p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

BẢO TỒN ĐẶC TRƯNG ÂM VỊ CHO ĐÁNH GIÁ PHÁT ÂM TỰ ĐỘNG

Nguyễn Thị Bích Nhật[1]

TÓM TẮT

Nghiên cứu này giới thiệu một phương pháp cải tiến trong đánh giá phát âm tự động (Automatic Pronunciation Assessment - APA) thông qua cơ chế điều chuẩn tương phản thứ bậc. Phương pháp này tập trung vào việc bảo tồn đặc trưng âm vị trong biểu diễn ngữ âm bằng cách kết hợp ba hàm mất mát: tương phản, âm vị và thứ bậc. Mô hình phân tầng gồm ba cấp độ (âm vị, từ và câu) được triển khai nhằm tối ưu hóa khả năng phát hiện và định lượng lỗi phát âm. Kết quả thực nghiệm trên bộ dữ liệu tiếng Anh của người học tiếng Việt cho thấy phương pháp đề xuất đã cải thiện đáng kể độ chính xác của hệ thống đánh giá phát âm. Phương pháp nghiên cứu mở ra tiềm năng ứng dụng trong các hệ thống học ngôn ngữ thông minh.

Từ khóa: *Đánh giá phát âm tự động, điều chuẩn phân tầng, biểu diễn ngữ âm, lỗi âm vị.*

DOI: *https://doi.org/10.70117/hdujs.79.09.2025.953*

1. ĐẶT VẤN ĐỀ

Phát âm là một thành phần thiết yếu trong năng lực giao tiếp ngôn ngữ. Đối với người học tiếng Anh như một ngôn ngữ thứ hai (L2), phát âm không chuẩn không chỉ làm suy giảm độ lưu loát mà còn gây cản trở đáng kể trong quá trình truyền đạt và tiếp nhận thông tin [1]. Trong bối cảnh giáo dục ngôn ngữ ngày càng ứng dụng công nghệ, nhu cầu về các công cụ hỗ trợ phát hiện, đánh giá và cải thiện phát âm ngày càng trở nên cấp thiết. Điều này đã thúc đẩy sự phát triển của các hệ thống đánh giá phát âm tự động (Automatic Pronunciation Assessment - APA), với khả năng cung cấp phản hồi tức thì, khách quan và chi tiết đến người học [2].

Tuy nhiên, phần lớn các hệ thống APA hiện tại vẫn chỉ tập trung vào việc dự đoán điểm số tổng quát mà chưa tận dụng triệt để các đặc trưng ngữ âm để nhận diện và phản ánh chính xác lỗi sai cụ thể ở cấp độ âm vị [3][4]. Các biểu diễn mà mô hình học được thường không đủ phân biệt giữa các âm vị gần nhau, khiến phản hồi trở nên mơ hồ, thiếu hiệu quả, và khó triển khai trong thực tiễn giảng dạy cá nhân hóa [5]. Nhằm khắc phục hạn chế đó, nghiên cứu này đề xuất một phương pháp học biểu diễn định hướng ngữ âm, kết hợp ba cơ chế điều chuẩn: học tương phản, phân biệt âm vị và đánh giá theo thứ bậc. Mục tiêu là xây dựng một không gian đặc trưng ngữ âm giàu tính phân biệt và có khả năng phản ánh chính xác thứ tự chất lượng phát âm, từ đó cải thiện cả độ chính xác lẫn khả năng phản hồi chi tiết của hệ thống đánh giá [6]).

2. TÌNH HÌNH NGHIÊN CỨU

2.1. Khái quát về đánh giá phát âm tự động (APA)

Automatic Pronunciation Assessment (APA) là lĩnh vực giao thoa giữa nhận dạng tiếng nói, ngôn ngữ học và công nghệ giáo dục. Các hệ thống APA đã được ứng dụng rộng rãi trong nhiều nền tảng học ngôn ngữ như Duolingo, ELSA Speak, Google Pronunciation Tool và nhiều ứng dụng khác [7][8].

[1]*Khoa Kỹ thuật, Công nghệ và Truyền thông, Trường Đại học Hồng Đức; Email: nguyenthibichnhat@hdu.edu.vn*

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Phần lớn các mô hình APA hiện nay dựa trên kỹ thuật máy học sâu, đặc biệt sử dụng các mô hình biểu diễn ngữ âm tự giám sát như Wav2vec 2.0, HuBERT và WavLM để trích xuất đặc trưng từ tín hiệu âm thanh [5][9][10][12][13]. Tuy nhiên, đa số các hệ thống này chỉ đánh giá phát âm dựa trên điểm đồng nhất (holistic score) cho toàn bộ câu nói mà chưa có khả năng phân tích sâu các lỗi sai chi tiết theo từng âm vị hay âm tiết.

2.2. Vấn đề bảo tồn đặc trưng âm vị

Automatic Pronunciation Assessment (APA) là lĩnh vực giao thoa giữa nhận dạng tiếng nói, ngôn ngữ học và công nghệ giáo dục. Các hệ thống APA trên thực tế đã được ứng dụng rộng rãi trong nhiều phần mềm và nền tảng học ngôn ngữ như Duolingo, ELSA Speak, Google Pronunciation Tool... [7][8].

Đa số các mô hình APA hiện nay sử dụng các kỹ thuật máy học sâu, đặc biệt là các mô hình biểu diễn ngữ âm tự giám sát như Wav2vec 2.0, HuBERT và WavLM để trích xuất đặc trưng từ tín hiệu âm thanh [5][9][10]. Tuy nhiên, phần lớn các hệ thống này chỉ xây dựng dựa trên điểm đồng nhất (holistic score) để đánh giá toàn bộ câu nói mà chưa phân tích sâu các lỗi sai ở mức độ âm vị hay âm tiết. Do đó, việc bảo tồn và phân tích đặc trưng âm vị một cách chi tiết vẫn còn là một thách thức lớn trong nghiên cứu APA hiện nay.

2.3. Những nghiên cứu liên quan

Trong những năm gần đây, các phương pháp học biểu diễn phát âm dựa trên chiến lược học tương phản đã thu hút sự quan tâm đáng kể trong lĩnh vực đánh giá phát âm tự động. Các nghiên cứu tiêu biểu tập trung vào việc học không gian đặc trưng mà trong đó, các biểu diễn âm thanh chính xác được kéo gần về phía biểu diễn ngữ âm lý tưởng, trong khi các lỗi phát âm bị đẩy xa [6]. Đặc biệt, phương pháp học tương phản kết hợp với đặc trưng âm vị được xem là hướng đi triển vọng nhằm tăng cường khả năng phân biệt tinh vi giữa các lỗi phát âm khó như /s/-//, /t/-//, vốn phổ biến ở người học tiếng Anh như ngôn ngữ thứ hai. Một tiền đề gần đây là mô hình ConPCO (Contrastive Phonemic and Ordinal regularization), được đề xuất bởi [6], trong đó cấu trúc huấn luyện bao gồm ba thành phần: contrastive loss, phonemic loss và ordinal loss. Mô hình này được chứng minh là hiệu quả trong việc học biểu diễn ngữ âm giàu tính phân biệt và thứ bậc. Tuy vậy, các nghiên cứu hiện tại phần lớn mới chỉ áp dụng trong ngữ cảnh tiếng Anh chuẩn hoặc các bộ dữ liệu tổng quát như LibriSpeech, chưa đánh giá đầy đủ trong bối cảnh người học tiếng Anh bản ngữ không phải tiếng Anh, ví dụ như người Việt [14]. Đồng thời, cũng còn thiếu nghiên cứu về khả năng thích ứng của các phương pháp học sâu này đối với dữ liệu có yếu tố văn hoá đặc thù - như nội dung giáo dục, đọc văn bản học thuật hoặc phát âm trong bối cảnh học đường.

3. KIẾN TRÚC MÔ HÌNH CẢI TIẾN

3.1. Kiến trúc mô hình cải tiến

Mô hình được xây dựng theo kiến trúc phân tầng ba cấp độ, phản ánh ba mức đánh giá phổ biến trong hệ thống đánh giá phát âm: âm vị (phoneme-level), từ (word-level) và câu (utterance-level). Cấu trúc này cho phép mô hình khai thác đặc trưng ngữ âm từ đơn vị nhỏ nhất đến toàn câu, đảm bảo tính chi tiết và toàn diện trong đánh giá.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Tầng âm vị (Phoneme-level)

Đầu vào gồm các đặc trưng phát âm như: GOP (Goodness of Pronunciation), năng lượng, độ dài, MFCC, cùng các biểu diễn từ mô hình tự giám sát (HuBERT, Wav2vec 2.0, WavLM).

Biểu diễn văn bản âm vị được mã hóa qua embedding để tạo vector tương ứng.

Hai nguồn đặc trưng (âm thanh và văn bản) được kết hợp, sau đó đưa qua khối Branchformer nhằm tăng khả năng học phụ thuộc thời gian và mối quan hệ ngữ âm cục bộ.

Tầng từ (Word-level)

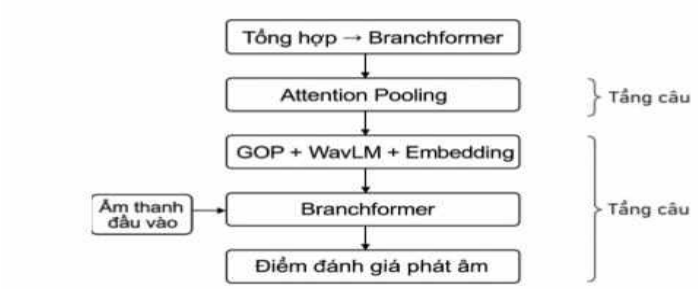
Sử dụng attention pooling để gom nhóm các biểu diễn âm vị thành biểu diễn từ.

Các vector từ tiếp tục được xử lý qua Branchformer, giúp đánh giá trọng âm, độ rõ ràng và lỗi nối âm ở cấp độ từ.

Tầng câu (Utterance-level)

Các biểu diễn từ được kết hợp và mã hóa lại thành biểu diễn toàn câu.

Một khối Branchformer cuối cùng xử lý biểu diễn này để đưa ra các đánh giá ở mức độ lưu loát, ngữ điệu, độ tự nhiên và điểm tổng quát. Tổng quan kiến trúc mô hình được minh họa trong Hình 1 dưới đây.



Hình 1. Kiến trúc phân tầng của mô hình đánh giá phát âm

3.2. Điều chuẩn không gian biểu diễn

Mô hình được huấn luyện với ba hàm mục tiêu chính: Hàm tương phản (contrastive loss), hàm âm vị (phonemic loss) và hàm thứ bậc (ordinal loss). Mỗi hàm đảm nhận một vai trò riêng trong việc điều chỉnh không gian biểu diễn đặc trưng. Cụ thể, hàm tương phản được áp dụng tại tầng âm vị, nhằm tối ưu hóa khoảng cách giữa biểu diễn âm thanh và văn bản tương ứng. Hàm âm vị được sử dụng tại tầng âm vị và tầng từ, giúp tăng cường khả năng phân biệt giữa các lớp âm vị. Trong khi đó, hàm thứ bậc được triển khai tại tầng câu, đảm bảo sự tương ứng giữa khoảng cách biểu diễn và chênh lệch điểm số đánh giá. Các nội dung này sẽ được tích hợp trực tiếp vào phần mô tả kiến trúc từng tầng nhằm tăng tính gắn kết giữa thiết kế hệ thống và chiến lược huấn luyện.

3.2.1. Hàm mất mát tương phản (Contrastive Loss)

Hàm mất mát tương phản được sử dụng để tối ưu khoảng cách giữa biểu diễn âm thanh $z_a \in R^d$ và biểu diễn văn bản $z_t \in R^d$, sao cho cặp đúng gần nhau và cặp sai cách xa trong không gian đặc trưng. Công thức như sau:

$$L_{contrastive} = -log \frac{\exp\left(sim(z_\alpha, z_t)/\tau\right.}{\sum_{i=1}^{N} \exp\left(sim(z_\alpha, z_j)/\tau\right.}$$

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Trong đó, sim(z_i,z_j) $\left(\frac{z_i{}^t\, z_j}{\|z_i\|.\|z_j\|}\right)$là độ tương đồng cosine, τ là hệ số nhiệt độ và N là số lượng biểu diễn văn bản (tức là số cặp z_j) được sử dụng trong quá trình tính toán hàm mất mát tại mỗi bước cập nhật trọng số.

3.2.2. Hàm mất mát âm vị (Phonemic Loss)

Hàm mất mát âm vị bao gồm hai thành phần: mất mát nội bộ và mất mát liên lớp. Trung tâm biểu diễn của lớp c được tính bởi: $\mu_c = \frac{1}{|S_c|}\sum_{i\epsilon S_c}(z_i)$

Mất mát nội bộ (intra-class): $L_{intra} = \frac{1}{N}\sum_{i=1}^{N} \| z_i - \mu_c||^2$

Mất mát liên lớp (inter-class): $L_{inter} = \sum_{c_1 \neq c_2} exp(-\frac{\|\mu_{c_1} - \mu_{c_2}||^2}{\sigma^2})$

Tổng hợp hàm mất mát âm vị: $L_{phonemic} = L_{intra} + \lambda L_{inter}$

3.2.3. Hàm mất mát thứ bậc (Ordinal Loss)

Hàm mất mát thứ bậc được thiết kế để đảm bảo khoảng cách biểu diễn tương ứng với sự chênh lệch điểm số đánh giá. Công thức như sau:

$$L_{ordinal} = \frac{1}{|P|}\sum_{(i,j\epsilon P)}(\| z_i - z_j||_2 - \alpha \| s_i - s_j||)^2$$

Trong đó, P là tập các cặp điểm trong batch, và α là hệ số điều chỉnh tỷ lệ. Hàm này duy trì sự tương ứng tuyến tính giữa độ lệch điểm số và khoảng cách trong không gian đặc trưng.

3.3. Thực nghiệm và đánh giá

3.3.1. Dữ liệu

Nghiên cứu sử dụng một tập dữ liệu gồm 1.500 câu đọc tiếng Anh được ghi âm bởi 75 sinh viên Trường Đại học Hồng Đức, thuộc hai nhóm: 30 người ngoài ngành (không chuyên ngôn ngữ) và 45 người chuyên ngành (đang theo học chuyên ngành tiếng Anh hoặc ngôn ngữ). Mỗi câu được chấm điểm thủ công theo thang 0-10 cho từng tiêu chí phát âm. Các đặc trưng đầu vào bao gồm GOP (Goodness of Pronunciation), WavLM và MFCC. Mô hình được huấn luyện theo kiến trúc phân tầng mô tả ở Mục 3, sử dụng thuật toán Adam (learning rate 1e-4, dropout 0.2), trong 100 epoch với cơ chế early stopping.

3.3.2. Chỉ số đánh giá

Hiệu quả của hệ thống được đánh giá thông qua ba chỉ số chính:

MSE (Mean Squared Error): đại diện cho sai số trung bình bình phương giữa điểm dự đoán $\hat{y}$ và điểm thực tế y, được tính theo công thức:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

PCC (Pearson Correlation Coefficient): đo lường mức độ tương quan tuyến tính giữa điểm dự đoán và điểm thực, được tính bằng:

$$\text{PCC} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})\ \ (\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}$$

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Accuracy: tỉ lệ phân loại đúng các mức điểm phát âm (ví dụ: Tốt, Trung bình, Kém) sau khi lượng hóa thành các mức rời rạc.

3.3.3. Kết quả thực nghiệm

Kết quả trên cho thấy mô hình đề xuất đạt hiệu quả vượt trội trên cả ba chỉ số. Đặc biệt, chỉ số Accuracy phản ánh rõ khả năng phân loại đúng các mức chất lượng phát âm, điều này có ý nghĩa thực tiễn quan trọng trong việc đưa ra phản hồi hiệu quả cho người học.

Bảng 1. So sánh kết quả giữa mô hình baseline và mô hình đề xuất

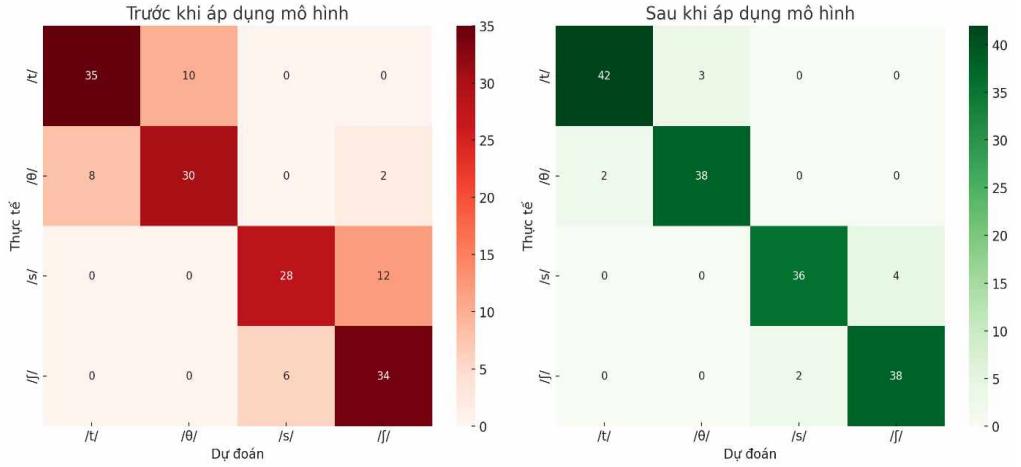
Tiêu chí	Baseline (GOP-only)	Mô hình cải tiến	Cải thiện
MSE	0.078	0.057	-26.9%
PCC	0.754	0.873	+15.7%
Accuracy	71.3%	84.5%	+13.2%

