and supporting multilingual communication. Traditional speech recognition systems requires large amounts of labeled data and extensive fine-tuning, which can be time consuming and resource-intensive.

Whisper AI represents a significant progress in speech recognition technology. By leveraging a large-scale weakly supervised dataset comprising 680,000 hours of multilingual and multitask audio data, Whisper AI aims to provide robust speech-to-text and translation capabilities without the need for extensive fine-tuning. This model's encoder-decoder transformer architecture allows it to generalize effectively across various tasks and languages, making it a versatile tool for real-world applications.

B. Problem Statement

Despite the good amount of progress in speech recognition technologies, many challenges remain. Traditional models often require vast amounts of labeled data, which can be costly and laborious to obtain. Moreover, existing systems may struggle with generalization, particularly in noisy environments or with diverse accents and languages. Deploying these models on mobile platforms like iOS presents additional challenges related to computational efficiency and resource constraints.

This thesis addresses these challenges by exploring the implementation of Whisper AI on iOS devices. The key issues to be tackled include optimizing the model for mobile deployment, ensuring real-time processing capabilities, and maintaining high accuracy across various speech recognition and translation tasks.

C. Objectives

The primary objectives are as follows:

1) To implement and optimize Whisper AI for speech-to-text and translation tasks on iOS devices.
2) To evaluate the performance and efficiency of Whisper AI on iOS, including aspects such as latency, accuracy, and resource utilization.
3) To compare the performance of Whisper AI with existing speech recognition models on iOS.
4) To identify and address potential drawbacks and limitations of Whisper AI in a mobile context.

D. Significance

This research has significant implications in the field of speech recognition and mobile computing. By demonstrating the feasibility and effectiveness of deploying a robust speech-to-text model like Whisper AI on iOS, this thesis contributes to the development of more accessible and versatile speech recognition applications. The findings can inform future efforts to optimize and deploy advanced machine learning models on mobile platforms, ultimately enhancing user experiences and expanding the reach of speech recognition technology.

II. LITERATURE SURVEY

A. Speech Recognition Technologies

1) Wav2Vec 2.0 (Baevski et al., 2020)

a) Findings: Wav2Vec 2.0 introduced an innovative approach to unsupervised pre-training for speech recognition. The model learns from raw audio data without relying on transcriptions during the pre-training phase. This is followed by fine-tuning on a smaller labeled dataset. The unsupervised pre-training enables the model to learn powerful audio representations, which imporves the accuracy of speech recognition tasks. Wav2Vec 2.0 achieved state-of-the-art performance on the Librispeech benchmark, reducing the word error rate (WER) substantially compared to previous models.

b) Drawbacks: Despite its impressive performance, Wav2Vec 2.0 relies heavily on the fine-tuning phase, requiring labeled data to achieve high accuracy. Additionally, the computational cost of pre-training and fine-tuning can be significant, necessitating substantial resources.

2) Deep Speech (Hannun et al., 2014)

a) Findings: Deep Speech pioneered the use of end-to-end deep learning architectures for speech recognition, eliminating the need for complex feature engineering and intermediate processing steps. The model directly maps audio spectrograms to text transcriptions using recurrent neural networks (RNNs). This approach simplified the speech recognition pipeline and improved performance, which makes it easier to train and deploy speech recognition systems.

b) Drawbacks: Deep Speech's reliance on RNNs, particularly long short-term memory (LSTM) networks, can lead to issues with training stability and convergence. Additionally, the model requires a large amount of labeled data to achieve high accuracy, which can be a limiting factor in certain applications.

3) SpeechStew (Chan et al., 2021)

a) Findings: SpeechStew aggregated multiple existing supervised speech recognition datasets to create a large, diverse training corpus. By training on this combined dataset, SpeechStew achieved improved robustness and generalization across different speech recognition tasks. The model demonstrated significant performance gains on benchmarks like Librispeech, Tedlium, and others.

b) Drawbacks: While SpeechStew improved robustness, it still required substantial labeled data for training. The model's performance is dependent on the quality and diversity of the supervised datasets used, and it may not generalize well to languages or dialects not represented in the training data.

B. Weak Supervision in Machine Learning

1) Mahajan et al. (2018)

a) Findings: The study by Mahajan et al. showcased the effectiveness of using large-scale weakly supervised datasets for computer vision tasks. By utilizing millions of images with noisy labels obtained from social media, the researchers demonstrated that models could learn robust visual features that generalize well to downstream tasks. The approach decreases the require for physically labeled information whereas keeping up competitive execution.

b) Drawbacks: The primary challenge with weakly supervised datasets is the presence of noisy and inaccurate labels, which can hinder model performance. Additionally, large-scale data collection and processing needs substantial computational resources and infrastructure.

