WJAETS

World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Research Article)

Check for updates

A study on the application of deep learning in Vietnamese speech recognition

Van Khoi Nguyen *

Faculty of Electrical and Electronic Engineering, University of Transport and Communications, HaNoi, Vietnam.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 2894–2898

Publication history: Received on 20 April 2025; revised on 27 May 2025; accepted on 30 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0877

Abstract

Speech recognition has become increasingly important in various real-world applications. However, Vietnamese presents unique linguistic challenges such as tones, syllabic structures, and complex morphology, which make speech recognition for this language significantly different from that of languages like English. In this paper, we propose a deep learning approach that combines Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks to recognize Vietnamese speech using the VIVOS dataset. The CNN component is employed to extract spatial features from audio spectrograms, while the BiLSTM captures the bidirectional temporal dependencies in speech signals. Experimental results show that the proposed CNN-BiLSTM model achieves a competitive Word Error Rate (WER) of 14.7%. These results highlight the potential of deep learning techniques in effectively recognizing tonal languages such as Vietnamese.

Keywords: Speech Recognition; Vietnamese; VIVOS; CNN; BiLSTM

1. Introduction

Automatic Speech Recognition (ASR) is one of the key fields in Artificial Intelligence (AI) and Natural Language Processing (NLP), enabling computers to automatically convert audio signals into text. This technology has made remarkable progress in recent years thanks to the rapid development of deep learning models, machine learning algorithms, and the increasing availability of high-quality training data. Modern speech recognition systems have achieved high accuracy across major languages such as English, Chinese, Spanish, and French.

However, for Vietnamese, the development of ASR systems continues to face significant challenges due to its unique linguistic characteristics. Vietnamese is a monosyllabic language with a complex tonal system, where a word's meaning can change entirely depending on its tone. Furthermore, regional dialect diversity (Northern, Central, Southern), speaking rates, and pronunciation variations across regions introduce additional difficulties in building a robust and accurate Vietnamese ASR system. Nevertheless, with the ongoing advancements in speech processing technologies, Vietnamese speech recognition is becoming a promising research area with many important practical applications.

Several studies have explored deep learning applications in Vietnamese ASR. The PhoWhisper model, proposed by the authors in [1], includes five versions with increasing training complexity. It was trained on the CMV-Vi dataset, the VIVOS dataset, and a proprietary dataset compiled from 26,000 speakers across 63 provinces and ethnic groups in Vietnam. The combined dataset was enhanced with environmental sound augmentation using Piczak data and the audiomentations library, resulting in a complete training set of 844 hours of audio. PhoWhisper is based on the multilingual Whisper model, fine-tuned using the Transformers library. The authors report strong performance compared to wav2vec2, with the largest version—PhoWhisper-large—achieving the lowest Word Error Rate (WER) score of 4.67%. Due to the challenges in collecting domain-specific speech data in healthcare, the authors in [2] introduced the VietMed dataset, comprising 16 hours of labeled medical speech, 1,000 hours of unlabeled medical-

* Corresponding author: Van Khoi Nguyen.

domain speech, and 1,200 hours of general Vietnamese speech encompassing various intonations and dialects. They also released two large-scale ASR models, w2v2-Viet and XLSR-53-Viet, along with a fine-tuned medical ASR model. These models achieved higher accuracy than standard models in the medical domain, with the XLSR-53-Viet model achieving the lowest WER at 29.6%. In [3], the authors proposed a high-quality Vietnamese speech corpus of 100.5 hours, used to train two prominent ASR models—Listen, Attend and Spell (LAS) and Speech-Transformer. The dataset includes a wide range of intonations (domestic and foreign), age groups, and topics, and underwent extensive preprocessing to remove noise. SpecAugment was used to improve model generalization. LAS and Speech-Transformer achieved WERs of 8.53% and 7.24%, respectively. Aiming to support accurate and fast communication in medical settings across language barriers, the authors in [4] introduced the HYKIST project, which applies ASR models and machine learning techniques to aid interpreters in medical dialogue. They evaluated models such as Transformer, MLP, HMM-based architectures, and ASR systems including wav2vec2 and XLSR-53. Through preprocessing, training strategies, and hyperparameter tuning, they emphasized the importance of diverse datasets and suggested better practices, such as using Kaiming initialization over Xavier for wav2vec2 models. In [5], the authors proposed techniques to improve the Whisper model's accuracy in low-resource languages like Vietnamese, particularly in specialized domains like military communication. By leveraging publicly available and domain-specific data, and applying hardware-efficient techniques like Low-Rank Adaptation (LoRA), they achieved a 20% improvement in WER and a 32% reduction in Character Error Rate (CER) across three Whisper variants: small, base, and tiny. The study in [6] proposed LingWav2Vec2, a linguistically enhanced ASR model combining language-aware modules with wav2vec2 for handling pronunciation errors. The model includes phoneme embeddings, sinusoidal positional encoding, and RMS normalization. With components like cross-attention, feed-forward, and pre-norm residual blocks, the model improves robustness. Evaluated on VLSP 2023's Vietnamese Mispronunciation Detection datasets, it achieved an F1-score of 59.68%, surpassing previous baselines by 9.72% while adding only 4.3 million parameters. The paper by Zhang et al. (2018) proposed a CNN-BiLSTM architecture to extract acoustic features and process bidirectional contextual information, achieving significantly reduced WER on the English TIMIT dataset. They showed that CNNs effectively learn spatial features from spectrograms, while BiLSTM captures context from both past and future time steps, improving sequence prediction accuracy [7]. Another study by Li et al. (2019) employed a similar combined model for Mandarin speech recognition, using Mel-spectrogram preprocessing and training with the CTC algorithm. This model not only improved accuracy but also enhanced generalization capability across different dialects [8]. Moreover, Nguyen et al. (2021) focused on Vietnamese speech recognition using CNN-BiLSTM. This research expanded the Vietnamese multilingual dataset and applied data augmentation techniques. The results demonstrated lower WER compared to traditional models, confirming the effectiveness of the CNN-BiLSTM architecture in handling the complex tonal characteristics of Vietnamese speech [9].

In this paper, we propose a CNN-BiLSTM hybrid model to optimize Vietnamese speech recognition performance. The model leverages CNN for feature extraction and BiLSTM for sequence processing to reduce error rates and improve accuracy. Experimental results demonstrate that the proposed model achieves a WER of 14.7%.

The remainder of this paper is organized as follows: Section 2 presents the methodology; Section 3 discusses experimental results; and Section 4 concludes the study.

2. Research methodology

2.1. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are widely applied in audio signal processing due to their ability to automatically extract features from input data without relying on manually designed feature engineering, as seen in traditional techniques. In this study, the CNN operates directly on spectrogram representations, where each convolutional layer acts as a filter to detect important patterns in the signal such as frequency characteristics and their temporal variations thereby enhancing recognition and classification capabilities. When processing spectrograms typically visualized as images with time on the x-axis, frequency on the y-axis, and pixel intensity representing amplitude CNNs can leverage their spatial pattern recognition capabilities to detect crucial spectral structures. As a result, CNNs have demonstrated superior performance in various critical applications such as Automatic Speech Recognition (ASR), sound source separation, emotion recognition from speech, environmental sound classification, anomaly detection in industrial sounds, speaker recognition, music analysis, and numerous other audio-related tasks. These advantages significantly improve the accuracy and generalization capability of modern audio signal processing systems.

2.2. Bidirectional Long Short-Term Memory (Bi-LSTM)

The Bidirectional Long Short-Term Memory (Bi-LSTM) model is an extension of the standard LSTM network, designed to learn and process information in both forward and backward temporal directions. This bidirectional structure enables the model to capture more comprehensive contextual information in time-series data, which is particularly useful for speech recognition tasks. In this study, the input audio is represented as a spectrogram—a time-dependent feature—so the use of Bi-LSTM allows the model to understand the relationships between time frames by incorporating information from both the past and the future. This enhances the model's ability to identify important acoustic patterns. One of Bi-LSTM's key advantages is its ability to retain long-term dependencies in audio signals. It overcomes the limitations of traditional RNNs by employing gating mechanisms, including the input gate, forget gate, and output gate, to regulate the flow of information being stored and passed through the network. This helps prevent the loss of critical information in long sequences. Thanks to these strengths, Bi-LSTM is widely used in various audio processing tasks such as ASR, speech separation, emotion recognition, environmental sound classification, speaker recognition, speech synthesis, and more.

