Ubiquitous Computing Framework for Reducing Ambiguity in the Lanna Thai Dialect Using Transformer Models and Fuzzy Logic

Wongpanya S. Nuankaew[1], Pathapol Jomsawan[2], Pratya Nuankaew[3]*
Department of Computer Science-School of Information and Communication Technology,
University of Phayao, Phayao, Thailand[1, 2]
Department of Digital Business-School of Information and Communication Technology, University of Phayao, Phayao, Thailand[3]

***Abstract*—This research focuses on developing a speech-recognition model that can better handle the unique sounds and vocabulary of the Lanna Thai dialect while supporting translation between Lanna and Standard Thai. A dataset of spoken Lanna Thai was collected from native and fluent speakers between 2023 and 2025, refined from 200 selected words to 100 terms that were consistently difficult to interpret. This dataset was used to train several supervised models, including HuBERT, Wav2Vec2 (baseTH), Wav2Vec2, and WavLM, with HuBERT showing the strongest overall performance and Wav2Vec2 (baseTH) offering a balanced vocabulary response. A companion web application was also created to convert Lanna speech to text and provide two-way translation, improving access to local language resources despite some limitations with rare terms and diverse accents. Early user feedback indicates that the system is practical and helpful, supporting the broader goal of preserving Lanna Thai in a modern digital environment.**

***Keywords*—Ambiguities in Lanna Thai vocabulary; Lanna Thai vocabulary; pervasive computing; Speech Transformer; Thai speech recognition; ubiquitous computing**

I. Introduction

Thailand is experiencing a decline in the use of the Lanna Thai language, also known as "Kham Muang", in daily life. Although the Lanna Thai dialect has long been an important part of the identity and traditions of the Northern people, today's young people are growing up surrounded by the Central Thai language, both in school and in the media. This has given Lanna Thai a place to flourish, and its use is decreasing [1], [2]. What was once part of everyday conversation is now being heard only within families and among the elderly.

One of the most significant causes of this decline is the structure of the local education system, or the concept of national integration. Most schools rely on the Central Thai language. Even in regions where Lanna Thai is still spoken at home, students do not encounter the Lanna Thai language in class, and it is not officially recognized in the national curriculum, limiting its visibility and status [1]. Outside the classroom, Central Thai influences television, social media, and entertainment. As children and students absorb this influence, Lanna Thai often feels alienated or inappropriate in practice. Some young people view Lanna Thai as outdated, embarrassing to speak, or irrelevant to their personal feelings. Over time, this shift in attitude has eroded the language's traditional vocabulary and distinctive tone.

*Corresponding author.

The decline in the use of the Lanna Thai language is not solely a result of educational and media factors; it also reflects profound and ever-changing social changes. The migration of people from rural areas to the capital or major cities, coupled with the emergence of multilingual societies from abroad, has led to a decline in the number of people who consistently use the Lanna language. The language has been replaced by standardized language, modern languages, transliteration, or the language used in the workplace and modern environments. This has resulted in the Lanna language being neglected and pushed out of daily use in many communities.

Furthermore, in today's digital world, the Lanna language faces limitations due to rapidly changing technologies that cannot keep up with the needs of the general public. For example, the Lanna Tham script, a key part of local wisdom, is not fully supported on digital platforms, making typing, writing, using, and communicating online difficult [3]. Without modern technology, transmitting the language to the next generation in a way that is contemporary is increasingly neglected and challenging.

With limited support from the central government, coming from key administrative institutions, and existing conservation programs being insufficient, the risk of the Lanna language being destroyed and limited to the elderly is becoming increasingly apparent. Collaboration between local communities, schools, government agencies, and higher education institutions is crucial. This is a crucial approach to ensure that the Lanna language continues to be spoken, heard, and passed down. These factors are the driving force behind this research project, which aims to use technology to help sustain the local language and enable it to move forward in the new world.

A. Research Objectives

This research aims to preserve the Lanna dialect by developing a tool that can operate via spoken language. The first objective is to create a ubiquitous computing model that can easily connect to Lanna dialect data and support practical applications. The second objective focuses on applying transformer-based deep learning to analyse and interpret Lanna dialect words and sentences, particularly those that are difficult or confusing.

The third objective is to address the uncertainty inherent in the dialect by applying fuzzy logic techniques to reduce

ambiguity and improve the system's accuracy in understanding dialect expressions. The final objective is to evaluate the model's performance to ensure its practical application.

B. Research Hypotheses

This research is based on three main hypotheses that outline the technical goals of the project and the practical limits related to the Lanna Thai language. The first hypothesis (H1) suggests that integrating ubiquitous computing into the system will enable the model to operate smoothly in real time. This means the technology must respond quickly and intuitively, allowing seamless communication between Lanna Thai and Standard Thai without feeling slow or disconnected from everyday use.

The second hypothesis (H2) explores the use of Transformer-based deep learning. It proposes that a Transformer model will better understand the complexities of the Lanna dialect, which includes many words and idioms whose meanings change depending on context, pronunciation, or tone, compared to traditional methods. If correct, this should reduce misunderstandings and improve the accuracy and clarity of translations or interpretations.

The third hypothesis (H3) states that uncertainty is a natural part of regional languages. Variations in vocabulary, phrasing, and sentence structure can make it harder for the system to identify the correct meaning. This hypothesis suggests that fuzzy logic analytics can provide a more flexible way to handle such uncertainties, helping the model produce more accurate results when dealing with unclear or ambiguous data.

C. Research Framework

The Lanna Thai, also known as the Kham Muang language translation application, is thoughtfully aimed and designed for various applications. This application's operational concept involves receiving spoken or typed text for translation purposes. The model streamlines the translation process and then presents the translated text. Additionally, it retains some user data from language translations to enhance the development of the translation model, as shown in Fig. 1.

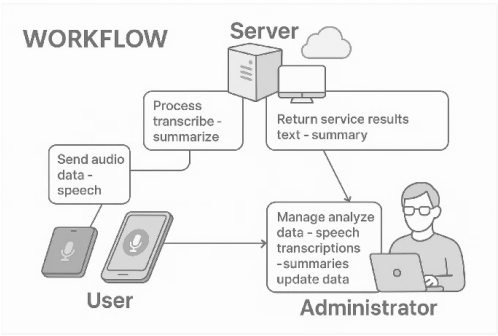

Fig. 1. Research framework.

The foundational structure of the research is composed of three essential components: the initial component involves the user, who can either verbalize or input the text intended for translation. Verbal input will be converted into text format before translation. The outcome will encompass the translated text along with a classification of the sentence's elements or individual words, identifying categories such as nouns, verbs, adjectives, pronouns, etc.

The second component concerns the administrator, who holds the capability to examine and analyze the translation statistics across various sections, using this information to improve and refine the translation model. Furthermore, the administrator is equipped to review the accuracy metrics obtained from the translation history for each instance.