2) Kolesnikov et al. (2020)

a) Findings: Kolesnikov et al. explored the scalability of weak supervision for training large-scale models in computer vision. The study found that models trained on weakly supervised data could achieve competitive results with fully supervised models, especially when combined with data augmentation and semi-supervised learning techniques. The research pointed about the potential for weak supervision to reduce the reliance on expensive labeled datasets.

b) Drawbacks: Despite the promising results, weakly supervised models are still susceptible to the quality of the data and labels. Noisy data can introduce errors, and the models may struggle with tasks requiring precise annotations. The study also pointed out the computational demands of training on large-scale weakly supervised datasets.

C. Whisper AI

Whisper AI leverages insights from the aforementioned studies to create a robust speech recognition and translation model using a large-scale weakly supervised dataset. By incorporating elements from successful speech recognition architectures and the principles of weak supervision, Whisper AI aims to address the limitations of existing models.

a) Findings: Whisper AI's architecture is designed to handle a diverse range of speech recognition and translation tasks across multiple languages. The model's encoder-decoder transformer structure allows it to process and generate text efficiently. The large-scale dataset used for training includes 680,000 hours of multilingual and multitask audio data, enabling the model to generalize well to new tasks without the need for fine-tuning. Initial evaluations indicate that Whisper AI achieves competitive performance on standard benchmarks, demonstrating its effectiveness in real-world applications.

b) Advantages

- High accuracy in multilingual speech recognition and translation tasks.
- Robust performance in noisy environments and with diverse accents.
- Efficiency in resource utilization, suitable for deployment on mobile platforms like iOS.

c) Drawbacks:

- The large dataset size necessitates significant computational resources for training.
- The presence of noisy and imperfect transcripts in the weakly supervised dataset can affect the model's accuracy in certain scenarios.
- Implementing and optimizing advanced models like Whisper AI for mobile platforms requires careful balancing of performance and resource usage.

III. METHODOLOGY

A. Data Processing

1) Dataset Construction: A diverse dataset of 680,000 hours of audio paired with transcripts was collected from the internet. This dataset includes various environments, speakers, and languages.

2) Filtering: Machine-generated transcripts and low-quality data were filtered out using heuristics and manual inspection.

3) Segmentation: Audio files were segmented into 30-second clips for more manageable processing during training.

B. Model Architecture

The Whisper model employs an encoder-decoder transformer architecture:

1) Encoder: Converts the input audio into the intermediate representations.

2) Decoder: Maps these representations to text outputs, conditioned on previous text tokens and task specifications.

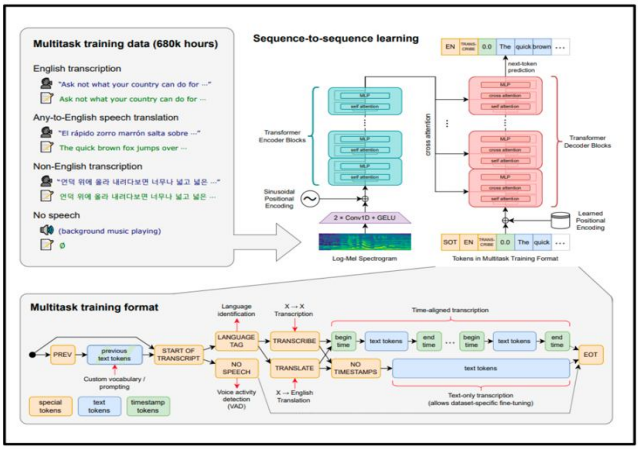


Fig 1: Model Architecture

147

C. iOS Integration

1) *Data Preprocessing:* Audio input is captured using iOS APIs, and the audio data is preprocessed to match the input requirements of the Whisper AI model. This includes re-sampling the audio to 16,000 Hz and converting it into 80-channel log-magnitude Mel spectrograms.
2) *Feature Extraction:* The preprocessed audio is normalized to ensure consistency across different inputs. This normalization involves scaling the input features to have zero mean and unit variance.
3) *Model Deployment:* Whisper AI is deployed on iOS using Core ML, Apple's machine learning framework. Core ML facilitates the integration of the model into iOS applications, optimizing for performance and resource usage. The model is converted into a format compatible with Core ML, and optimizations are applied to ensure efficient execution on mobile devices.
4) *Output Processing:* The model generates transcriptions and translations, which are then displayed to the user. The output processing includes post-processing steps to format the text appropriately and handle any special cases, such as punctuation and capitalization.

