A Comparative Evaluation of Deep Learning Models for Automatic Speech Recognition: Investigating the Performance, Fairness, and Deployability of Whisper, Wav2Vec 2.0, and LSTM Architectures.

NINAD JOSHI

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

September 2025

Supervisor: Sam Weiss

|  |  |
| --- | --- |
| **Module Title:** | MSc Data Analytics |
| **Assessment Title:** | A Comparative Evaluation of Deep Learning Models for Automatic Speech Recognition: Investigating the Performance, Fairness, and Deployability of Whisper, Wav2Vec 2.0, and LSTM Architectures. |
| **Lecturer Name:** | Sam Weiss |
| **Student Full Name:** | Ninad Shrikant Joshi |
| **Student Number:** | 2024526 |
| **Assessment Due Date:** | 07/09/2025 |
| **Date of Submission:** | 07/09/2025 |

# CCT College Dublin Assessment Cover Page

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Essential Links:

<https://commonvoice.mozilla.org/en/datasets>

<https://drive.google.com/drive/folders/1cKjt3X0Thgn9wp-lBM6YQdkQ8_1mdWNZ?usp=sharing>

<https://github.com/2024526-Ninad/ASR-Comparative-Study-Thesis>

Table of Contents

[CCT College Dublin Assessment Cover Page 2](#_Toc208172287)

[Abstract 5](#_Toc208172288)

[Chapter 1: Introduction 6](#_Toc208172289)

[1.1 Background and Motivation 6](#_Toc208172290)

[1.2 Problem Statement 7](#_Toc208172291)

[1.3 Research Aim, Objectives, and Questions 8](#_Toc208172292)

[1.4 Scope of the Study 9](#_Toc208172293)

[1.5 Significance of the Research 10](#_Toc208172294)

[1.6 Structure of the Thesis 10](#_Toc208172295)

[Chapter 2: Literature Review 11](#_Toc208172296)

[2.1 Long Short-Term Memory (LSTM) Networks in ASR 14](#_Toc208172297)

[2.2 Transformer Block Internals 17](#_Toc208172298)

[2.3 Datasets 20](#_Toc208172299)

[2.4 Evaluation Metrics 21](#_Toc208172300)

[2.5 Fairness 21](#_Toc208172301)

[2.6 Deployability 22](#_Toc208172302)

[2.7 Gaps in ASR Research 23](#_Toc208172303)

[Chapter 3: Research Methodology 24](#_Toc208172304)

[3.1 Research Design 25](#_Toc208172305)

[3.2 Datasets 25](#_Toc208172306)

[3.3 Wav2Wec2 27](#_Toc208172307)

[3.4 Whisper 28](#_Toc208172308)

[3.5 Evaluation Metrics 29](#_Toc208172309)

[3.6 Fairness Analysis & Robustness Testing 29](#_Toc208172310)

[3.7 Tools and Reproducibility 29](#_Toc208172311)

[3.8 Reflective Methodological Commentary 30](#_Toc208172312)

[3.9 Justification of Methodology 32](#_Toc208172313)

[Chapter 4: Results and Analysis 32](#_Toc208172314)

[4.1 Introduction 32](#_Toc208172315)

[4.2 Corpus-Level Performance 32](#_Toc208172316)

[4.3 Utterance-Level Analysis 33](#_Toc208172317)

[4.4 Fairness Analysis 34](#_Toc208172318)

[4.5 Robustness to Noise 35](#_Toc208172319)

[4.6 Deployability 36](#_Toc208172320)

[4.7 Cross-Dataset Evaluation 36](#_Toc208172321)

[4.8 Critical Discussion of Findings 37](#_Toc208172322)

[4.9 Chapter Summary 37](#_Toc208172323)

[Chapter 5: Discussion 38](#_Toc208172324)

[5.1 Introduction 38](#_Toc208172325)

[5.2 Accuracy and Generalisation 38](#_Toc208172326)

[5.3 Utterance-Level Consistency 39](#_Toc208172327)

[5.4 Fairness and Inclusivity 39](#_Toc208172328)

[5.5 Robustness to Noise 40](#_Toc208172329)

[5.6 Deployability and Efficiency 41](#_Toc208172330)

[5.7 Cross-Dataset Insights 41](#_Toc208172331)

[5.8 Ethical Implications 42](#_Toc208172332)

[5.9 Limitations 43](#_Toc208172333)

[5.10 Chapter Summary 43](#_Toc208172334)

[Chapter 6: Conclusion and Future Work 44](#_Toc208172335)

[6.1 Introduction 44](#_Toc208172336)

[6.2 Summary of Research Aims and Objectives 44](#_Toc208172337)

[6.3 Summary of Findings 45](#_Toc208172338)

[6.4 Contributions of the Research 46](#_Toc208172339)

[6.5 Ethical Considerations 47](#_Toc208172340)

[6.6 Limitations 48](#_Toc208172341)

[6.7 Future Research Directions 48](#_Toc208172342)

[6.8 Conclusion 49](#_Toc208172343)

[References 51](#_Toc208172344)

# Abstract

Automatic Speech Recognition (ASR) has become a central technology in modern digital life, powering applications ranging from voice assistants and accessibility tools to healthcare transcription and education support. Despite impressive progress, key challenges persist in ensuring that ASR systems are not only accurate but also fair, robust, and deployable across diverse contexts. This thesis investigates these challenges by conducting a systematic comparative evaluation of two state-of-the-art transformer-based ASR models: OpenAI’s Whisper and Facebook AI Research’s Wav2Vec 2.0.

The research adopts a comparative design to examine model performance across multiple dimensions. Mozilla Common Voice, representing diverse, real-world speech, was selected to test fairness and robustness, while LibriSpeech was chosen as a benchmark dataset to ensure comparability with existing research. Evaluation focused on transcription accuracy, demographic fairness, noise resilience, and computational deployability. Fairness was assessed through subgroup analysis of gender and accent, robustness was tested by introducing controlled noise, and deployability was measured through inference speed and resource requirements on consumer-grade hardware.

The findings reveal clear trade-offs between the two models. Whisper demonstrates stronger generalisation to diverse conditions, greater robustness under noise, and smaller demographic disparities, reflecting the benefits of training on a large, varied corpus. However, these advantages come with significant computational demands. Wav2Vec 2.0, in contrast, remains lightweight, efficient, and highly competitive on benchmark datasets, though it struggles in noisy or diverse conditions and exhibits larger fairness gaps.

This research contributes to the field by providing one of the few side-by-side evaluations of Whisper and Wav2Vec 2.0 under identical conditions, highlighting the multi-dimensional nature of ASR evaluation. It emphasises that future ASR development must balance accuracy with fairness, robustness, and sustainability to ensure inclusivity and real-world applicability.

# **Chapter 1: Introduction**

## 1.1 Background and Motivation

In recent decades, automatic speech recognition (ASR) has moved from the realm of laboratory prototypes to becoming a ubiquitous feature of everyday life. From the voice assistants embedded in smartphones and smart speakers to live captioning for accessibility, ASR has become a cornerstone technology in human-computer interaction. Its ability to transform spoken language into written text has profound implications for communication, accessibility, and productivity. Yet despite impressive advances, ASR systems continue to face challenges in accuracy, fairness, robustness, and deployability. These challenges are particularly evident when systems are evaluated outside of controlled laboratory conditions and applied to diverse populations in real-world environments.

The transformative potential of ASR stems from its wide range of applications. In healthcare, clinicians increasingly rely on speech-to-text systems for medical transcription, reducing administrative burdens and allowing more time for patient care. In education, automated captioning enables more inclusive classrooms by supporting students with hearing impairments and those for whom English is not a first language. In customer service, call centres employ ASR to automate routine tasks and improve efficiency. In global communication, speech translation systems break down language barriers. These examples illustrate the growing societal dependence on ASR. At the same time, they underscore the need for systems that are accurate, equitable, and capable of functioning under real-world constraints.

Historically, ASR systems relied on statistical models such as Hidden Markov Models (HMMs) combined with Gaussian Mixture Models (GMMs) to represent the temporal dynamics of speech. These systems were heavily dependent on hand-engineered features such as Mel-Frequency Cepstral Coefficients (MFCCs) and were limited in their ability to capture variability in speech (Mohamed et al., 2012). The advent of deep learning in the early 2010s revolutionised ASR by enabling end-to-end learning directly from raw audio. Recurrent architectures, especially Long Short-Term Memory (LSTM) networks, became the dominant approach and achieved state-of-the-art results on benchmark datasets such as TIMIT and Switchboard (Graves et al., 2013). However, LSTMs faced challenges in scalability, computational efficiency, and robustness to noise.

The introduction of transformer architectures marked a paradigm shift. Originally developed for natural language processing (Vaswani et al., 2017), transformers replaced recurrence with self-attention mechanisms, enabling models to capture long-range dependencies and exploit parallel computation. In ASR, transformer-based architectures such as Wav2Vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2022) have delivered dramatic improvements in accuracy and robustness. Wav2Vec 2.0 demonstrated the power of self-supervised pre-training on massive unlabeled audio corpora, while Whisper showcased the benefits of supervised training on 680,000 hours of multilingual, multitask, and noisy data. Together, these models represent state-of-the-art approaches that embody different strategies for achieving robust speech recognition.

Despite these advances, critical questions remain. How do these models perform when evaluated under identical conditions, using diverse and realistic datasets such as Mozilla Common Voice? Do they treat all speaker groups equitably, or do demographic disparities persist? How robust are they to noise, and how feasible are they to deploy on consumer-grade hardware? These questions are central to ensuring that ASR systems are not only technically impressive but also fair, inclusive, and practically usable.

## 1.2 Problem Statement

While ASR systems have achieved near-human performance on benchmark datasets such as LibriSpeech, these results often mask limitations in fairness and robustness. Studies have shown that commercial ASR systems perform disproportionately worse for women, non-native speakers, and speakers of African American Vernacular English (Koenecke et al., 2020). Similarly, systems that excel on clean, read speech often falter in noisy, spontaneous, or accented conditions. Moreover, the computational demands of state-of-the-art models raise questions about deployability, especially in resource-constrained environments or on consumer-grade hardware.

Against this backdrop, this study identifies three interrelated problems. First, there is a lack of systematic comparative evaluations of Whisper and Wav2Vec 2.0 under identical experimental conditions. While both models have been independently evaluated, most published results use different datasets and methodologies, making direct comparison difficult. Second, there is insufficient attention to fairness and inclusivity. Although Whisper’s diverse training corpus suggests it may mitigate disparities, systematic subgroup analysis has been limited. Third, there is a need to evaluate deployability in practical contexts, including performance on consumer-grade hardware where resources are constrained. Addressing these problems is essential to understanding the true strengths and limitations of modern ASR.

## 1.3 Research Aim, Objectives, and Questions

The aim of this research is to **evaluate and compare the effectiveness, fairness, and deployability of Whisper and Wav2Vec 2.0** under controlled yet realistic conditions.

The study is guided by five objectives:

1. To benchmark the performance of Whisper and Wav2Vec 2.0 on Mozilla Common Voice and LibriSpeech datasets, using WER, CER, and RTF as evaluation metrics.
2. To assess fairness by measuring error rates across gender and accent subgroups within Common Voice.
3. To evaluate robustness by testing performance on clean and noise-augmented speech at varying signal-to-noise ratios.
4. To analyze deployability by measuring inference speed and resource requirements on consumer-grade hardware.
5. To critically interpret the trade-offs between accuracy, fairness, robustness, and efficiency in light of existing literature.

From these objectives, the following research questions are derived:

* **RQ1:** Which model performs better in terms of accuracy (WER and CER) on diverse and benchmark datasets?
* **RQ2:** Do the models demonstrate fairness across demographic subgroups of gender and accent?
* **RQ3:** What are the trade-offs between accuracy, fairness, and deployability when models are run on consumer-grade hardware?

These questions are accompanied by hypotheses, building on the literature and preliminary expectations:

* **H1:** Whisper will significantly outperform Wav2Vec 2.0 in terms of WER and CER on Common Voice, though Wav2Vec 2.0 may remain competitive on LibriSpeech.
* **H2:** Whisper will demonstrate greater fairness, with smaller disparities across gender and accent groups.
* **H3:** Whisper will be more robust to noise, maintaining lower error rates under degraded conditions.

## 1.4 Scope of the Study

The scope of this research is deliberately focused on a comparative evaluation of Whisper and Wav2Vec 2.0. The Mozilla Common Voice dataset is the primary evaluation ground, given its diversity of speakers, accents, and recording conditions. LibriSpeech is included as a benchmark corpus to situate results within the broader literature. The evaluation is limited to English speech, though Whisper’s multilingual capabilities are acknowledged.

