LingWav2Vec2: Linguistic-augmented wav2vec 2.0 for Vietnamese Mispronunciation Detection

Anonymous submission to Interspeech 2024

Abstract

Pronunciation error detection algorithms rely on both acoustic and linguistic information to identify errors. However, these algorithms face challenges due to limited training data, often just a few hours, insufficient for building robust phoneme recognition models. This has led to the adoption of self-supervised learning models like wav2vec 2.0. We propose an innovative approach that combine canonical text and audio inputs to balance the trade-off between accurate phoneme recognition performance and pronunciation scoring. This is done by feeding audio-encoded and normalised canonical phoneme embedding into a linguistic encoder including multi-head attention (MHA) layer and specifically designed feed forward module (FFN). Our system, with only 4.3 million parameters on top of pretrained wav2vec 2.0, achieved top-1 performance at the VLSP Vietnamese Mispronunciation Detection 2023 challenge with 9.72% relative improvement over the previous state-of-the-art.

Index Terms: Mispronunciation detection and diagnosis, wav2vec 2.0, pronunciation assessment, phoneme recognition

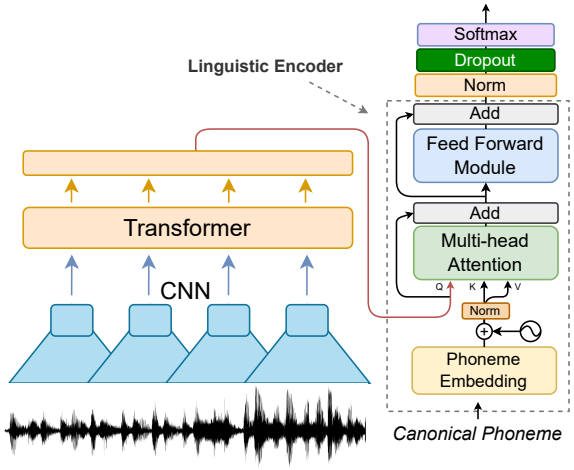

Figure 1: *LingWav2Vec2*

1. Introduction

Mispronunciation can impede effective communication and present obstacles in language learning. To combat this, the development of automatic Mispronunciation Detection and Diagnosis (MD&D) systems has become a key area of focus in linguistic research. These systems are designed to provide learners with feedback that aids in the improvement of their pronunciation abilities.

While traditional methods like Goodness of Pronunciation (GOP) [1, 2] and the Extended Recognition Network (ERN) [3, 4] have been explored for pinpointing and correcting mispronunciations, their effectiveness has been limited. GOP fails to provide the specific feedback needed for training improvement, while ERN struggles with mispronunciation patterns not included in its training data, often misclassifying errors. The introduction of CTC [5] and the rise of end-to-end ASR frameworks paved the way for the success of CNN-RNN-CTC [6] in phoneme recognition task. This step plays a crucial role in the overall performance of MD&D task.

The naturalness of MD&D task is we have the right text by default, which mean we can use this to guiding for the phoneme recognition while the audio means to score the pronunciation. SED-MDD [7] improve of CNN-RNN-CTC by use information of characters of canonical setence provide, the results help reducing the error of recognize phoneme compared to CNN-RNN-CTC and start of a promising framework for MD&D task. The [8] improve performance by using canonical phoneme sequence instead of canonical characters, and final representation learned through attention mechanism between canonical phoneme information and mispronunced audio, showing the canonical phoneme may help align better with the mispronunced phoneme output we need to predict. From that view, APL [9], introduce phonetic encoder, complement with acoustic encoder, which using the phonetics embeddings from pretrained native ASR models, to help the learning of the model on limited amounts of dataset, showing better results compare to [8] without using any augmentation methods. PAPL [10], improve APL by introducing pitch encoder, replace the normal attention mechanism with multi-headed attention mechanism, and use powerful wav2vec 2.0 [11] SSL pretrained model, experiments on Vietnamese Mispronunciation Detection dataset showing it's improvement, set promising results and approach for Vietnamese MD&D task.

