Mispronunciation detection and diagnosis model for tonal language, applied to Vietnamese

Huu Tuong Tu[1], *Pham Viet Thanh*[1], *Dao Thai Lai*[2], *Nguyen Thi Thu Trang**[1]

[1]Hanoi University of Science and Technology, Vietnam
[2]Vietnam Psycho-Pedagogical Association, Vietnam

huutu12312vn@gmail.com, thanh.pv.ds@gmail.com, daothailai2015@gmail.com, trangntt@soict.hust.edu.vn

Abstract

A tonal language is a language in which the meaning of words is not only determined by the sounds of the consonants and vowels, but also by the pitch or tone used to pronounce them. Mispronunciation Detection and Diagnosis (MD&D) of tonal languages is challenging since tone presentation is difficult to be detected correctly. There has been relatively little research conducted on tonal languages, with most focusing on Mandarin. Furthermore, there are no publicly available datasets and source codes for the task. This work constructs and publishes a Vietnamese dataset for experimenting with MD&D, as well as proposes an end-to-end model that utilizes pitch analysis to detect and diagnose mispronunciations for tonal languages, especially focusing on Vietnamese. Experiments show that the proposed model achieved a relative improvement in phone error rate of 7.1% and detection accuracy of 7.4% compared to a state-of-the-art baseline.

Index Terms: Mispronunciation Detection and Diagnosis, Phoneme Recognition, Computer Assisted Pronunciation Training, Vietnamese, tonal language

1. Introduction

Mispronunciation is a widespread issue that can negatively impact communication skills and lead to misunderstandings. The consequences of mispronunciation are not limited to impeding language learning progress, but also hindering effective communication. As a result, researchers have become increasingly interested in developing automatic Mispronunciation Detection and Diagnosis (MD&D) systems to provide feedback to learners and help them improve their pronunciation skills.

Over the years, there have been various approaches proposed for addressing the problem of detecting and diagnosing mispronunciations. One such approach is Goodness of Pronunciation (GOP) [1, 2, 3], which employs acoustic models to calculate scores and phone-dependent criteria to identify mispronunciations. GOP can identify pronunciation errors, but it does not offer enough details to allow for correction. To overcome this limitation, the Extended Recognition Network (ERN) [4, 5, 6, 7] was created to get around this restriction and uses pre-established phonological rules to gather more diagnostic data. However, ERN has limitations in addressing mispronunciation patterns that are not present in the training data. Moreover, it is difficult to construct ERNs that combine multiple mispronunciation paths, which raises the false accept rate and causes the model to fail to detect mispronunciation.

Recently, the ASR end-to-end structure has shown good promise for the MD&D task. A CNN-RNN-CTC [8] was proposed, has shown an outperformed result compared with previous approaches [9]. However, the aforementioned research does not make use of the pre-existing prior text information in the case of reading text that is already known. If models can include linguistic data from canonical text, it can enhance MD&D performance. As a result, SED-MDD [10] was proposed to incorporate acoustic features and canonical sequences. To enhance the power of linguistic features, K. Fu et al. [11] proposed a model that combines acoustic features with phoneme encoder features, demonstrating the impact of sentence-to-phoneme features.

Despite the additional information linguistic embedding provides, the input of the acoustic encoder solely consists of low-level features. This presents certain challenges during model training, as low-level features are sensitive to noise and variations and may not adequately capture specific characteristics or unique attributes required for this task. APL approach [12] has shown that phonetic features extracted from a well-trained Automatic Speech Recognition (ASR) model may represent phonetic information in a noise-robust and speaker-independent manner. The added phonetic embedding to the model has significantly improved the performance of the mispronunciation detection and diagnosis task and achieved state-of-the-art results in this task.

Tone refers to the utilization of pitch within a language in order to differentiate between lexical or grammatical significance, thereby distinguishing and modifying words. Languages possessing this characteristic are known as tonal languages. In such languages, alterations in tone can lead to variations in the meanings of words that would otherwise sound identical, emphasizing the crucial role of tonal pronunciation. There are some MD&D techniques have been proposed for tonal languages, particularly in relation to Mandarin. These approaches typically revolve around utilizing ASR models to recognize phonemes [13]. However, accurately detecting phonemes in tonal languages can be challenging, primarily because of the difficulty in correctly identifying tone. Consequently, researchers have been exploring various methods to address this issue, such as developing context-aware models [14] that take into account the linguistic context when identifying tones and phonemes. While progress has been made in improving MD&D for tonal languages, further research is needed to address these challenges fully.