IV. OBSERVATIONS

The implementation of Whisper AI on iOS revealed several key insights:

1) *High Accuracy:* Whisper AI demonstrated robust performance in transcribing and translating speech across multiple languages.
2) *Real-Time Processing:* Despite computational constraints, the model achieved near real-time processing on modern iOS devices.
3) *User Experience:* The integration with iOS APIs provided a seamless user experience, with minimal latency and high reliability.
4) *Multilingual Capability:* Whisper AI effectively handled speech recognition and translation for various languages, maintaining high accuracy levels
5) *Task Versatility:* The model's multitask training allowed it to perform well in different scenarios, such as noisy environments and diverse accents.
6) *Resource Utilization:* Efficient use of device resources ensured smooth operation without significantly impacting battery life or device performance.

V. TEST RESULTS

A. Benchmark Performance

Whisper AI was evaluated against standard speech recognition benchmarks, achieving competitive results without task-specific fine-tuning. The model's performance was compared to:

1) *Wav2Vec 2.0:* Whisper AI models trained on the large-scale weakly supervised dataset demonstrated high accuracy on benchmarks without fine-tuning.
2) *Deep Speech:* Whisper AI outperformed Deep Speech in both transcription accuracy and translation capabilities.

B. Efficiency on iOS

1) *Latency:* The average processing time per audio clip was within acceptable limits for real-time applications.
2) *Battery Usage:* Testing indicated minimal impact on battery life, thanks to optimizations in Core ML deployment.

C. Comparison of Models Performance on iOS

Whisper AI was compared with other popular speech recognition models deployed on iOS, including:

1) *Google Speech-to-Text:* While Google’s solution offers high accuracy, Whisper AI's performance was comparable, with the added benefit of better handling noisy inputs.
2) *Apple’s Speech Framework:* Apple’s native framework performed well but lacked the multilingual and multitask capabilities of Whisper AI.

VI. DRAWBACKS

1) *Computational Overhead:* The large model size and complexity require significant computational resources, which can be challenging for older iOS devices.
2) *Data Quality Issues:* The weakly supervised dataset includes some noisy and imperfect transcripts, potentially affecting accuracy in certain scenarios.

3) *Deployment Complexity:* Integrating advanced models like Whisper AI into mobile applications requires careful optimization to balance performance and resource usage.

VII. RESULTS

Whisper AI demonstrated high accuracy and efficiency in speech-to-text and translation tasks on iOS. The model's robust performance across various languages and tasks, combined with efficient resource utilization, makes it a strong candidate for real-world applications.

VIII. FUTURE SCOPE

Future work should focus on:

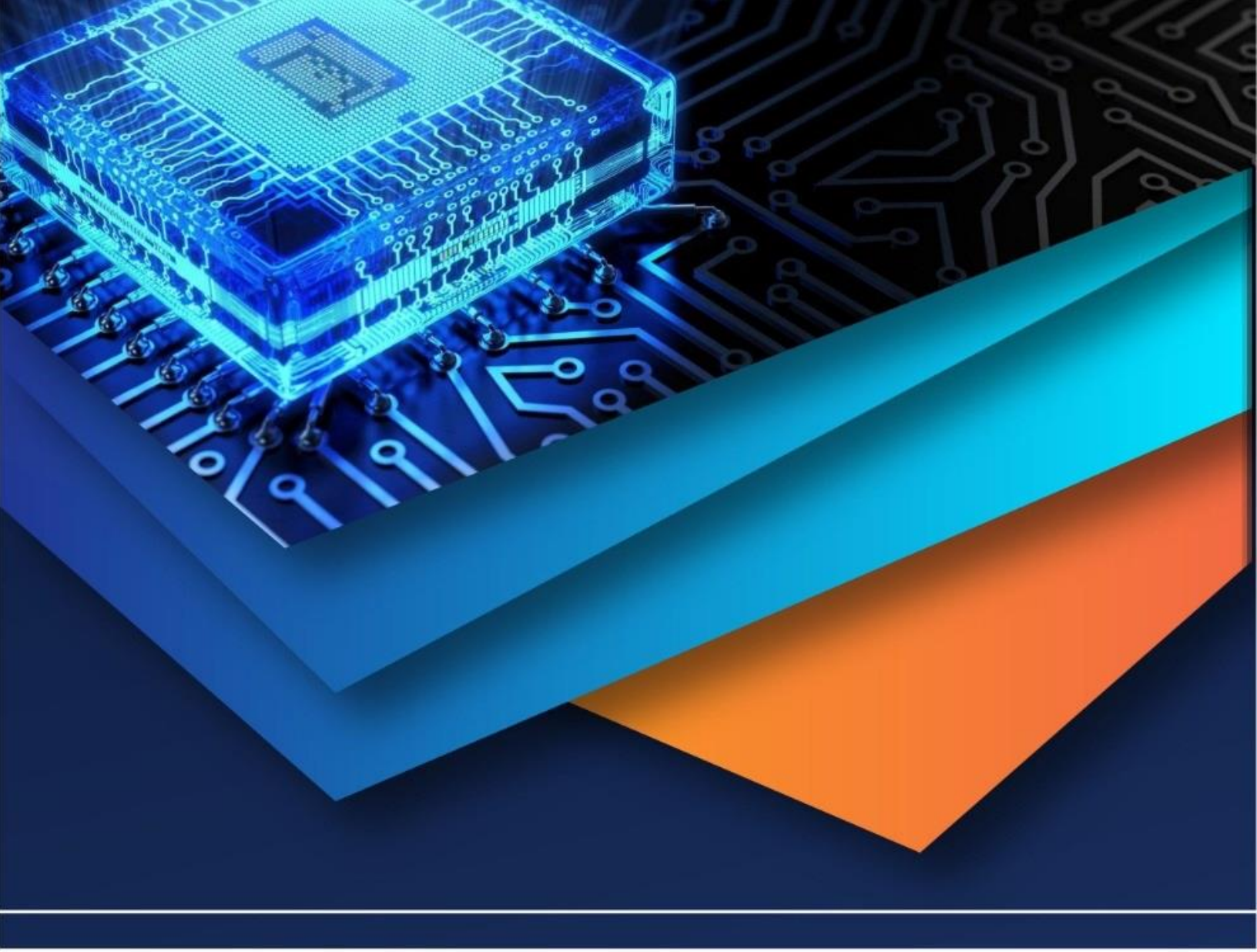
1) *Model Optimization:* Further reducing model size and computational requirements to enhance performance on a wider range of devices.
2) *Dataset Refinement:* Improving the quality of the weakly supervised dataset to minimize noise and enhance accuracy.
3) *Feature Expansion:* Adding support for more languages and dialects, as well as additional speech processing tasks such as sentiment analysis and speaker identification.