Our solution for MD&D task is inspired by both the works presented in [9], [10] and [12] but with the key differences described as follows. First, we keep the single wav2vec 2.0 pretrained encoder for the audio input. Secondly, the canonical phoneme embedding is normalized before processing which helps to stabilize the linguistic encoder learning. Furthermore, we feed both encoded audio and normalised text embedding to a multi-head attention (MHA) layer before passing them to a FFN module. This play a crucially important role in providing a balanced trade-off between guiding a correct phoneme sequence output from canonical phoneme and having right scores for pronunciation scoring which is the main purpose of the system. An additional novelty of our system is the design of the feed forward module (FFN) illustrated in Figure 2, combining RMS norm, linear layers, SwiGLU activation function with dropout.

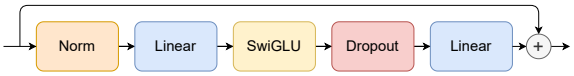

Figure 2: *FFN Module*

The new design of FFN has greatly contributed to the performance of the final system.

Our proposed system has officially achieved the top-1 position in the Vietnamese Mispronunciation Detection (VMD) task at VLSP 2023 challenge [13].

The structure of the paper is as follows: Section 2 details the methodology behind LingWav2Vec2. Section 3 elaborates on our experimental procedures and results. Section 4 concludes with a discussion on the implications of our work and suggestions for future research directions.

2. LingWav2Vec2

The LingWav2Vec2 model utilizes a pre-trained wav2vec 2.0 model [11] for audio processing. The resulting contextual representation of the audio is then integrated with the Linguistic Encoder. Subsequently, the model learns a representation through cross-attention between the audio and canonical phoneme, enabling phoneme recognition via CTC Loss [5]. Our model depicted in Figure 1.

2.1. Linguistic Encoder

Building upon prior research that highlights the benefits of incorporating linguistic information alongside acoustic data for improved model robustness in MD&D tasks (e.g., [9, 10, 7, 12]), we also leverage linguistic information in our approach.

Specifically, the linguistic encoder, depicted on the right-hand side of Figure 1, utilizes phonetic embedding, sinusoidal positional encoding, and RMS normalization [14] for these embeddings. This is followed by a cross-attention block and a feed-forward module, both employing pre-norm residual units for improved training stability (as detailed in [15, 16]).

2.1.1. Feed Forward Module

Inspired by [17], we have also designed a feed-forward module. It consists of an RMS norm layer [14], a linear layer, SwiGLU (which performs better than Swish [18]), a dropout layer for regularization, and a final linear layer. As mentioned, we employ pre-norm residual units for the feed-forward module. Figure 2 illustrates the Feed Forward (FFN) module.

2.2. LingWav2Vec2

As shown in Figure 1, our model leverages wav2vec 2.0 as an acoustic encoder to process raw audio waveforms and a separate linguistic encoder to handle the provided canonical phonemes.

The model receives raw audio input denoted as $\mathcal{X}$. Wav2vec 2.0 extracts encoded acoustic features, represented mathematically as $f : \mathcal{X} \mapsto Q$. Simultaneously, the canonical phonemes $\mathcal{P}$ are embedded using a learned embedding layer, resulting in E. This embedded representation is then fed into a pre-norm cross-attention mechanism (MHA) that combines Q (acoustic features) and E (linguistic information) as key and value inputs, respectively. This is followed by a pre-norm feed-forward module (FFN) for further processing. Finally, an RMS normalization layer is applied, and the output is transformed into a phoneme distribution using a final Softmax layer. The CTC criterion is employed to optimize the model during training.

The overall process for LingWav2Vec2 can be summarized as follows:

$$Q = \text{wav2vec2.0}(\mathcal{X}) \tag{1}$$
$$H = \text{LinguisticEncoder}(\mathcal{P}, Q) \tag{2}$$
$$O = \text{Softmax}(\text{RMSNorm}(H)) \tag{3}$$

Breaking down the Linguistic Encoder:

$$E = \text{PosEnc}(\text{Embedding}(\mathcal{P})) \tag{4}$$
$$K = V = \text{RMSNorm}(E) \tag{5}$$
$$H = Q + \text{MHA}(Q, K, V) \tag{6}$$
$$H = H + \text{FFN}(H) \tag{7}$$

The integration of wav2vec 2.0 with the linguistic encoder maximizes the use of both acoustic and phonemic information. Cross-attention mechanisms guide the model by connecting acoustic and linguistic data, while the feed-forward module further refines this integration, leading to more precise phoneme sequence predictions.

