Speech and Vision for Deaf and Blind on Edge Devices

https://ai4deafblind.org/

Abstract: This paper presents an edge-AI system for automatic real-time translation of spoken language into Braille, designed to improve accessibility for individuals with visual and dual sensory impairments. The proposed solution integrates an end-to-end speech processing and translation pipeline deployed on an NVIDIA Jetson Orin Nano platform [1], operating within a 15 W power envelope. Leveraging optimized deep learning models for speech recognition, language processing, and Braille encoding, the system achieves high accuracy and low latency on continuous audio streams. Unlike cloud-based approaches, the edge deployment ensures privacy preservation, reduced response time, and reliable operation in bandwidth-limited environments. Experimental results demonstrate that the Jetson-based implementation can sustain real-time streaming performance while maintaining translation fidelity comparable to higher-power systems. This work highlights the feasibility of deploying assistive AI technologies on low-power IoT hardware and provides a scalable framework for accessible communication applications.

1. Introduction

Communication is a fundamental human capability that underpins social participation, access to education, employment, and overall quality of life. For individuals who are both deaf and blind - commonly referred to as persons with deafblindness or dual sensory loss—communication presents a uniquely profound challenge. Deafblindness is not merely the coexistence of hearing and vision impairments; rather, it represents a distinct disability in which the loss of both senses compounds communication, information access, and environmental awareness in ways that cannot be mitigated by compensatory mechanisms from either modality alone.

Globally, deafblindness is estimated to affect between 0.2% and 2% of the world's population [2][3], corresponding to approximately 160 million individuals worldwide. Severe deafblindness alone is estimated at around 0.2% of the global population, while milder and moderate forms raise the upper bound significantly. Despite these numbers, deafblindness remains one of the least recognized and least understood disability groups among the general public and policymakers. Many individuals with deafblindness are misclassified solely as "blind" or "deaf" in official statistics, which obscures their distinct needs and leads to systematic under-representation in disability services and data collection.

This invisibility is particularly pronounced in low- and middle-income countries, where limited diagnostic infrastructure and inadequate disability reporting mechanisms result in large segments of the deafblind population going unidentified and unsupported.

Access to effective communication remains the most significant challenge for people with deafblindness. Traditional auditory and visual communication channels—such as spoken language, sign language, printed text, and digital media—are often inaccessible or only partially usable. As a result, individuals with deafblindness may rely on tactile communication methods, including tactile sign language, Braille, or assisted communication through trained interpreters, which are not always available in everyday environments.

These persistent communication barriers commonly lead to social isolation, reduced autonomy, and limited participation in community life. Empirical studies across multiple countries consistently report high levels of isolation, frustration, and dependency among deafblind individuals, even in regions with relatively advanced disability services.

The communication gap experienced by deafblind individuals has far-reaching socio-economic consequences. Research

indicates that persons with deafblindness are more likely to experience unemployment, poverty, and lower educational attainment compared to individuals with single sensory impairments. Barriers to education emerge early, as inaccessible instructional materials and the lack of specialized communication support hinder academic progress. In adulthood, these challenges translate into restricted employment opportunities and economic marginalization.

Moreover, societal misconceptions and low public awareness of deafblindness often result in attitudinal barriers, discrimination, and exclusion from mainstream social and economic activities, further reinforcing cycles of disadvantage [4].

Given the scale of deafblindness and the persistent shortage of trained human interpreters and assistive services, there is a critical need for scalable, low-power, and real-time assistive technologies that can facilitate independent communication. Advances in artificial intelligence, edge computing, and embedded systems offer new opportunities to bridge this gap. In particular, the automatic translation of speech into Braille using AI models deployed on power-efficient edge platforms holds promise for enabling real-time access to spoken information while preserving privacy and portability.

This work is motivated by the need to address the communication challenges of an often overlooked population and to demonstrate that high-accuracy, real-time assistive solutions can be deployed effectively on low-power IoT platforms.

While cloud-based speech recognition and translation systems have demonstrated high performance, their reliance on persistent network connectivity introduces critical limitations for assistive communication technologies intended for deafblind users. Dependence on cloud processing raises concerns related to latency, privacy, reliability, and accessibility, particularly in environments with limited or unreliable connectivity. For individuals who rely on assistive devices as a primary communication channel, interruptions in service or delayed responses can significantly degrade usability and independence.