The third component of the system is the host network, which integrates a translation model between Lanna Thai and Central Thai. The methodologies and artificial intelligence techniques employed in the processing encompass speech recognition, data mining, and machine learning, among other approaches.

II. Literature Review

A. Lanna Thai Dialect

Lanna Thai, also known as Northern Thai or *Kham Mueang*, is a regional language spoken primarily in Northern Thailand, covering areas that historically belonged to the Lanna Kingdom, including Chiang Mai, Chiang Rai, Lampang, Lamphun, Phrae, Nan, Phayao, and Mae Hong Son [4], [5]. Although Lanna Thai shares historical roots with Central Thai, it has evolved independently, resulting in distinctive phonological patterns, vocabulary usage, and grammatical structures that differ substantially from Standard Thai [1]. These linguistic characteristics contribute to frequent ambiguity in pronunciation and meaning, particularly when Lanna Thai is processed using models originally designed for Central Thai.

From a historical perspective, Lanna Thai belongs to the Tai–Kadai language family and developed during the Lanna Kingdom period (18th–24th Buddhist centuries), with its own writing system used extensively in religious texts and legal records [1]. However, the gradual replacement of the Lanna script by Central Thai during educational reforms has limited its presence in formal communication and reduced opportunities for systematic digital representation.

In the contemporary digital environment, Lanna Thai faces additional challenges related to technological compatibility and language standardization. Digital media platforms, streaming services, and online communication tools predominantly support Central Thai, reinforcing its dominance among younger users and marginalizing regional dialects. As a result, Lanna Thai is underrepresented in speech corpora, language models, and commercial speech recognition systems, which constrains the performance of existing technologies when applied to dialectal speech.

These limitations highlight a critical gap between the linguistic characteristics of Lanna Thai and the assumptions embedded in mainstream speech recognition and natural language processing systems. Addressing this gap requires not only language preservation efforts but also the development of dialect-aware computational approaches. Supporting technologies such as dialect-specific speech datasets, adapted language models, and translation systems tailored to Lanna Thai are therefore essential. Such developments provide the foundation for applying artificial intelligence techniques to reduce ambiguity, improve recognition accuracy, and enhance the practical usability of Lanna Thai in digital contexts, which directly motivates the focus of this research.

B. Speech Recognition Technology

Automatic Speech Recognition (ASR) is a core research area within computational linguistics that focuses on converting spoken language into textual representations using computational models [6], [7]. In practical applications, ASR systems are expected to operate reliably across different accents, pronunciation variations, speaking styles, and acoustic conditions. However, most state-of-the-art ASR systems are trained on large-scale, standardized speech corpora, which limit their effectiveness when applied to regional dialects or low-resource languages with distinctive phonological characteristics and limited training data.

From a system perspective, ASR can be viewed as a pipeline that transforms raw speech signals into linguistic units that can be further processed by downstream language applications. This process typically involves an initial speech-to-text conversion stage, where acoustic signals are mapped to textual representations, followed by a text processing stage that handles normalization, segmentation, and error correction. For dialectal speech, inaccuracies introduced in the early stages often propagate through the pipeline, resulting in higher error rates when encountering unfamiliar vocabulary or accent variations.

Traditional ASR systems are commonly categorized based on the nature of speech input they handle, including isolated-word recognition, continuous speech recognition, and spontaneous speech recognition. While these categorizations are useful for system design, dialectal speech such as Lanna Thai poses challenges across all categories due to phonetic overlap, contextual ambiguity, and inconsistent pronunciation patterns. These issues are further amplified in spontaneous speech, where informal expressions and local variations frequently occur.

Most contemporary ASR architecture consists of several interconnected components, including acoustic preprocessing, pronunciation modeling, acoustic modeling, language modeling, and decoding. Each component plays a critical role in determining overall system performance. In the context of regional dialects, limitations in pronunciation dictionaries, language models trained on standard language forms, and insufficient dialect-specific acoustic data often lead to mismatches between the model assumptions and actual speech patterns. These constraints motivate the adoption of advanced speech models and adaptation strategies that can better capture dialectal variability, as illustrated in Fig. 2, and form the basis for the model selection and experimental design employed in this study.

C. Natural Language Processing

Natural Language Processing (NLP) is a subfield of artificial intelligence concerned with enabling computers to analyze, interpret, and generate human language in both spoken and written forms [8], [9]. NLP techniques provide the foundation for applications such as machine translation, information retrieval, question-answering systems, and text analysis. In the context of regional dialects, NLP systems must cope with linguistic variability, informal expressions, and limited textual resources, which are often not adequately represented in standard language corpora.

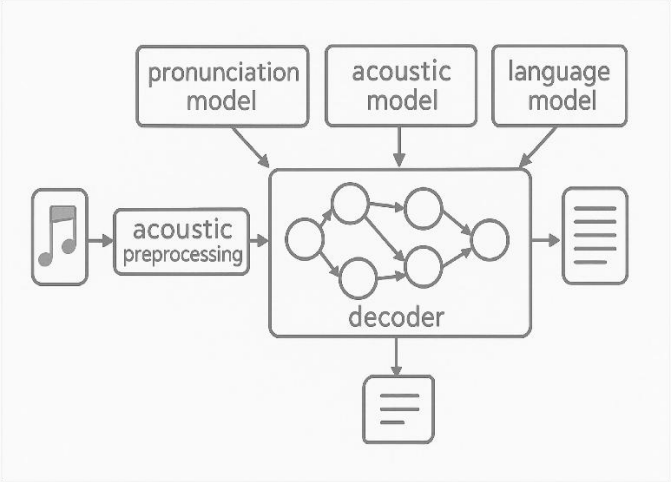

Fig. 2. Five-submodule voice recognition system.

Language modelling (LM) is a fundamental component of NLP that focuses on estimating the probability of word sequences by capturing syntactic structure and semantic relationships within language data [10], [11]. Effective language models allow systems to predict upcoming words, resolve ambiguities, and generate coherent text. However, for dialectal languages such as Lanna Thai, conventional language models trained on Standard Thai or general-purpose corpora frequently fail to represent dialect-specific vocabulary and contextual usage, leading to higher ambiguity and translation errors.

The typical operation of a language model involves training on large-scale text data to learn statistical and contextual relationships among words. Modern language models often adopt an encoder–decoder architecture, in which the encoder transforms input text into a latent representation, while the decoder generates output sequences based on this representation, as illustrated in Fig. 3. This architecture is particularly relevant for translation and interpretation tasks, where capturing contextual dependencies is essential for reducing semantic ambiguity.

Language models can be broadly categorized into n-gram models, statistical language models, and neural language models. N-gram models estimate word sequence probabilities based on fixed-length context windows, while statistical language models extend this approach using probabilistic frameworks such as Hidden Markov Models and Conditional Random Fields. Neural language models, including those based on recurrent neural networks and Transformer architectures, leverage deep learning to capture long-range dependencies and contextual nuances. Although neural models have significantly improved language understanding, their performance on low-resource and dialectal languages remains constrained by data availability and linguistic diversity. These limitations motivate the integration of domain-specific language resources and adaptive modelling strategies to improve NLP performance for Lanna Thai and similar regional dialects.