3. Experiments

3.1. VLSP VMD Shared Task and Data

The VLSP 2023 competition included a challenge called Vietnamese Mispronunciation Detection (VMD) [13], inspired by Computer-Assisted Pronunciation Training (CAPT) systems. In this challenge, participants were given two Vietnamese pronunciation datasets containing utterances with deliberately introduced errors. The goal was to develop a model that can identify these mispronunciations at the individual sound unit (phoneme) level within the data. Table 1 summarizes the datasets, which are divided into two main subsets.

The first dataset, provided by MachinaX [19], consists of augmented recordings of adults pronouncing pairs of single-syllable Vietnamese words. These word pairs may not necessarily form a coherent meaning, but they are phonetically pronounceable. Each pair is modified to reflect all six Vietnamese tonal variations (mid-level, high-rising, low-falling, high-rising-glottal, low-falling-rising, and low-falling-glottal) and then recorded by native speakers.

The second dataset, compiled by SoICT-HUST and the Vietnam Psycho-Pedagogical Association [10], features recordings of children aged 5 to 7 speaking or reading Vietnamese sentences. The dataset includes contributions from both kindergarten and primary school children. Kindergarteners read aloud from Vietnamese school textbooks, while primary school children speak extemporaneously. The recordings were then annotated by 20 trained professionals to identify and mark any mispronunciations.

We analyzed the prevalence of mispronounced sounds (phonemes) in spoken sentences across the training, public testing, and private testing datasets. This involved comparing the correct pronunciations (canonical phonemes) to the actual pronunciations (transcribed phonemes) in each utterance. The findings are visualized in Figure 3. As the graph shows, most sentences have no mispronounced sounds. The number of sentences with errors gradually decreases as the number of mispronunciations increases.

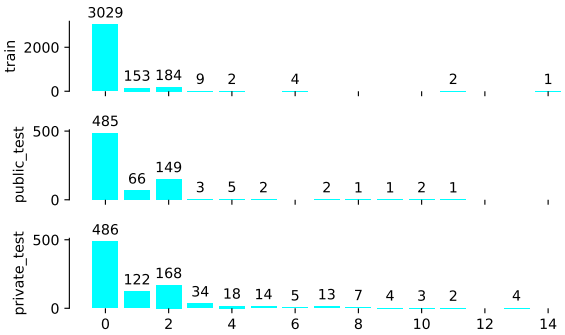

Figure 3: *Number of mispronounced per utterances on train/public test/private test*

3.2. Pretrained wav2vec 2.0

We choose [20] as our wav2vec 2.0 variant due to its strong performance on Vietnamese speech datasets [21, 22, 23] and robust encoded information. The model comprises 7 CNN layers, 12 Transformer layers, and a LayerNorm layer. We retain the CNNs and Transformers (94.3M parameters) and remove the ASR-specific layers. We explore freezing/unfreezing the CNN feature extraction layers (details in Section 3.7).

3.3. Evaluation Metrics

In the context of MD&D tasks, it is essential to evaluate both ASR and MD&D performance metrics. ASR performance pertains to phoneme recognition Accuracy and Correctness. On the other hand, MD&D performance primarily focuses on the F1-Score, which encompasses Precision, Recall, and additional metrics such as FRR, FAR, Detection Rate, and Diagnosis Error Rate. In our Ablation studies (3.7), we will primarily emphasize reporting the F1-Score. When comparing with the other methods, we will provide a comprehensive overview by reporting all these metrics. For detailed instructions on calculating these metrics, one can refer to [9].