To address these challenges, this work adopts an edge-AI approach, performing all speech recognition, language processing, and Braille translation locally on the device without the need for cloud interaction. Edge processing enables deterministic real-time performance, ensures data privacy by keeping sensitive audio information on-device, and allows the system to function autonomously in diverse real-world scenarios.

Portability is a critical design requirement for meaningful daily use. Assistive devices for deafblind individuals must be small in physical size, low in power consumption, and wearable or easily carried, enabling continuous operation throughout the day without reliance on large batteries or tethered power sources. Consequently, the target hardware platform must operate within a constrained power envelope while still supporting computationally intensive AI models. In this work, the system is designed to operate on a compact embedded platform with a power budget of approximately 15 W, balancing energy efficiency with computational capability.

Equally important, the system must achieve accuracy and robustness suitable for real-world deployment, not merely proof-of-concept performance. Errors in speech recognition or translation directly impact comprehension and trust, particularly for users who cannot easily cross-check spoken content via alternative sensory channels. Therefore, the proposed solution prioritizes high transcription fidelity, low latency, and stable performance on continuous audio streams, demonstrating that edge-based AI systems can meet practical accessibility requirements without sacrificing portability or usability.

By combining low-power embedded hardware with optimized AI models, this work aims to show that cloud-independent, real-time speech-to-Braille translation is feasible, reliable, and suitable for everyday assistive use, supporting greater autonomy and social participation for people with deafblindness.

2. Hardware Platform

To support real-time, cloud-independent speech-to-Braille translation in a portable form factor, this work employs the NVIDIA Jetson Orin Nano Super Developer Kit, a low-power embedded AI computing platform designed for edge inference applications. The platform delivers high computational capability within a compact footprint and constrained power envelope, making it well suited for wearable and assistive technologies.

2.1 System-on-Module Architecture

At the core of the platform is the Jetson Orin Nano 8 GB system-on-module (SoM). The module integrates a heterogeneous computing architecture that combines CPU, GPU, AI accelerators, memory, and I/O subsystems onto a single module, capable of efficient execution of multiple AI pipelines concurrently.

The processing subsystem consists of a 6-core Arm Cortex-A78AE 64-bit CPU, optimized for deterministic, and energy-efficient embedded workloads, paired with an NVIDIA Ampere-architecture GPU featuring 1,024 CUDA cores and 32 Tensor Cores. This configuration enables accelerated inference for deep neural networks used in speech recognition, natural language processing, and Braille translation tasks.

2.2 AI Compute Performance

The Jetson Orin Nano Super platform provides up to 67 trillion operations per second (TOPS) of AI performance, representing approximately a 1.7× performance increase over the original Jetson Orin Nano configuration through the introduction of the "Super" mode enabled via updated JetPack software. This level of performance enables real-time execution of modern deep learning models, including transformer-based architectures, on edge devices without cloud assistance.

Tensor cores within the Ampere GPU natively accelerate mixed-precision computation (INT8, FP16), which is critical for maintaining high inference throughput while minimizing latency and energy consumption. These characteristics are particularly relevant for continuous audio streaming and low-latency speech processing in assistive communication systems.

2.3 Memory and Bandwidth

The platform integrates 8 GB of 128-bit LPDDR5 memory shared between the CPU and GPU, offering a sustained memory bandwidth of approximately 102 GB/s in Super mode. This high memory bandwidth is essential for real-time streaming workloads, where concurrent audio buffering, neural network inference, and Braille encoding must operate without contention or bottlenecks. [nvidia.com], [cdn.sparkfun.com]

Support for external high-speed storage is provided through M.2 NVMe interfaces, allowing efficient access to model weights, language models, and auxiliary data without impacting run-time performance.

2.4 Power Envelope and Portability

A key advantage of the Jetson Orin Nano Super platform is its flexible power-performance scaling. The system supports multiple operating modes, including a 15 W power configuration, which is the target operating point for the proposed assistive device. At this power level, the platform maintains sufficient computational headroom to sustain real-time AI inference while remaining suitable for battery-powered, portable operation.