In 2014, Ghahremani et al. [15] proposed an algorithm to improve pitch extraction for ASR systems. Their approach involves tuning the pitch extraction algorithm to better suit the requirements of ASR. The resulting algorithm, dubbed Kaldi pitch, has demonstrated remarkable performance improvements for tonal languages in ASR systems.

The application of pitch information can go beyond speech

*Corresponding author

10.21437/Interspeech.2023-364

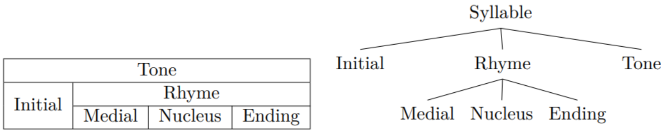

Figure 1: *The hierarchical structure of Vietnamese syllables [16]*

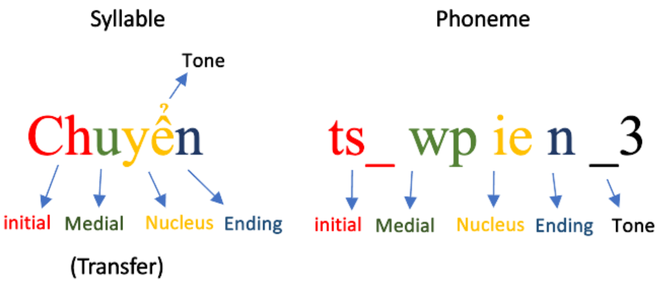

Figure 2: *Example of a Vietnamese phoneme presentation.*

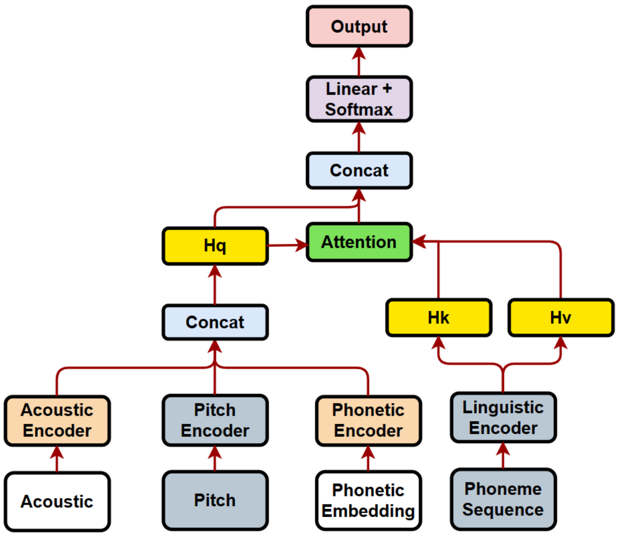

Figure 3: *Proposed MDD model architecture for tonal languages.*

recognition tasks. It has also been shown to enhance the performance of MD&D systems. By leveraging pitch extraction algorithms, these systems can detect and analyze subtle differences in pronunciation and provide more accurate feedback to language learners.

This work presents an end-to-end model for tonal language which combines pitch analysis with a state-of-the-art MD&D model. Additionally, we also introduce the first-ever Vietnamese dataset for MD&D.

The remainder of this document is structured as follows: Section 2 outlines our methods, while Section 3 provides an overview of the dataset. Our proposed approach's experimental setup and evaluation results are presented in Section 4. Finally, we conclude the paper and suggest potential areas for future research in Section 5.

2. Proposed Method

2.1. Overview of Vietnamese phonemes

An overview of Vietnamese syllables is illustrated in Figure 1, where the hierarchical structure is proposed [16]. 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 (.). We also follow this study to map each syllable of Vietnamese to phoneme, which covers 52 phonemes where six tones of syllables are split into eight tones phoneme presentation: Level Tone (A1), Slightly Falling Tone (A2), Falling Tone (C1), Falling-Rising Tone (C2), Rising Tone (B1), Sharply Rising Tone (D1), Dropping Tone (B2) and Sharply Dropping Tone (D2). The example of syllable conversion to phoneme is presented in Figure 2. Because the phoneme can keep the original hierarchical structure of syllables, this mapping can allow us to detect all types of mispronunciation.