3.4. LingWav2Vec2 CTC

We have structured the model as shown in Figure 1, incorporating wav2vec2.0 with hyperparameters detailed in Section 3.2. Additionally, we introduced an embedding layer for a phoneme vocabulary of size 123, outputting to the model dimension of wav2vec 2.0, which is 768. For simplicity, fixed-length sinusoidal positional encoding is utilized. RMS Norm [14] is the primary normalization layer applied in our model, particularly for the Linguistic Encoder, as our tests show it outperforms Layer Norm [24]. The dropout rate in the Feed Forward Module is consistently set at 0.1, but the dropout for the last layer defaults to 0; our experiments (Section 3.7) indicate that an increased dropout rate leads to suboptimal results. Result the model with additional 4.3M parameters. The model employs the phoneme recognition task and is optimized using the CTC criterion [5] with vocabulary of phonemes equal to 123.

For optimization, the AdamW optimizer is used with β_1 and β_2 set to $(0.9, 0.98)$. The learning rate follows a linear warmup for the first 10% of total training steps and then transitions to a cosine decay for the remaining steps, with a peak learning rate of $2e^{-5}$. The training is conducted over 100 epochs, with a batch size of 8, on $4\times$ NVIDIA A40 gpus for all experiments.

Table 1: *Summary of the VLSP VMD 2023 shared task dataset, detailing the minimum and maximum utterance lengths (in seconds), total duration (in hours), and the proportion of the first dataset in the entire dataset (%).*

Name	Min	Max	Total	1st data
Train	0.54	19.81	3.91	7.27
Public Test	1.2	18.46	0.65	34.12
Private Test	1.11	18.45	0.93	25.57

3.5. Result on private test of VLSP VMD task

We have achieved top-rank result on VLSP private test leaderboard VMD shared-task. Table 2 showing the result (final leaderboard) before we further experiments (Section 3.7).

Table 2: *Competition Results on Private Test*

#	Team Name	F1	Precision	Recall
1	**LossKhongGiam (our)**	**57.55**	**55.52**	59.73
2	SpeechHust98	55.19	41.37	82.86
3	DaNangNLP	52.02	38.34	80.89
4	TruongNguyen	49.27	34.51	**86.07**
5	TranTuanBinh	14.90	12.88	17.68

We compared our best model with previous works on the MD&D task. To ensure a fair comparison, we meticulously reproduced the reported results by following the exact methodologies and configurations described in the original studies ([9, 10, 12]), additional with that, we also employ SpecAugment [25] with same configurations as [17, 11]. The training hyperparameters differ for each model: APL-MHA [9] uses a learning rate of $2e^{-5}$, a batch size of 64, and is trained for 200 epochs. PAPL-KALDI-MHA [10] uses a learning rate of $1e^{-5}$, a batch size of 4, and is trained for 101 epochs. Finally, the TextGateContrast [12] model is trained for 200 epochs with a learning rate of $5e^{-5}$ and a batch size of 64.

Table 3 shows that even though APL-MHA and PAPl-KALDI-MHA use handcrafted acoustic mel-spectrograms, automatic phonetic embeddings, and pitch information, their performance falls short of the fully optimized wav2vec 2.0 models (TextGateContrast and LingWav2Vec2). TextGateContrast and LingWav2Vec2 significantly outperform the other models, demonstrating the power of pretrained wav2vec 2.0 models on limited amount of data. Our model even surpasses the more complex and larger TextGateContrast model, achieving a relative improvement of 9.72% in F1-Score. This highlights the effectiveness of carefully designing and arranging model components with a simple CTC loss function (compared to the combination of CTC Loss and Triple Margin Loss used in TextGateContrast).