2.3. CNN and Bi-LSTM Hybrid Model

Figure 1 illustrates the overall processing pipeline proposed in this paper. First, audio data from the VIVOS dataset is subjected to a preprocessing phase, where the continuous speech signal $x(t)$ is transformed into the frequency domain using the Short-Time Fourier Transform (STFT). This transformation divides the signal into overlapping windows and computes the frequency spectrum over time, resulting in a spectrogram representation $X \in R^{T\times F}$, where T is the number of time frames and F is the number of frequency features. Next, the spectrogram is fed into the CNN to extract spatial features from the data. The CNN employs filters $W \in R^{k_t\times k_f}$ that scan across the spectrogram to detect sound patterns, enabling the model to identify key features in the speech signal. The output of the CNN is then passed to a Bidirectional LSTM (Bi-LSTM) network, which captures long-term dependencies by considering both past and future information within the sequence. This helps the model better learn the temporal context of Vietnamese speech. Subsequently, the model maps the sequence of audio signals to a sequence of output characters using the Connectionist Temporal Classification (CTC) algorithm. CTC calculates the probability of all valid character sequences that can be formed from the model's predicted output sequence, as defined in Equation (1).

$$P(y|X) = \sum_{\pi\in B^{-1}(y)} P(\pi|X) \quad \text{..........(1)}$$

Here, B is the mapping function that removes repeated characters and blank tokens. CTC then optimizes the model by maximizing the probability of the correct character sequence given the training data. Finally, the model outputs a predicted text sequence corresponding to the content of the input speech signal.

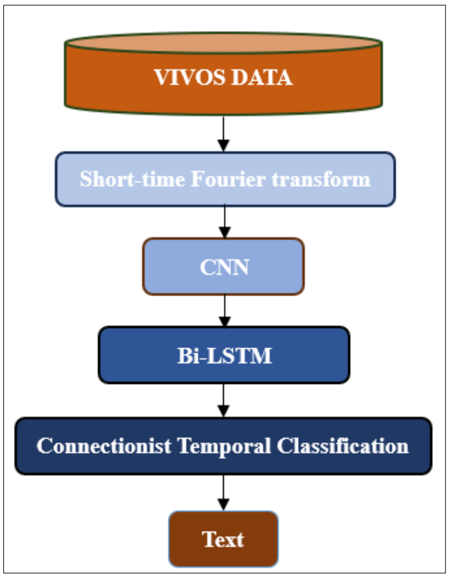
VIVOS DATA
Short-time Fourier transform
CNN
Bi-LSTM
Connectionist Temporal Classification
Text

Figure 1 Overview of the processing pipeline proposed in this study

3. Experimental results

3.1. Dataset

In this study, we utilized the VIVOS dataset for training and evaluating the model [10]. VIVOS is a Vietnamese speech corpus with a total duration of approximately 15 hours, consisting of 15,000 utterances from 116 speakers (58 male, 58 female) in the training set and 760 utterances from 20 speakers in the test set. The audio signals were standardized with a sampling rate of 16kHz and stored in 16-bit WAV format, ensuring high accuracy. The utterances were selected from natural Vietnamese text and represent a wide range of regional accents (Northern, Central, and Southern), enabling the model to learn language-specific characteristics of Vietnamese and improving its generalization ability when applied in real-world scenarios.

3.2. Experimental Results

This section presents the training and testing results of the speech recognition models. Training was conducted on a system with an Intel Xeon CPU 2.20GHz, NVIDIA Tesla T4 GPU with 16GB VRAM, 16GB RAM, and Python 3.6. The training parameters included a learning rate of 0.001, batch size of 32, and dropout rate of 0.3. The model was trained for 50 epochs on the VIVOS dataset.

To clearly demonstrate the effectiveness of the deep learning model, we used Word Error Rate (WER) as the evaluation metric. WER measures the discrepancy between the predicted transcription and the reference transcript and is defined as follows (2).

$$WER = \frac{S+D+I}{N} \ldots\ldots\ldots\ldots \quad (2)$$

Where, S is the number of substitution errors, D is the number of deletions, I is the number of insertions, N is the total number of words in the reference transcript. Figure 2 shows the WER values over 50 training epochs.