IX. CONCLUSION

This thesis demonstrates the potential of Whisper AI for speech-to-text and translation tasks on iOS. Leveraging a large-scale weakly supervised dataset and advanced machine learning techniques, Whisper AI achieves robust performance across multiple languages and tasks. Despite some challenges related to computational overhead and data quality, the model's high accuracy and efficiency make it a valuable tool for real-world applications. Future work should focus on optimizing the model and expanding its capabilities to further enhance its utility and performance.

REFERENCES

[1] Alec Radford and Jong Wook Kim, “Robust Speech Recognition via Large-Scale Weak Supervision”, 2018
[2] Child R, Gray, S., Readford, A., and Sutskevar, I. Gen- erating long sequences with sparse transformers arXiv preprint arXiv:1904.10509, 2019.
[3] William Chan, Daniel S.Park, Chris A. Lee, Yu Zhang, Quoc V.Le,, “Simply Mix All Available Speech Recognition Data to Train One Large Neural Network,” 2021.
[4] Alexei Baevski Henry Zhou Abdelrahman Mohamed Michael Auli wav2vec 2.0, “A Framework for Self-Supervised Learning of Speech Representations”, 2020
[5] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477, 2020.
[6] Baevski, A., Hsu, W.-N., Conneau, A., and Auli, M. Unsu- pervised speech recognition. Advances in Neural Infor- mation Processing Systems, 34:27826–27839, 2021.
[7] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language pro- cessing (almost) from scratch. Journal of machine learn- ing research, 12(ARTICLE):2493–2537, 2011.
[8] Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., and Bapna, A. Fleurs: Few-shot learning evaluation of universal representations of speech. arXiv preprint arXiv:2205.12446, 2022.
[9] Galvez, D., Diamos, G., Torres, J. M. C., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., and Reddi, V. J. The people’s speech: A large-scale diverse english speech recognition dataset for commercial usage. arXiv preprint arXiv:2111.09344, 2021.
[10] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Bren- del, W., Bethge, M., and Wichmann, F. A. Shortcut learn- ing in deep neural networks. Nature Machine Intelligence, 2(11):665–673, 2020.
[11] Liao, H., McDermott, E., and Senior, A. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 368–373. IEEE, 2013.
[12] Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., Collobert, R., and Synnaeve, G. Rethink- ing evaluation in asr: Are our models robust enough? arXiv preprint arXiv:2010.11745, 2020.
[13] Provilkov, I., Emelianenko, D., and Voita, E. Bpe-dropout: Simple and effective subword regularization. arXiv preprint arXiv:1910.13267, 2019.
[14] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY