DaNangVMD: Vietnamese Speech Mispronunciation Detection

Nguyen Ket Doan[1], Tran Nguyen Anh[1], Vo Van Nam[1], Nguyen Tran Tien[1], Le Pham Tuyen[2], Nguyen Quoc Vuong[3] and Nguyen Huu Nhat Minh[1]

[1] The University of Danang, Vietnam - Korea University of Information and Communication Technology, Vietnam

[2] Industrial University of Ho Chi Minh City, Ho Chi Minh, Vietnam

[3] Dong A University, Danang, Vietnam

Correspondence: Nguyen Huu Nhat Minh, nhnminh@vku.udn.vn

Received: 04/2/2024, revised: 19/5/2024, accepted: 27/5/2024

Digital Object Identifier: 10.32913/mic-ict-research-vn.v2024.n1.1271

Abstract: Automatic Speech Recognition, also known as ASR, has grown exponentially over the past decade and is used to recognize and translate human speech into readable text automatically. However, Vietnamese Speech Recognition faces critical challenges such as frequent mispronunciations as well as a huge variant in Vietnamese speech. In this work, we dive into the difficult challenge of Mispronunciation Detection (MD) in the Vietnamese language. As such a tonal language, Vietnamese is not only based on consonants and vowels but also on variations in pitch or tone during pronunciation. In this paper, we propose DaNangVMD model for detecting mispronunciations in Vietnamese speech based on the audio speech and canonical transcript. By leveraging multi-head attention-based multimodal representation from the embeddings of the phonetic encoder and linguistic encoder, DaNangVMD aims to provide a robust solution for accurate mispronunciation detection and diagnosis. Throughout the extensive evaluation, the proposed DaNangVMD exhibits superior performances rather than that of the PAPL baseline models by 15% in F1 score and 13% in accuracy.

Keywords: *mispronunciation detection, multimodal embeddings, Vietnamese speech recognition.*

Tiêu đề: **DaNangVMD: Nhận diện phát âm sai tiếng Việt**

Tóm tắt: Nhận diện giọng nói tự động, còn được gọi là ASR, đã phát triển mạnh mẽ trong thập kỷ qua và được sử dụng để nhận diện và dịch giọng nói của con người thành văn bản đọc được một cách tự động. Tuy nhiên, nhận diện giọng nói tiếng Việt gặp phải những thách thức lớn như lỗi phát âm thường xuyên cũng như sự biến đổi lớn trong giọng nói tiếng Việt. Trong công trình này, chúng tôi giải quyết thách thức của việc phát hiện lỗi phát âm (MD) trong tiếng Việt. Là một ngôn ngữ có thanh điệu, tiếng Việt không chỉ dựa trên các phụ âm và nguyên âm mà còn phụ thuộc vào sự thay đổi trong cao độ hoặc thanh điệu trong quá trình phát âm. Trong bài báo này, chúng tôi đề xuất mô hình DaNangVMD để phát hiện lỗi phát âm trong giọng nói tiếng Việt dựa trên âm thanh giọng nói và văn bản đúng. Bằng cách tận dụng biểu diễn đa mô hình dựa trên sự chú ý đa đầu từ các mã hóa của bộ mã hóa ngữ âm và bộ mã hóa ngôn ngữ, DaNangVMD nhằm cung cấp một giải pháp mạnh mẽ cho việc phát hiện và chẩn đoán lỗi phát âm chính xác. Qua đánh giá mở rộng, DaNangVMD đề xuất cho thấy hiệu suất vượt trội so với các mô hình PAPL với mức tăng 15% trong điểm F1 và 13% trong độ chính xác.

Từ khóa: *nhận diện phát âm sai, biểu diễn đa phương thức, nhận diện giọng nói tiếng Việt*

I. INTRODUCTION

Nowadays, organizations, individuals, and businesses increasingly embrace AI as a supportive tool, voice input assumes a pivotal role in the operations. Voice input facilitates tasks such as searching, and popular social media platforms utilize automatic speech recognition to generate automatic subtitles for videos posted on platforms like YouTube, Facebook, TikTok, among others. However, there are still many errors in recognition due to the system not yet fully optimized for Vietnamese speech. Furthermore, there more and more foreigners working in Vietnam and preschool children who have a huge demand to study Vietnamese language. However, current language learning applications only focus on English pronunciation and do not focus on Vietnamese pronunciation.

Mispronunciation Detection (MD) is an crucial application of phoneme recognition which is a core of Computer-Aided Pronunciation Training (CAPT). This module plays

an important role in speech recognition for tonal languages such as Vietnamese and has recently attracted significant attention from researchers and many methods have been proposed to address these issues. However, up to now, results have not been effective due to the complex and variable nature of the target being recognized as human speech. A multitude of factors, including linguistic differences, syntactical, phonetic structures, pronunciation, and intonation contribute to this complexity. In the initial approach, a promising ASR end-to-end structure for the mispronunciation detection task, called CNN-RNN-CTC [1] was proposed with the phoneme representation by associating raw acoustic features with the corresponding pronounced phoneme sequences, facilitated by the utilization of the Connectionist Temporal Classification (CTC) loss function [1], showed superior performance, surpassing the outcomes of Kun Li et al [2]. However, researchers were not utilizing the pre-existing prior texts from canonical sequences.

In the MD task, because the canonical phoneme sequences are available, Linguistic data from the canonical text can be helpful to integrate into models to improve the performance. SED-MDD [3] was proposed by leveraging the attention mechanism to combine acoustic features and corresponding canonical character sequences. Besides, Kaiqi et al [4] proposed the model by combining the acoustic feature with the encoded phoneme features to improve the linguistic embedding, but the low-level features of the acoustic encoder might be sensitive to noise and affect the model performance. Thus, the Acoustic-Phonetic-Linguistic (APL) [5] approach has dramatically improved the performance of MD and achieved state-of-the-art results by including the phonetic embedding to the model. Moreover, due to the tonal nature of the Vietnamese speech, the APL architecture alone is inadequate for addressing its complexity and variety. Identifying tone is also a challenge for MD tasks in Vietnamese. In fact, the pitch feature, which represents the high and low frequency, is measured in the time and frequency domain which can provide useful features to identify the tone. An APL-based architecture, called PAPL leverages the pitch feature, extracts pitch by utilizing the Kaldi toolkit [6], and achieves promising results for VMD. In this work, we propose the DaNangVMD architecture for Vietnamese Speech mispronunciation detection by customizing the APL [5], and PAPL [6] architecture by utilizing Phonetic and linguistic embeddings for a better and more robust MD solution. Different from APL, PAPL we introduce Phoneme type and multi-head attention across modals data (i.e., speech and transcript), and eliminate unnecessary components in the models to produce the smaller model with a better performance in VMD by a large margin.