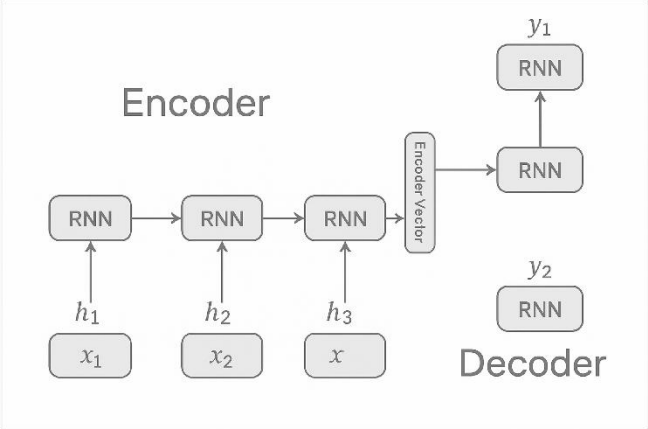

Fig. 3. The encoder-decoder sequence-to-sequence model.

Summary of Literature Review: The reviewed literature highlights that Lanna Thai is a linguistically rich regional dialect with distinctive phonological, lexical, and grammatical characteristics that are insufficiently supported by existing digital language technologies. Prior studies indicate that mainstream speech recognition and natural language processing systems, which are largely trained on standardized language corpora, struggle to handle dialectal variability, resulting in increased ambiguity and reduced performance for low-resource languages. Advances in ASR and NLP—particularly transformer-based and neural language models—have significantly improved language understanding; however, their effectiveness remains constrained when dialect-specific data and adaptive modeling strategies are absent. Collectively, the literature underscores the necessity of dialect-aware computational approaches, specialized datasets, and adapted language models to bridge the gap between Lanna Thai's linguistic properties and current technological assumptions, thereby providing a clear rationale for the methodological choices and research focus of this study.

III. Materials and Methods

A. Research Scope

This research adopts a data science–driven methodology to design and evaluate a dialect-aware speech recognition and translation framework specifically tailored for the Lanna Thai language. Rather than treating Lanna Thai as a direct variant of Standard Thai, the study explicitly addresses its phonological ambiguity, dialect-specific vocabulary, and limited digital resources. The research process encompasses problem formulation, dialect-focused data preparation, model design and adaptation, performance evaluation, and system deployment within a real-world application context.

The research workflow is structured into two main stages. The first stage focuses on the development of a speech recognition model that targets phonetically similar and semantically ambiguous Lanna words, which are a primary source of misrecognition in conventional ASR systems. Instead of relying solely on generic fine-tuning, this stage emphasizes a dialect-sensitive workflow, including curated vocabulary filtering, controlled word selection based on ambiguity frequency, and character-level handling aligned with Lanna phonetic patterns. These design choices aim to reduce ambiguity propagation during recognition and to improve robustness when processing dialectal speech.

To support this objective, the study integrates advanced Transformer-based speech models—namely Wav2Vec2 [12], [13], HuBERT [14], [15], and WavLM [16], [17] —as foundational architectures rather than end-to-end solutions. These models are adapted within a customized pipeline that emphasizes dialect-specific data preparation, targeted training on ambiguous vocabulary, and structured evaluation across accent variants and unseen words. In addition, fuzzy logic mechanisms are incorporated at the post-recognition stage to manage uncertainty arising from phonetic overlap and contextual ambiguity, enabling more flexible interpretation of competing recognition outputs. The overall conceptual design of this speech recognition workflow is illustrated in Fig. 4.

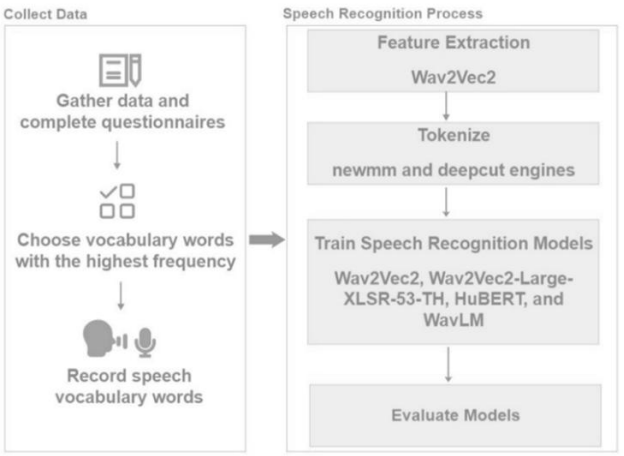

Fig. 4. Speech recognition model's conceptual framework.

Fig. 4 presents the conceptual framework of the proposed speech recognition model, which consists of two primary components: dialect-oriented data preparation and adaptive model development. During data preparation, high-quality audio data are collected from native Lanna speakers, focusing on frequently confused lexical items. Recordings are stored in uncompressed ".wav" format to preserve acoustic fidelity, and all data collection procedures adhere to established research ethics to protect participant privacy. The resulting dataset is structured to support both model training and ambiguity analysis.

In the model development phase, Transformer-based architectures are employed within the proposed workflow to enhance recognition performance for phonetically overlapping words. The models are trained on carefully curated dialect-specific audio data, enabling improved discrimination among similar-sounding lexical items. This approach demonstrates how artificial intelligence can be adapted to resource-constrained and dialectal languages by prioritizing linguistic characteristics rather than relying exclusively on large-scale generic corpora.

For translation, the research adopts a hybrid strategy that combines word-level translation with a dialect-specific dictionary explicitly constructed for Lanna Thai and Central Thai. The dictionary stores translation pairs enriched with

contextual information, while phrase-based translation rules are applied when literal word-for-word translation fails to convey the intended meaning. Unrecognized or ambiguous terms are systematically logged to support iterative dictionary expansion and future model refinement.

The second stage of the research concerns system implementation and evaluation. The proposed speech recognition and translation models are integrated into a web-based application through an API-based architecture, enabling real-time speech input and bidirectional translation output. User interaction data and empirical usage assessments are collected from target user groups to evaluate system usability, performance, and practical applicability, as well as to inform recommendations for future improvements.

B. Data Collection

Data collection is executed in agreement with established statistical principles and methodologies, which effectively mitigate bias in both data acquisition and the development of applications. This study comprised three primary sample groups. The first group consisted of vocabulary informants with expertise in the Lanna languages. They selected 200 words characterized by ambiguous meanings or easily confused meanings sourced from dictionaries and literature on the Lanna language. These words included nouns, verbs, pronouns, and adjectives.

The second group consisted of the audio informants' group, which included eight volunteers who spoke the "Kham Mueang" language, also known as Lanna, as their native language. Consent for the audio recordings was obtained for research purposes and as a case study for a student project conducted by the School of Information and Communication Technology at the University of Phayao.

The third group consists of 30 system evaluators who are familiar with both Lanna Thai and Standard Thai.

C. Ambiguous Terminology in Data Collection

The confusing data collection for this research is divided into three main sections. It includes the specified sample group and the analysis of unclear vocabulary, sound recording for model development, and audio recording for system testing.

The examination of perplexing Lanna Thai vocabulary comprises four principal components. The initial component involves the development of a "Lanna-Thai" dictionary intended to serve as a reference source for word meanings. The subsequent component entails the selection of linguistic experts tasked with identifying 200 words with ambiguous meanings derived from Lanna language texts and documents, encompassing nouns, verbs, pronouns, and adverbs. The third component consists of formulating and disseminating a questionnaire directed at 70 participants from the School of Information and Communication Technology at the University of Phayao, aimed at uncovering unfamiliar words and those with confused meanings. Finally, the fourth component encompasses the frequency analysis of unidentified and easily confounded words, selecting the 100 most frequently occurring terms based on the survey results for utilization in audio recordings.

The audio recording process for model development is comprised of four distinct components. The first component involves the organization of eight voice informants tasked with recording one hundred target words, each recorded five times per vocabulary. The second component specifies that audio recordings must occur in a low-noise environment to guarantee high-quality data acquisition. The third component encompasses storing the audio data in ".wav" files, accompanied by the relevant textual identifiers. Lastly, the fourth component delineates partitioning the audio data into a training set, which constitutes eighty percent, and a test set, comprising twenty percent.

The final objective of developing an efficient model concerns recording audio for system testing, which encompasses two primary components. The first component necessitates selecting two male and two female volunteers from a pool of eight audio contributors identified during the initial phase. The second component involves recording three sets of audio data for testing purposes, utilizing the precise terminology employed during the model training, the original words articulated in various accents, and supplementary words not incorporated into the training dataset.

The methodology utilized for the analysis of the intricate vocabulary of Lanna Thai is illustrated in Fig. 5.

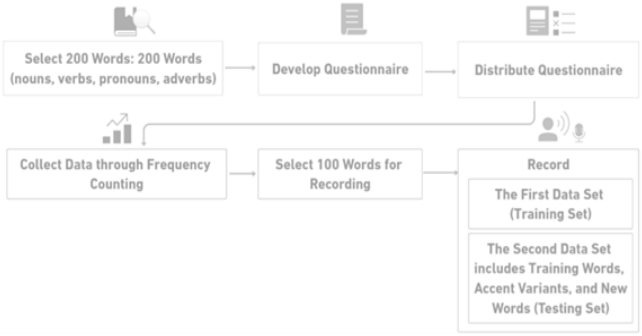

Fig. 5. An examination of ambiguous Lanna Thai terminology.

Fig. 5 illustrates the methodology used to analyze the complex vocabulary related to the collection of audio data from native speakers of Northern Thai. This audio data collection is crucial for developing an effective language processing model. The researchers successfully acquired 12 hours, 8 minutes, and 37 seconds of recordings from native speakers of Northern Thai. The participation of numerous volunteers enhances the diversity of the data and reduces bias during the model's training and testing phases. Repeated recordings help identify variations in individual pronunciation patterns, while using ".wav" files ensures superior audio quality. Throughout the data collection process, there was a strong emphasis on privacy and research ethics, safeguarding the confidentiality of personal data, and ensuring data collection in a controlled environment.

D. Fundamental Data Management Protocols

The accompanying figure illustrates four-degree steps in data preprocessing, which are critical for enhancing the system's overall performance, as detailed in Fig. 6.

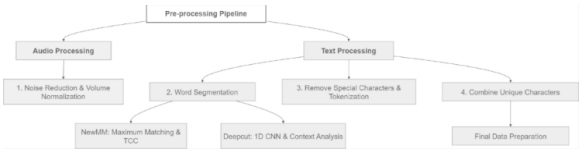

Fig. 6. Schematic representation of processing.

Fig. 6 shows a schematic representation of processing consisting of four steps. The first part involves audio signal processing, which includes reducing noise and normalizing the data to enhance the model's generalization. The second part involves word segmentation and text processing. It employs PyThaiNLP's newmm engine or deepcut to segment Thai words, which caters to the characteristics of the Thai language. This process has two components: newmm utilizes Maximum Matching with Thai Character Clusters (TCC) by dividing the text into character sequences and searching for the longest word in the matching dictionary. If a match is not found, other methods, such as root and extension analysis, come into play. The second component, deepcut, applies deep learning with a 1D CNN to predict correct word boundaries by considering letters and context. The third part involves removing special characters and performing tokenization on the text to convert it into a standard data set. The fourth part combines unique characters into a vocabulary set for comprehensive character management.

In conclusion, the aforementioned preprocessing steps significantly enhance the quality and diversity of the dataset. They also ensure a balanced representation of each word and effectively prepare both the audio and text data for the development of highly accurate and reliable models.

E. Development of Speech Recognition Models and Translation Systems

Speech recognition refers to the process of transforming human speech into readable text using a computer or an automated system. This process employs speech signal processing and artificial intelligence to analyze and interpret spoken language, effectively turning it into meaningful text.

This research investigates the adaptation of several recent ASR models for specific tasks, focusing on adjusting the Speech Transformer model to pre-trained models, specifically Wav2Vec2-large-xlsr-53 and wav2vec2-large-xlsr-53-th, which are tailored for the Thai language, HuBERT, and WavLM. The models possess the following features and functionalities:

Wav2Vec2 [13] is a self-supervised model designed to extract speech representations from raw audio data without needing labelled datasets. The model undergoes pre-training that involves predicting masked speech segments and is then fine-tuned for specific tasks, such as Automatic Speech Recognition (ASR) using Connectionist Temporal Classification (CTC) loss.

Wav2Vec2-Large-XLSR-53-TH, commonly referred to as Wav2Vec2-TH, has been optimized utilizing Thai language data sourced from Common Voice 7.0, thereby enhancing the efficacy of Thai Automatic Speech Recognition (ASR) through the meticulous capture of linguistic nuances. The associated notebook, script, and training model are accessible at vistec-ai/wav2vec2-large-xlsr-53-th and airesearch/wav2vec2-large-xlsr-53-th.

HuBERT [14] (Hidden-Unit BERT) utilizes the principles of Bidirectional Encoder Representations from Transformers (BERT) to analyze speech data by aggregating features into units and then predicting these from masked inputs, thus producing a strong representation of speech.

WavLM [17] is an advanced model that builds upon the HuBERT framework by introducing a universal speech representation model. This enhancement significantly improves recognition performance by implementing gated relative position bias within the Transformer architecture. Additionally, it provides refined speaker discrimination via speech fusion techniques, rendering it exceptionally practical for tasks associated with speaker identification.

