APPLICATION OF ARTIFICIAL INTELLIGENCE IN RECOGNIZING AND CORRECTING STUTTERING ERRORS IN CHILDREN: DEVELOPING A SYSTEM SPECIFICALLY FOR VIETNAMESE

ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG NHẬN DIỆN VÀ SỬA LỖI NÓI LẮP Ở TRẺ EM: PHÁT TRIỂN HỆ THỐNG DÀNH RIÊNG CHO TIẾNG VIỆT

Huy-Hung Huynh[1]*, Tuan-Kiet Nguyen Nhat[1], Huy-Hoang Huynh[2]

[1]Le Quy Don High School For The Gifted, Danang city, Vietnam
[2]University Institute of Technology in Troyes, France

*Corresponding author: hhuyhung2209@gmail.com

(Received: January 08, 2025; Revised: January 23, 2025; Accepted: May 12, 2025)

DOI: 10.31130/ud-jst.2025.013E

Abstract - The paper presents a study on the application of artificial intelligence (AI) and natural language processing (NLP) to detect and correct stuttering in Vietnamese children aged 7–12. The optimal speech recognition algorithm for Vietnamese identifies errors through a keyword database. The system provides feedback through audio and images of mouth movements. The interface supports learning by tracking progress and displaying original and corrected sentences, helping children detect and correct errors. This method improves the ability to correct speech errors, improve pronunciation and communication confidence. The results show the potential for AI applications in education, speech therapy and supporting children's language development.

Key words - Artificial intelligence; stuttering; speech education and therapy; language processing; Vietnamese lip-shaping

Tóm tắt - Bài báo trình bày nghiên cứu ứng dụng trí tuệ nhân tạo (AI) và xử lý ngôn ngữ tự nhiên (NLP) để phát hiện và sửa lỗi nói lắp ở trẻ em Việt Nam từ 7–12 tuổi. Thuật toán nhận dạng giọng nói tối ưu cho tiếng Việt xác định lỗi qua cơ sở dữ liệu từ khóa. Hệ thống cung cấp phản hồi thông qua âm thanh và hình ảnh khẩu hình miệng. Giao diện hỗ trợ học tập bằng cách theo dõi tiến độ và hiển thị câu gốc cùng câu chỉnh sửa, giúp trẻ phát hiện và khắc phục lỗi. Phương pháp này giúp cải thiện khả năng sửa lỗi nói, nâng cao phát âm và sự tự tin giao tiếp. Kết quả cho thấy, tiềm năng ứng dụng AI trong giáo dục, trị liệu ngôn ngữ và hỗ trợ phát triển ngôn ngữ trẻ em.

Từ khóa - Trí tuệ nhân tạo; nói lắp; giáo dục và trị liệu ngôn ngữ; xử lý ngôn ngữ; khẩu hình tiếng Việt

1. Introduction

Language plays a fundamental role in the comprehensive development of children, serving not only as a primary means of communication but also as a critical medium for acquiring knowledge, shaping cognitive processes, and building social relationships [1-3]. In Vietnam, it is estimated that approximately 374,000 children experience language disabilities, with stuttering constituting a significant proportion of these cases [4]. This stuttering disability not only hinders communication ability but also seriously affects children's psychology, causing self-consciousness, feelings of social isolation and reduced confidence in learning as well as daily activities.

Several countries have successfully leveraged artificial intelligence (AI) and natural language processing (NLP) to develop solutions aimed at improving children's language skills, effectively supporting the detection and treatment of language disorders [5-7]. These solutions encompass error identification and the provision of corrective feedback to learners. However, in Vietnam, the application of such technologies remains limited, particularly in language-related solutions, which have yet to be optimized to accommodate the unique structural and phonetic characteristics of Vietnamese. Currently, most language support solutions in Vietnam are implemented directly by educators, often involving high treatment costs, extended therapy durations, and the need for specialized resources. These factors render such services inaccessible to many families, creating a significant gap in the provision of effective language care and therapy, particularly for children who stutter. This situation highlights the critical question: How can a support system be developed for children who stutter that is cost-effective, easy to implement, and tailored to the linguistic and phonetic characteristics of Vietnamese, while maintaining effectiveness?

