Received Date: 7th November, 2025
Revision Date: 15th December, 2025
Accepted Date: 26th January, 2026

Speech Recognition-Driven Language Learning: The Case of Tiny Talks for Nepali Using Wav2Vec2

Archana Mahat[1]*, Nush Ojha[2], Aarati Acharya[3], Anjali Sapkota[4], Sudeep Shakya[5]

[1] Dept. of Computer Engineering, Kathmandu Engineering College, Nepal, Email: archanamahat7@gmail.com
[2] Dept. of Computer Engineering, Kathmandu Engineering College, Nepal, Email: nushojha@gmail.com
[3] Dept. of Computer Engineering, Kathmandu Engineering College, Nepal, Email: aarati59acharya@gmail.com
[4] Dept. of Computer Engineering, Kathmandu Engineering College, Nepal, Email: sapkotaanjali123@gmail.com
[5]Assoc. Professor, Dept. of Computer Engineering, Kathmandu Engineering College, Nepal, Email: sudeep.shakya@kecktm.edu.np

Abstract— The TinyTalks project presents a mobile based system designed to support Nepali language learning among children aged 4 to 8. The application addresses the challenge of language preservation by offering an engaging and interactive learning environment tailored to young learners. TinyTalks integrates automatic speech recognition using a fine-tuned Wav2Vec2 model to evaluate children's pronunciation in real time. Nepali speech samples collected from multiple speakers were used to train the model through supervised learning, resulting in a validation accuracy of about 75 percent. The final system provides interactive lessons, pronunciation feedback and simple quizzes that support early language acquisition. The study demonstrates the feasibility of combining mobile technology and speech recognition to assist foundational Nepali language learning for young children.

Keywords — Nepali language learning, automatic speech recognition (ASR), speech accuracy, early language acquisition, language preservation

Introduction

In today's global environment many children grow up in multilingual settings where the language spoken at home differs from the language used in school or the wider community. This creates unique challenges for families who want their children to stay connected to their linguistic and cultural roots. As a result, there is increasing interest in approaches that can support early language learning outside traditional classroom settings. Digital learning tools, especially those that allow children to listen, speak and interact with language have become an important part of this shift. These tools offer flexible ways for young learners to practice their mother tongue even when they are far from their home country using technology as a valuable support for heritage language maintenance.

* *Corresponding Author*

Language learning in early childhood plays an important role in cognitive development. Research shows that young children have a natural ability to learn languages because their brains are highly flexible and sensitive to sound patterns during early years. This ability is especially strong between the ages of four and eight when children can learn new languages with ease and accuracy [1]. The Critical Period Hypothesis also supports this idea and explains that early childhood is the most effective time for natural language learning [2].

Interactive and meaningful activities help children learn languages more effectively. Methods like storytelling, songs and play allow children to understand new words and ideas in a natural and enjoyable way. These approaches support children's curiosity and help them engage with language through daily experiences [3].

For Nepali children living abroad, maintaining fluency in the Nepali language can be challenging. Limited exposure to Nepali around them and the lack of suitable learning resources often result in reduced confidence and weaker language skills. Digital platforms offer a practical solution by giving children access to engaging language materials that can be used anytime and from any place [4].

The use of technology in language learning has a long history. In the 1950s language laboratories introduced recorded native speech that allowed learners to listen and practice speaking independently [5]. Over time technology continued to grow and support new methods of learning making educational content more accessible to learners of different ages and backgrounds.

Digital storytelling is an effective approach for young children. It combines pictures, audio and simple videos to help learners connect ideas, remember new words and

express themselves more confidently. Research shows that digital storytelling encourages creativity and improves understanding and makes learning more meaningful for children [6].

This study examines how speech recognition can support early language learning for Nepali children living abroad with a focus on Nepali as a low resource language. Speech recognition systems can aid language learning by helping learners hear clear examples of words, receive consistent pronunciation input and interact with spoken language in a more natural way. These systems can also identify spoken words and provide feedback that encourages repeated exposure and practice, which is important for maintaining language skills among children who have limited contact with Nepali outside the home. Nepali has many regional variations and limited publicly available speech datasets, which creates challenges for developing reliable tools that can process Nepali speech correctly.

To address this gap, the research explores the use of Wav2Vec2 for recognizing Nepali speech at the character and word level. Wav2Vec2 is a transformer-based model that learns speech patterns directly from raw audio without relying on hand crafted features and has shown strong performance for automatic speech recognition across multiple languages [7]. This makes it suitable for low resource languages because the model can learn useful representations even from smaller datasets. The model is trained on selected Nepali speech recordings so it can better recognize how Nepali is spoken across different voices. This study focuses on developing a speech recognition-based system that can support Nepali language learning by providing pronunciation input and spoken word recognition. The aim is to build a system that offers a foundation for future tools designed to help children engage with Nepali and support ongoing research on Nepali language learning for communities abroad.