2.2. Baseline model

We choose a novel method for detecting and diagnosing mispronunciations - APL [12] - as our baseline model. The method combines three types of features: acoustic, phonetic, and linguistic into APL embeddings, which are then used to train an end-to-end model with CTC loss [17]. The results of the study show that the APL approach outperforms existing methods and achieves state-of-the-art results in MD&D for English.

2.3. Proposed model

Because pitch information is very important in tonal languages, we propose PAPL, a model in which we leverage the APL model [12] by incorporating a pitch encoder to improve the performance in tone detection. A detailed illustration of our model architecture is shown in Figure 3. The sections below will discuss all the components appearing in our proposed model.

2.3.1. Acoustic Encoder

For acoustic encoder, the input is an 81-dimensional acoustic feature, which comprises 80-dim fbanks and 1-dim energy. Once the audio features are extracted, they undergo processing in the CNN-RNN module, which consists of 2 CNN stacks and 2 RNN stacks. The CNN and RNN architectures are similar to APL [12]. The output of the acoustic encoder is denoted as $\mathbf{H}^a$, where $\mathbf{H}^a = [\mathbf{h}^a_1, ..., \mathbf{h}^a_{t'}, ..., \mathbf{h}^a_{T'}]$.

2.3.2. Pitch Encoder

Pitch extraction provides valuable information about the underlying acoustic properties of speech. It is especially important for tonal languages, such as Vietnamese.

Ghahremani et al. [15] presented a method to enhance pitch extraction in ASR systems. Their strategy involves adjusting the pitch extraction algorithm to better align with the needs of ASR. The method has demonstrated significant improvements in ASR performance for tonal languages. So, we opt to utilize this algorithm for pitch extraction (Kaldi pitch). However, in order to provide a comprehensive comparison of the results, we also extract pitch using another approach known as the normalized cross-correlation function and median smoothing (NCCF pitch).

The pitch feature after extraction is passed to the CNN-RNN module, which has the same architecture and number of RNN and CNN stacks as the acoustic encoder. The input of the pitch encoder is a vector: $\mathbf{Pi} = [\mathbf{pi}_1, ..., \mathbf{pi}_t, ..., \mathbf{pi}_T]$, which has an identical number of frames as the acoustic features. The output of the pitch encoder is referred to as $\mathbf{H}^{pi}$, to derive its representations: $\mathbf{H}^{pi} = [\mathbf{h}^{pi}_1, ..., \mathbf{h}^{pi}_{t'}, ..., \mathbf{h}^{pi}_{T'}]$.

2.3.3. Phonetic Encoder

The input of the phonetic encoder is phonetic embeddings denoted as $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_t, ..., \mathbf{p}_T]$, which are extracted by a pre-trained ASR model. Like acoustic encoder and pitch encoders, phonetic encoder is also constructed using CNN and RNN stacks. Since these embeddings represent higher-level information compared to the raw acoustic features, the encoder includes one RNN stack and one CNN stack. The resulting output of this encoder is denoted as $\mathbf{H}^{p} = [\mathbf{h}^{p}_1, ..., \mathbf{h}^{p}_{t'}, ..., \mathbf{h}^{p}_{T'}]$.

2.3.4. Linguistic Encoder

In the MD&D task, prior texts are available, so the phoneme sequences of the prior text can be fed into the model. Phoneme sequences are embedded into a vector and passed through a Bi-LSTM. The output of the Bi-LSTM is then passed into two branches, each with one linear layer. Two outputs of the two branches are referred to as $\mathbf{H}_K = [\mathbf{h}^K_1, ..., \mathbf{h}^K_n, ..., \mathbf{h}^K_N]$ and $\mathbf{H}_V = [\mathbf{h}^V_1, ..., \mathbf{h}^V_n, ..., \mathbf{h}^V_N]$, where N is the number of phonemes of the sentence.