Thus, this study aims to address this gap by presenting a research proposal to develop a system applying AI and NLP to support Vietnamese-speaking children who stutter. The system not only recognizes and corrects stuttering errors in real time but also integrates the feature of converting sound into Vietnamese mouth shape through the face of a Robot. We expect this to be a new step forward, allowing children to visualize correct pronunciation, thereby easily adjusting their mouth shape and improving their speaking skills. In addition, the system also focuses on personalizing users through the ability to store learning history, analyze error trends and practice programs, meeting the individual learning needs of each child. Additionally, the system emphasizes user personalization by incorporating features such as learning history tracking, error trend analysis, and customized practice programs, catering to the individual learning needs of each child. The research findings aim to enhance the

communication abilities of children who stutter while providing broader opportunities for application in education and speech therapy. This solution addresses practical needs, contributes to the development of a Vietnamese language support technology platform, and establishes a foundation for future research and applications.

2. Methodology

2.1. Data collection

To ensure the system's accuracy and reliability, the data collection process was conducted systematically and meticulously, prioritizing representativeness and quality. A total of 250 speech samples were collected from 25 children aged 7 to 12 in 15 days with each samples recorded in 60 seconds for 10 sentences, including many from the Da Nang Center for Supporting the Development of Inclusive Education, as well as others obtained through direct community surveys conducted by the research team, with particular attention to regional characteristics in the Central region of Vietnam.

The speech samples were categorized into groups based on the severity of stuttering. The first dataset comprised approximately 100 samples of grammatically accurate utterances without stuttering errors, serving as a baseline for comparison. The second group, consisting of 100 samples, contained mild stuttering errors characterized by the repetition of words once or twice within a sentence (e.g., "I I want want to to eat"). The third group included 100 samples with severe stuttering errors, where multiple repetitions of words or phrases significantly disrupted the fluency of speech (e.g., "I I I want want want to to to eat eat eat"). These datasets effectively captured a range of stuttering severities, providing a robust basis for system evaluation and development.

The data were designed to reflect real-life communication scenarios, featuring content familiar to children, such as “You always study well” or “I want to go to school.” These selected sentences not only ensured relevance to everyday contexts but also targeted common stuttering patterns in Vietnamese.

By carefully selecting data from multiple sources and validating its relevance in various contexts, this approach significantly broadened the scope of the research and heightened its applicability. By encompassing a wider range of scenarios and participant profiles, the resulting data set became more representative of the real-world environment, thereby enhancing the practical implications of the study. At the same time, the inclusion of diverse and well-structured data points provided a more comprehensive corpus for in-depth analysis, ultimately laying a solid foundation for robust findings and actionable insights.

Data collection was conducted in a controlled environment to minimize noise, a critical factor in ensuring the quality of input data. Voice samples were recorded in a soundproof chamber, and each sample was recorded three times to introduce variations in tonality, pace, and intonation.

After collection, the audio samples were transcribed into text using the SpeechRecognition tool. The transcribed content was then compared against a predefined list of keywords and common phrases, which included words and phrases frequently used in children's daily communication.

The collected data samples will be used to count duplicate phrases that are not stuttering, phrases that children often stutter, and put into tests to test the children's communication progress later. In addition, it will be used for the Sound to lip library built based on video samples to set the robot's mouth shape.

That is, the robot will have ability to convert sound into Vietnamese mouth shapes, visualized using a human face model (robot). The model is equipped with 11 servos controlling the mouth and jaw, 6 servos for facial expressions, 5 servos for eyebrow movements, and 2 servos for neck movement, enabling dynamic interaction with users. This functionality utilizes animation technology-manipulating lines, ovals, and circles-and audio processing to accurately reproduce mouth shapes, eye expressions, and eyebrow movements corresponding to the corrected sentence. When the system detects a repetition error in a sentence, it automatically corrects it and simultaneously displays a visual representation of the mouth shape illustrating the correct pronunciation. For example, when a child says, “I I want want to to eat,” the system corrects it to “I want to eat” and displays the corresponding mouth shape pronouncing the corrected sentence. This visual aid allows children to easily observe and practice proper pronunciation. This conversion capability not only helps children visualize correct pronunciation but also enhances their ability to adjust mouth shapes during communication. This feature is particularly beneficial for children who struggle with connecting sounds to their physical production. By displaying mouth shapes in real-life contexts and with familiar content, the system fosters a user-friendly and effective environment, enabling children to improve their language and communication skills.