When higher performance is required for development or benchmarking, the platform can scale up to a maximum power mode of approximately 25 W; however, all real-time experiments in this work focus on the 15 W configuration to reflect realistic deployment constraints for wearable or handheld devices.

Physically, the developer kit occupies a compact footprint (approximately 100 mm × 80 mm), enabling integration into small enclosures without active cooling complexity, further supporting portability and unobtrusive usage in daily environments.

2.5 I/O and Peripheral Support

The Jetson Orin Nano Super reference carrier board provides a rich set of I/O interfaces suitable for multimodal assistive systems, including:

- Multiple USB 3.2 ports for audio interfaces, Braille displays, and peripherals
- Dual MIPI CSI-2 camera connectors for optional vision-based context awareness
- Gigabit Ethernet and Wi-Fi support for development and optional connectivity
- DisplayPort and GPIO expansion headers for debugging and system integration

These interfaces allow seamless integration with microphones, tactile output devices, and auxiliary sensors without requiring external controllers.

3. Software Platform and C++/CUDA

The platform runs the NVIDIA JetPack SDK [5], which includes CUDA, TensorRT, cuDNN, and optimized multimedia frameworks, enabling efficient development and deterministic, real-time AI inference on edge devices. All software components are executed locally on the Jetson Orin Nano Super to meet strict real-time and power constraints while preserving user privacy. This section outlines the Jetson software environment, system architecture, and the C++-based porting and optimization of the Whisper speech recognition model for edge deployment.

3.1 Jetson Software Stack

The system is built on the NVIDIA JetPack SDK, which provides a complete Linux-based development environment optimized for Jetson platforms. JetPack includes:

- Ubuntu-based Linux distribution
- CUDA for GPU acceleration
- cuDNN for deep neural network primitives
- TensorRT for high-performance inference optimization
- Multimedia APIs for efficient audio and streaming I/O

These components enable low-latency, hardware-accelerated execution of AI workloads while minimizing CPU overhead. In particular, TensorRT plays a key role in compiling neural network models into optimized execution graphs that leverage the Ampere GPU tensor cores for mixed-precision inference.

The application software follows modular pipeline architecture, where audio acquisition, speech recognition, text processing, and Braille encoding are implemented as loosely coupled components. This design allows individual modules to be optimized independently and facilitates future upgrades or model substitutions without architectural changes.

3.2 Overview of the Software Flow

Whisper is an automatic speech recognition (ASR) model based on an encoder–decoder transformer architecture [6], trained on a large and diverse multilingual speech dataset. The model is well suited for assistive applications due to its robustness to noise, accents, and varied speaking styles. However, the original implementation is designed for desktop or cloud-class hardware and relies heavily on Python, PyTorch, and dynamic memory allocation.

To enable deployment on the Jetson Orin Nano Super within a 15 W power envelope, the C++ based Whisper code based in used. The decoded text output from Whisper is passed directly to a table based text-to-Braille translator for displaying in Braille avoiding intermediate file storage or format conversions.

3.3 Rationale for C++ Implementation

Although many modern speech recognition models are developed and distributed using Python-based frameworks, Python introduces non-deterministic memory management and execution latency that are undesirable for real-time edge systems. To ensure predictable performance, reduced runtime overhead, and tighter integration with CUDA and TensorRT, we adopted a version of speech recognition model implemented in C++.

C++ provides:

- Lower inference latency through direct memory control
- Reduced CPU overhead and improved power efficiency

- Simplified integration with Jetson multimedia and GPU APIs
- Easier deployment as a standalone embedded application

For these reasons, the C++ ported OpenAI Whisper model was used for this project [7].

3.4 Real-Time Performance Considerations

Achieving real-time performance on a low-power edge device requires careful balancing of accuracy, latency, and computational load. Several techniques are employed to meet these constraints:

Model Size Selection: Smaller Whisper model variants are used to fit within memory limits while maintaining acceptable transcription accuracy for conversational speech.

Mixed-Precision Inference: FP16 precision is employed where supported, reducing memory bandwidth usage and improving GPU throughput.

Pipeline Parallelism: Audio capture, preprocessing, and inference are overlapped to minimize end-to-end latency.

Memory Reuse: Static buffer allocation is used to avoid runtime memory fragmentation and reduce allocation overhead.