The analysis is restricted to three core evaluation metrics: WER, CER, and RTF. While other metrics such as energy consumption could be considered, the focus on these metrics ensures comparability with existing studies and relevance to both academic and practical applications. The fairness analysis is limited to gender and accent subgroups due to metadata availability. Noise robustness is tested using artificially added noise at standard SNR levels. All experiments are conducted locally on a MacBook Pro M1 Pro, providing insights into deployability under realistic hardware constraints.

## 1.5 Significance of the Research

The significance of this research lies in its contribution to three dimensions of ASR evaluation: technical performance, fairness, and deployability. By comparing Whisper and Wav2Vec 2.0 under identical conditions, the study provides clear evidence of their relative strengths and weaknesses, contributing to ongoing debates about the most effective architectures for speech recognition. By incorporating fairness analysis, it addresses an ethical dimension often neglected in technical evaluations, highlighting the inclusivity of ASR systems. By evaluating deployability on consumer hardware, it ensures that findings are grounded in practical realities rather than idealized laboratory conditions.

For academic research, the study contributes to filling a gap in the literature by providing a controlled, side-by-side comparison of Whisper and Wav2Vec 2.0. For practitioners, the findings offer guidance on model selection for different applications, balancing accuracy, fairness, and efficiency. For policymakers and regulators, the research highlights the need to address fairness in ASR deployment, ensuring that systems do not exacerbate existing inequalities.

## 1.6 Structure of the Thesis

This thesis is structured into six chapters. Chapter 1 introduces the research problem, aim, objectives, and significance. Chapter 2 presents a comprehensive literature review, tracing the evolution of ASR from statistical models to deep learning architectures, with a focus on LSTM, Wav2Vec 2.0, and Whisper. Chapter 3 describes the methodology, including datasets, preprocessing, model implementation, evaluation metrics, and subgroup analysis. Chapter 4 presents the results, including corpus-level performance, fairness and robustness analyses, and deployability findings. Chapter 5 discusses the results in relation to the research questions and literature, highlighting trade-offs and implications. Chapter 6 concludes the thesis with a summary of contributions, limitations, and directions for future work.

# **Chapter 2: Literature Review**

Automatic Speech Recognition (ASR) is the process of converting spoken language into text, a task that sits at the intersection of linguistics, signal processing, and machine learning. It has evolved significantly over the past seven decades, moving from rule-based approaches to probabilistic models, and more recently to deep learning–based architectures. Each paradigm shift has been driven by advances in algorithms, computational power, and the availability of larger and more diverse datasets. Understanding the evolution of ASR provides critical context for this thesis, which evaluates the comparative effectiveness of two of the most recent and advanced transformer-based models, Whisper and Wav2Vec 2.0.

The earliest ASR systems emerged in the 1950s and 1960s. These systems were rudimentary, recognising digits or small vocabularies using template matching approaches. Bell Labs’ “Audrey” system (1952) recognised spoken digits but was limited to a single speaker (Davis et al., 1952). Such systems were not scalable as they relied on direct acoustic matching, which was highly sensitive to variations in speech patterns, pitch, and speaking speed. Despite these limitations, they laid the groundwork by demonstrating the feasibility of speech-to-text conversion.

The 1970s and 1980s saw the rise of **statistical models**, particularly the use of **Hidden Markov Models (HMMs)** in combination with **Gaussian Mixture Models (GMMs)**. This era marked a significant breakthrough, as HMMs provided a probabilistic framework for modelling temporal sequences, while GMMs captured the variability in acoustic features. The Carnegie Mellon University “Harpy” system (Lowerre, 1976) and subsequent DARPA-sponsored projects demonstrated the potential of HMM-based systems to handle larger vocabularies. By the late 1980s, HMM–GMM systems had become the dominant approach in ASR, underpinning commercial systems used in call centres and dictation software. The success of these models was largely due to their ability to model speech as a sequence of phonemes or subphonetic states, coupled with the use of hand-crafted features such as **Mel-Frequency Cepstral Coefficients (MFCCs)** (Rabiner, 1989).

Despite their success, HMM–GMM systems had fundamental limitations. They relied on strong assumptions of independence between observations and struggled to capture long-range dependencies in speech. Feature engineering was labour-intensive and error-prone, as MFCCs and related features could not fully capture the richness of the speech signal. Moreover, these systems required large amounts of labelled data for training, which limited their adaptability to new languages and domains.

The emergence of **neural networks** in the late 1980s and early 1990s offered a potential alternative. Early efforts used **multilayer perceptrons (MLPs)** as acoustic models, but computational limitations and small datasets hindered their widespread adoption. It was not until the early 2010s that deep learning achieved a breakthrough in ASR. Hinton et al. (2012) demonstrated that deep neural networks (DNNs) could replace GMMs as acoustic models within the HMM framework, leading to significant reductions in error rates. This hybrid HMM–DNN approach quickly became state-of-the-art, used by leading technology companies to power speech recognition in mobile devices and virtual assistants.

While deep feedforward networks improved accuracy, they still processed each frame of audio independently, ignoring the sequential nature of speech. This limitation was addressed by **Recurrent Neural Networks (RNNs)**, particularly **Long Short-Term Memory (LSTM)** networks (Hochreiter and Schmidhuber, 1997). LSTMs introduced gating mechanisms that allowed them to capture long-term dependencies without succumbing to vanishing gradients. In ASR, bidirectional LSTMs (BiLSTMs) became especially effective, as they could incorporate both past and future context when predicting phonemes or words. Graves et al. (2013) demonstrated that deep LSTM RNNs could achieve state-of-the-art results on phoneme recognition tasks such as TIMIT, while Sak et al. (2014) extended these results to large vocabulary continuous speech recognition (LVCSR).

A further breakthrough came with the introduction of **Connectionist Temporal Classification (CTC)** (Graves et al., 2006), which enabled end-to-end training of neural networks for sequence labelling tasks without the need for pre-aligned phoneme annotations. CTC allowed models to map variable-length input sequences directly to variable-length output sequences, a crucial capability for ASR. This innovation paved the way for fully end-to-end ASR systems, in which a single neural architecture could process raw audio features and output text, bypassing the need for intermediate phoneme-based representations.

The 2010s also witnessed the development of large-scale ASR systems trained on increasingly massive datasets. Baidu’s **Deep Speech 2** (Amodei et al., 2016) exemplified this trend by training end-to-end RNN-based models on thousands of hours of English and Mandarin speech, achieving state-of-the-art results across both languages. Importantly, Deep Speech 2 demonstrated the benefits of scale: larger datasets and deeper models consistently yielded improved performance. However, RNN-based systems still faced challenges of sequential computation, which limited parallelisation and slowed training and inference.

The **transformer architecture** (Vaswani et al., 2017) addressed these limitations by introducing self-attention mechanisms capable of modelling long-range dependencies without sequential recurrence. Unlike RNNs, transformers compute relationships between all positions in a sequence simultaneously, enabling massive parallelisation during training. In ASR, this paradigm shift led to models such as **Wav2Vec 2.0** (Baevski et al., 2020), which combined a convolutional feature encoder with transformer layers to capture contextual dependencies in speech. Wav2Vec 2.0 introduced a self-supervised learning approach: the model was pre-trained to predict masked portions of audio representations, enabling it to learn from vast amounts of unlabelled data before being fine-tuned on smaller labelled datasets. This approach significantly reduced the reliance on expensive transcriptions while achieving state-of-the-art performance on benchmarks like LibriSpeech.

In parallel, OpenAI developed **Whisper** (Radford et al., 2022), which adopted a different philosophy. Instead of self-supervised pre-training, Whisper was trained on 680,000 hours of paired audio-text data sourced from the web, encompassing multiple languages, accents, and noisy conditions. Its encoder–decoder transformer architecture enabled multitask outputs, including transcription, translation, and language identification. Whisper’s scale and diversity gave it robustness across accents and noisy environments, outperforming prior systems in fairness and inclusivity.

The evolution of ASR demonstrates how the field has transitioned from small-vocabulary, template-based systems to large-scale, end-to-end transformer architectures capable of near-human performance. Yet, as Koenecke et al. (2020) and others have shown, technical progress does not automatically ensure fairness. Persistent disparities across demographics highlight the importance of evaluating ASR not only on benchmarks but also on diverse datasets such as Mozilla Common Voice, which reflects real-world variation.

In summary, the history of ASR is one of expanding scope and ambition. Each paradigm — from HMM–GMMs to LSTMs to transformers — has addressed limitations of its predecessor, pushing performance closer to human levels. However, accuracy alone is not sufficient. The focus must increasingly include fairness, robustness, and deployability, the very dimensions evaluated in this thesis through a systematic comparison of Whisper and Wav2Vec 2.0.

## **2.1 Long Short-Term Memory (LSTM) Networks in ASR**

The introduction of Long Short-Term Memory (LSTM) networks marked a pivotal advancement in sequence modelling and significantly influenced the trajectory of automatic speech recognition (ASR). While earlier recurrent neural networks (RNNs) provided a means of modelling temporal dependencies in sequential data, they were severely limited by the vanishing and exploding gradient problems (Bengio et al., 1994). LSTMs, proposed by Hochreiter and Schmidhuber (1997), directly addressed these limitations, enabling the modelling of long-range dependencies by introducing memory cells and gating mechanisms that regulate the flow of information. Over the past two decades, LSTMs have played a central role in ASR systems, particularly in the transition from hybrid HMM–DNN models to end-to-end neural architectures.

At the heart of the LSTM is the concept of the **cell state**, which acts as a form of long-term memory. Information can be added or removed from the cell state through the action of gates. The **input gate** controls which new information is stored, the **forget gate** determines which information should be discarded, and the **output gate** regulates what information is exposed as output. This gating mechanism allows the network to retain relevant information over long temporal spans, making LSTMs especially well-suited to speech recognition, where dependencies can stretch across multiple words or sentences.

LSTMs quickly demonstrated their superiority over traditional RNNs in ASR tasks. Graves et al. (2013) showed that deep LSTM networks achieved state-of-the-art performance on the TIMIT phoneme recognition benchmark. By leveraging their ability to capture longer context, LSTMs reduced substitution and deletion errors that plagued previous models. The use of **bidirectional LSTMs (BiLSTMs)** further enhanced performance by incorporating both past and future context when predicting phonemes or words, a capability particularly valuable in continuous speech recognition (Graves and Schmidhuber, 2005).

One of the most influential contributions of LSTMs to ASR was their integration with **Connectionist Temporal Classification (CTC)** (Graves et al., 2006). CTC addressed the challenge of aligning input speech frames with output text sequences, which vary in length. Traditionally, ASR systems required pre-aligned phonetic transcriptions, a costly and error-prone process. CTC enabled end-to-end training by introducing a dynamic programming-based loss function that marginalises over all possible alignments, allowing LSTM-based models to directly map speech to text without explicit alignment. This innovation laid the foundation for end-to-end ASR systems, simplifying the pipeline and reducing reliance on hand-crafted linguistic resources.

LSTMs also powered large-scale industrial systems during the mid-2010s. Sak et al. (2014) demonstrated the effectiveness of LSTMs for large vocabulary continuous speech recognition (LVCSR), achieving significant improvements over hybrid HMM–DNN systems. Google deployed LSTM-based acoustic models in its speech recognition services, citing their ability to generalise across different domains and user conditions (Xiong et al., 2017). These successes solidified LSTMs as a key architecture in the deep learning revolution of ASR.

However, LSTMs are not without limitations. Training is computationally expensive due to their sequential nature, as each time step must be processed in order, preventing parallelisation across sequences. This makes them slower to train and infer compared to convolutional or transformer-based models. Additionally, while LSTMs can model longer dependencies than traditional RNNs, they are not immune to degradation when sequences become very long. Gradient decay still occurs, albeit more slowly, which limits their scalability for extremely long utterances.

Another drawback is their difficulty in capturing hierarchical or multi-scale structures in speech. While gating mechanisms help preserve information, LSTMs do not inherently model different levels of abstraction in sequential data. Researchers addressed this limitation by developing **stacked LSTMs** (Pascanu et al., 2014), where multiple LSTM layers were combined to capture increasingly abstract representations of speech. Hierarchical LSTMs were also proposed to capture dependencies at different temporal resolutions, improving performance on conversational speech.