Để làm rõ hơn vai trò của từng thành phần trong kiến trúc mô hình, chúng tôi tiến hành thực nghiệm như sau:

So sánh theo nhóm người học: Mô hình đạt Accuracy 87.1% đối với nhóm chuyên ngành, và 81.2% đối với nhóm ngoài ngành, cho thấy khả năng thích ứng với sự đa dạng về nền tảng ngôn ngữ.

Phân tích đóng góp của từng đặc trưng đầu vào: Khi huấn luyện mô hình với từng loại đặc trưng riêng biệt, WavLM cho thấy mức cải thiện PCC lớn nhất (+11.2%), tiếp theo là GOP (+6.8%) và MFCC (+4.1%).

Đánh giá vai trò của hàm mất mát thứ bậc: Khi loại bỏ hàm ordinal loss khỏi quá trình huấn luyện, Accuracy giảm từ 84.5% còn 79.4%, khẳng định vai trò quan trọng của cơ chế học thứ bậc trong việc phản ánh chính xác chất lượng phát âm.



Hình 2. Kiến trúc phân tầng của mô hình đánh giá phát âm

Ngoài các đánh giá tổng thể, chúng tôi tiến hành thêm một phân tích định lượng chi tiết ở Bảng 2 về khả năng phân biệt lỗi giữa các cặp âm vị dễ gây nhầm lẫn, đặc biệt phổ biến ở người học tiếng Việt như /t/–/θ/ và /s/–/ʃ/. Kết quả thu được từ confusion matrix ở 2 cho thấy mô hình đề xuất giúp giảm đáng kể sai số nhận diện giữa các cặp âm này.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Bảng 2. So sánh kết quả giữa mô hình baseline và mô hình đề xuất

Cặp âm	Sai số trước mô hình	Sai số sau mô hình	Giảm sai số (%)
/t/–/θ/	18	5	72.2%
s/–/ʃ/	18	6	66.7%

Đánh giá này cho thấy rằng việc học tương phản đóng vai trò quan trọng trong việc đẩy biểu diễn của các lỗi phát âm ra xa khỏi vùng biểu diễn chuẩn trong không gian đặc trưng. Đồng thời, kiến trúc phân tầng của mô hình cũng hỗ trợ mô tả rõ nét hơn mối liên hệ ngữ âm từ cấp độ âm vị đến toàn câu, từ đó nâng cao độ nhạy đối với các lỗi âm vị tinh vi. Các biểu đồ trực quan sẽ được bổ sung trong phiên bản mở rộng để minh họa rõ hơn cho kết luận này.

4. KẾT LUẬN

Nghiên cứu này đã đề xuất một phương pháp mới trong lĩnh vực đánh giá phát âm tự động, dựa trên sự kết hợp giữa học tương phản và điều chuẩn thứ bậc để bảo tồn đặc trưng âm vị trong biểu diễn. Phương pháp tận dụng kiến trúc mô hình phân tầng cùng chiến lược huấn luyện tích hợp ba thành phần hàm mất mát (tương phản, âm vị và thứ bậc), từ đó không chỉ nâng cao độ chính xác trong việc chấm điểm phát âm mà còn cho phép hệ thống đưa ra phản hồi chi tiết ở cấp độ âm vị. Thực nghiệm trên bộ dữ liệu tiếng Anh của người học cho thấy phương pháp đề xuất đạt hiệu quả vượt trội so với các tiếp cận truyền thống, thể hiện qua mức sai số thấp hơn, tương quan cao hơn và khả năng phân loại rõ ràng các mức chất lượng phát âm. Trong tương lai, hướng phát triển tiếp theo sẽ bao gồm: Mở rộng áp dụng phương pháp sang các tình huống giao tiếp tự nhiên thay vì chỉ dựa trên đọc văn bản; Tích hợp hệ thống vào các ứng dụng học phát âm thời gian thực; Bổ sung các yếu tố ngữ điệu và trọng âm để tăng tính toàn diện của đánh giá; Thử nghiệm trên nhiều ngôn ngữ và đối tượng người học khác nhau nhằm kiểm định năng lực khái quát hóa của mô hình.