In the target 15 W configuration, the resulting system sustains continuous audio transcription with latency suitable for real-time assistive use, without reliance on network connectivity or external compute resources.

3.6 Integration with Edge-Based Assistive Output

The C++ Whisper inference module serves as the front end of the full assistive communication pipeline. Its output text is immediately converted into Braille encoding and transmitted to a tactile display or Braille output device. By keeping the entire pipeline on-device, the system ensures:

Immediate response to spoken input

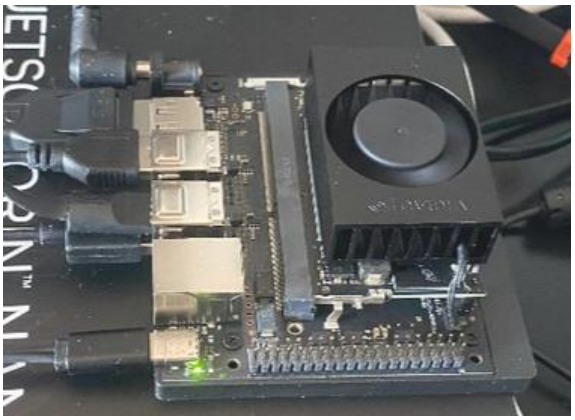
Privacy-preserving processing of sensitive conversations

Reliable operation in offline and mobile environments

This software architecture demonstrates that sophisticated transformer-based speech recognition models can be deployed effectively on compact, low-power embedded platforms, enabling practical edge-AI solutions for real-world accessibility applications.

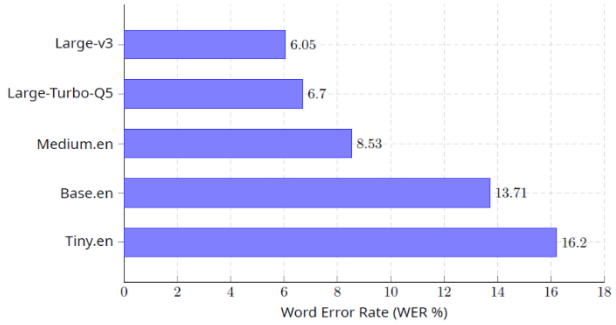
4. Results and Demonstrations

A Jetson Orin Nano Super is used for the demonstration. The current Jetpack is loaded onto a MicroSD card where the standard Ubuntu is the OS.



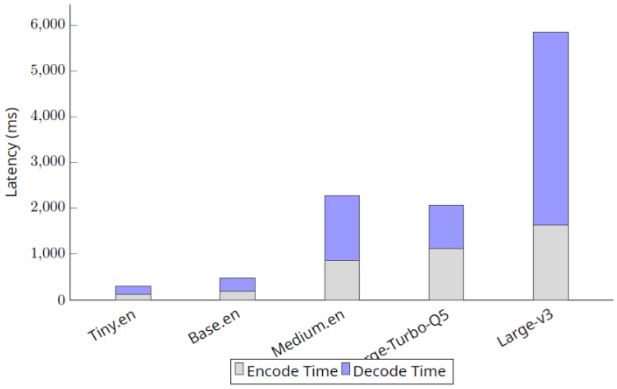
4.1 Accuracy with Whisper Models

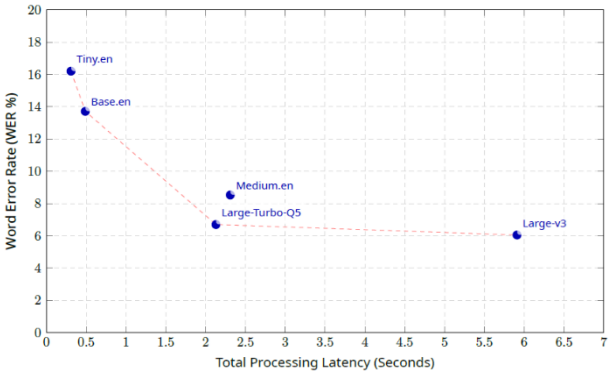
We experimented with different existing available Whisper models (sizes). They are Tiny, Base, Medium, Large, and Large with quantization (Turbo-Q5). For each model we run with 100 test audio clips from Common Language database. We used the Australian dataset [8] to test the accuracy. Below is the average accuracy.