These models have been optimized for application with Thai and Northern Thai dialects, employing specialized word segmentation tools, namely newmm and deepcut, to facilitate the efficient processing of model training data and speech recognition.

In the translation process of this research, a systematic word-by-word translation method was employed alongside a specialized dictionary developed for the Lanna and Thai languages. This dictionary includes translation pairs to clarify the meaning and context of each word. The phrase-based translation method gathers Lanna-specific phrases and expressions to enhance accuracy when word-by-word translations fail to convey the correct meaning. Additionally, unfamiliar words are meticulously recorded for future updates to the dictionary, and contextual meaning estimation techniques are applied to interpret words not found in the dictionary.

F. Evaluation of the System Prototype and Data Analysis

The evaluation of the performance of the system and the data analysis conducted in this research were divided into three primary components in alignment with the defined research objectives: an assessment of the performance of the speech recognition model (Speech-to-Text), an analysis of the efficiency of the language translation system, and an investigation into user satisfaction.

1) Performance analysis of speech recognition models: The performance analysis of the speech recognition model is categorized into two principal segments. The initial segment comprises the model assessment, which is further subdivided into two components based on the defined research objectives. The first component involves the selection of an efficient model utilizing an initial dataset of 4,000 recorded entries, partitioned into 80% (3,200 entries) for model training and 20% (800 entries) for testing purposes. The metrics evaluated in this segment include Training Loss, Validation Loss, Character Error Rate (CER), and Character Accuracy.

The subsequent segment pertains to evaluating the model's usability through data derived from a second audio recording aimed at assessing the selected model. This evaluation utilizes previously unseen data to validate the model's reliability. The evaluation dataset is comprised of three sets: "Training Words",

"Accent Variant", and "New Words". The metrics assessed in this segment encompass Training Loss, Validation Loss, Character Error Rate (CER), Character Accuracy, and processing time (in seconds) for the test set.

2) Assessment of language translation efficiency: The translation system's analysis and performance evaluation follow the Software Testing Process (STP) [18], [19], ensuring that the software functions correctly and is error-free. Consequently, the analysis is divided into two distinct parts. The first part focuses on evaluating the performance of the translation system, which consists of two subcomponents. The first subcomponent assesses the accuracy of the translation from Lanna (Kham Muang) into standard Thai. In contrast, the second subcomponent pertains to the performance analysis of the word-by-word translation using a dictionary method. The second part encompasses the analysis of user satisfaction, further divided into two subcomponents. The first subcomponent evaluates the system's ease of use, speed, accuracy, and convenience, conducted through a questionnaire administered to 30 testers. The second subcomponent collects issues identified by users, along with their recommendations for future enhancements to the system.

G. Research Tools

The research tools are classified into two primary components: the first component relates to the system design, whereas the second component addresses the system user interface. The essence of system design is categorized into six distinct sub-sections: use case diagram, class diagram, sequence diagram, activity diagram, entity-relation diagram, and data dictionary.

User Interface Design: The researchers developed a translation system for the Lanna Thai language, utilizing predominantly web application functionalities. The design of the web application interface was user-friendly, and the researcher meticulously structured each section to encompass various primary topics pertinent to the voice reception system, language translation capabilities, functionalities for segmenting individual elements of each word, and features for generating reports for administrators.

H. Satisfaction Assessment and Interpretation

A comprehensive usability evaluation and satisfaction study of the system was meticulously designed to collect data from a sample comprising 30 users, with the objective of assessing the system's effectiveness, user satisfaction, and ease of use. The evaluation procedure was conducted utilizing both desktop and mobile devices, in conjunction with UX/UI testing tools, while qualitative data was obtained through a structured questionnaire. The criteria for evaluating the system prototype and the satisfaction results are systematically presented in tabular format, as encapsulated in Tables I and II.

TABLE I. PERFORMANCE SCORING CRITERIA

Performance Scoring Criteria		**Description**
Quantitative	*Qualitative*	
Points at Level 5	Strongly agree	The performance of the prototype significantly surpasses established standards.
Points at Level 4	Agree	The performance of the prototype exceeds established standards.
Points at Level 3	Neither agree nor disagree	The prototype's performance aligns with the required standards.
Points at Level 2	Disagree	The performance of the prototype does not meet the established standards.
Points at Level 1	Strongly disagree	The performance of the prototype is significantly below established standards.

TABLE II. INTERPRETATION CRITERIA

Performance Scoring Criteria		**Description**
Quantitative	*Qualitative*	
4.51 - 5.00	Excellent	The product demonstrates high performance, is user-friendly, and incorporates outstanding features.
3.51 - 4.50	Very good	The product performs well, is user-friendly, and has comprehensive features.
2.51 - 3.50	Average	The performance is deemed satisfactory, functional, and encompasses fundamental features.
2.51 - 2.50	Adequate	The prototype exhibits inefficiency poses challenges with usability and is deficient in significant features.
1.00 - 1.50	Requires enhancement	The product exhibits terrible performance, is challenging to utilize, and fails to satisfy user requirements.

The researchers calculated the average score of each sub-item using the formula (sum of scores × number of raters at each level) ÷ total number of raters. Next, the average of the sub-items was derived as the average of the main topic using the formula Sum of the sub-item averages ÷ Number of sub-items in that main topic. This process was repeated for each main topic. The average value was compared with the specified interpretation criteria to determine the level of quality or satisfaction in each topic, which would lead to the conclusion of the evaluation results and recommendations for further development or improvement of the system.

IV. RESULTS

The research findings are categorized into three primary topics: the development of the speech transcription model and the translation system, the evaluation results of the speech

transcription model, and the study results related to the performance and satisfaction of the application prototype.

A. Speech Transcription Model and Translation System

The team successfully developed a speech-to-text model in conjunction with a translation system designed to facilitate the conversion of Lanna Thai into Standard Thai. In the initial phase of training the transcription model, the team selected 200 words from a reference dictionary. Subsequently, they implemented further screening of words utilizing a questionnaire aimed at local native speakers. Consequently, less common words and those identified as challenging to comprehend were eliminated from consideration, yielding a refined selection of 100 words for training the transcription model. This meticulous methodology enabled the model to incorporate vocabulary pertinent to the daily experiences of Northern speakers, thereby significantly enhancing the accuracy of translating Lanna Thai speech into text.

In the context of Lanna to Thai translation, the team has devised a translation system that employs rule-based translation techniques alongside a reference dictionary. This system meticulously analyses sentence structures and the rules of language usage in accordance with the principles of Lanna Thai to ensure precise translation of text into standard Thai. The utilization of this approach enhances the system's capability to effectively translate words and expressions with specific meanings inherent in the northern dialect, while simultaneously mitigating errors that may arise from general translations.