TÀI LIỆU THAM KHẢO

[1] Derwing, T. M., & Munro, M. J. (2005), *Second language accent and pronunciation teaching: A research-based approach*, TESOL Quarterly, 39(3) 379-397.

[2] Witt, S. M. (2012), *Automatic error detection in pronunciation training: Where we are and where we need to go*, Proceedings of the ISADEPT Workshop.

[3] Zhang, G., Song, K., Tan, X., Tan, D., Yan, Y., Liu, Y., Wang, G., Zhou, W., Qin, T., Lee, T., & Zhao, S. (2022), *Mixed-Phoneme BERT: Improving BERT with mixed phoneme and sup-phoneme representations for text to speech*, Proceedings of Interspeech, 456-460.

[4] Gong, Y., Chen, Z., Chu, I.-H., Chang, P., & Glass, J. (2022), *Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment*, Proceedings of ICASSP, 7262-7266.

[5] Chen, G., Qian, Y., & Yu, K. (2018), *Pronunciation assessment based on phonological features and acoustic model likelihood scores*, Proceedings of Interspeech, 2182-2186.

[6] Yan, B.-C., Chao, W.-C., Li, J.-T., Wang, Y.-C., Wang, H.-W., Lin, M.-S., & Chen, B. (2025), *ConPCO: Phonemic characteristic preserving for automatic pronunciation assessment via hierarchical contrastive ordinal regularization*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

[7] Liang, Y., Song, K., Mao, S., Jiang, H., Qiu, L., Yang, Y., Li, D., Xu, L., & Qiu, L. (2023), *End-to-end word-level pronunciation assessment with MASK pre-training.* Proceedings of Interspeech, 969-973.
[8] Elimat, A. K., & AbuSeileek, A. F. (2014), *Automatic speech recognition technology as an effective means for teaching pronunciation,* The JALT CALL Journal, 10(1) 21-47.
[9] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020), *wav2vec 2.0: A framework for self-supervised learning of speech representations.* Advances in Neural Information Processing Systems, 33, 12449-12460.
[10] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021), *HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451-3460.
[11] Chen, S., Huang, Z., & Huang, J. (2022), *WavLM: Large-scale self-supervised pretraining for full stack speech processing.* Proceedings of ICASSP 2022-IEEE International Conference on Acoustics, Speech and Signal Processing, 3537-3541. IEEE.
[12] Chung, Y.-A., Hsu, W.-N., Tang, H., & Glass, J. (2020), *An unsupervised autoregressive model for speech representation learning.* Proceedings of Interspeech.
[13] Kim, E., Jeon, J.-J., Seo, H., & Kim, H. (2022), *Automatic pronunciation assessment using self-supervised speech representation learning,* Proceedings of Interspeech 2022.
[14] Yassine El Kheir, A., Ali, A., & Chowdhury, S. A. (2023), *Automatic pronunciation assessment - A review.* Findings of the Association for Computational Linguistics: EMNLP, 8304-8324.

PRESERVING PHONETIC FEATURES FOR AUTOMATIC PRONUNCIATION ASSESSMENT

Nguyen Thi Bich Nhat

ABSTRACT

This paper proposes an enhanced approach to Automatic Pronunciation Assessment (APA) through a hierarchical contrastive regularization mechanism. The method emphasizes the preservation of phonetic features in speech representation by integrating three loss functions: contrastive, phonetic, and hierarchical. A stratified modeling framework with three levels-phoneme, word, and sentence-is employed to optimize the detection and quantification of pronunciation errors. Experimental results on an English dataset of Vietnamese learners demonstrate that the proposed approach significantly improves the accuracy of pronunciation assessment systems. The study highlights the potential of this method for integration into intelligent language learning systems.

Keywords: *Automatic Pronunciation Assessment, hierarchical phonetic representation, phoneme-based pronunciation, error analysis.*

** Ngày nộp bài: 06/6/2025; Ngày gửi phản biện: 09/6/2025; Ngày duyệt đăng: 25/9/2025*

** Bài báo là kết quả nghiên cứu từ đề tài NCKH cấp sơ sở (mã số ĐT-2023-19) của Trường Đại học Hồng Đức.*