Materials and Methods

A. *System Design*

The development process began with early discussions with educators, parents and children to understand common pronunciation difficulties, learning preferences and the need for a tool that provides clear pronunciation support for Nepali. These insights guided the design of TinyTalks and shaped decisions related to system architecture, data collection and the speech recognition approach used in this study. The system architecture is organized into modular components that work together to support speech-based learning. It includes the mobile user interface, automatic speech recognition integration, adaptive quiz mechanisms and a central database for storing lesson content and user activity. These modules interact through a structured flow that delivers lessons, collects speech input and provides feedback to the learner. The overall system structure is illustrated in Fig. 1.

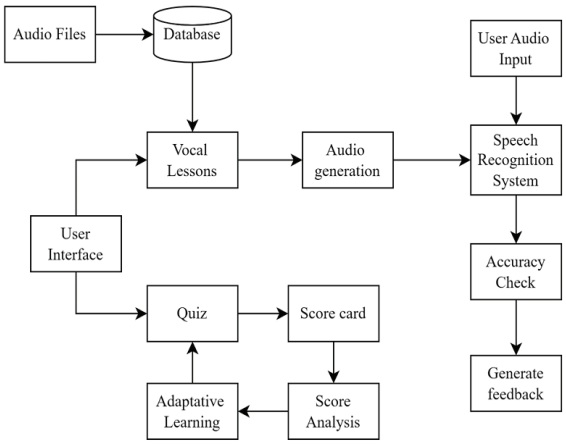

Fig.1 Block Diagram of the System

Fig. 1 presents the block diagram of the system. It shows how audio files, vocal lessons, user speech input, speech recognition, feedback generation and adaptive learning modules work together to support Nepali language learning.

B. *Data Collection and Cleaning*

Audio data were collected from GitHub repositories, Montessori students and family contributors. All audio files were standardized to the WAV format and cleaned using tools such as Clipchamp. To enhance model robustness, recordings were captured at slow, normal and fast speaking speeds. The preprocessing steps included converting all files to WAV format, trimming recordings to one-second clips and applying noise reduction techniques.

C. *Fine-Tuning the Wav2Vec2 Model*

1) Model Loading: The pre-trained *facebook/wav2vec2-base-960h* model was loaded along with its processor. This model works directly with raw audio and does not require manual feature extraction. The processor was responsible for converting waveforms into the format needed for the model to generate predictions.

2) Preprocessing: The Nepali audio files were loaded using Torchaudio and resampled to 16 kHz to match the input requirements of the model. Normalization was applied to maintain consistent amplitude levels across all samples. Feature extraction was performed directly from the waveform using Torchaudio, allowing the model to learn speech patterns without the use of handcrafted features. Fig.2 shows the original and resampled waveform of the character (शुन्य) demonstrating that key features were preserved after resampling while reducing the data size.

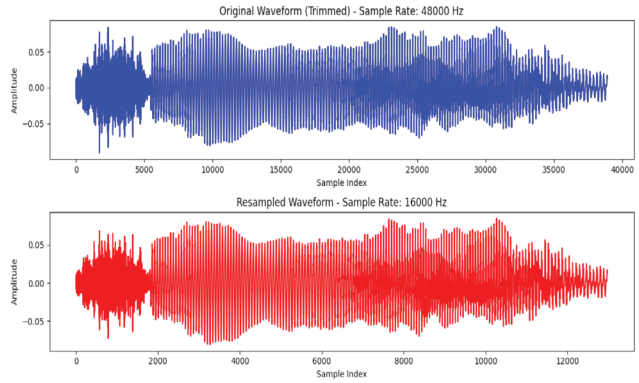

Fig.2 Plot of original and resampled audio (शुन्य)

3) Fine Tuning: Each audio sample was assigned a label based on its spoken content. Folder names were mapped to integer labels using the following mapping:

{'Eight': 0, 'Five': 1, 'Four': 2, 'Nine': 3, 'One': 4, 'Seven': 5, 'Six': 6, 'Three': 7, 'Two': 8, 'Zero': 9, 'अ': 10, 'अं': 11, 'अः': 12, 'आ': 13, 'इ': 14, 'ई': 15, 'उ': 16, 'ऊ': 17, 'ए': 18, 'ऐ': 19, 'ओ': 20, 'औ': 21, 'क': 22, 'ख': 23, 'ग': 24, 'घ': 25, 'ङ': 26}

4) Data Division: The dataset contained 2593 audio samples. It was divided into 2074 training samples and 519 validation samples using an 80 to 20 split. This division allowed the model to learn from the majority of the data while preserving a separate subset for evaluating generalization on unseen samples. This approach ensured a reliable assessment of model performance.