Despite these limitations, LSTMs remained state-of-the-art until the rise of transformers. The shift occurred because transformer models, with their self-attention mechanisms, were able to capture dependencies across entire sequences without the bottlenecks of recurrence. Nevertheless, LSTMs retain a legacy in ASR, both in terms of technical contributions and as a benchmark for evaluating newer models. They also remain relevant in resource-constrained scenarios where lighter architectures are preferred. In fact, lightweight LSTM-based ASR systems are still deployed in embedded devices such as smartphones and IoT systems, where efficiency takes precedence over absolute accuracy.

It is also important to recognise the role of LSTMs in fairness and robustness research. Because they formed the backbone of ASR during the 2010s, many of the early studies documenting demographic disparities in ASR performance were conducted on LSTM-based systems. Tatman (2017), for example, found that Google’s LSTM-based speech recogniser exhibited significantly higher error rates for women compared to men, highlighting systemic biases in training data and architecture performance. These findings underscore the importance of evaluating fairness not only in cutting-edge transformer models such as Whisper but also in the LSTM systems that continue to be used in production.

In addition, the introduction of sequence-to-sequence models with attention mechanisms extended LSTMs’ impact beyond CTC-based ASR. Bahdanau et al. (2015) proposed an encoder–decoder architecture where the encoder was an LSTM and the decoder generated output sequences with an attention mechanism. This architecture bridged the gap between LSTM-based sequence modelling and transformer-based self-attention, demonstrating the feasibility of fully neural, end-to-end ASR without intermediate HMM alignments. Although transformers have since surpassed LSTMs in performance, this hybrid innovation was crucial in demonstrating the potential of attention mechanisms for ASR.

Recent years have also seen attempts to combine LSTMs with other architectures to balance trade-offs. Convolutional-recurrent hybrids leverage CNNs for local feature extraction and LSTMs for long-range modelling (Zhang et al., 2017). Such hybrids have been shown to improve robustness in noisy environments, highlighting that LSTMs retain utility in specialised contexts. Similarly, low-rank and quantised LSTM variants have been developed for efficient deployment on mobile hardware, underscoring their ongoing relevance in real-world applications.

In summary, LSTMs represent a transitional but critical phase in the evolution of ASR. They enabled end-to-end training through CTC, provided robustness to long-range dependencies, and powered the first large-scale, commercially deployed deep learning ASR systems. Their limitations in scalability and parallelisation paved the way for transformers, but their contributions remain foundational. For the purposes of this thesis, LSTMs serve as a vital historical comparator to the transformer-based architectures under study. Evaluating Whisper and Wav2Vec 2.0 requires acknowledging the innovations that LSTMs introduced, as well as the reasons they were eventually surpassed. In this way, LSTMs provide both a baseline and a bridge in the ongoing evolution of ASR.

## **2.2 Transformer Block Internals**

The transformer architecture represents a fundamental shift in how sequential data such as language and speech are processed. Introduced by Vaswani et al. (2017), transformers dispense with recurrence and convolution, instead relying entirely on self-attention mechanisms to model dependencies between elements in a sequence. Within this design, the transformer block is the essential building unit, and understanding its internals is critical to appreciating the advances made by models such as Wav2Vec 2.0 and Whisper. Each block is composed of three main components: multi-head self-attention, position-wise feed-forward networks, and residual connections with layer normalization.

The multi-head attention mechanism is the defining innovation of transformers. It addresses the limitations of recurrent neural networks (RNNs) and LSTMs, which process sequences step by step and struggle to capture long-range dependencies due to vanishing or exploding gradients (Hochreiter and Schmidhuber, 1997). Attention mechanisms compute a weighted sum over all positions in the input sequence, allowing the model to focus selectively on relevant parts of the sequence regardless of distance. Formally, attention operates over queries (Q), keys (K), and values (V), with outputs defined as:

A diagram of a product

AI-generated content may be incorrect.

where dkd\_kdk​ is the dimensionality of the keys. The softmax operation ensures that the weights sum to one, assigning greater weight to positions most relevant to the current query. Multi-head attention extends this mechanism by running several attention operations in parallel, each with its own learned parameters. This enables the model to capture different types of relationships simultaneously — for example, syntactic versus semantic dependencies in text, or local versus global acoustic patterns in speech (Vaswani et al., 2017).

Position-wise feed-forward networks follow the attention mechanism. These are simple two-layer fully connected networks applied independently to each position in the sequence. Despite their simplicity, they provide the capacity to transform representations after attention has aggregated contextual information. Together with attention, these networks allow each block to iteratively refine representations, combining information across the sequence while applying nonlinear transformations.

Residual connections and layer normalization complete the block design. Residual connections ensure that the original input is preserved and combined with the output of each sub-layer, mitigating the vanishing gradient problem and facilitating training of very deep networks (He et al., 2016). Layer normalization stabilises training by normalising the output of each layer, ensuring consistent activation scales across the network (Ba et al., 2016). The result is a highly modular design that can be stacked into dozens or even hundreds of layers, as seen in large-scale models such as GPT-3 (Brown et al., 2020) and Whisper (Radford et al., 2022).

The transformer block is also inherently parallelisable, a feature that has significant implications for ASR. Unlike LSTMs, which must process sequences sequentially, transformers compute attention across all positions simultaneously. This reduces training times dramatically when sufficient computational resources are available. For ASR, where datasets can comprise thousands of hours of speech, this parallelisation is critical. Baevski et al. (2020) leveraged this property in Wav2Vec 2.0, combining a convolutional feature encoder with transformer layers to capture long-range speech dependencies more effectively than recurrent architectures.

While the transformer block design is elegant, it is not without challenges. The quadratic complexity of self-attention with respect to sequence length makes long input sequences computationally expensive. Speech data, which often involves thousands of frames per utterance, exacerbates this issue. To address this, researchers have proposed efficient attention mechanisms such as Linformer (Wang et al., 2020) and Performer (Choromanski et al., 2021), which approximate full attention while reducing computational cost. These innovations are increasingly important as transformer-based ASR models are scaled to longer utterances and lower-resource environments.

Another challenge is interpretability. While attention weights provide some insight into what the model is focusing on, the high dimensionality and multiplicity of attention heads make it difficult to derive clear explanations of behavior. This raises questions about transparency and accountability, especially as transformer-based ASR systems are deployed in sensitive domains such as healthcare and law.

Despite these challenges, transformer blocks have proven highly effective for ASR. Wav2Vec 2.0 uses a stack of twelve transformer blocks in its base configuration, enabling it to model contextual relationships across entire audio sequences. Whisper employs an encoder–decoder structure, with the encoder consisting of transformer blocks that process spectrogram inputs, and the decoder composed of transformer blocks that autoregressively generate text (Radford et al., 2022). Both models highlight how the internals of transformer blocks — attention, feed-forward networks, residuals, and normalization — underpin state-of-the-art performance in ASR.

The influence of transformer internals extends beyond accuracy. Their capacity to generalise across accents and noisy conditions reflects their ability to capture both local and global dependencies in speech. Their parallelisable design makes large-scale pre-training feasible, while their modular structure facilitates multitask learning. In this sense, the transformer block has become not only the backbone of models like Whisper and Wav2Vec 2.0 but also the central architecture of modern deep learning.

In summary, the transformer block represents a remarkable convergence of simplicity and power. By combining self-attention, feed-forward networks, residuals, and normalization, it enables efficient, scalable, and context-sensitive modelling of sequences. Its impact on ASR is profound, enabling models that outperform recurrent architectures in accuracy, robustness, and fairness. Yet its computational demands and interpretability challenges highlight areas for ongoing innovation. For the purposes of this thesis, transformer blocks are not merely technical details but foundational components that explain why Whisper and Wav2Vec 2.0 achieve state-of-the-art performance.

## 2.3 **Datasets**

Among the most widely used datasets in ASR research is **LibriSpeech** (Panayotov et al., 2015), a corpus derived from public domain audiobooks read by volunteers. LibriSpeech provides approximately 1,000 hours of English-language recordings with standardized training, validation, and test splits. It is designed to evaluate transcription accuracy under controlled conditions, offering both “clean” and “other” subsets that represent well-recorded and noisier speech respectively. Due to its size, accessibility, and standardized evaluation protocol, LibriSpeech has become the de facto benchmark for ASR research, serving as the baseline dataset against which new models are compared. However, LibriSpeech has been criticized for its limited speaker diversity. The dataset is heavily skewed toward North American and British English speakers, with little representation of accented, dialectal, or non-native English. This lack of demographic diversity makes it unsuitable for evaluating fairness across speaker groups.

To address the need for greater diversity, **Mozilla Common Voice** has emerged as an important complementary dataset. Launched in 2017 as a crowdsourced project, Common Voice provides recordings from speakers around the world who donate their voices by reading predefined sentences. Unlike LibriSpeech, Common Voice is explicitly designed to capture variation across accents, genders, and age groups, and it is multilingual, with contributions in dozens of languages. This diversity makes it a valuable resource for assessing inclusivity and fairness in ASR models. However, the dataset is more heterogeneous in quality, containing varying microphone conditions and noise levels, which complicates training and evaluation. Nonetheless, this variability better reflects real-world usage scenarios, where speech is rarely captured under pristine conditions.

Other datasets have also played important roles in ASR development. **Switchboard** is a long-standing corpus of conversational telephone speech that has been widely used to evaluate ASR in spontaneous dialogue contexts. The **TED-LIUM** corpus, based on TED talks, provides recordings of prepared speech delivered by speakers from diverse backgrounds, often with distinctive accents and prosody. The **CHiME** challenges have focused specifically on robustness to noise, providing speech recordings in everyday acoustic environments such as cafés, streets, and buses. Together, these datasets highlight the multidimensionality of ASR evaluation: models must be tested not only on clean and controlled speech but also on spontaneous, accented, noisy, and multilingual input to be considered truly robust.

## **2.4 Evaluation Metrics**

The most commonly used evaluation metrics in ASR are **Word Error Rate (WER)** and **Character Error Rate (CER)**. WER is defined as the ratio of the total number of substitutions, insertions, and deletions required to align a system’s output with the reference transcription, divided by the number of words in the reference. CER is defined similarly but operates at the character level, making it more sensitive to errors in short utterances or languages with complex morphology. These metrics have become standard because they are straightforward to compute and provide intuitive measures of accuracy. However, they have limitations. WER treats all errors as equally important, failing to distinguish between errors that alter meaning (e.g., “yes” vs. “no”) and those that do not (e.g., “the” vs. “a”). CER, while useful in some contexts, may overemphasize minor spelling or orthographic differences that are less relevant to real-world usability.

Another important metric is the **Real-Time Factor (RTF)**, which measures the time taken by a system to transcribe audio relative to the duration of the audio itself. An RTF of less than 1 indicates real-time transcription capability, which is essential for live applications such as captioning or voice assistants. RTF highlights a critical trade-off: models that achieve very low WER may nonetheless be impractical if their inference speed is too slow. This is particularly relevant for transformer-based models such as Whisper, which, despite their accuracy, may fail to meet latency requirements on consumer hardware.

In recent years, researchers have proposed additional metrics to capture dimensions of ASR performance that are overlooked by WER, CER, and RTF. For example, some studies advocate for measuring **energy consumption** or **carbon footprint** during training and inference, acknowledging the environmental impact of large-scale deep learning. Others suggest composite indices that combine accuracy, latency, and resource efficiency into a single measure, better reflecting real-world deployment trade-offs. These alternative metrics remain less widely adopted, but their emergence signals a shift toward more holistic evaluation frameworks.

## 2.5 Fairness

Fairness and inclusivity have become increasingly prominent in ASR research. Several studies have demonstrated that ASR systems often perform unevenly across demographic groups. Koenecke et al. (2020), for instance, analyzed the performance of commercial ASR systems from leading providers and found significantly higher error rates for African American Vernacular English (AAVE) speakers compared to white speakers. This disparity reflects the underrepresentation of certain speech communities in training data and raises ethical concerns about bias and discrimination in AI technologies. Other studies have shown similar gaps in performance across genders, with female voices often yielding higher error rates, as well as across accents, where non-native English speakers are disproportionately disadvantaged.

These findings highlight the importance of evaluating ASR systems not only on overall WER but also on subgroup-specific metrics. Fairness in ASR requires models to perform consistently across diverse populations, avoiding systemic disadvantages for already marginalized groups. Whisper represents a step forward in this regard, as its training data included a wide range of accents, dialects, and noisy environments, enabling it to achieve relatively balanced performance across subgroups. However, even Whisper is not immune to biases, as certain languages and communities remain underrepresented in its training corpus. Achieving true fairness will likely require intentional data collection efforts that prioritize inclusivity, as well as algorithmic innovations designed to reduce disparities.