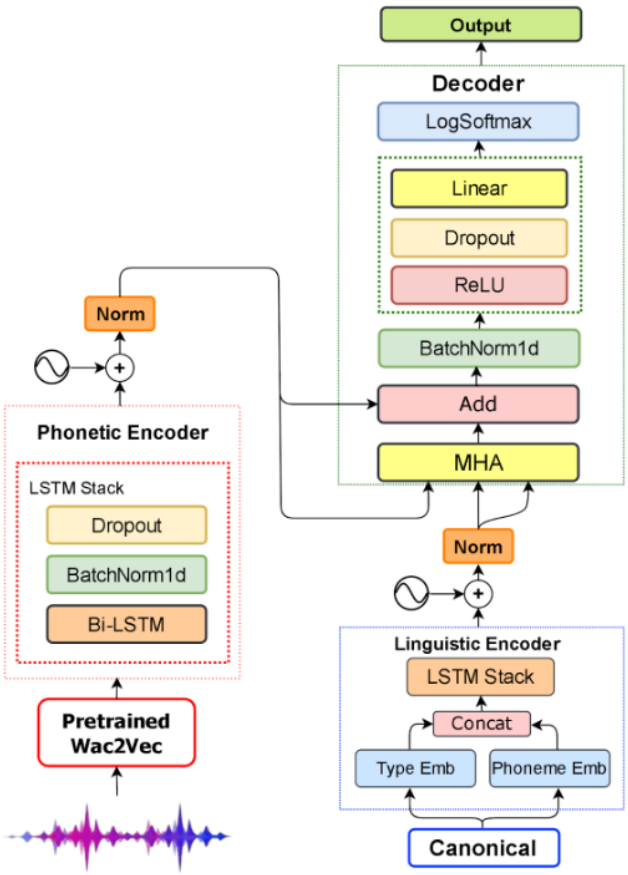

Figure 1. The proposed DaNangVMD for Vietnamese Speech mispronunciation detection

II. THE PROPOSED DANANGVMD ARCHITECTURE

1. Overview of Vietnamese phonemes

An overview of Vietnamese syllables is illustrated in Figure 2, where the hierarchical structure [7] is proposed. Each syllable of Vietnamese is divided into 3 parts: Initial, Rhyme, and Tone. Rhyme is broken down into Medial, Nucleus, and Ending. There are six tones in Vietnamese: Mid-Level Tone (no tone mark), Low Falling Tone (‘), High Rising Tone (´), Low Rising Tone (?), High Broken Tone (~), and Heavy Tone (.). This study maps each Vietnamese syllable to a phoneme, encompassing 53 phonemes. This mapping preserves the hierarchical structure of syllables, facilitating the detection of various types of mispronunciations. The example of syllable conversion to phoneme is presented in Figure 3. Because the phoneme can keep the original hierarchical structure of syllables, this mapping can allow us to detect all types of mispronunciation. In

the DaNangVMD, we will incorporate the phoneme type to better represent the canonical transcript to preserve the basic structures of Vietnamese speech.

2. DaNangVMD model

a) 2.2.1. Baseline model

We have chosen the recent PAPL architecture [4] which has been considered the state-of-the-art for Vietnamese MD as the baseline model for detecting and diagnosing mispronunciations. The PAPL approach integrates pitch, acoustic, phonetic, and linguistic features into the common embeddings for the end-to-end training with a CTC loss function. This enables the model to effectively learn and extract patterns, enhancing its performance in addressing Vietnamese mispronunciations. While conducting the experiments, we discovered that extracted acoustic and pitch features are sensitive to noise and have the potential to yield a negative impact on the overall output, especially in the small training dataset. On the other hand, the phoneme and tone features are already represented in the pre-trained model Wav2Vec2 for Vietnamese speech recognition using a large speech dataset [11]. This pre-trained model brings an incredible result in extracting phonetic embedding even when the data includes noises. Therefore, we decide to eliminate the acoustic and pitch encoder from PAPL architecture and retain the phonetic encoder for learning the relationships between audio frames. Additionally, we customize the linguistic encoder with the extra information from the given transcript and the decoder for phoneme recognition. The sections below will discuss all the components appearing in our proposed model.

b) 2.2.2 Detailed architecture

Phonetic Encoder

In the phonetic encoder, we utilize Wav2Vec2.0 [9] one of the most powerful ASR models for extracting the phonetic embedding which was fine-tuned on a large Vietnamese Speech dataset [11]. The phonetic embedding is fed into an encoder designed with a Bi-LSTM layer, a BatchNorm layer, and a Dropout layer. Additionally, we also leverage the Positional Encoding [14] - this technique provides the position of each audio frame, and this information enhances the output of the LSTM stack. Finally, we normalize this result by the Norm layer and achieve a matrix $H_q = [h_1^q, ..., h_T^q]$. This process is illustrated in Figure 1. We detected the redundancy of CNN layer of the phonetic encoder due to their already existed in Wav2Vec2.0 and produce low performance. As the results, we can reduce the model size significantly.

Linguistic Encoder

In the MD task, the language encoder is also a necessary component in identifying incorrectly pronounced phonemes. The input of this encoder will be canonical phoneme sequences (i.e., transcript) embedded into one hot embedding vector. Additionally, utilizing the genre of canonical phonemes such as Initial, Medial, Nucleus, Ending, and Tone is also crucial to enhance the robustness of canonical representation. It helps the model grasp the structure of phoneme sequences and prior information on identifying phonemes, as shown in Figure 1. Canonical phoneme embedding and phoneme type embedding will be concatenated and passed through a Bi-LSTM. The output of the Bi-LSTM $H_k = H_v = [h_1^{k,v}, ..., h_N^{k,v}]$ will be used as keys and values in the cross attention of the decoder. The N symbol represents the number of phonemes in the sentence.

Decoder

For the decoder, we implement Multi-Head cross Attention 1 (MHA) instead of the single-head attention mechanism utilized in the baseline model. The vector H_q from the Phonetic decoder and H_k, H_v from the Linguistic decoder are transmitted to the Multi-Head Attention as follows:

$$\begin{aligned} &\textbf{MHA}(H_q, H_k, H_v) = \textbf{Concat}(head_1, ..., head_h)W^O \\ &\text{and } head_i = \textbf{Attention}(H_q W_j^{H_q}, H_k W_j^{H_k}, H_k W_j^{H_k}) \end{aligned} \quad (1)$$

This mechanism enables the model to match the relevant linguistic features corresponding to specific segments of the extracted phonetic features. Thus, multiple cross-relations were built upon MHA to strengthen the correlation between phonetics and linguistic features which implicitly represented in the attention score and helped to boost pronunciation recognition performance. In the mispronounced cases, the phonetic feature needs to be emphasized to provide additional information for the decoder, hence the output of MHA is added to H_q. Thereby enhancing the model's capacity to attend to mispronounced phonemes and phonetic features. Subsequently, the combined output undergoes BatchNorm, Relu, Linear, Dropout, and finally, the LogSoftmax layer to compute probabilities based on audio frames. This entire process is depicted in Figure 1 to illustrate the stages of our model architecture.

III. DATASETS

In this work, we conduct the training and testing for the MD task of Vietnamese speech with two MD datasets [13] for detecting mispronunciation. The first dataset [10] contains augmented recordings of adults speaking pairs of

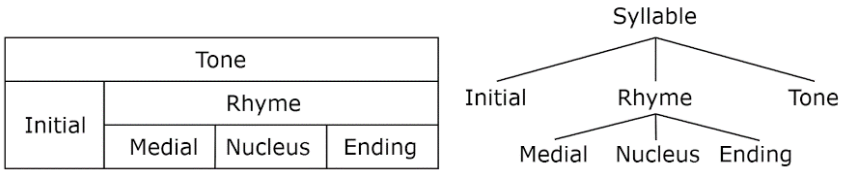

Figure 2. The hierarchical structure of Vietnamese syllables

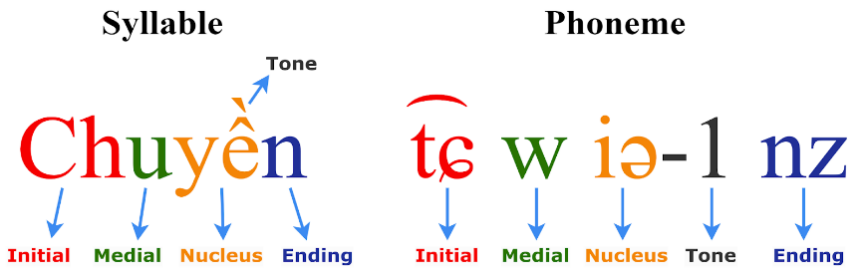

Figure 3. Example of a Vietnamese phoneme presentation.

single-syllable Vietnamese words (released by MachinaX). The second dataset [4] features recordings of children aged five to seven, either speaking or reading Vietnamese sentences in passages or dialogues (released by SoICT-HUST and the Vietnam Psycho-Pedagogical Association). For the first dataset, it is created by a set of pairs of single-syllable words. Therefore, a pair may or may not have a meaning but be easier to pronounce. Each pair will be generated into five other variations, corresponding to the overall six tones in Vietnamese: mid-level, high-rising, low-falling, high-rising-glottal, low-falling-rising, low-falling-glottal. These pairs of single-syllable words will later be recorded by a native speaker. For the second dataset, a group of native children who come from both kindergarten and primary schools participated in creating this dataset. The primary school children were recorded reading sample sentences that were collected from various Vietnamese schoolbooks. There are 20 annotators who were trained to evaluate the entire dataset and mark any instances of incorrect or defective pronunciation. There are a total of 4983 audio, of which 4263 are recordings of native children's voices, the rest are adult voices as shown in Table I.

Children aged from five to seven whose pronunciation varied from very poor and adults deliberately mispronouncing pairs of single-syllable Vietnamese words (such as tone, initial, etc.) as the testing set as described in Table II.

Table I
DETAILS OF THE COLLECTED DATASET [6]

Properties	Children	Adult
# of Recording	4263	746
# of Hours	5.68	0.36

Table II
DETAILS OF THE VIETNAMESE DATASET USED IN THE EXPERIMENTS

Properties	Training	Testing
# of Recording	4101	880

IV. EXPERIMENTS

1. Experimental Setups

In this work, we validate nine models for evaluation and incrementally find out based on the ablation study the best architecture for DaNangVMD:

- **PAPL Baseline:** Pitch - Acoustic - Phonetic - Linguistic model [6].
- **PPL**: Using Pitch features (Kaldi library [15]), Phonetic embeddings, and Linguistic embeddings (canonical phoneme sequences) as input and not using Acoustic features.
- **APL**: Taking Acoustic features (mel spectrogram), Phonetic embeddings, and Linguistic embeddings as input and not using Pitch features.
- **Customized PAPL (C-PAPL)**: Similar to the baseline PAPL model [6], but without using CNN in the Phonetic Encoder.
- **C-PL**: Use only Phonetic embeddings and Linguistic embeddings as input and don't use CNN in the Phonetic Encoder.
- **C-PL1**: Similar to the C-PL model, but using additional Phoneme Type Embedding in the Linguistic

Encoder.

- **C-PL2**: Similar to the C-PL model, but using an LSTM Layer (LSTM → Norm → Dropout) instead of LSTM in the Linguistic Encoder.
- **C-PL3**: Similar to the C-PL model, but using Multi-Head Attention (MHA) in Decoder.
- **DaNangVMD (Proposed architecture)**: Taking Phonetic embeddings and Linguistic embeddings as input; without using CNN in the Phonetic Encoder; using additional Phoneme Type Embedding and an LSTM Layer in the Linguistic Encoder; using Multi-Head cross Attention (MHA) in Decoder as shown in Figure 1.

The audio files were uniformly sampled at a rate of 16,000Hz, and we derived mel-spectrogram, pitch, and phonetic embeddings using a frameshift of 20*ms* and a frame length of 25*ms*. For model training, we employed the AdamW optimizer, setting a learning rate of 0.00005 with a maximum of 100 epochs.

2. Phoneme Recognition

To align the ASR predictions with human annotations, we utilized the Needleman-Wunsch algorithm [14]. We conducted our evaluation metrics using the formula 2, where “I" represents insertions, “D” represents deletions, “S” represents substitutions, and “N” represents the total number of phonetic units as follows:

$$\begin{aligned} Correctness &= \frac{N-S-D}{N}, \\ Accuracy &= \frac{N-S-D-I}{N} \end{aligned} \tag{2}$$

As a result, Table III shows the outcomes of our phoneme recognition performance from all models. The accuracy for phone recognition of PAPL baseline architecture stands at 79.46%. However, with PPL, the accuracy rises to 80.95%, APL achieves 81.41%, and C-PL sees an improvement to 82.87%. Therefore, this ablation study indicates that we can enhance the model accuracy by omitting Acoustic and Pitch features and excluding CNN in the Phonetic Encoder enhances the model accuracy. This improvement could be attributed to the good quality pre-trained Wav2Vec2.0 model in extracting Acoustic and Pitch features instead of training from the scratch acoustic, pitch encoder from a small mispronunciation dataset. In addition, since the Wav2Vec2.0 model already incorporated CNN layers, hence utilizing CNN in the Phonetic Encoder becomes redundant. Based on this observation, we validated this C-PAPL model without CNN layers in the Phonetic encoder. Consequently, we produced the **C-PL** model, resulting in a better accuracy of 82.87% and a correctness of 84.86%. These findings demonstrate the substantial improvement of the **C-PL** architecture over the baseline PAPL for more precise phoneme recognition as well as substantially reducing the model size.

Building upon the **C-PL** model, we iteratively refined its variations. **C-PL1** exhibited an accuracy of 83.55%, **C-PL2** reached 84.46%, and notably, **C-PL3** with multi-head cross attention achieved an impressive 90.68% accuracy, signifying a substantial improvement. This highlights the effectiveness of incorporating additional Phoneme Type Embedding, employing an LSTM Layer in the Linguistic Encoder, and Multi-Head Attention (MHA) in the Decoder in the architecture. Based on all of the aforementioned observations and enhancements, we built the **DaNangVMD** model, resulting in an accuracy of 92.3% and correctness of 93.53%. This marks a noteworthy increase of 12.84% and 11.39% compared to the original PAPL baseline, respectively. DaNangVM exhibits superior performance compared to the other baselines, particularly in timbre recognition. Moreover, **DaNangVMD** achieves a significant reduction in the number of model parameters, diminishing from 64.6M to 10.4M. This substantial reduction substantially improves the speed of phoneme recognition in the Vietnamese language and training time.

3. Mispronunciation Detection Result

The hierarchical evaluation structure is used to measure the MD system performance. With correct pronunciations of phones, *TA (True Acceptance)* means the recognized phone sequence is matched with the reference text, and *FR (False Rejection)* means the recognized phone sequence differs from the reference. With mispronounced phones, *TR (True Rejection)* means the mispronounced phones are detected, and *FA (False Acceptance)* means failing to recognize the mispronounced phones.

$$\begin{aligned} Recall &= \frac{TR}{FA+TR}, \; Precision = \frac{TR}{FR+TR} \\ F1Score &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{aligned} \tag{3}$$

As shown in Table III, the PAPL baseline model achieved the highest Recall of 87.95% but the Precision score was very low at only 29.72%, leading to a low F1 score of only 44.43%. We analyzed the PAPL model results which have a high Recall but low Precision score due to the heavy dependence on the canonical labels and could not detect the mispronounced phoneme. On the other hand, the proposed DaNangVMD model overcomes the above problems and balances the recall and precision performance by leveraging the refined Linguistic Encoder, and Multi-Head Attention in the Decoder of the model. The refined structure reduces

Table III
EXPERIMENTAL RESULTS OF VMD MODELS FOR VIETNAMESE MD TASK. NOTE THAT, WE DIDN'T COUNT THE NUMBER PARAMETERS OF THE PRE-TRAINED VIETNAMESE WAVE2VEC2.0 ENCODER FOR ALL OF THE MODELS

VMD Model	**Params**	**Phoneme Recognition performance (%)**		**Mispronunciation Detection and Diagnosis (%)**		
		Accuracy	**Correctness**	**Recall**	**Precision**	**F1**
PAPL *(Baseline)*	64.6M	79.46	80.91	**87.95**	29.72	44.43
PPL *(without Acoustic)*	46.2M	80.95	82.59	86.79	31.76	46.51
APL *(without Pitch)*	61M	81.41	82.89	85.37	32.26	46.83
C-PAPL *(without CNN in Phonetic)*	29.2M	81.38	82.8	86.25	32.03	46.71
C-PL	8M	82.87	84.96	68.39	31.48	43.12
C-PL1 *(Using Phoneme Type)*	8.6M	83.55	85.34	82.5	35.66	49.8
C-PL2 *(Using LSTM Layer)*	8M	84.46	86.13	83.75	37.39	51.69
C-PL3 *(MHA in Decoder)*	9.6M	90.68	92.9	32.68	47.41	38.69
DaNangVMD *(Proposed)*	10.4M	**92.3**	**93.53**	55.54	**63.73**	**59.35**

the over-dependency on the canonical via MHA and adds a skip connection after MHA from phoneme representation. Hence, the acoustic features impact more on the predictive capabilities of the model. As the results, DaNangVMD outperforms PAPL baseline 14,92% and 34,01% in terms of F1 and Precision, respectively. These results highlight the outstanding performance of our DaNangVMD model, particularly in the realm of mispronounced phoneme recognition within the Vietnamese language.

V. CONCLUSIONS

In this paper, we propose the DaNangVMD architecture designed for detecting mispronunciations in Vietnamese speech. Our findings in removing the CNN layers in the phonetic encoder, integrating Phoneme type embeddings in the Phonetic Encoder, and utilizing MHA in the Decoding, prove significant enhancement of phoneme recognition in the context of the Vietnamese language. The experimental results demonstrate the effectiveness of the proposed DaNangVMD system, by significantly improving Recall, Precision, and F1 scores by 55.54%, 63.73%, and 59.35%, respectively. The overall accuracy reached an impressive 92.3%, representing a substantial advancement compared to the baseline model. In future works, we involve further development of the DaNangVMD system into a more robust and effective speech recognition tool by leveraging diverse and extensive Vietnamese datasets, encompassing regional and ethnic minority voices.

ACKNOWLEDGEMENT

This work is under the support of Vietnam - Korea University of Information and Communication and the dataset is provided by Association for Vietnamese Language and Speech Processing.

REFERENCES

[1] Wai-Kim Leun. "CNN-RNN-CTC BASED END-TO-END MISPRONUNCIATION DETECTION AND DIAGNOSIS." *Department of Systems Engineering and Engineering Management*, https://www1.se.cuhk.edu.hk/ hccl/publications/pub/ICASSP2019-mdd.pdf. Accessed 22 November 2023.

[2] Kun Li, and Helen Meng. "Mispronunciation Detection and Diagnosis in L2 English Speech Using Multi-Distribution Deep Neural Networks." *Human-Computer Communications Laboratory Department of System Engineering and Engineering Management The Chinese University of Hong Kong, Hong Kong SAR*, China, 16 June 2023, https://www1.se.cuhk.edu.hk/ hccl/publications/pub/2014 PID3298385 LK.pdf. Accessed 22 November 2023.

[3] Yiqing Feng, et al. "SED-MDD: Towards Sentence Dependent End-To-End Mispronunciation Detection and Diagnosis." 16 June 2023, https://ieeexplore.ieee.org/document/9052975. Accessed 24 November 2023.

[4] Kaiqi Fu. Jones Lin, Dengfeng Ke, Yanlu Xie, Jinsong Zhang, Binghuai Lin "A Full Text-Dependent End to End Mispronunciation Detection and Diagnosis with Easy Data Augmentation Techniques." arXiv, 17 April 2021, https://arxiv.org/abs/2104.08428. Accessed 24 November 2023.

[5] Wenxuan Ye, et al. "An Approach to Mispronunciation Detection and Diagnosis with Acoustic, Phonetic and Linguistic (APL) Embeddings." arXiv, 14 October 2021, https://arxiv.org/abs/2110.07274. Accessed 24 November 2023.

[6] Huu Tuong Tu, et al. "Mispronunciation detection and diagnosis model for tonal language, applied to Vietnamese." [Dublin, Ireland], 20-24 8 2023, https://www.isca-speech.org/archive/pdfs/interspeech 2023/huu23 interspeech.pdf.