3.6. Trade-off analysis on using canonical phoneme sequence

To evaluate the models's ability to recognize mispronounced phoneme sequences, we report accuracy results on both the target (mispronounced) and canonical sequences in Table 4. A larger relative difference (Rel. Diff.) between these accuracies indicates a better model. Simpler models like single-layer Bi-LSTMs struggle to leverage the canonical sequence for improved accuracy (Rel. Diff. of 3.9% for APL-MHA, 2.11%

Table 3: *Results of phoneme recognition and MD&D on Private Test*

Methods	#P (M)	Phoneme Recognition		Mispronunciation Detection and Diagnosis						
		Accuracy	Correctness	FRR	FAR	Detection Rate	Diagnosis Error Rate	Recall	Precision	F1
APL-MHA [9]	26	59.68	81.05	17	**11.79**	83.39	38.56	**88.21**	29.70	44.43
PAPL-KALDI-MHA [10]	36	75.63	84.94	12.41	19.64	87.04	36.78	80.36	34.51	48.28
TextGateContrast [12]	124	68.37	90.78	**5.45**	37.77	92.12	48.48	62.23	48.17	54.3
LingWav2Vec2 (our)	98	**79.25**	**91.27**	5.77	27.32	**92.61**	**32.92**	72.68	**50.62**	**59.68**

for PAPL-KALDI-MHA), even have multi-head attention. Conversely, overly complex models like TextGateContrast can overemphasize the canonical sequence, leading to better canonical accuracy but worse mispronounced accuracy (Rel. Diff. of -2.18%). Our model strikes a balance by learning cross-attention between the mispronounced audio and the canonical sequence. This allows the canonical sequence to guide the alignment without harming the target phoneme recognition, resulting in a significant Rel. Diff. of 27.63%.

Table 4: *Phoneme Recognition Accuracy when compare with Mispronunced phoneme sequence and Canonical phoneme sequence (%)*

	Mispronunced Phoneme	Canonical Phoneme	Rel. Diff.
APL-MHA [9]	**59.68**	57.35	3.9
PAPL-KALDI-MHA [10]	**75.63**	74.03	2.11
TextGateContrast [12]	68.37	**69.86**	(2.18)
LingWav2Vec2 (our)	**79.25**	57.35	**27.63**

3.7. Ablation Studies

To further understand the LingWav2Vec2, we conduct experiments on different aspects of model, include: freeze/non-freeze the wav2vec 2.0 (Table 5), SpecAugment policies (Table 6), performance of integrate Linguistic Encoder (Table 7), and why our FFN Module is better (Table 8).

Our experiments on this dataset suggest that fine-tuning wav2vec 2.0 without freezing the CNN layers is optimal for learning task-specific features for phoneme recognition. Additionally, using SpecAugment with specific hyperparameters ($p_T = 0.025$, $l_T = 10$, $p_F = 0.001$, $l_F = 16$) achieved the best F1-score (50.28%) based solely on audio encoder, demonstrating its effectiveness in improving data robustness. Furthermore, incorporating a Linguistic Encoder significantly improved performance on both public and private tests (Table 7), highlighting its potential to leverage linguistic knowledge and enhance phoneme recognition, which aligns with previous research [8, 9, 10, 12]. Notably, our FFN Module outperforms Transformer [26] and Conformer [17] FFNs with fewer parameters. This is achieved by doubling the hidden layer size instead of quadrupling it, and using SwiGLU [18] instead of ReLU or Swish (Table 8).

Table 5: *F1-scores on public and private test of the two way to finetuning wav2vec 2.0: freeze and non-freeze.*

	Public	Private
Freeze	36.83	45.45
Non-Freeze	**38.26**	**49.58**

Table 6: *F1-scores on public and private tests for various SpecAugment policies.*

p_T	l_T	p_F	l_F	Public	Private
0.05	5	0.001	64	38.67	50.21
0.025	10	0.008	64	37.52	49.51
0.05	10	0.001	64	37.42	49.73
0.05	5	0.008	16	38.23	49.74
0.025	**10**	**0.001**	**16**	38.05	**50.28**
0.025	10	0.008	16	38.41	50.06
0.05	5	0.008	64	**38.87**	49.21
0.05	10	0.008	64	37.94	50.13

Table 7: *Result on public/private test for baseline with Linguistic Encoder (LingEnc) with different dropout rate (D).*

Name	Public Test		Private Test	
	PER	F1	PER	F1
Baseline (SpecAugment)	14.07	38.05	21.09	50.28
+ LingEnc (0D)	16.47	**49.51**	20.75	**59.68**
+ LingEnc (0.05D)	9.25	40.25	**14.27**	56.51
+ LingEnc (0.1D)	**9.09**	39.63	14.54	55.88