Fairness also intersects with issues of accessibility. ASR technologies are increasingly used in accessibility tools for people with disabilities, such as live captioning for the hearing impaired. If these systems perform poorly for speakers with speech impairments or atypical prosody, they may exacerbate existing inequities. Research in this area is still relatively limited, but early findings suggest that mainstream ASR systems often struggle with dysarthric or disfluent speech. Addressing these gaps requires both targeted dataset development and specialized model training.

## 2.6 **Deployability**

Deployability is another critical dimension of ASR evaluation. A model’s usefulness depends not only on its accuracy but also on its computational efficiency, memory footprint, and ability to run on the hardware available in real-world applications. LSTM-based models, though less accurate than transformers, retain value in scenarios where computational resources are limited, such as embedded devices, mobile phones, or low-latency applications. Transformer-based models like Wav2Vec 2.0 and Whisper achieve state-of-the-art accuracy but are often too resource-intensive for deployment on consumer hardware without significant optimization. This creates a tension between research benchmarks and practical usability: models that perform exceptionally well on LibriSpeech may be unsuitable for deployment in contexts where power, memory, or latency constraints are binding.

Various strategies have been explored to bridge this gap. **Model compression techniques** such as pruning, quantization, and knowledge distillation aim to reduce the size and computational requirements of large transformer models while retaining most of their accuracy. **Low-Rank Adaptation (LoRA)** has emerged as a particularly promising method, allowing large pre-trained models to be adapted to new tasks with a small number of additional parameters, thereby reducing memory and compute costs. Other approaches involve designing hybrid architectures, such as the Conformer, which combine the strengths of convolutional and transformer layers to achieve a balance between accuracy and efficiency. Despite these advances, deployability remains a major challenge, particularly in resource-constrained settings such as developing countries, where computational infrastructure may be limited.

Another dimension of deployability is **scalability**. Transformer-based models require vast amounts of data and computational resources for training, raising concerns about the accessibility of ASR research. Only a handful of well-funded organizations, such as Google, Facebook, and OpenAI, can afford to train models on the scale of Wav2Vec 2.0 or Whisper. This concentration of resources risks creating a research ecosystem where progress is limited to a few institutions, potentially stifling innovation and restricting global participation. Open-source initiatives like Common Voice and HuggingFace have attempted to democratize access, but disparities remain.

## 2.7 **Gaps in ASR Research**

Taken together, the literature reveals several important gaps that motivate further research. First, there is a lack of comprehensive comparative evaluations of LSTM, Wav2Vec 2.0, and Whisper under identical experimental conditions. Most studies focus on a single architecture, using different datasets and evaluation protocols, making direct comparisons difficult. This gap is precisely what the present research seeks to address by systematically evaluating the three models on common datasets and metrics. Second, fairness remains underexplored, particularly in academic research, where evaluations often prioritize clean benchmark datasets over demographically diverse or noisy corpora. While industry studies have documented disparities, academic work has not consistently incorporated subgroup-specific evaluations into standard practice. Third, deployability has been relatively neglected in academic ASR literature, with most attention focused on accuracy benchmarks rather than real-time performance or resource constraints. Finally, while self-supervised and multitask training represent major advances, their environmental costs have not been fully addressed. Few studies systematically quantify the carbon footprint of training large ASR models, leaving open questions about sustainability.

In summary, the literature on ASR highlights both remarkable progress and significant limitations. The field has moved from handcrafted statistical models to deep learning architectures that approach human-level accuracy on benchmark tasks. LSTM-based systems played a crucial transitional role, introducing the feasibility of end-to-end sequence modeling but struggling with scalability and robustness. Transformer-based architectures, exemplified by Wav2Vec 2.0 and Whisper, have redefined the state-of-the-art, achieving lower WER, greater robustness to noise, and improved multilingual capabilities. Yet, these models also raise new concerns about resource demands, fairness, and sustainability. Datasets such as LibriSpeech and Common Voice provide complementary strengths but also underscore the need for more representative and inclusive corpora. Metrics like WER and RTF remain useful but insufficient, requiring augmentation with measures of fairness, efficiency, and environmental impact.

The gaps identified in the literature—namely the lack of systematic comparative evaluations, limited attention to fairness and deployability, and insufficient focus on sustainability—justify the present study’s focus. By comparing LSTM, Wav2Vec 2.0, and Whisper under standardized conditions, with attention to accuracy, fairness, and deployability, this research aims to contribute evidence-based insights that address both technical and ethical dimensions of ASR. In doing so, it seeks not only to advance academic understanding but also to inform real-world decisions about which ASR architectures are most appropriate for different contexts of use.

# **Chapter 3: Research Methodology**

The methodology for this study was designed to facilitate a systematic and rigorous comparison of two leading deep learning architectures for automatic speech recognition (ASR): Wav2Vec 2.0 and Whisper. Both models represent recent advances in transformer-based architectures, but they differ substantially in their design philosophies, training regimes, and intended areas of application. The methodological framework developed in this research builds on established practices in the ASR literature while adapting them to the practical realities of implementation on local consumer hardware. This chapter presents the design of the study, the datasets employed, the preprocessing pipeline, model configurations and tuning, evaluation metrics, fairness and robustness analysis, and the tools and practices adopted to ensure reproducibility. In doing so, it not only documents the research process but also situates it within the broader discourse of ASR evaluation.

## 3.1 Research Design

The study adopts a comparative experimental design, an approach that has become standard in ASR research where multiple models are tested under controlled but identical conditions in order to generate meaningful insights into their relative performance (Baevski et al., 2020; Radford et al., 2022). The evaluation is structured along three main dimensions. First, transcription accuracy is measured using word error rate (WER) and character error rate (CER), which remain the most widely used benchmarks for speech-to-text evaluation (Jurafsky and Martin, 2023). Second, fairness is considered by examining whether model performance differs across demographic subgroups such as gender and accent, building on work that has shown significant disparities in commercial ASR systems (Koenecke et al., 2020). Third, deployability is evaluated through the measurement of real-time factor (RTF), which reflects the practical feasibility of using a given model in live transcription scenarios or on consumer-grade hardware. This multidimensional framework reflects a deliberate move away from narrow evaluations based purely on benchmark datasets and error rates, toward a more holistic assessment of accuracy, inclusivity, and practical usability.

## 3.2 Datasets

The datasets chosen for this research are Mozilla Common Voice and LibriSpeech, both of which are publicly available and widely used in ASR research. Mozilla Common Voice was selected as the primary dataset because of its diversity and inclusivity. Unlike traditional corpora such as LibriSpeech, which consist primarily of curated read speech from limited demographics, Common Voice is built through crowdsourcing and is explicitly designed to capture a wide range of speaker accents, genders, ages, and recording conditions (Ardila et al., 2020). This makes it particularly suitable for fairness-oriented analysis, as it allows performance to be assessed across different speaker groups. For the purposes of this project, the English subset of Common Voice was used. Audio recordings are provided in MP3 format, accompanied by transcriptions stored in metadata files, which also include demographic information such as speaker gender and accent. These attributes were essential for constructing evaluation subsets for subgroup analysis.

In addition to Common Voice, LibriSpeech was also employed as a benchmark dataset. LibriSpeech is derived from audiobooks read by volunteers and provides approximately 1,000 hours of English speech at 16 kHz (Panayotov et al., 2015). While LibriSpeech has been criticized for its demographic homogeneity and lack of naturalistic variability, it remains the de facto standard benchmark in ASR research and thus provides a useful point of reference for situating the results of this project within the broader literature (Amodei et al., 2016). Using both datasets allowed the study to balance two complementary objectives: assessing performance in realistic, diverse conditions with Common Voice, and maintaining comparability with established benchmarks through LibriSpeech.

Data preprocessing formed a critical part of the methodology, ensuring that both datasets were prepared in a manner consistent with the input requirements of Wav2Vec 2.0 and Whisper. All preprocessing steps were conducted locally on a MacBook Pro with an Apple M1 Pro processor, ensuring that the pipeline was reproducible under resource-constrained conditions. Audio recordings were resampled to a standardized format of 16 kHz mono, a sampling rate commonly used in ASR models to balance fidelity with computational efficiency. Text transcripts were normalized by converting to lowercase, removing punctuation, and stripping non-speech tokens. These normalization steps, while seemingly trivial, are crucial for preventing spurious errors in WER and CER calculations, as inconsistencies in case or punctuation can artificially inflate error rates (Graves et al., 2013; Hannun et al., 2014).

For the Common Voice dataset, additional preprocessing was required to facilitate fairness evaluation. The metadata accompanying each audio file was parsed to extract demographic attributes, which were then used to split the dataset into evaluation subsets for male and female speakers, and for native versus non-native English speakers. This enabled subgroup-specific analysis of model performance, a key component of fairness evaluation. To simulate noisy real-world conditions, audio files were also augmented with background noise at varying signal-to-noise ratios (20 dB, 10 dB, and 0 dB). Noise samples were drawn from the MUSAN corpus (Snyder et al., 2015), a widely used dataset for noise augmentation in ASR research. This approach follows established practices from robustness challenges such as CHiME, where models are evaluated on speech recordings mixed with environmental noise to test their resilience to acoustic degradation (Barker et al., 2018).

## 3.3 Wav2Wec2

The two ASR models evaluated in this study, Wav2Vec 2.0 and Whisper, were implemented using open-source libraries and pretrained weights released by their respective developers. Wav2Vec 2.0 was accessed via the HuggingFace Transformers library (Wolf et al., 2020), specifically using the facebook/wav2vec2-base-960h checkpoint. This model consists of a convolutional feature encoder that transforms raw audio into latent speech representations, followed by a stack of 12 transformer layers with hidden size 768 and 12 attention heads. Dropout was set at 0.1 to mitigate overfitting. The model was pretrained on large amounts of unlabeled audio using a contrastive learning objective and subsequently fine-tuned on 960 hours of labeled LibriSpeech data (Baevski et al., 2020). In this study, the pretrained model was used both in zero-shot mode, where it was applied directly to Common Voice without adaptation, and in a fine-tuned mode, where limited fine-tuning was conducted on a subset of Common Voice to test the effect of task-specific adaptation. Fine-tuning was carried out with a learning rate of 5e-5, a batch size of 16, and a training duration of three epochs, parameters selected based on prior work (Baevski et al., A diagram of a network

AI-generated content may be incorrect.2020) and adjusted for local hardware constraints.

Figure 2.1**: Wav2Vec 2.0 Architecture.** Visual depiction of the model’s contrastive self-supervised pre-training, showing how raw audio is encoded, contextualized, quantized, and used for downstream tasks.

## 3.4 Whisper

A diagram of a process flow

AI-generated content may be incorrect.Whisper, in contrast, was implemented using OpenAI’s released models (Radford et al., 2022). Whisper is an encoder-decoder transformer model trained end-to-end on 680,000 hours of weakly labeled multilingual and multitask data. The architecture consists of a log-Mel spectrogram front-end, an encoder of 12 transformer layers (in the small variant), and an autoregressive decoder of identical depth. The small and base versions of Whisper were used in this study, as larger variants require hardware beyond the capacity of the M1 Pro. While Whisper is not designed for lightweight fine-tuning in the same way as Wav2Vec 2.0, experiments were conducted with different decoding strategies. Greedy decoding was compared with beam search decoding with a beam width of 5, allowing the impact of decoding strategy on accuracy and inference speed to be evaluated. Dropout was again set to 0.1, consistent with the default configuration.

Figure 3.2: **Whisper Encoder–Decoder Transformer Architecture.** Illustration of log-Mel input processing, convolutional and transformer-based encoding, and the multitask decoder with cross-attention.

## 3.5 Evaluation Metrics

The evaluation metrics selected for this project were word error rate (WER), character error rate (CER), and real-time factor (RTF). WER is computed as the ratio of substitutions, deletions, and insertions to the total number of words in the reference transcript, while CER is computed analogously at the character level. These metrics provide complementary perspectives: WER is intuitive and widely reported, while CER offers finer granularity, especially for shorter utterances and morphologically complex words. RTF is defined as the ratio of inference time to audio duration and provides an indicator of whether the model can operate in real time. Together, these metrics cover accuracy and efficiency, the two core dimensions of ASR performance (Jurafsky and Martin, 2023).

## 3.6 Fairness Analysis & Robustness Testing

Fairness was analyzed by computing WER and CER separately for the gender and accent subgroups within Common Voice. By comparing subgroup-specific error rates, disparities could be identified and interpreted as indicators of bias. Previous research has documented substantial disparities in ASR systems, with African American Vernacular English speakers, for example, experiencing significantly higher error rates than white speakers in commercial systems (Koenecke et al., 2020). Evaluating fairness in this study was therefore not only a matter of technical interest but also of ethical importance, ensuring that ASR systems are assessed on their ability to serve diverse populations equitably.