B. Speech Transcription Model Testing Results

The developers constructed a speech-to-text transcription model and conducted assessments to evaluate the model's accuracy using audio data collected and prepared by them, totaling 4,000 audio files. The terms used for the audio recordings were meticulously selected from reference dictionaries, covering 100 categories of words, including verbs, nouns, pronouns, and adverbs. These terms served as examples in creating a dataset for training the model. The models used for training and evaluation included four primary models: Wav2Vec2, WavLM, and HuBERT, with the training process involving a total of 50 iterations (epochs). The results of the analysis and testing are outlined as follows (Fig. 7):

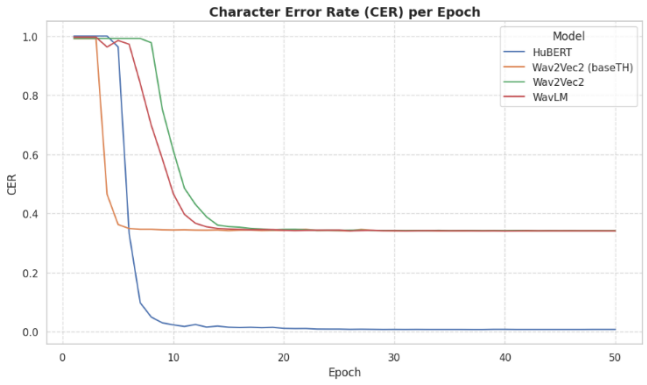

Fig. 7. Display of character error rate (CER) for four models.

Over the span of fifty epochs, HuBERT exhibits the pinnacle of performance, succeeded by Wav2Vec2 (baseTH) and WavLM, while Wav2Vec2 registers the least favorable performance metrics. All models illustrate significant progress during the preliminary ten epochs and subsequently begin to stabilize after the twentieth epoch, reflecting consistent learning processes.

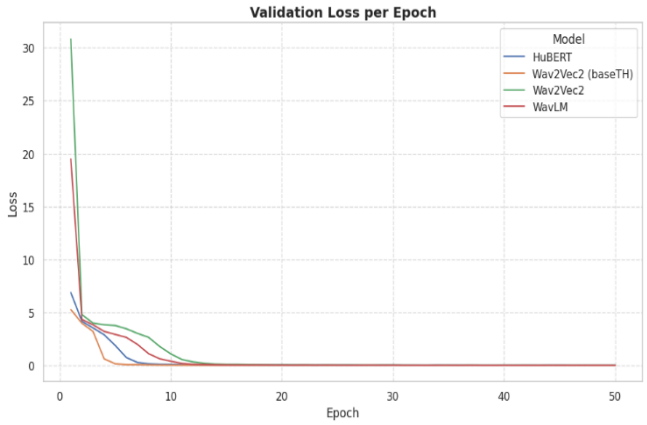

Fig. 8. Presentation of validation loss for four models.

From Fig. 8, we can see that the validation loss for the four models decreased a lot in the first five epochs, with HuBERT and Wav2Vec2 (baseTH) doing the best, followed by WavLM and Wav2Vec2. After 20 epochs, the loss for all models is nearly zero and steady, showing they are effective and reliable in predicting the validation data. After Epoch 20, the loss of all models is close to zero and stable, indicating their efficiency and stability in predicting the validation data.

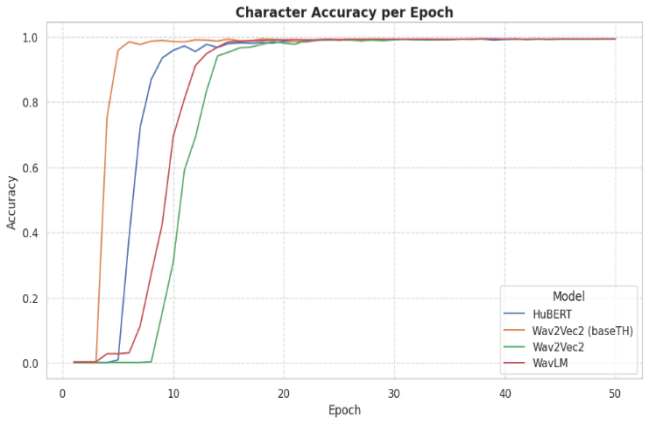

Fig. 9. Graph illustrating the loss value during the inspection process.

In Fig. 9, the character accuracy graph illustrates the performance of the four models over 50 epochs, with HuBERT achieving the highest accuracy, followed by Wav2Vec2 (baseTH), WavLM, and Wav2Vec2. All models showed significant performance improvements during the first 10 epochs; however, HuBERT notably excelled, surpassing the others starting at epoch 10 and nearing 100% accuracy by epoch 20. The other models also exhibited rapid enhancement during the initial epochs, but HuBERT particularly stood out as it exceeded its counterparts from epoch 10 and approached an accuracy of 100% by epoch 20. Meanwhile, the other models produced comparable results in subsequent epochs, despite starting their improvements at a later stage. Given its remarkable

performance at epoch 30, with an accuracy of 99.295%, we have chosen HuBERT for further evaluation with new data.

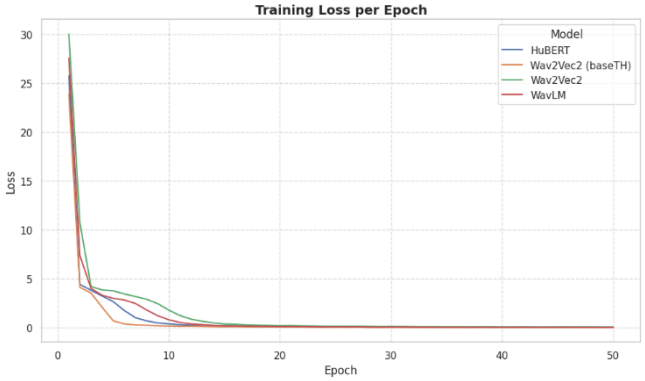

Fig. 10. Training loss indicates the loss of the four models.

Fig. 10 illustrates the training loss of the four models over 50 epochs, resembling the validation loss curve. The Wav2Vec2 (baseTH) model demonstrates superior performance, closely followed by HuBERT, WavLM, and Wav2Vec2. All models experience a sharp drop in loss during the first five epochs and then stabilize near zero after epoch 20, indicating their effectiveness and stability in learning from the training dataset.

C. Summary of the Training Results for all Four Models

Based on the performance comparison of the four models, HuBERT stands out as the best regarding the lowest Character Error Rate (CER), followed by Wav2Vec2 (baseTH), WavLM, and Wav2Vec2. The trend is similar when analyzing validation loss and training loss, with HuBERT and Wav2Vec2 (baseTH) demonstrating lower losses since the early epochs. At the same time, WavLM and Wav2Vec2 exhibit slower development, but adapt well over several epochs.