Table 8: *Performance of our proposed FFN Module compared to Transformer and Conformer FFN*

Methods	#P (M)	Public Test		Private Test	
		PER	F1	PER	F1
Transformer FFN [26]	4.72	**11.95**	41.06	**18.78**	52.49
Conformer FFN [17]	4.72	12.19	41.22	18.93	51.96
Our FFN	1.77	16.47	**49.51**	20.75	**59.68**

4. Conclusion

We introduce LingWav2Vec2, a system that combines a pre-trained speech model (wav2vec 2.0) with a linguistic encoder for automatic mispronunciation detection and diagnosis. This inclusion of the linguistic encoder significantly improves performance. Our model achieved the best results on the VLSP Vietnamese Mispronunciation Detection challenge, achieving a private test F1-Score of 59.68%, a 9.72% improvement over the previous state-of-the-art, and showing effective balance use of canonical linguistic information, with only 4.3 million additional parameters. Apart from things we have done, we're still lack of applying MD&D tailor data augmentation methods, and study how pitch information affect the performance on Vietnamese, we consider that's the future direction based on this approach.

5. References

[1] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Commun.*, vol. 67, pp. 154–166, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:205223190

[2] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," 2020.

[3] A. Harrison, W.-K. Lo, X.-J. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," *SLaTE*, 01 2009.

[4] X. Qian, F. Soong, and H. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (capt)," 09 2010, pp. 757–760.

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[6] W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8132–8136.

[7] Y. Feng, G. Fu, Q. Chen, and K. Chen, "Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3492–3496.

[8] K. Fu, J. Lin, D. Ke, Y. Xie, J. Zhang, and B. Lin, "A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques," 2021.

[9] W. Ye, S. Mao, F. Soong, W. Wu, Y. Xia, J. Tien, and Z. Wu, "An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings," 2022.

[10] T. T. Huu, V. T. Pham, T. T. T. Nguyen, and T. L. Dao, "Mispronunciation detection and diagnosis model for tonal language, applied to Vietnamese," in *Proc. INTERSPEECH 2023*, 2023, pp. 1014–1018.

[11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[12] L. Peng, Y. Gao, R. Bao, Y. Li, and J. Zhang, "End-to-end mispronunciation detection and diagnosis using transfer learning," *Applied Sciences*, vol. 13, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/11/6793

[13] "Vlsp 2023 challenge on vietnamese mispronunciation detection," accessed: November 24, 2023. [Online]. Available: https://vlsp.org.vn/vlsp2023/eval/vmd

[14] B. Zhang and R. Sennrich, "Root mean square layer normalization," 2019.

[15] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, "Learning deep transformer models for machine translation," 2019.

[16] T. Q. Nguyen and J. Salazar, "Transformers without tears: Improving the normalization of self-attention," 2019. [Online]. Available: https://zenodo.org/record/3525484

[17] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[18] N. Shazeer, "Glu variants improve transformer," 2020.

[19] N. Minh and P. Hung, *The System for Detecting Vietnamese Mispronunciation*, 11 2021, pp. 452–459.

[20] T. B. Nguyen, "Vietnamese end-to-end speech recognition using wav2vec 2.0," 09 2021. [Online]. Available: https://github.com/vietai/ASR

[21] "Papers with code - common voice vietnamese benchmark (speech recognition)," accessed: November 24, 2023. [Online]. Available: https://paperswithcode.com/sota/speech-recognition-on-common-voice-vietnamese

[22] H.-T. Luong and H.-Q. Vu, "A non-expert Kaldi recipe for Vietnamese speech recognition system," in *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 51–55. [Online]. Available: https://aclanthology.org/W16-5207

[23] "Papers with code - vivos benchmark (speech recognition)," accessed: November 24, 2023. [Online]. Available: https://paperswithcode.com/sota/speech-recognition-on-vivos

[24] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, ser. interspeech$_2$019. $ISCA, Sep.$2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.