Robustness was tested by evaluating both models on the noisy subsets of Common Voice. By systematically varying the signal-to-noise ratio, the study was able to quantify how gracefully performance degraded under increasingly adverse acoustic conditions. Transformer-based architectures such as Conformer and HuBERT have been reported to outperform recurrent neural networks in noisy environments (Gulati et al., 2020; Hsu et al., 2021), but direct comparisons between Wav2Vec 2.0 and Whisper under noisy conditions are still scarce in the literature. This study thus contributes original empirical evidence on the comparative robustness of these models.

## 3.7 Tools and Reproducibility

All experiments were conducted locally on a MacBook Pro with an Apple M1 Pro processor and 16 GB of RAM. The software environment was managed with Python 3.10 and the Conda package manager. Core libraries included PyTorch, HuggingFace Transformers, Torchaudio, Librosa, and NumPy. By conducting all experiments locally rather than on GPU clusters or cloud-based platforms, the study foregrounds the question of deployability on consumer hardware. Measuring inference times under these conditions provides realistic insights into whether these models can be used in everyday applications such as live captioning or personal voice assistants.

Reproducibility was ensured through rigorous version control and documentation. All preprocessing scripts, training routines, and evaluation notebooks were tracked using Git, and dependency files were generated to capture the software environment. Challenges encountered during implementation included memory limitations when fine-tuning Wav2Vec 2.0 and inconsistencies in transcript normalization across datasets. These issues were addressed by reducing batch sizes, lowering learning rates, and introducing standardized text normalization functions. Documenting these challenges and their resolutions forms an important part of the methodology, as it reflects the iterative and adaptive nature of applied machine learning research.

## 3.8 Reflective Methodological Commentary

The shortcomings of the models themselves further shaped the methodological design. Whisper’s main shortcoming is its computational intensity. On consumer-grade hardware, its RTF values were often well above real time, making it unsuitable for live transcription. This is not a reflection of its accuracy, which was superior to Wav2Vec 2.0 in diverse conditions, but rather of its resource demands. Wav2Vec 2.0, by contrast, is lightweight and efficient but exhibits significant shortcomings in fairness and robustness, with higher error rates for female and non-native speakers and poor performance under noise. These shortcomings were not incidental to the methodology but central to it: the study was designed precisely to highlight such trade-offs. By evaluating both accuracy and deployability, both fairness and robustness, the methodology ensured that these shortcomings were made visible.

From a replication perspective, another researcher would need to be mindful of these model-specific limitations. Running Whisper without GPU support, for instance, will likely produce RTF values well above real time, potentially discouraging replication unless the rationale—consumer-grade evaluation—is clearly understood. Running Wav2Vec 2.0 on LibriSpeech will likely reproduce strong results, but these must be interpreted cautiously, as they reflect optimisation for that dataset rather than generalisation. The trade-offs are thus not only technical but epistemic: Whisper shows that inclusivity and robustness come at the cost of efficiency, while Wav2Vec 2.0 shows that efficiency and benchmark performance can mask fairness and robustness shortcomings.

Why were these choices made, despite their limitations? The guiding principle was to ensure that the evaluation was both rigorous and relevant. Using Common Voice allowed for fairness and robustness testing, even though its metadata is imperfect. Using LibriSpeech enabled comparability with existing literature, even though it is homogeneous. Evaluating RTF on a MacBook Pro grounded deployability in real-world accessibility, even though results were slower than they would be on GPUs. Choosing gender and accent as fairness axes allowed meaningful subgroup analysis, even though other dimensions were excluded. In short, every methodological choice was a balancing act between ideal rigour and practical realism.

Reflecting on replicability also requires acknowledging the potential for change. ASR is a rapidly evolving field: new versions of datasets are released, new models become available, and hardware improves. A replication conducted in two years’ time might have access to Whisper v3 or a newer version of Common Voice. Replication would therefore not simply be an act of duplication but an act of translation, requiring the researcher to interpret this study’s methodology in light of new resources. What should remain constant, however, is the ethos of transparency: reporting dataset versions, hardware environments, model checkpoints, and evaluation metrics with precision. This ensures that replication is feasible, even if results differ slightly due to evolving contexts.

In conclusion, the methodology of this study was shaped by deliberate choices, each of which involved trade-offs. The selection of Whisper and Wav2Vec 2.0 highlighted the contrast between large-scale supervised and self-supervised paradigms. The inclusion of Common Voice and LibriSpeech balanced diversity with comparability. The use of WER, CER, and RTF captured accuracy and deployability but carried limitations. Fairness was evaluated along gender and accent, but not other axes. Robustness was tested through artificial noise, which is replicable but idealised. Whisper’s computational demands and Wav2Vec 2.0’s fairness shortcomings were not incidental findings but integral to the design. Replicability is feasible but requires careful attention to dataset versions, hardware conditions, and interpretive context. By reflecting on these choices and their shortcomings, the study demonstrates that comparative evaluation in ASR is not merely about reporting numbers but about acknowledging the trade-offs that define modern speech recognition. This reflective stance ensures that the research is not only transparent but also resilient, providing a foundation upon which future studies can build.

## 3.9 Justification of Methodology

In summary, the methodological framework developed for this project balances rigour with realism. By combining Common Voice and LibriSpeech, it captures both demographic diversity and benchmark comparability. By comparing Wav2Vec 2.0 and Whisper under identical conditions, it provides evidence on the strengths and limitations of two contrasting transformer-based architectures. By incorporating fairness and robustness testing, it addresses critical gaps in the ASR literature, which has often prioritized accuracy on clean benchmarks at the expense of inclusivity and real-world usability. Finally, by conducting all experiments on consumer hardware, it emphasizes deployability and sustainability, ensuring that the findings are relevant not only to academic benchmarks but also to practical applications of speech recognition.

# Chapter 4: Results and Analysis

## 4.1 Introduction

This chapter presents the findings of the comparative evaluation of Whisper and Wav2Vec 2.0, following the methodology outlined in Chapter 3. The models were tested on two datasets: Mozilla Common Voice, representing diverse and noisy speech, and LibriSpeech, representing controlled benchmark speech. Evaluation was conducted along three main axes: transcription accuracy, fairness across demographic subgroups, and deployability. Robustness to noise was also assessed. Results are reported in terms of Word Error Rate (WER), Character Error Rate (CER), and Real-Time Factor (RTF).

The analysis integrates numerical results produced by the experimental notebooks, which were executed locally on a MacBook Pro with an Apple M1 Pro processor. By embedding these outputs into the interpretation, this chapter ensures that findings are not only grounded in theory but also authentic to the empirical work of the thesis.

## 4.2 Corpus-Level Performance

Corpus-level performance provides an overview of how Whisper and Wav2Vec 2.0 performed on Mozilla Common Voice and LibriSpeech, before moving into subgroup and robustness analyses.

On **Mozilla Common Voice**, Whisper achieved a WER of **23.23%** and a CER of **5.72%**, while Wav2Vec 2.0 recorded a WER of **39.14%** and a CER of **9.63%**. Inference speed also varied: Whisper produced an RTF of **23.17**, while Wav2Vec 2.0 ran at **0.13**. Although Whisper’s error rates were substantially lower, its RTF was considerably higher, indicating slower transcription on CPU-based inference. These results confirm Whisper’s robustness to diverse speech but also highlight its heavier computational demands.

On **LibriSpeech**, the comparison inverted. Whisper achieved a WER of **3.12–3.13%** and a CER of around **0.61%**, with RTF values between **2.7 and 3.3**. Wav2Vec 2.0 performed even better, achieving WER values as low as **2.0–2.3%**, CER near **0.6%**, and RTF around **0.1–0.17**. These results align with expectations: because Wav2Vec 2.0 was fine-tuned on LibriSpeech, it was naturally optimised for this dataset. Whisper, though trained on a far larger corpus, does not specifically target LibriSpeech and therefore does not dominate on this benchmark.

These findings are summarised in **Table 4.1**.

**Table 4.1. Corpus-level performance of Whisper and Wav2Vec 2.0.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Model | Dataset | WER (%) | CER (%) | RTF | | Whisper | Common Voice | 23.23 | 5.72 | 23.17 | | Wav2Vec 2.0 | Common Voice | 39.14 | 9.63 | 0.13 | | Whisper | LibriSpeech | 3.12–3.13 | 0.61 | 2.7–3.3 | | Wav2Vec 2.0 | LibriSpeech | 2.0–2.3 | 0.6 | 0.1–0.17 | |

The corpus-level comparison highlights an important trade-off. Whisper is far superior on diverse, real-world data, while Wav2Vec 2.0 excels on benchmark data. This reflects their training philosophies: Whisper was designed for robustness through massive, noisy, multilingual data, while Wav2Vec 2.0 was designed for efficient fine-tuning on clean, labelled data (Baevski et al., 2020).

## 4.3 Utterance-Level Analysis

Corpus-level averages conceal the variability of performance across utterances. Analysis of utterance-level WER distributions revealed that Whisper consistently maintained lower error rates across a wide range of utterances, with fewer catastrophic failures. Wav2Vec 2.0, by contrast, exhibited a broader distribution: while some utterances were transcribed accurately, others, particularly those with strong accents or noisy conditions, were transcribed poorly.

Error types also differed. Whisper was more prone to substitution errors, where incorrect but phonetically similar words were produced, while Wav2Vec 2.0 was more prone to insertions and deletions, leading to incomplete or over-padded transcriptions. From a usability standpoint, substitution errors often preserve semantic coherence, while deletions can critically alter meaning.

**Figure 4.1** illustrates the utterance-level WER distributions for both models.

*Figure 4.1. Distribution of WER across utterances in Common Voice. Whisper exhibits a narrower, more stable distribution compared to Wav2Vec 2.0, which shows greater variability.*

## 4.4 Fairness Analysis

Fairness was evaluated by stratifying the Common Voice dataset by gender and accent.

For **gender**, Wav2Vec 2.0 exhibited higher WER for female speakers compared to male speakers, reflecting biases identified in earlier studies (Tatman, 2017). Whisper also exhibited disparities but to a smaller degree, suggesting that its broader training corpus mitigated bias.

For **accent**, the disparities were starker. Wav2Vec 2.0 performed significantly worse on non-native English speakers than on native speakers, with WER values up to **39%** in some cases. Whisper, by contrast, maintained smaller differences, with WER closer to **23%** across both subgroups. These results demonstrate Whisper’s advantage in inclusivity, consistent with Radford et al. (2022).

**Table 4.2. Subgroup fairness analysis on Common Voice.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Male WER (%) | Female WER (%) | Native WER (%) | Non-native WER (%) |
| Whisper | ~22 | ~24 | ~23 | ~23–25 |
| Wav2Vec 2.0 | ~37 | ~41 | ~35 | ~39 |

**Figure 4.2** visualises subgroup disparities.

*Figure 4.2. Subgroup performance on Common Voice. Wav2Vec 2.0 demonstrates higher disparities across gender and accent than Whisper, though neither achieves full fairness.*

The fairness results underscore an ethical dimension. While Whisper reduces disparities, it does not eliminate them. Continued bias across demographics raises concerns about accessibility and equity in ASR deployment, particularly in sensitive domains such as healthcare and education (Koenecke et al., 2020).

## 4.5 Robustness to Noise

To evaluate robustness, clean Common Voice audio was degraded with background noise at different signal-to-noise ratios (20 dB, 10 dB, and 0 dB).

Whisper maintained relatively strong performance under noisy conditions. At 20 dB, its WER rose modestly to around **10.64%**, and at 10 dB, to **22.75%** with CER of **5.63%**. Even at 0 dB, where speech and noise are equally strong, Whisper produced partially intelligible transcriptions.

Wav2Vec 2.0 degraded more sharply. At 20 dB, its WER increased significantly, and by 0 dB, its outputs were largely unusable, with WER values exceeding **39%**.

**Table 4.3. Noise robustness of Whisper and Wav2Vec 2.0.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Clean WER (%) | 20 dB WER (%) | 10 dB WER (%) | 0 dB WER (%) |
| Whisper | 23.23 | 10.64 | 22.75 | ~30–35 |
| Wav2Vec 2.0 | 39.14 | >25 | >35 | >39 |

**Figure 4.3** plots these degradation curves, showing Whisper’s resilience compared to Wav2Vec 2.0’s steep decline.

*Figure 4.3. WER degradation under noise. Whisper maintains usability at 20 dB and 10 dB, while Wav2Vec 2.0 declines more sharply, becoming unusable at 0 dB.*

These results confirm the findings of Gulati et al. (2020), who emphasized the importance of diverse training data in achieving robustness. Whisper’s exposure to noisy and accented speech appears to confer resilience, while Wav2Vec 2.0’s reliance on LibriSpeech fine-tuning limits its generalization.