2.3.5. Decoder

For decoder, the attention mechanism[18] is used. The cross attention mechanism allows the model to identify relevant parts of the linguistic features that correspond to specific segments of the audio features. The output of the acoustic encoder, pitch encoder, and phonetic encoder are concatenated together to obtain the query, denoted as $\mathbf{H}_Q$. For a given frame t', we have:

$$\mathbf{h}^Q_{t'} = (\mathbf{h}^a_{t'}; \mathbf{h}^p_{t'}; \mathbf{h}^{pi}_{t'}) \tag{1}$$

To compute the normalized attention weight, the following formula can be used:

$$\alpha_{t',n} = \frac{exp(\mathbf{h}^Q_{t'} \mathbf{h}^{K^T}_n)}{\sum_{n=1}^{N} exp(\mathbf{h}^Q_{t'} \mathbf{h}^{K^T}_n)} \tag{2}$$

Finally, the context vector is computed as:

$$\mathbf{c}_{t'} = \sum_{n}^{N} \alpha_{t',n} \mathbf{h}^V_n \tag{3}$$

After applying attention and obtaining the context vector, it is possible that the information captured may not be sufficient to represent mispronounced phonemes that are not present in the canonical phoneme sequence. This limitation arises because attention mechanisms typically rely on the alignment between input elements, such as phonemes, and the corresponding context vector. If a mispronounced phoneme is not present in the canonical sequence, it may not receive significant attention, and its representation in the context vector might be insufficient. Hence, both $\mathbf{c}_{t'}$ and $\mathbf{h}^Q_{t'}$ are used to calculate framewise probability $\mathbf{y}_{t'}$, the formula of $\mathbf{y}_{t'}$:

$$\mathbf{y}_{t'} = softmax(W[\mathbf{c}_{t'}; \mathbf{h}^Q_{t'}] + \mathbf{b}) \tag{4}$$

Beam search is then applied to generate phoneme sequences with probability output.

3. Datasets

We present the first Vietnamese dataset for the task of MD&D. A total of 84 native children, 53 from kindergarten, and 31 from primary schools participated in creating the dataset. Table 1 shows the statistics of each subset. All primary school children were recorded speaking spontaneously, while all kindergarten children were recorded reading sample sentences that we collected from various Vietnamese schoolbooks.

Table 1: *Details of the collected subsets*

Properties	Primary school	Kindergarten
# of Speakers	31	53
# of Utterances	3,818	448
# of Hours	4.58	0.31

To identify instances of mispronunciation, 20 annotators were trained to evaluate the entire dataset and mark any instances of incorrect or defective pronunciation. All texts were annotated at the phoneme level using the mapping provided by [16], which covers 52 phonemes. 25 children, including 22 from kindergarten, whose pronunciation varied from very poor, were reserved as the testing set. The sets of speakers in the training, testing, and development sets are mutually exclusive, and each set covers all the phonemes. The dataset has been released and published for the community [1]. The data split is described in Table 2.

Table 2: *Details of Vietnamese dataset used in the experiments*

Properties	Train	Test	Dev
# of Speakers	50	25	9
# of Utterances	3,181	612	473

4. Experiments

4.1. Pretrained-ASR

Since the quality and quantity of data used in automatic speech recognition (ASR), training can affect phonetic embeddings, selecting an appropriate acoustic model is crucial. One of the most advanced ASR models is wav2vec2.0 [19], which is capable of handling a range of data variations, including noisy and diverse voice signals. As a result, we will utilize wav2vec2.0 to extract phonetic embeddings. We will use a variant of wav2vec2.0 that has been fine-tuned on the VLSP2020 ASR dataset[20].

4.2. Experimental Setups

All audio files were at a 16000 sampling rate, and we computed filter banks, pitch, and phonetic embeddings with a 20ms

[1] https://github.com/VietMDDDataset/VietMDD

Table 3: *Results of phoneme recognition and MD&D in Vietnamese. Note that MHA refers to Multi-Head Attention, while PAPL is a term we have coined as our proposed approach*