Interestingly, all models significantly improve during the first 10 epochs and gradually stabilize after epoch 20. The ability of all models to reduce the loss value to near zero and achieve an accuracy of over 99% illustrates the potential of deep learning techniques to address the challenge of speech recognition in local languages such as Northern Thai, as shown in Table III.

TABLE III. TRAINING RESULTS FOR ALL FOUR MODELS

Epoch / Model	Training Loss	Validation Loss	CER	Accuracy
Epoch 10				
HuBERT	0.39830	0.09751	0.02245	0.95896
Wav2Vec2	1.81560	1.09326	0.61201	0.31060
Wav2vec2(baseTH)	0.17650	0.03425	0.34352	0.98568
WavLM	0.82190	0.40436	0.34352	0.69581
Epoch 20				
HuBERT	0.13660	0.04116	0.01069	0.98675
Wav2Vec2	0.22270	0.04932	0.34545	0.98119
Wav2vec2(baseTH)	0.08110	0.02013	0.34203	0.99124
WavLM	0.12790	0.03238	0.34267	0.99081
Epoch 30				
HuBERT	0.12790	0.02993	0.00705	0.99295
Wav2Vec2	0.13330	0.02987	0.34160	0.99145
Wav2vec2(baseTH)	0.05500	0.01881	0.34053	0.99295
WavLM	0.07000	0.02706	0.34117	0.99316
Epoch 40				
HuBERT	0.05410	0.01982	0.00727	0.99230
Wav2Vec2	0.07400	0.02573	0.34117	0.99337
Wav2vec2(baseTH)	0.03790	0.01590	0.34053	0.99295
WavLM	0.04510	0.02030	0.34010	0.99337
Epoch 50				
HuBERT	0.05080	0.01842	0.00705	0.99316
Wav2Vec2	0.06230	0.02437	0.34096	0.99380
Wav2vec2(baseTH)	0.02850	0.01343	0.34032	0.99316
WavLM	0.03790	0.02152	0.34053	0.99359

TABLE IV. THE PERFORMANCE TABLE FOR THE EVALUATION TEST SET

Speech Transformer Models	CER by newmm (%)			CER by deepcut (%)		
	Training Words	*Accent Variant*	*New Words*	*Training Words*	*Accent Variant*	*New Words*
HuBERT	0.344	3.615	57.491	0.344	3.615	57.491
Wav2Vec2	0.172	2.926	54.704	0.172	2.926	54.704
Wav2Vec2(baseTH)	0.172	1.721	40.244	0.172	1.721	40.244
WavLM	0.172	4.303	61.150	0.172	4.303	61.150

Table IV summarizes the recognition performance of the evaluated models under three testing conditions: training words, which were included in the training set; accent variants, which represent pronunciation variations of known words; and new words, which were entirely excluded from model training. This evaluation design was intended to assess not only in-domain performance but also the models' generalization capability under dialectal variation and out-of-vocabulary conditions.

Although the models achieved high accuracy for training words and accent variants—particularly at epoch 30—this performance does not fully translate to unseen lexical items. While Wav2Vec2 (baseTH) demonstrated comparatively better results across all test categories, the Character Error Rate (CER) for new words exceeded 40% for all models. This indicates a substantial degradation in performance when the models encounter unfamiliar vocabulary, despite near-saturated training and validation accuracy. Such behavior suggests a tendency toward overfitting to the limited training vocabulary, where the models effectively memorize known word patterns but struggle to generalize to unseen linguistic forms.

A closer examination of the error patterns reveals that misrecognitions are primarily associated with phonetic overlap between Lanna words, accent-dependent pronunciation shifts, and the absence of lexical representations for out-of-vocabulary terms. These factors collectively amplify ambiguity during decoding, particularly when dialectal variations deviate from the acoustic distributions observed during training. The limited size of the dataset—restricted to 100 target words and a small number of speakers—further constrains the models' exposure to phonetic and accentual diversity, thereby limiting generalization performance.

To mitigate these limitations, expanding both the lexical coverage and speaker diversity of the audio dataset is a necessary direction for future work. Incorporating data augmentation strategies and collecting recordings from speakers with a broader range of accents may help reduce overfitting and improve robustness to unseen vocabulary. This observation is consistent with the findings of Suwanbandit et al. (2023), who reported that increased accent diversity in training data significantly enhances recognition performance in dialectal speech recognition tasks.

TABLE V. RESULTS OF WORKING TIME ASSESSMENT

Speech Transformer Models	Processing Time (s)		
	Model Inference	*newmm*	*deepcut*
HuBERT	0.01812	0.000024	0.063662
Wav2Vec2	0.01801	0.000022	0.062519
Wav2Vec2(baseTH)	0.01741	0.000022	0.061175
WavLM	0.06316	0.000032	0.063185

In Table V, the performance demonstrates considerable consistency across all models, with an inference time of approximately 0.018 seconds. In contrast, WavLM exhibits a slightly longer inference duration of 0.06316 seconds when compared to the other models.

In conclusion, the majority of the models exhibit comparable processing times, approximately 0.018 seconds, which is regarded as satisfactory performance. An exception is WavLM, which requires a marginally longer duration. Nonetheless, the discrepancy in processing time among these models is minimal and does not impact the system's overall performance.

D. Results of the Evaluation of the Translation System

The automatic translation system's performance test revealed that it can accurately translate fundamental words and sentences from Thai Lanna to Central Thai.

However, the system still struggles to translate new words not found in its dictionary and has difficulty with complex sentences that don't follow standard grammar, like slang or local expressions. Future improvements might require using a mix of different translation methods, such as dictionaries with a broader range of words, better grammar rules, and machine learning techniques to learn new rules and sentence structures.

E. Comprehensive Evaluation Results of the System

The system's usability was comprehensively evaluated by administering a questionnaire to 30 users, aimed at assessing the system's efficiency, user satisfaction, and ease of use. This evaluation involved testing on desktop and mobile devices, utilizing UX/UI testing tools, and collecting qualitative data via a questionnaire. Table II outlines the criteria, while Table VI presents the evaluation results. Meanwhile, Fig. 11 displays a chart depicting satisfaction data towards the application.

TABLE VI. WEB APPLICATION EVALUATION RESULTS

Evaluation Topic	Evaluation Results
1. User Experience and User Interface (UX/UI) Design	Very good
2. Performance of Web Applications	Medium
3. Usability Assessment of the Core Functions of the Web Application	Very good
4. Satisfaction with the Application	Very good

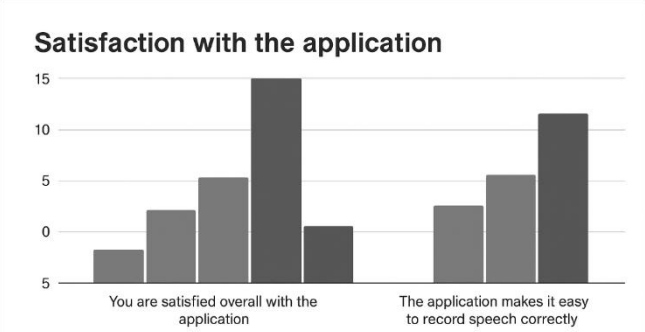

Fig. 11. Satisfaction data towards the application.