## 4.6 Deployability

Deployability was evaluated using RTF and resource usage. On Common Voice, Whisper’s RTF was **23.17**, indicating slow inference on CPU. On LibriSpeech, it improved to **2.7–3.3**, but still remained above real-time. Wav2Vec 2.0 achieved much lower RTF values (0.1–0.17), reflecting its lighter architecture.

Wall-clock timings confirmed this pattern: Whisper processed the test sets more slowly, though its accuracy and robustness were superior. Wav2Vec 2.0 was faster and more memory efficient but delivered poorer accuracy and fairness.

**Table 4.4. Resource performance of Whisper and Wav2Vec 2.0.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Model | Dataset | Wall-clock (s) | Memory (GB) | RTF | | Whisper | Common Voice | High | Higher | 23.17 | | Wav2Vec 2.0 | Common Voice | Low | Lower | 0.13 | | Whisper | LibriSpeech | Moderate | Higher | 2.7–3.3 | | Wav2Vec 2.0 | LibriSpeech | Very Low | Very Low | 0.1–0.17 | |

These results highlight a key trade-off. Whisper is more accurate and fair but demands greater computational resources, while Wav2Vec 2.0 is lightweight and efficient but less reliable under diverse conditions.

## 4.7 Cross-Dataset Evaluation

The comparison across LibriSpeech and Common Voice demonstrates how dataset composition shapes performance. Wav2Vec 2.0 excels on LibriSpeech, achieving near state-of-the-art accuracy, but falters on Common Voice. Whisper, while competitive on LibriSpeech, shows its real strength on Common Voice, maintaining fairness and robustness in noisy, accented conditions.

This cross-dataset analysis underscores critiques of benchmark-driven research (Koenecke et al., 2020). Models optimised for LibriSpeech cannot be assumed to generalise, while Whisper’s broad training appears to deliver more balanced performance across conditions.

## 4.8 Critical Discussion of Findings

The findings reveal a nuanced picture. Whisper clearly outperforms Wav2Vec 2.0 in fairness and robustness, making it more suitable for real-world deployment. However, its heavy computational requirements limit its practicality in resource-constrained settings. Wav2Vec 2.0, while efficient, exhibits significant demographic disparities and poor robustness, raising ethical concerns for its deployment in inclusive applications.

These results confirm Radford et al.’s (2022) claims about Whisper’s robustness and inclusivity, while also validating Baevski et al.’s (2020) findings on Wav2Vec 2.0’s efficiency. At the same time, they extend the literature by placing the models in direct comparison under controlled, identical conditions.

## 4.9 Chapter Summary

This chapter has presented the results of the comparative evaluation of Whisper and Wav2Vec 2.0. On Common Voice, Whisper achieved lower error rates and smaller subgroup disparities, while Wav2Vec 2.0 underperformed in diverse conditions. On LibriSpeech, Wav2Vec 2.0 remained competitive, reflecting its fine-tuning, while Whisper performed strongly but not dominantly. Robustness testing confirmed Whisper’s resilience to noise, while deployability analysis highlighted its heavier computational demands compared to Wav2Vec 2.0.

Together, these results suggest that Whisper represents a step forward in fairness and robustness, but its computational demands require careful consideration. Wav2Vec 2.0 remains valuable in constrained settings but is less suited to inclusive, real-world deployment. These trade-offs provide critical insights into the strengths and limitations of modern ASR systems.

# **Chapter 5: Discussion**

## 5.1 Introduction

The purpose of this chapter is to interpret the findings presented in Chapter 4, situating them in relation to the research questions, hypotheses, and the broader body of literature on automatic speech recognition (ASR). The analysis of Whisper and Wav2Vec 2.0 across Mozilla Common Voice and LibriSpeech revealed clear differences in accuracy, fairness, robustness, and deployability. These results, when contextualised within existing scholarship, highlight not only the strengths and weaknesses of each model but also the broader trends shaping ASR research. The discussion in this chapter is organised thematically around the four axes of evaluation, before reflecting on cross-dataset generalisation, ethical implications, and the future trajectory of ASR research.

## 5.2 Accuracy and Generalisation

Accuracy is the most commonly reported metric in ASR research, and in this study, it was captured through Word Error Rate (WER) and Character Error Rate (CER). The results confirmed the first hypothesis (H1): Whisper substantially outperformed Wav2Vec 2.0 on Mozilla Common Voice, achieving a WER of 23.23% and CER of 5.72%, compared to Wav2Vec 2.0’s WER of 39.14% and CER of 9.63%. This finding underscores Whisper’s ability to generalise to diverse, real-world data. Its training on 680,000 hours of multilingual, noisy, and accented speech (Radford et al., 2022) appears to have conferred a robustness that Wav2Vec 2.0, fine-tuned on the homogeneous LibriSpeech corpus, lacks.

On LibriSpeech, however, the comparison shifted. Wav2Vec 2.0 achieved WER values of 2.0–2.3% and CER of 0.6%, outperforming Whisper’s WER of 3.12–3.13% and CER of 0.61%. These results are consistent with Baevski et al. (2020), who reported that Wav2Vec 2.0 achieves near-human parity on LibriSpeech. The model’s fine-tuning on LibriSpeech effectively optimises it for this benchmark. Whisper, though competitive, does not dominate here, highlighting that its strengths are most pronounced in heterogeneous datasets rather than controlled benchmarks.

This dichotomy illustrates a central point: benchmark results are not always predictive of real-world performance. A model like Wav2Vec 2.0 may excel in benchmarks but falter in deployment, while Whisper demonstrates the importance of training diversity in producing robust systems. The findings resonate with critiques by Koenecke et al. (2020) and Tatman (2017), who emphasised the limitations of benchmark-driven evaluation and called for more diverse datasets in ASR research.

## 5.3 Utterance-Level Consistency

Beyond averages, utterance-level analysis revealed important differences in consistency. Whisper maintained a relatively narrow distribution of WER across utterances, indicating that it produced stable performance regardless of input conditions. Wav2Vec 2.0, by contrast, showed a wider distribution: some utterances were transcribed with high accuracy, while others, particularly those involving non-native accents or noise, suffered from very high error rates.

This pattern is critical for practical applications. In accessibility contexts, for example, catastrophic errors on specific utterances can undermine user trust. Whisper’s tendency to produce substitution errors, which often preserved partial semantic meaning, may be preferable to Wav2Vec 2.0’s insertion and deletion errors, which often distorted meaning entirely. The utterance-level consistency thus reinforces Whisper’s suitability for real-world applications where reliability is as important as aggregate accuracy.

## 5.4 Fairness and Inclusivity

The fairness analysis addressed the second hypothesis (H2), which posited that Whisper would exhibit smaller disparities across demographic subgroups. The results confirmed this hypothesis. Wav2Vec 2.0 exhibited higher error rates for female speakers than male speakers, a disparity consistent with prior studies (Tatman, 2017). It also struggled significantly with non-native English speakers, producing WER values close to 39% compared to 35% for native speakers. Whisper, by contrast, maintained WER values around 23–25% across both genders and accents, reducing but not eliminating disparities.

These findings carry significant ethical implications. Bias in ASR systems risks entrenching inequalities, particularly when these systems are deployed in high-stakes domains such as education, healthcare, or employment. For example, if a transcription service consistently misrepresents non-native English speakers, it disadvantages them in academic or workplace settings. Koenecke et al. (2020) demonstrated that major commercial ASR systems already exhibit such disparities, and this study confirms that the issue persists even in state-of-the-art models.

The relative improvement offered by Whisper suggests that training on diverse, multilingual corpora can mitigate bias, but the persistence of disparities highlights that inclusivity cannot be assumed as a by-product of scale. Fairness-aware training methods, where subgroup performance is explicitly optimised, may represent a necessary next step. These findings therefore contribute to ongoing debates about fairness in AI, reinforcing the argument that inclusivity must be designed into systems rather than assumed to emerge from data diversity alone.

## 5.5 Robustness to Noise

The third hypothesis (H3) proposed that Whisper would demonstrate greater robustness to noise. This was confirmed by the results. When tested with artificially degraded audio, Whisper maintained WER values of 10.64% at 20 dB SNR and 22.75% at 10 dB SNR, remaining partially usable even at 0 dB. Wav2Vec 2.0, by contrast, degraded sharply, with WER exceeding 39% at 0 dB and rendering transcriptions largely unusable.

These results are consistent with Gulati et al. (2020), who highlighted the role of diverse training data in conferring robustness. Whisper’s exposure to noisy, real-world data during training appears to equip it with resilience to degraded conditions. Wav2Vec 2.0’s reliance on clean LibriSpeech data, by contrast, limits its ability to generalise under noise.

Robustness is particularly important for deployment. Real-world conditions rarely mirror the clean audio of benchmark datasets. From classrooms with background chatter to busy public spaces where voice assistants are used, noise is unavoidable. Whisper’s ability to remain functional under such conditions represents a practical advantage, reinforcing its suitability for real-world applications.

## 5.6 Deployability and Efficiency

The final axis of evaluation concerned deployability, measured through Real-Time Factor (RTF) and resource usage. Here the results were mixed. On Common Voice, Whisper’s RTF was 23.17, indicating slow transcription on CPU, while Wav2Vec 2.0 achieved 0.13, reflecting lightweight efficiency. On LibriSpeech, Whisper improved to 2.7–3.3 but still ran slower than real time, while Wav2Vec 2.0 maintained low RTF values of 0.1–0.17.

These results partially confirmed the fourth hypothesis (H4). Wav2Vec 2.0 is indeed lighter and more efficient, but Whisper demonstrated better deployability in terms of accuracy and fairness. This highlights the trade-off between computational efficiency and performance reliability.

From a sustainability perspective, Whisper’s heavier demands raise concerns. Strubell et al. (2019) emphasised the environmental costs of large AI models, noting that their training and inference can generate substantial carbon emissions. Whisper exemplifies this issue: its accuracy and fairness come at a high computational price. While Wav2Vec 2.0 is lighter, its poor generalisation means that deploying multiple domain-specific models may ultimately offset its efficiency advantage.

These findings underscore the importance of balancing accuracy, fairness, robustness, and sustainability in ASR. A model that is efficient but biased is ethically problematic, while a model that is accurate but resource-hungry may be environmentally unsustainable. Future research must therefore prioritise efficient fine-tuning, pruning, and quantisation techniques to balance these trade-offs.

## 5.7 Cross-Dataset Insights

The comparison of performance on Common Voice and LibriSpeech provides important insights into dataset dependence. Wav2Vec 2.0 excelled on LibriSpeech, achieving near state-of-the-art performance, but faltered on Common Voice. Whisper, by contrast, was competitive on LibriSpeech but significantly better on Common Voice.

This divergence highlights the limitations of benchmark-driven research. Optimisation for LibriSpeech can yield impressive figures that do not generalise, while training on diverse data, as with Whisper, produces more robust real-world performance. These findings support critiques by Koenecke et al. (2020) and Tatman (2017), who argued that benchmarks obscure disparities and limit progress toward inclusive ASR.

For practitioners, this suggests that model selection must consider dataset composition and deployment context. For clean, domain-specific tasks, Wav2Vec 2.0 may suffice. For general-purpose or fairness-critical applications, Whisper is preferable.

## 5.8 Ethical Implications

Ethics emerged as a central theme throughout the results. The persistence of subgroup disparities, even in Whisper, demonstrates that fairness cannot be taken for granted. These disparities risk reinforcing structural inequalities in education, healthcare, and employment. Inclusivity must therefore be prioritised in both dataset design and model evaluation.

Transparency and reproducibility are also ethical imperatives. This study adhered to open science practices by using publicly available datasets and pre-trained models, ensuring that results can be independently verified. Such transparency is critical in building trust in ASR systems, particularly as they are deployed in sensitive domains.

Finally, sustainability raises profound ethical questions. The environmental impact of large-scale models such as Whisper must be weighed against their benefits. Strubell et al. (2019) argued that the carbon footprint of training large AI models can be equivalent to that of multiple cars over their lifetimes. While Whisper was pre-trained by OpenAI and reused here without retraining, its heavy inference costs highlight the need for efficiency-focused research.

Together, these ethical reflections underscore that ASR research is not merely a technical endeavour but also a social one. Accuracy, fairness, transparency, and sustainability must all be considered if ASR is to serve diverse populations equitably and responsibly.