5) Training Arguments: The model was trained for 60 epochs with a batch size of 16. A learning rate of 3×10^{-5} was used with a weight decay of 0.01 to prevent overfitting. A linear scheduler-controlled updates to the learning rate during training. The best model checkpoint was selected based on validation accuracy to ensure that the most effective version was saved. A seed value of 42 was used to keep results consistent across training runs.

Results

The performance of the Wav2Vec2 model was evaluated using standard accuracy metrics and a separate validation dataset. The fine-tuned model achieved a training accuracy of 83.99 percent and a validation accuracy of 75.53 percent, showing that it was able to generalize to new data with reasonable effectiveness. The difference between training and validation accuracy indicates mild overfitting which is expected given the dataset size and pronunciation variability. The trained model and its processor were saved for future use and further refinement

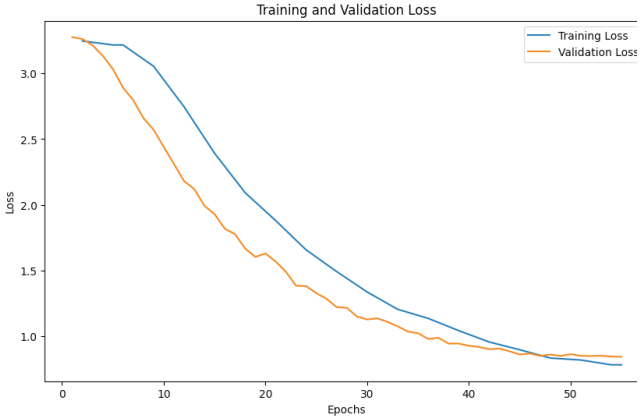

Fig.3. Training and Validation Accuracy

The accuracy curves in Fig. 3 show a steady upward trend during training, confirming that the model learned consistently across epochs. The gap of about 8.46 percent between training and validation accuracy indicates overfitting where the model performed better on the training data than on unseen samples. This behavior likely reflects the limited size of the dataset and uneven distribution of certain characters. Expanding the dataset and incorporating augmentation may help reduce this gap.

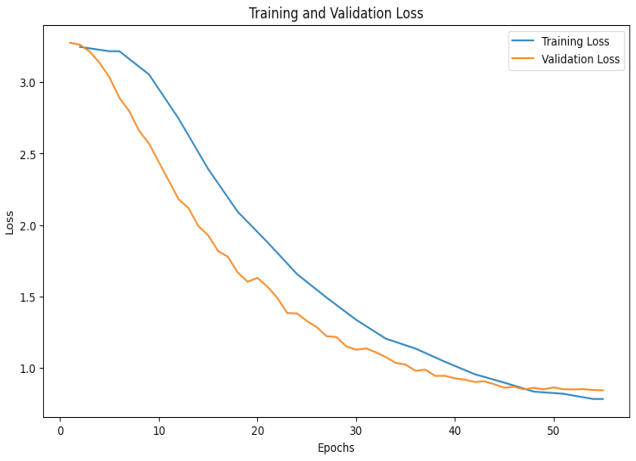

Fig.4. Training and Validation Loss

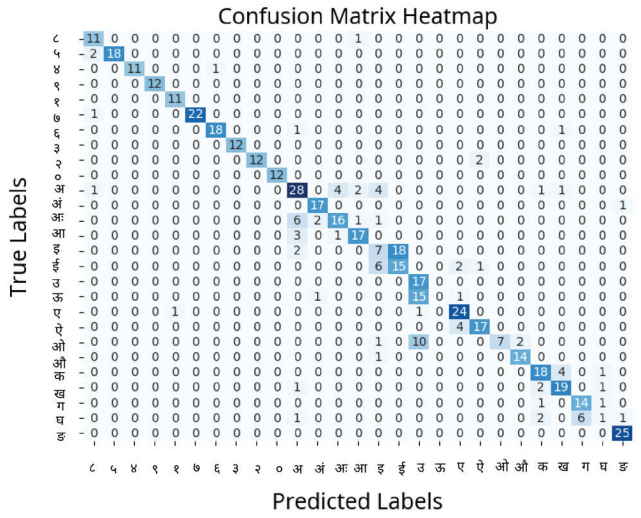

Fig.5. Confusion Matrix

The loss curves in Fig. 4 show a stable and continuous decrease from an initial value near 3.2. Both training and validation loss decline over time and level off around 0.9 at approximately the fiftieth epoch. This indicates that the model reached a stable convergence point and maintained reasonable generalization despite the observed overfitting.