From Fig. 11 and Table VI, the summary of the satisfaction survey results indicates that most users provided high satisfaction scores, particularly regarding the user-friendly UX/UI design and smooth operation of functions such as language translation and voice transcription. However, while the overall performance remains acceptable, some minor issues with translation speed were noted. Several users offered suggestions for enhancing the clarity of specific work steps and improving efficiency when multiple users engage with the system simultaneously. The analysis of the results shows that the system effectively meets user needs, yet there is still potential for further development to enhance usability.

V. Conclusion and Discussion

Language constitutes a fundamental mechanism for human communication and social participation; however, linguistic variability poses substantial challenges for computational systems, particularly when regional dialects are involved. Even among speakers of the same nationality, non-standard varieties such as Northern Thai (Lanna Thai) introduce phonological ambiguity, accent variation, and vocabulary divergence that are not adequately handled by mainstream language technologies. In response to these challenges, this research set out to develop a dialect-aware speech recognition and translation framework tailored specifically to Lanna Thai, with the dual objectives of reducing pronunciation ambiguity and enabling practical translation between Lanna Thai and Standard Thai through a web-based application.

To support this objective, a dedicated speech dataset was constructed from recordings collected between 2023 and 2025 from native and fluent Lanna Thai speakers. An initial set of 200 dialectal words was derived from authoritative Northern Thai linguistic resources and systematically categorized by grammatical function. Through expert screening and empirical analysis of ambiguity, this set was refined to 100 words that exhibited high levels of phonetic similarity and translation difficulty. While this focused dataset enabled controlled experimentation on ambiguity-prone vocabulary, its limited lexical scope and speaker diversity inevitably constrained the models' exposure to broader dialectal variation.

Experimental results obtained from Transformer-based speech models—namely HuBERT, Wav2Vec2(baseTH), Wav2Vec2, and WavLM—demonstrate that all evaluated architectures can effectively learn Lanna Thai speech patterns within the training domain. HuBERT consistently achieved the strongest overall performance across evaluation metrics, while Wav2Vec2(baseTH) exhibited balanced behavior across different test conditions. However, the observed near-saturated accuracy on training and validation sets, contrasted with Character Error Rate (CER) values exceeding 40% for previously unseen words, indicates a clear limitation in generalization capability. This discrepancy suggests a degree of overfitting to the restricted training vocabulary and highlights the difficulty of recognizing out-of-vocabulary items and accent-driven pronunciation shifts in low-resource dialect settings.

These findings underscore that high in-domain accuracy alone is insufficient as an indicator of robustness for dialectal speech recognition. The elevated error rates for new words reveal structural challenges related to limited dataset size, narrow vocabulary coverage, and insufficient accent diversity. Addressing these limitations will require expanding the speech corpus to include a wider range of lexical items and speaker profiles, as well as incorporating data augmentation techniques to improve phonetic variability. In addition, future research should explore dialect-specific modeling strategies—such as character-level representations, subword modeling, or hybrid linguistic–statistical approaches—to better capture the intrinsic properties of Lanna Thai.

Despite these constraints, the study successfully demonstrates the feasibility of integrating a dialect-focused speech recognition model into a functional web application that supports speech-to-text conversion and bidirectional translation between Lanna Thai and Standard Thai. User evaluations indicate that the system is practically usable and meets core functional requirements, although performance efficiency was assessed at a moderate level, leaving room for further optimization. Overall, this research contributes empirical evidence and methodological insights into the challenges of dialectal speech recognition and provides a foundation for future advancements in applying artificial intelligence to the preservation, accessibility, and digital integration of regional languages such as Lanna Thai.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this study.

Acknowledgment

This research project was supported by the Thailand Science Research and Innovation Fund and the University of Phayao. It also received support from many advisors, academics, researchers, students, and staff. The authors thank everyone for their support and cooperation in completing this research.

References

[1] P. Lekuthai, "Lanna culture and social development: a case study of Chiangmai Province in Northern Thailand," Cancer Epidemiology, Biomarkers & Prevention, vol. 168, no. 27–35, 2008.

[2] A. Johnson, "Rebuilding Lanna: Constructing and consuming the past in urban Northern Thailand," 2010.

[3] S. Chakraborty, "Problem and Prospects of Lanna Community, Chiang Mai, Northern Thailand: Issue of Human Rights and Social Justice," Afro Asian Journal of Anthropology and Social Policy, vol. 5, no. 1, pp. 1–7, 2014.

[4] K. Unthanon, "The Northern Thai Dialect Used in Chiang Mai, Thailand," International Journal of Education and Social Science Research (IJESSR), vol. 5, no. 4, pp. 262–270, 2022.

[5] N. Phetsuriya, "Formal language of Lanna Shop House's Façade in Lampang Old city, Thailand," in IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2017, p. 052015.

[6] J. Meng, J. Zhang, and H. Zhao, "Overview of the speech recognition technology," in 2012 fourth international conference on computational and information sciences, IEEE, 2012, pp. 199–202.

[7] M. Johnson et al., "A systematic review of speech recognition technology in health care," BMC medical informatics and decision making, vol. 14, pp. 1–14, 2014.

[8] J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science, vol. 349, no. 6245, pp. 261–266, 2015.

[9] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 544–551, 2011.

[10] S. M. Katz, "Distribution of content words and phrases in text and language modelling," Natural language engineering, vol. 2, no. 1, pp. 15–59, 1996.

[11] H. Erdogan, R. Sarikaya, Y. Gao, and M. Picheny, "Semantic structured language models.," in INTERSPEECH, 2002, pp. 933–936.

[12] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," arXiv preprint arXiv:2012.12121, 2020.

[13] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition," IEEE Access, vol. 11, pp. 46938–46948, 2023.

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM transactions on audio, speech, and language processing, vol. 29, pp. 3451–3460, 2021.

[15] T. Hattori and S. Tamura, "Speech Recognition for Minority Languages Using HuBERT and Model Adaptation.," in ICPRAM, 2023, pp. 350–355.

[16] S. Chen et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022.

[17] H. Song et al., "Exploring wavlm on speech enhancement," in 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2023, pp. 451–457.

[18] M. A. Jamil, M. Arif, N. S. A. Abubakar, and A. Ahmad, "Software Testing Techniques: A Literature Review," in 2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M), Nov. 2016, pp. 177–182. doi: 10.1109/ICT4M.2016.045.

[19] L. Baresi and M. Pezzè, "An Introduction to Software Testing," Electronic Notes in Theoretical Computer Science, vol. 148, no. 1, pp. 89–111, Feb. 2006, doi: 10.1016/j.entcs.2005.12.014.