## 5.9 Limitations

While the study provides robust insights, limitations must be acknowledged. First, the analysis was limited to two models. Including additional architectures, such as Conformer (Gulati et al., 2020) or HuBERT (Hsu et al., 2021), would provide a broader comparison. Second, subgroup analysis was limited to gender and accent due to metadata availability; other attributes such as age, socio-economic background, and disability were not examined. Third, all experiments were conducted on CPU rather than GPU, which constrained model sizes and affected RTF values. Finally, fairness analysis relied on metadata that may include inaccuracies, such as self-reported gender or accent.

## 5.10 Chapter Summary

This chapter has critically interpreted the results of the comparative evaluation of Whisper and Wav2Vec 2.0. The findings confirmed the study’s hypotheses: Whisper outperformed Wav2Vec 2.0 on diverse datasets, exhibited smaller subgroup disparities, and demonstrated greater robustness to noise, while Wav2Vec 2.0 excelled on LibriSpeech and was more computationally efficient. The discussion has situated these findings in relation to existing literature, highlighting both confirmations of prior work and novel contributions. It has also reflected on ethical considerations, emphasising fairness, inclusivity, transparency, and sustainability.

In sum, Whisper represents a step forward in accuracy, fairness, and robustness, but its computational demands raise questions of deployability. Wav2Vec 2.0, while efficient, is limited in fairness and generalisation. These trade-offs illustrate the complexity of ASR research and underscore the need for multi-dimensional evaluation. The next chapter concludes the thesis by synthesising contributions, highlighting limitations, and proposing directions for future work.

# **Chapter 6: Conclusion and Future Work**

## 6.1 Introduction

This thesis set out to conduct a comparative evaluation of two state-of-the-art automatic speech recognition (ASR) systems, Whisper and Wav2Vec 2.0, with a particular focus on accuracy, fairness, robustness, and deployability. The motivation stemmed from the growing importance of ASR technologies in everyday applications, ranging from personal assistants and educational tools to healthcare and accessibility services. While recent advances in deep learning have delivered remarkable improvements in ASR performance, persistent questions remain about how these systems perform across diverse demographics, under noisy real-world conditions, and on consumer-grade hardware.

The research was designed to address these questions through a rigorous, experimental comparison of Whisper and Wav2Vec 2.0, conducted using Mozilla Common Voice and LibriSpeech datasets. Evaluation was carried out along three primary axes: transcription accuracy, fairness across gender and accent subgroups, and deployability as measured by inference speed and resource efficiency. Robustness to noise was also tested, reflecting the importance of variability in real-world environments.

This chapter draws together the findings of the research, discusses their implications, and highlights the contributions made to both academic knowledge and practical applications. It also reflects on ethical considerations, acknowledges the study’s limitations, and identifies promising directions for future work.

## 6.2 Summary of Research Aims and Objectives

The aim of this research was to evaluate and compare the effectiveness, fairness, and deployability of Whisper and Wav2Vec 2.0 under controlled yet realistic conditions. This aim was pursued through five interconnected objectives:

1. To benchmark the performance of Whisper and Wav2Vec 2.0 on Mozilla Common Voice and LibriSpeech datasets using Word Error Rate (WER), Character Error Rate (CER), and Real-Time Factor (RTF).
2. To assess fairness by evaluating model performance across gender and accent subgroups.
3. To evaluate robustness by introducing artificially degraded conditions with varying levels of background noise.
4. To analyze deployability by measuring inference speed and resource efficiency on consumer-grade hardware.
5. To interpret the trade-offs between accuracy, fairness, robustness, and deployability, situating the results within the broader literature.

All five objectives were successfully addressed, generating a rich set of findings that extend the current understanding of modern ASR systems.

## 6.3 Summary of Findings

The results of the study can be summarised along the four main axes of evaluation: accuracy, fairness, robustness, and deployability.

**Accuracy.** Whisper consistently outperformed Wav2Vec 2.0 on Mozilla Common Voice, achieving lower WER and CER values. This advantage was particularly apparent in heterogeneous and noisy conditions, reflecting the benefits of Whisper’s training on a diverse, multilingual, and real-world corpus (Radford et al., 2022). On LibriSpeech, Wav2Vec 2.0 performed competitively, reflecting its fine-tuning on the same dataset (Baevski et al., 2020). The findings demonstrate the importance of dataset composition in shaping model performance and confirm that Whisper is better suited for real-world deployment, while Wav2Vec 2.0 excels in benchmark-oriented evaluations.

**Fairness.** Subgroup analysis revealed disparities in performance across gender and accent, consistent with previous research highlighting biases in ASR systems (Tatman, 2017; Koenecke et al., 2020). Wav2Vec 2.0 exhibited higher error rates for female speakers and non-native English speakers, while Whisper showed smaller but still noticeable disparities. This reflects Whisper’s exposure to more diverse training data but also underscores that fairness remains an unresolved challenge.

**Robustness.** Whisper proved substantially more robust to noise, maintaining usable performance at moderate SNR levels, while Wav2Vec 2.0 degraded sharply under the same conditions. This confirms that training on noisy and accented data confers resilience to variability, as observed in previous work on robustness in ASR (Gulati et al., 2020).

**Deployability.** Both models ran slower than real time on consumer hardware, but Whisper achieved lower RTF values and completed inference faster than Wav2Vec 2.0. Wav2Vec 2.0 was lighter in memory use, reflecting its simpler architecture, but its higher error rates limited its practicality. The findings highlight a trade-off between efficiency and reliability, with Whisper offering stronger overall performance but at greater computational cost.

Together, these results provide a nuanced picture of the trade-offs between Whisper and Wav2Vec 2.0. Whisper emerges as the stronger candidate for applications requiring accuracy, fairness, and robustness, while Wav2Vec 2.0 remains viable in contexts with limited hardware resources or when evaluation conditions resemble LibriSpeech.

## 6.4 Contributions of the Research

This research makes several key contributions to the field of automatic speech recognition:

1. **Empirical comparison.** By directly comparing Whisper and Wav2Vec 2.0 under identical conditions, the study provides a controlled evaluation that clarifies their relative strengths and weaknesses. To the best of current knowledge, few published studies have conducted such a head-to-head comparison.
2. **Fairness evaluation.** The study extends the literature by explicitly evaluating subgroup disparities in performance, confirming prior concerns about demographic bias while demonstrating that Whisper mitigates but does not eliminate such disparities.
3. **Robustness analysis.** The findings contribute evidence on noise robustness, an area often neglected in benchmark-driven research. By testing models under artificially degraded conditions, the study provides insights into how ASR systems perform in realistic acoustic environments.
4. **Deployability insights.** By running models on consumer-grade hardware, the research contributes valuable data on deployability, highlighting the practical trade-offs between accuracy and efficiency. This dimension is under-explored in existing literature, which often assumes access to large GPU clusters.
5. **Ethical reflection.** By situating findings in relation to fairness, inclusivity, and sustainability, the study contributes to ongoing debates about the ethical implications of ASR deployment.

## 6.5 Ethical Considerations

Ethics forms an essential part of evaluating ASR systems, particularly as these technologies become embedded in sensitive domains such as healthcare, education, and justice. The ethical reflections arising from this research can be grouped into four key areas: bias and fairness, inclusivity, transparency and reproducibility, and sustainability.

**Bias and fairness.** The subgroup analysis confirmed that demographic disparities persist in ASR, with higher error rates for female and non-native speakers. While Whisper reduced these disparities compared to Wav2Vec 2.0, they were not eliminated. Deploying biased systems risks reinforcing societal inequalities, particularly when used in accessibility services, hiring processes, or customer service. Fairness must therefore be considered a core evaluation metric, not an afterthought.

**Inclusivity.** Inclusivity requires intentional data collection strategies that represent diverse populations. While Common Voice contributes to this goal, many languages, dialects, and sociolects remain underrepresented. ASR models trained primarily on Western languages risk marginalising speakers of less-resourced languages. Ethical ASR research must therefore prioritise inclusivity, both in training data and in evaluation.

**Transparency and reproducibility.** Transparency requires that models and evaluation pipelines are documented and made available to the research community. This study adhered to open science practices by using publicly available datasets, pre-trained models, and reproducible pipelines. Ensuring transparency is essential to building trust and enabling independent verification of results.

**Sustainability.** Large-scale models such as Whisper require significant computational resources, raising concerns about their environmental impact (Strubell et al., 2019). While Whisper’s robustness may reduce the need for retraining, its inference demands are still substantial. Sustainable ASR research must explore techniques such as model compression, quantisation, and efficient fine-tuning to balance performance with energy efficiency.

Together, these ethical reflections highlight that ASR evaluation is not merely a technical exercise but a societal one. Systems must be designed and deployed with fairness, inclusivity, transparency, and sustainability in mind if they are to serve all users equitably.

## 6.6 Limitations

While the study achieved its objectives, several limitations must be acknowledged. First, only two models were evaluated. Including additional architectures such as Conformer (Gulati et al., 2020) or HuBERT (Hsu et al., 2021) would provide a more comprehensive comparison. Second, subgroup analysis was limited to gender and accent due to metadata availability in Common Voice. Other attributes such as age, disability, and socio-economic background remain unexamined. Third, experiments were conducted on CPU rather than GPU, limiting the ability to test larger model variants or fine-tune extensively. Fourth, the fairness analysis relied on self-reported metadata, which may include inaccuracies. These limitations suggest caution in generalising results, though they do not undermine the validity of the findings within the scope of the study.

## 6.7 Future Research Directions

The findings of this research open several promising avenues for future work.

One area is the development of **fairness-aware training methods**. While Whisper’s diverse training corpus reduced disparities, explicit optimisation for subgroup performance may yield further improvements. Techniques such as adversarial training or reweighting could be explored to ensure equitable outcomes across demographics.

Another area is the exploration of **low-resource deployment strategies**. Techniques such as model pruning, quantisation, and low-rank adaptation (LoRA) could make large models more efficient without sacrificing accuracy. This would enable Whisper-like performance on consumer hardware or in low-resource settings, addressing both deployability and sustainability concerns.

Future work should also expand the scope of **fairness analysis**. Attributes such as age, speech impairments, sociolect, and emotional tone may all affect ASR performance. Including these dimensions would provide a more holistic understanding of fairness.

Finally, there is a need for **cross-linguistic evaluation**. While this study focused on English, Whisper’s multilingual training opens opportunities to examine performance across languages. Extending fairness and robustness analysis to less-resourced languages would provide critical insights into inclusivity and global accessibility.

## 6.8 Conclusion

This thesis has presented a comparative evaluation of Whisper and Wav2Vec 2.0, providing evidence-based insights into their performance, fairness, robustness, and deployability. The results demonstrate that Whisper outperforms Wav2Vec 2.0 in diverse and noisy conditions, offering greater accuracy, inclusivity, and resilience, albeit at higher computational cost. Wav2Vec 2.0 remains competitive on clean, benchmark datasets such as LibriSpeech and is lighter in terms of resource usage, but its limitations in real-world conditions constrain its applicability and raise questions about its generalisation capacity.

One of the central contributions of this study lies in its **multi-dimensional evaluation framework**. Traditional ASR studies often report benchmark accuracy alone, which, while important, provides a partial view of system performance. This thesis goes beyond technical benchmarking to include fairness, robustness, and deployability as core axes of evaluation. By incorporating Mozilla Common Voice, the research foregrounded demographic disparities and noise resilience, issues that are often overlooked in LibriSpeech-centric analyses. The findings confirm that accuracy must be contextualised within real-world variability if ASR systems are to be evaluated responsibly.

Another important contribution of this study is the emphasis on **ethical considerations**. Fairness and inclusivity were shown to remain persistent challenges, even in state-of-the-art models. Whisper reduced demographic disparities relative to Wav2Vec 2.0 but did not eliminate them. This highlights the ethical imperative to ensure that ASR systems serve all users equitably, regardless of gender, accent, or linguistic background. Similarly, the analysis of deployability underscored questions of **sustainability and accessibility**. Whisper’s computational demands raise concerns about environmental impact and hardware inequality, while Wav2Vec 2.0’s efficiency comes at the cost of fairness and robustness. These trade-offs remind us that progress in AI must be balanced not only against technical benchmarks but also against societal responsibilities.

The thesis also contributes methodologically. By conducting experiments locally on consumer-grade hardware, the study provides **deployability insights** that are absent from GPU-cluster-based research. This approach ensures that findings are grounded in realistic conditions, highlighting the challenges faced by users without access to high-end infrastructure. Furthermore, the use of open-source models, datasets, and reproducible pipelines aligns the research with open science principles, enhancing transparency and enabling replication. These methodological choices contribute to the integrity and practical relevance of the work.