4.2 Latency and Pareto Plots

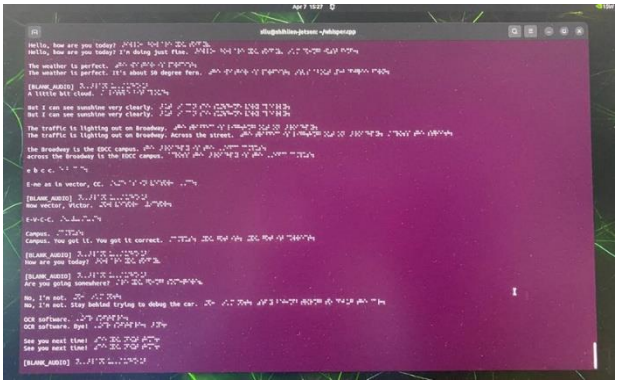
We also collected time used for each model. Below we plot the decode/encode time and the Pareto comparing accuracy against latency.





4.3 Actual deployment in Streaming Fashion

We also built a streaming version of the Whisper code and experimented with different model sizes. Subjective experience finds the base model the most usable model to deploy. Below we show a screen capture of the actual running of the program on the Jetson Orin Nano Super.



5. Discussion and Future Work

We found the approach we have taken usable. Jetson Orin is powerful enough to run multiple versions of the Whisper model and give reasonably acceptable real time experience. As edge hardware continues to improve, we believe a portable device is completely doable to assist and enhance the ability of deafblind.

We are currently working on a C/C++ version of the OCR model. Our goal is to give deafblind the capability of reading actual scene and translate the scene through a video camera (such as glasses) to Braille in real time.

6. Conclusion

This work demonstrates the practical feasibility of delivering real-time, privacy-preserving speech-to-Braille translation on a low-power edge platform, addressing a critical communication gap faced by the deafblind community. By deploying an end-to-end AI pipeline entirely on the NVIDIA Jetson Orin Nano within a 15 W power envelope, the system achieves a balance between accuracy, latency, and portability that is essential for everyday assistive use. The results confirm that modern transformer-based speech recognition models, when carefully optimized and implemented in C++/CUDA, can provide continuous, reliable performance without reliance on cloud connectivity. This edge-based approach not only safeguards sensitive personal conversations but also ensures consistent operation in bandwidth-limited or mobile environments, making it suitable for real-world deployment.

More broadly, this research underscores the potential of edge-AI technologies to empower individuals with deafblindness by enabling more independent, immediate access to spoken information through tactile channels. By translating speech directly into Braille in real time, the proposed system supports greater autonomy, social participation, and inclusion—core goals of assistive technology design. As edge hardware and embedded AI software continue to advance, solutions like the one presented here can scale beyond laboratory prototypes into practical devices that meaningfully enhance communication and quality of life for an often-overlooked population. Ultimately, this work contributes toward a future in which accessible, portable, and

energy-efficient AI systems help ensure that communication is a right enjoyed by all, regardless of sensory ability.

References

[1] NVIDIA, "Jetson Orin Nano Series Modules Data Sheet," DS-11105-001_v1.5, Dec. 2024 (http://www.plink-ai.com/Uploads/download/68259b8680fcf.pdf)

[2] 2021 National Deaf-Blind Child Count Report (https://deafblindprogram.wa.gov/wp-content/uploads/2023/01/2021_NCDB_Child_Count_FINALa.pdf)

[3] Second global report on the situation of persons with deafblindness (https://wfdb.eu/wp-content/uploads/2023/03/ENG_WFDB-2nd-Global-Report_-FINAL-V6.pdf)

[4] https://wfdb.eu/

[5] NVDIA, NVIDIA JetPack SDK versin 6.2 (https://developer.nvidia.com/embedded/jetpack)

[6] A. Radford, et. Al., "Robust Speech Recognition via Large-Scale Weak Supervision," Proceedings of the 40th International Conference on Machine Learning (ICML'23), 1182, Pages 28492 – 28518

[7] G. Gerganov, https://github.com/ggml-org/whisper.cpp

[8] R. Ardila et. Al., "Common Voice: A Massively-Multilingual Speech Corpus," Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4218–4222, Marseille, 11–16 May 2020