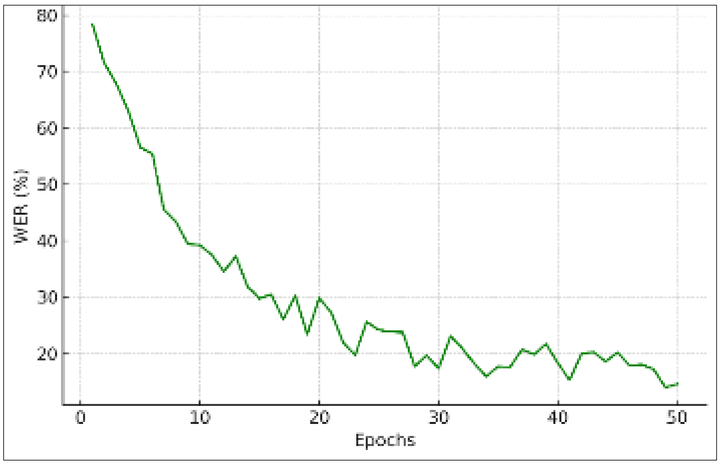

Figure 2 Word Error Rate over 50 training epochs

In this figure, the x-axis represents the number of training epochs, indicating the number of times the entire training set was passed through the model. The y-axis shows the WER as a percentage, reflecting the error rate in converting speech to text — the lower the value, the better the model's performance. In the early epochs, the WER was high (around 80%), indicating poor performance as the model had not yet been optimized. However, the WER rapidly decreased to below 40%, demonstrating that the model quickly learned important patterns from the input data and significantly improved recognition performance.

This stage reflects the model’s fast acquisition of speech patterns, phonetics, and the mapping between audio signals and corresponding text. After the 10th epoch, the rate of WER reduction slowed down, though it continued to decline with slight fluctuations. These fluctuations could result from data complexity, speaker variability, or optimization methods used during training. From the 20th epoch onward, WER continued to decrease at a slower pace, showing an overall downward trend as the model gradually converged, reaching a final WER of 14.7% by the end of training. The period from epoch 30 to 50 marks the model's convergence phase. Minor fluctuations in this range may be attributed to dataset characteristics, optimization techniques, or factors such as overfitting — where the model begins to memorize training data rather than learning generalized features. Nevertheless, WER maintained a stable downward trend, indicating that the model successfully learned the necessary features for effective speech recognition.

4. Conclusion

This paper presents a Vietnamese speech recognition approach using a deep learning model that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM). By extracting features from Mel-Spectrograms with CNN and capturing bidirectional temporal context with Bi-LSTM, our model achieved a Word Error Rate (WER) of 14.7%. In future work, we plan to extend our research using advanced techniques such as Transformers or Wav2Vec to further improve accuracy and evaluate performance on more diverse datasets to enhance the model’s generalizability.

References

[1] Thanh Thien Le, Linh The Nguyen, Dat Quoc Nguyen, “PhoWhisper: Automatic Speech Recognition for Vietnamese,” ICRL 2024, 2024.

[2] Khai Le Duc, “VietMed: A Dataset and Benchmark for Automatic Speech Recognition of Vietnamese in the Medical Domain,” Computation and Language, arXiv, 2024.

[3] Tran Linh Thi Thuc, Han-Gyu Kim, Hoang Minh La, Su Van Pham, “Automatic Speech Recognition of Vietnamese for a New Large-Scale Corpus,” Electronics, vol. 13, no. 5, 2024.

[4] Khai Le Duc, “Unsupervised Pre-Training for Vietnamese Automatic Speech Recognition in the HYKIST Project,” Computation and Language, arXiv, 2023.

[5] Nhu Hai Phung, Duc Thinh Dang, Khanh Duy Ta, Khac Tuan Anh Nguyen, Trung Kien Tran, Chi Thanh Nguyen, “Enhancing Whisper Model for Vietnamese Specific Domain with Data Blending and LoRA Fine-Tuning,” Proceedings of the International Conference on Intelligent Systems and Networks, ICISN 2024, Lecture Notes in Networks and Systems, vol. 1077, 2024.

[6] Tuan Nguyen, Huy Dat Tran, “LingWav2Vec2: Linguistic-augmented wav2vec 2.0 for Vietnamese Mispronunciation Detection,” Interspeech 2024, 2024.

[7] Y. Zhang, D. Wang, and Y. Gong, “CNN-BiLSTM with CTC for speech recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5559–5563, 2018.

[8] H. Li, W. Liu, and Y. Sun, “Mel-spectrogram based CNN-BiLSTM architecture for Mandarin speech recognition,” Speech Communication, vol. 112, pp. 58–67, 2019.

[9] T. T. Nguyen, H. T. Tran, and T. T. Le, “Vietnamese speech recognition using CNN-BiLSTM with data augmentation,” Journal of Electrical Engineering & Technology, vol. 16, no. 4, pp. 1743–1751, 2021.

[10] H. Q. Luong and H. Q. Vu, “A non-expert Kaldi recipe for Vietnamese Speech Recognition System,” 2016.