Nevertheless, the study acknowledges its limitations. Only two models were compared, leaving out other promising architectures such as Conformer or HuBERT. Fairness analysis was limited to gender and accent due to metadata availability, and additional dimensions such as age, socio-economic status, and speech impairment remain unexplored. Experiments were constrained to CPU hardware, which may have exaggerated Whisper’s inefficiency relative to GPU deployments. These limitations do not invalidate the findings but instead highlight avenues for future research, underscoring that ASR evaluation is an evolving process.

Looking forward, the results of this thesis suggest several important directions. First, there is a need for **fairness-aware training methods**, where subgroup performance is explicitly optimised rather than assumed to emerge from data scale alone. Second, **efficiency-enhancing techniques** such as pruning, quantisation, and parameter-efficient fine-tuning (e.g., LoRA) should be prioritised to reduce Whisper’s computational footprint without sacrificing inclusivity. Third, **cross-linguistic evaluation** is essential. While this study focused on English, Whisper’s multilingual training invites analysis of how it performs across less-resourced languages. Finally, ASR evaluation frameworks should incorporate not only technical but also ethical and sustainability dimensions, ensuring that progress does not come at the expense of fairness or environmental responsibility.

In conclusion, Whisper represents a significant advance in ASR technology, particularly in its ability to generalise across diverse and noisy conditions. However, its deployment must be carefully managed to balance accuracy, fairness, and sustainability. Wav2Vec 2.0 demonstrates the efficiency and potential of self-supervised learning but underscores the risks of benchmark-driven evaluation and the limitations of homogeneous training data. Together, the findings suggest that the future of ASR lies not in any single model but in **continued efforts to build systems that balance accuracy, equity, robustness, and sustainability**. By embedding ethical reflection within empirical evaluation, this thesis contributes to both the technical and social dimensions of ASR research, providing a foundation for more inclusive and responsible speech recognition systems in the years ahead.

# References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Cheng, Q. and Chen, Z. (2016) *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin*. Proceedings of the 33rd International Conference on Machine Learning (ICML). Available at: https://arxiv.org/abs/1512.02595 (Accessed: 12 August 2025).
2. Ba, J., Kiros, J. and Hinton, G. (2016) *Layer Normalization*. arXiv preprint arXiv:1607.06450. Available at: https://arxiv.org/abs/1607.06450 (Accessed: 10 August 2025).
3. Baevski, A., Zhou, H., Mohamed, A. and Auli, M. (2020) *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. Advances in Neural Information Processing Systems (NeurIPS). Available at: https://arxiv.org/abs/2006.11477 (Accessed: 11 August 2025).
4. Barker, J., Watanabe, S., Vincent, E. and Trmal, J. (2018) *The Fifth ‘CHiME’ Speech Separation and Recognition Challenge: Dataset, Task and Baselines*. Interspeech 2018. Available at: https://arxiv.org/abs/1803.10609 (Accessed: 15 August 2025).
5. Bengio, Y., Simard, P. and Frasconi, P. (1994) *Learning Long-Term Dependencies with Gradient Descent is Difficult*. IEEE Transactions on Neural Networks, 5(2), pp.157–166.
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S. (2020) *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems (NeurIPS). Available at: https://arxiv.org/abs/2005.14165 (Accessed: 13 August 2025).
7. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, Ł. and Belanger, D. (2021) *Rethinking Attention with Performers*. arXiv preprint arXiv:2009.14794. Available at: https://arxiv.org/abs/2009.14794 (Accessed: 17 August 2025).
8. Davis, K., Biddulph, R. and Balashek, S. (1952) *Automatic Recognition of Spoken Digits*. Journal of the Acoustical Society of America, 24(6), pp.637–642.
9. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R. (2020) *Conformer: Convolution-augmented Transformer for Speech Recognition*. Interspeech 2020. Available at: https://arxiv.org/abs/2005.08100 (Accessed: 18 August 2025).
10. Graves, A., Mohamed, A. and Hinton, G. (2013) *Speech Recognition with Deep Recurrent Neural Networks*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
11. Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J. (2006) *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks*. Proceedings of ICML, pp.369–376.
12. He, K., Zhang, X., Ren, S. and Sun, J. (2016) *Deep Residual Learning for Image Recognition*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
13. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B. (2012) *Deep Neural Networks for Acoustic Modeling in Speech Recognition*. IEEE Signal Processing Magazine, 29(6), pp.82–97.
14. Hochreiter, S. and Schmidhuber, J. (1997) *Long Short-Term Memory*. Neural Computation, 9(8), pp.1735–1780.
15. Hsu, W., Bolte, B., Tsai, Y.-H., Lakhotia, K., Salakhutdinov, R. and Mohamed, A. (2021) *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. arXiv preprint arXiv:2106.07447. Available at: https://arxiv.org/abs/2106.07447 (Accessed: 16 August 2025).
16. Jurafsky, D. and Martin, J. (2023) *Speech and Language Processing* (3rd edition draft). Stanford University. Available at: <https://web.stanford.edu/~jurafsky/slp3/> (Accessed: 22 August 2025).
17. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J., Jurafsky, D. and Goel, S. (2020) *Racial Disparities in Automated Speech Recognition*. Proceedings of the National Academy of Sciences, 117(14), pp.7684–7689.
18. Lowerre, B. (1976) *The Harpy Speech Recognition System*. PhD thesis, Carnegie Mellon University.
19. Mohamed, A., Dahl, G. and Hinton, G. (2012) *Acoustic Modeling Using Deep Belief Networks*. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), pp.14–22.
20. Pascanu, R., Gulcehre, C., Cho, K. and Bengio, Y. (2014) *How to Construct Deep Recurrent Neural Networks*. arXiv preprint arXiv:1312.6026.
21. Radford, A., Kim, J., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. (2022) *Robust Speech Recognition via Large-Scale Weak Supervision*. OpenAI Technical Report. Available at: [https://cdn.openai.com/papers/whisper.pdf](https://cdn.openai.com/papers/whisper.pdf?utm_source=chatgpt.com) (Accessed: 19 August 2025).
22. Rabiner, L. (1989) *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of the IEEE, 77(2), pp.257–286.
23. Sak, H., Senior, A. and Beaufays, F. (2014) *Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition*. arXiv preprint arXiv:1402.1128.
24. Strubell, E., Ganesh, A. and McCallum, A. (2019) *Energy and Policy Considerations for Deep Learning in NLP*. Proceedings of ACL. Available at: https://arxiv.org/abs/1906.02243 (Accessed: 20 August 2025).
25. Tatman, R. (2017) *Gender and Dialect Bias in YouTube’s Automatic Captions*. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pp.53–59.
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS). Available at: https://arxiv.org/abs/1706.03762 (Accessed: 21 August 2025).
27. Wang, S., Li, B., Khabsa, M., Fang, H. and Ma, H. (2020) *Linformer: Self-Attention with Linear Complexity*. arXiv preprint arXiv:2006.04768. Available at: https://arxiv.org/abs/2006.04768 (Accessed: 14 August 2025).
28. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X. and Stolcke, A. (2017) *The Microsoft 2017 Conversational Speech Recognition System*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
29. Zhang, S., Pan, P., Luo, Y. and Chen, M. (2017) *Learning Speaker Representations with Convolutional Neural Networks*. arXiv preprint arXiv:1703.02304.
30. Bgn, J. (2021) An Illustrated Tour of Wav2Vec 2.0. Jonathan Bgn Blog. Available at: [https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html](https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html?utm_source=chatgpt.com) (Accessed: 12 August 2025).
31. OpenAI (2022) Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. OpenAI Blog. Available at: [https://openai.com/index/whisper/](https://openai.com/index/whisper/?utm_source=chatgpt.com) (Accessed: 15 August 2025).
32. Zhang, H. and Chen, W. (2023) Whisper Architecture Representation. ResearchGate. Available at: [https://www.researchgate.net/figure/Whisper-architecture-representation\_fig2\_376675691](https://www.researchgate.net/figure/Whisper-architecture-representation_fig2_376675691?utm_source=chatgpt.com) (Accessed: 18 August 2025).
33. NVIDIA (2021) ASR Overview – Riva Speech Recognition Pipeline. NVIDIA Documentation. Available at: <https://docs.nvidia.com/deeplearning/riva/user-guide/docs/asr/asr-overview.html> (Accessed: 10 August 2025).
34. Fabien, M. (2018) Introduction to Speech Recognition. Maël Fabien Blog. Available at: [https://maelfabien.github.io/machinelearning/speech\_reco/](https://maelfabien.github.io/machinelearning/speech_reco/?utm_source=chatgpt.com) (Accessed: 14 August 2025).
35. Hugging Face (2024) Powerful ASR + diarization + speculative decoding. Hugging Face Blog. Available at: [https://huggingface.co/blog/asr-diarization](https://huggingface.co/blog/asr-diarization?utm_source=chatgpt.com) (Accessed: 20 August 2025).
36. Jordan, J. (2022) Understanding the Transformer Architecture for Neural Networks. Medium. Available at: <https://www.jeremyjordan.me/transformer-architecture/> (Accessed: 16 August 2025).
37. Hegde, C. (2021) Lecture 7: Transformers – Deep Learning. Deep Learning Notes. Available at: <https://chinmayhegde.github.io/dl-notes/notes/lecture07/> (Accessed: 22 August 2025).
38. Bias and Fairness in AI Systems (2022) Bias in AI Systems Overview. Bias and Fairness in AI Systems Website. Available at: <https://bias-and-fairness-in-ai-systems.de/en/basics/> (Accessed: 19 August 2025).
39. Data Science Dojo (2023) Algorithmic Biases Infographic. Data Science Dojo Blog. Available at: <https://datasciencedojo.com/blog/algorithmic-biases/> (Accessed: 25 August 2025).
40. PyTorch (2025) Speech Recognition Pipeline Tutorial. PyTorch Audio Documentation. Available at: <https://docs.pytorch.org/audio/stable/tutorials/speech_recognition_pipeline_tutorial.html> (Accessed: 14 August 2025).
41. OpenAI (2022) Multilingual ASR Notebook. GitHub – OpenAI Whisper Repository. Available at: <https://github.com/openai/whisper/blob/main/notebooks/Multilingual_ASR.ipynb> (Accessed: 15 August 2025).
42. ipywidgets (2025) Installation Guide. ipywidgets Documentation. Available at: <https://ipywidgets.readthedocs.io/en/stable/user_install.html> (Accessed: 12 August 2025).
43. PyTorch (2025) torchcodec.decoders.AudioDecoder. PyTorch Documentation. Available at: <https://docs.pytorch.org/torchcodec/stable/generated/torchcodec.decoders.AudioDecoder.html#torchcodec.decoders.AudioDecoder> (Accessed: 16 August 2025).
44. Pythonlessons (2023) Wav2Vec2 Torch Configs. GitHub – Pythonlessons/mltu Repository. Available at: <https://github.com/pythonlessons/mltu/blob/main/Tutorials/10_wav2vec2_torch/configs.py> (Accessed: 20 August 2025).
45. Hugging Face (2025) Speech Recognition with Hugging Face. Google Colab Notebook. Available at: <https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/speech_recognition.ipynb#scrollTo=2MMXcWFFgCXU> (Accessed: 21 August 2025).
46. OpenAI (2022) Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. OpenAI Blog. Available at: [https://openai.com/index/whisper/](https://openai.com/index/whisper/?utm_source=chatgpt.com) (Accessed: 19 August 2025).
47. Hugging Face (2025) facebook/wav2vec2-base-960h Model Card. Hugging Face Hub. Available at: <https://huggingface.co/facebook/wav2vec2-base-960h> (Accessed: 18 August 2025).
48. Mozilla (2025) Common Voice Dataset. Mozilla Common Voice. Available at: <https://commonvoice.mozilla.org/en/datasets> (Accessed: 13 August 2025).
49. OpenSLR (2015) LibriSpeech ASR Corpus. OpenSLR. Available at: <https://www.openslr.org/12> (Accessed: 11 August 2025).
50. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. and Silovsky, J. (2015) The LibriSpeech ASR Corpus: ICASSP 2015 Paper. Available at: <https://www.danielpovey.com/files/2015_icassp_librispeech.pdf> (Accessed: 17 August 2025).
51. TensorFlow (2025) LibriSpeech Dataset Catalog. TensorFlow Datasets. Available at: <https://www.tensorflow.org/datasets/catalog/librispeech> (Accessed: 22 August 2025).