[7] Thi Thu Trang Nguyen. HMM-based Vietnamese Text-To-Speech: Prosodic Phrasing Modeling, *Corpus Design System Design, and Evaluation*, 22 January 2016, https://theses.hal.science/tel-01260884/document. Accessed 24 November 2023.

[8] Maxprotect, https://www.maxprotect.com/. Last accessed 21 Feb. 2024

[9] M. Hammami, Y. Chahir and L. Chen, "WebGuard: Web based adult content detection and filtering system," *In Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, Halifax, NS, Canada, 2003, pp. 574-578.

[10] Hu, Weiming, Haiqiang Zuo, Ou Wu, Yunfei Chen, Zhongfei Zhang, and David Suter. "Recognition of adult images, videos, and web page bags." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 7, no. 1 (2011): 1-24.

[11] Sharma, Preeti, Manoj Kumar, and Hitesh Sharma. “Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: an evaluation." *Multimedia Tools and Applications* 82, no. 12 (2023): 18117-18150.
[12] Soliman, Mohamed Mostafa, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. “Violence recognition from videos using deep learning techniques." *In 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 80-85. IEEE, 2019.
[13] Human Action Recognition (HAR) Dataset. https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset. Last accessed 21 Feb. 2024
[14] Caruana, Rich. “Multitask learning." Machine learning 28 (1997): 41-75.
[15] M. Ravanelli, T. Parcollet, Y. Bengio, "The PyTorch-Kaldi Speech Recognition Toolkit", https://arxiv.org/abs/1811.07453. Accessed 19 Nov 2018.

Nguyen Ket Doan is pursuing the B.Eng. degree in Information Technology from the University of Danang, Vietnam - Korea University of Information and Communication Technology. His research interests include software development, machine learning and deep learning.
Email: nkdoan.20it7@vku.udn.vn

Tran Nguyen Anh is pursuing the B.Eng. degree in Information Technology from the University of Danang, Vietnam - Korea University of Information and Communication Technology. His research interests include software development, machine learning and deep learning.
Email: anhtn.21it@vku.udn.vn

Vo Van Nam is pursuing the B.Eng. degree in Information Technology from the University of Danang, Vietnam - Korea University of Information and Communication Technology. His research interests include software development, machine learning and deep learning.
Email: namvv.21it@vku.udn.vn

Nguyen Tran Tien is pursuing the B.Eng. degree in Information Technology from the University of Danang, Vietnam - Korea University of Information and Communication Technology. His research interests include software development, machine learning and deep learning.
Email: nttien.20it6@vku.udn.vn

Le Pham Tuyen received the B.S. degree in computer science from Ho Chi Minh City University of Technology, Vietnam, in 2013, and the Ph.D. degree in computer science and engineering from Kyung Hee University, South Korea, in 2019. He is currently an Principal Research Engineer at AgileSoDA Company and visiting lecturer at Industrial University of Ho Chi Minh City. His current research interests include machine learning, reinforcement learning, combinatorial optimization, and robotics.
Email: tuyen_01036033@iuh.edu.vn

Nguyen Quoc Vuong receive master degree computer science from the University of Da Nang, in 2011. His research interests include software development, machine learing and deep learning.
Email: voung@donga.edu.vn

Nguyen Huu Nhat Minh (M’20) received Ph.D. degree in Computer Science and Engineering from Kyung Hee University, South Korea, in 2020. He continued Post-Doc with Federated Learning and Democratized Learning at Intelligent Networking lab, Kyung Hee University, South Korea. He is Deputy Head of Department of Science, Technology, and International Cooperation, and In charge of Research Program at Digital Science and Technology Institute, The University of Danang – Vietnam - Korea University of Information and Communication Technology, Vietnam. He received the best KHU Ph.D. thesis award in engineering in 2020. He had publications in premier ACM/IEEE journals and conferences. His research interests include wireless communications, federated learning, NLP, and computer vision.
Email: nhnminh@vku.udn.vn