2.2. Algorithm Development

PhoWhisper large-v2, a speech recognition model for Vietnamese, was integrated into the system to enhance transcription accuracy. This model, fine-tuned specifically for Vietnamese language with its tonal and phonetic complexities, processes input audio to generate highly accurate transcriptions. These transcriptions are further refined using a custom stuttering correction algorithm. The workflow ensures the text output is fluent, grammatically correct, and ready for further processing in the Sound-to-Lip synchronization module.

2.2.1. Stuttering Detection

The stuttering detection module employs a combination of AI and NLP to analyze speech inputs. By leveraging a pre-trained Vietnamese speech recognition model, the system identifies stuttering patterns, such as word repetitions, elongated sounds, or abrupt pauses. Detected errors are then processed using a stuttering correction

algorithm to generate fluent output, which is synchronized with a visual mouth shape simulation. The following code snippet illustrates the detection process:

```python
def fix_stuttering(text):
    words = text.lower().split()
    corrected_words = []
    for i in range(len(words)):
        if i == 0 or words[i] != words[i - 1]:
            corrected_words.append(words[i])
    return ' '.join(corrected_words)
```

The flowchart below describes how the system works:

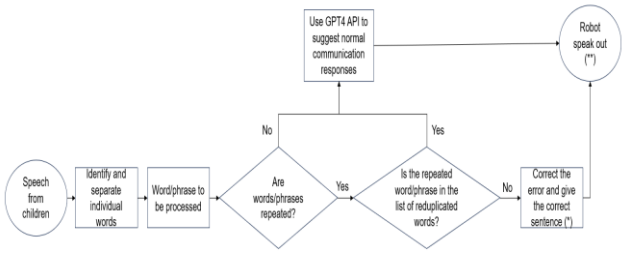

Figure 1. *Speech processing and response generation flowchart*

Besides some available libraries, we decided to build some new libraries for Robot to express pronunciation and expression:

Text2Mouth: Synchronizing pronunciation with mouth state.

***FixStuttering:** Eliminate stuttering errors.

Image and camera processing: Face recognition through the camera system to identify the child's face to adjust the neck servo motor, creating direct communication.

****Drawing faces,** manipulating facial state changes (FaceDrawer and FaceMover): Drawing and controlling the Robot face in sync with communication.

The stuttering correction algorithm extends beyond merely removing repetitions; it also addresses the replacement of incorrect phrases with contextually appropriate ones. When an inappropriate or semantically incorrect phrase is detected, the algorithm automatically substitutes it with a predefined phrase from a specialized phrase list. This approach ensures that the output is not only grammatically correct but also contextually appropriate, enhancing the overall coherence and quality of the communication. Here is a code snippet illustrating the algorithm with a list of phrases:

```
def fix_stuttering_with_phrases(text, allowed_phrases):
    words = text.lower().split()
    corrected_words = []
    i = 0
    while i < len(words):
        for phrase in allowed_phrases:
            phrase_words = phrase.split()
            if words[i:i + len(phrase_words)] == phrase_words:
                corrected_words.extend(phrase_words)
                i += len(phrase_words)
                break
        else:
            if i == 0 or words[i] != words[i - 1]:
                corrected_words.append(words[i])
            i += 1
    return ' '.join(corrected_words)
```

2.2.2. Sound-to-Lip Conversion

The Sound-to-Lip library provides a solution for synchronizing speech and mouth movements in robots, enhancing natural communication. The library works by converting text into speech while controlling the robot's mouth states corresponding to each phoneme, ensuring the robot simultaneously articulates and visually expresses speech accurately.

When receiving text input, Sound-to-Lip uses Google Text-to-Speech (gTTS) to generate an audio file of the spoken text. It then extracts the duration of the speech to divide it into segments corresponding to each mouth state. Each character or vowel in the text is mapped to a predefined mouth state code (e.g., "a" maps to "A0", "ô" maps to "OO"). The duration of each speech segment is calculated to ensure that the mouth state transitions are synchronized with the spoken words.

The library also integrates hardware control capabilities, where each mouth state is converted into a data string that can be sent to devices like Arduino. For instance, the "A0" state can be translated into control signals for mechanical structures simulating the robot's mouth movements. This synchronization is achieved through parallel threads, ensuring the robot not only delivers precise speech but also displays expressive and dynamic mouth movements, creating a more engaging interaction with users.