The confusion matrix in Fig. 5 provides deeper insight into classification performance. The model accurately recognized most digits and several Nepali characters, especially when the pronunciations were clear and noise free. Misclassifications occurred mostly among phonetically similar sounds such as उ and ऊ or ग and घ. These errors highlight the sensitivity of the model to subtle acoustic differences which require more balanced and diverse training samples. Background noise and variations in speaking style also caused occasional prediction errors.

During evaluation several practical observations were made. The model showed strong accuracy for clearly spoken numbers but performance declined when speech contained accent shifts or pronunciation variations. Noisy environments had a noticeable impact on accuracy because background sounds interfered with feature extraction. Training required significant computation time due to the size of the dataset and the complexity of the Wav2Vec2 architecture. Signs of overfitting were consistent across accuracy curves and validation behavior indicating that the model learned detailed properties of the training data more strongly than general speech characteristics. These observations suggest that more diverse data with larger sample sizes and regularization strategies will further improve model robustness.

The results demonstrate that the fine-tuned Wav2Vec2 model provides effective recognition of Nepali speech within the TinyTalks application context. At the same time the findings point to important areas for improvement including expansion of the dataset for better representation of phonetically similar classes and techniques to enhance robustness in noisy or variable speaking conditions.

Discussion

The findings of this study show that transformer-based ASR models can be adapted for low resource Nepali speech recognition. Even with a small dataset the fine-tuned Wav2Vec2 model learned clear patterns for Nepali digits and characters. This suggests that end to end models can understand useful Nepali speech features even when only a limited amount of labeled data is available.

A key challenge seen in this study was pronunciation variation. Nepali speakers use different accents and speaking styles, and these differences caused the model to confuse certain characters. The confusion between similar sounds such as उ and ऊ or ग and घ shows that the model needs more balanced data for sounds that are close in pronunciation. These findings highlight a common issue in Nepali ASR where subtle sound differences are hard to capture when the dataset is small.

Background noise also affected recognition. Speech that included noise or unclear recording conditions reduced accuracy because important parts of the sound were harder for the model to pick up. This indicates the need for noise handling methods and more varied training samples collected in real settings.

In early language learning, these results support the use of ASR in tools like TinyTalks. Even with moderate accuracy the model can give children clear pronunciation input and respond to their speech in real time. This can make learning more engaging than static alphabet apps. Although this study did not test learning outcomes directly, the technical results show that ASR can add meaningful interaction to Nepali language learning tools for young users

Conclusion

This study demonstrates the feasibility of using a Wav2Vec2 based speech recognition model to support early Nepali language learning through the TinyTalks system. The model showed effective recognition of basic Nepali digits and characters, indicating that transformer architectures can extract meaningful speech patterns in low resource settings.

Integrating ASR with interactive learning features provides a foundation for speech driven educational tools aimed at young Nepali learners.

Acknowledgements

We express our sincere gratitude to the Department of Computer Engineering at Kathmandu Engineering College for providing the support needed to complete this project. We also acknowledge Budhanilkantha Kids Montessori Academy for assisting with audio data collection which played an important role in developing and evaluating the model. Finally, we thank our peers and contributors who participated in speech recording and offered valuable feedback throughout the project. Their support and collaboration greatly improved the overall quality of this work.

References

[1] P. K. Kuhl, "Early language acquisition: Cracking the speech code," *Nat. Rev. Neurosci.*, vol. 12, no. 4, pp. 283–295, Apr. 2011.

[2] H. Lenneberg, *Biological Foundations of Language*. New York, NY, USA: Wiley, 1967.

[3] J. Bruner, *Child's Talk: Learning to Use Language*. Oxford, U.K.: Oxford Univ. Press, 1983.

[4] Patel. (2024) "The evolution of online language learning," *Training Industry*. [Online]. Available: https://www.trainingindustry.com/articles/content-development/the-evolution-of-online-language-learning

[5] Webster. (n.d.) "A brief history of language learning technology: Pre computers." [Online]. Available: https://languagemuseum.org/a-brief-history-of-language-learning-technology-pre-computers

[6] H. A. Alismail, "Integrate digital storytelling in education," *J. Educ. Pract.*, vol. 6, no. 9, pp. 126–129, 2015.

[7] Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477