2.2.3. Special phrases in Vietnamese

Stuttering in Vietnamese often involves word or phrase repetition, particularly in emphasized or repetitive structures. Representative examples like “luôn luôn,” “thường thường,” and “ngày ngày” may naturally convey repetition but risk being misclassified as stuttering. To address this, a curated list of such phrases was developed, ensuring linguistic accuracy and differentiation from genuine stuttering.

The stuttering correction algorithm segments sentences, compares segments with the list, and removes redundant repetitions while retaining grammatical integrity. This method reduces false corrections, enhances algorithm effectiveness, and allows customization for various user groups. Integrating this list boosts both scientific value and practical utility, improving stuttering correction and communication skills.

2.2.4. Illustration of results and user interface features of stuttering support

The stuttering correction software features an intuitive and interactive interface designed to help children improve their speaking and communication skills. It can be used independently or with a Robot face model.

Stuttering Error Display: When children speak into the microphone, the system processes their speech and displays the original and corrected sentences side by side. For example:

- Original: I I want want to to eat

- Corrected: I want to eat

This immediate feedback helps children recognize and correct their mistakes, fostering awareness and improving speech.

Corrected Sentences List: A "Corrected" list tracks frequent errors with their occurrence count. For instance:

- I want to eat [3]

- I want to hang out [2]

This feature helps users focus on commonly repeated mistakes for targeted practice.

Learning History Storage: The system saves all original and corrected sentences in a .HOC file, enabling parents and teachers to monitor progress and analyze trends. Example:

- Original: I want want to to eat

- Corrected: I want to eat

This data supports personalized training and reinforcement learning.

Performance Analysis: An "Analyze" button provides detailed reports on performance, including the number of correct and incorrect sentences and common mistakes. Example:

- You said 50 sentences; 34 were correct.

- Incorrect Sentences:

- I want to hang out (2 times);
- I want to eat (4 times).

Customizable Special Phrases: Users can add or modify special phrases (e.g., "luôn luôn," "thường thường") via the Options menu, enhancing correction accuracy for individual needs.

Positive Feedback: The system encourages children by providing motivational messages for consecutive correct sentences, such as "You did great!" or "Great, you've improved a lot!" This boosts confidence and engagement in practice.

Overall, the user interface is designed to support the learning process effectively, not only helping children identify errors easily but also motivating them to improve their speaking skills through intuitive and convenient features. Storing learning history, displaying error lists, and the ability to compare original sentences with corrected sentences all contribute to the practicality and effectiveness of the system. The interface is shown in Figure 2.

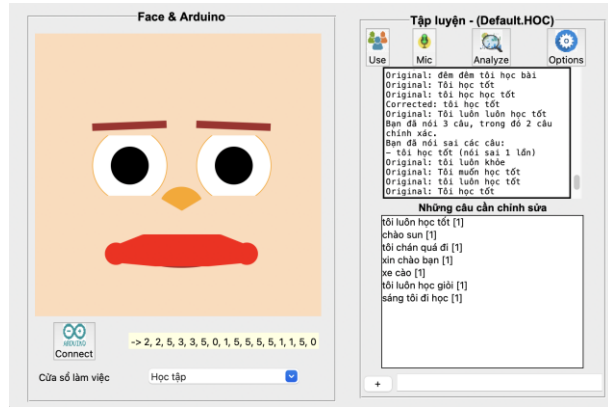

Figure 2. *The interface of the system*

3. Results and Discussions

This study demonstrates the effectiveness of personalizing user interactions through stuttering recognition, correction, and display using the Robot face model to help children with stuttering disabilities improves their Vietnamese communication skills. With an accuracy ranging from 86.47% to 96.47%, the research meets efficiency and stability criteria, particularly in processing the Vietnamese language, which has a complex grammatical structure and tone system.

The results of evaluating the effectiveness of the robot in providing responses to users are based on sample dialogue scenarios. The outcomes of testing conducted over 30 days with five children are as follows:

Table 1. *Accuracy of Robot responses across different dialogue scenarios with children*

Type of Dialogue	Number of Test Sentences	Number of Correct Responses	Number of Incorrect Responses	Accuracy (Acc %)
Simple Dialogue	509	491	18	96.47
Complex Dialogue	396	373	23	94.19
Dialogue with Mild Pronunciation Errors	621	550	71	88.57
Dialogue with Severe Pronunciation Errors	724	626	98	86.47
Total	2,250	2,040	210	92.67

Over a four-week period, the error rate was reduced from approximately 40% to around 15%, significantly enhancing communication skills and fostering more natural and confident speech habits.

A key highlight of this research is the conversion of sound into Vietnamese lip shapes displayed via the Robot face model. This feature provides visual feedback, allowing children to identify and replicate correct lip positions, thus improving pronunciation. The real-time lip shape simulation promotes active learning, which increases both engagement and the effectiveness of

practice. Additionally, the system integrates an intuitive interface with features such as parallel display of original and corrected sentences, learning history storage, and common error lists. These features support personalized learning and provide valuable data for parents and teachers, helping to improve children's learning performance.

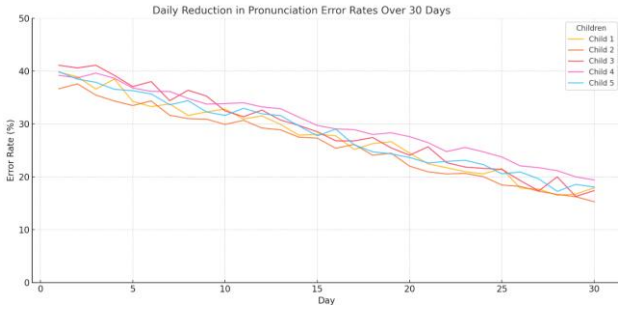

Figure 3. *Daily reduction in pronunciation error rates over 30 days*

Furthermore, the research contributes to the application of artificial intelligence in Vietnamese language processing, expanding the ability to optimize AI algorithms with different language characteristics. Compared to traditional methods, the system is not only fast and flexible but also highly applicable in education and speech therapy. The integration of the lip display function further enhances the value and wide applicability of the system in improving communication skills and quality of life.

4. Conclusion

Research shows that it is possible to use Robots. This study demonstrates the potential of AI-powered systems in addressing speech disabilities, particularly stuttering. By combining AI-driven stuttering detection with visual feedback through synchronized mouth shape simulations, the system provides an innovative solution for improving Vietnamese speech fluency. Furthermore, with the ability to personalize, express emotions, communicate using Vietnamese and provide accurate feedback, Robots not only help children access education more effectively but also create a comfortable and fun learning environment.

REFERENCES

[1] M. Nawar, M. Nizamani, Mehak, and R. Hameed, “Analyze How Children Acquire Language and the Cognitive Processes Involved, including the Role of Environmental and Social Factors”, *Bulletin of Business and Economics*, vol. 13, no. 3, pp. 239-247, 2024. DOI:10.61506/01.00483

[2] H. M. Feldman, “How young children learn language and speech: Implications of theory and evidence for clinical pediatric practice”, *Pediatrics in Review,* vol. 40, no. 1, pp. 398-411, 2019. doi: 10.1542/pir.2017-0325

[3] R. Riad, M. W. Allodi, E. Siljehag, and S. Bolte, “Language skills and well-being in early childhood education and care: a cross-sectional exploration in a Swedish context”, *Frontiers in Education*, vol. 8, 2023. https://doi.org/10.3389/feduc.2023.963180

[4] G. T. Pham *et al.,* “Identifying Developmental Language Disorder in Vietnamese Children, Journal of Speech”, *Language, and Hearing Research*, vol. 62, no. 5, pp. 1452-1467, 2019. doi: 10.1044/2019_JSLHR-L-18-0305

[5] K. Panesar and M. B. P. Cabello de Alba, “Natural language processing-driven framework for the early detection of language and cognitive decline”, *Language and Health*, vol. 1, pp. 20-35, 2023. https://doi.org/10.1016/j.laheal.2023.09.002

[6] F. Zafar *et al*., “The Role of Artificial Intelligence in Identifying Depression and Anxiety: A Comprehensive Literature Review”, *Cureus,* vol. 16, no. 3, 2024. doi: 10.7759/cureus.56472

[7] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, “Natural language processing applied to mental illness detection: a narrative review”, *NPJ Digital Medicine,*vol. 5, 2022. https://doi.org/10.1038/s41746-022-00589-7