MDD model	**Phoneme Recognition (%)**		**Mispronunciation detection and Diagnosis (%)**						
	Accuracy	Correctness	FRR	FAR	Detection Rate	Diagnosis Error Rate	Recall	Precision	F1
APL (Baseline)	76.52	78.23	20.74	27.85	79.01	39.87	72.15	11.44	19.74
APL-MHA (Baseline)	76.57	77.87	21.00	25.57	78.84	41.31	74.43	11.62	20.11
PAPL-NCCF	82.81	84.85	14.28	**20.85**	85.48	38.85	**79.15**	**17.05**	**28.06**
PAPL-KALDI	82.53	83.62	15.21	22.22	84.54	**36.01**	77.78	15.95	26.47
PAPL-NCCF-MHA	81.78	83.11	15.60	24.96	84.06	41.78	75.04	15.15	25.20
PAPL-KALDI-MHA	**83.67**	**85.52**	**13.22**	28.61	**86.23**	43.07	71.39	16.69	27.06

Table 4: *Correctness of tone recognition*

MDD model	**Tone correctness (%)**								**Phoneme (%)**
	A1	A2	C1	C2	B1	D1	B2	D2	
APL-MHA (Baseline)	71.08	68.32	51.85	33.36	61.46	54.92	46.99	55.02	83.73
PAPL-KALDI-MHA (Ours)	**92.85**	**84.90**	**79.01**	**70.55**	**81.76**	**90.16**	**82.24**	**84.34**	**85.53**

shift. To train our models, we utilized the AdamW optimizer, with a learning rate of 1e-5 and a maximum of 101 epochs. We employed PyTorch to implement the model.

4.3. Phoneme recognition

To align the ASR predictions with human annotations, we utilized the Needleman-Wunsch algorithm [21]. We computed our evaluation metrics using formula (5), where ”I” represents insertions, ”D” represents deletions, ”S” represents substitutions, and ”N” represents the total number of phonetic units.

$$Correctness = \frac{N-S-D}{N}, Accuracy = \frac{N-S-D-I}{N} \tag{5}$$

Table 3 presents the results of our phone recognition experiments. For free phone recognition in Vietnamese, which is a tonal language, the baseline accuracy was 76.57%. While adding NCCF pitch resulted in an accuracy of 82.81%, using Kaldi pitch extraction led to an accuracy of 83.67%. Table 4 shows more details of tone recognition. The correctness of tone increases in the range from 16.58% to 37.19%, while phoneme increases by 1.8%. These outcomes demonstrate that our pitch approach outperforms the baseline without pitch features, especially in tone recognition, confirming our hypothesis that pitch features are essential for tonal language phoneme recognition.

4.4. MD&D Result

Following previous works [8, 9, 10, 11, 12], the hierarchical evaluation structure is used to measure the MD&D system performance. The correctness of prediction is denoted by true accept (TA) and true rejection (TR), while false accept (FA) and false rejection (FR) indicate incorrect prediction. TR is further classified into correct diagnosis (CD) and diagnosis error (DE). The metrics of MD&D are calculated following the formula in [12].

As presented in Table 3, the PAPL-KALDI-MHA system appears to have the highest Detection Rate (+7.39%) and lowest FRR (-7.78%), and also performs well on F-measure (+6.95%). However, this system also has the highest diagnosis error rate and FAR. The PAPL-NCCF system has the lowest FAR (-4.72%), and also performs best on recall (+4.72%), precision (+5.43%), and F-measure (+7.95%) compared with the baseline. Some speccially that PAPL-KALDI-HMA performs the best in ASR, when PAPL-NCCF seems the best in MD&D. It because as the number of correct pronunciations constitutes the majority of our dataset, the FRR weight has a significant impact on ASR performance, leading to the superior performance of PAPL-KALDI-MHA in ASR. Recall and precision are based on both TR and FAR and since PAPL-NCCF has a much higher TR value than PAPL-KALDI-MHA (8%), PAPL-NCCF has higher recall and precision than KALDI-MHA.

5. Conclusions

In this paper, we published a Vietnamese dataset for the MD&D task and proposed an MD&D system for Vietnamese learners that incorporates pitch information. Our system uses a combination of acoustic, phonetic, linguistic, and pitch features. The testing results show that the proposed approach is effective in improving the FRR and F-measure over the baseline system by 7.78% and 6.95% absolutely. Moreover, our proposed approach outperformed the baseline system in tone detection, with an increasing range of 24.27% (A2) to 111.48% (C2) relative to the baseline. Furthermore, our approach increased the correctness of tone recognition, which demonstrates the effectiveness of incorporating pitch information in MD&D for tonal languages such as Vietnamese. In future works, we hope to come up with a more efficient architecture, as well as extend our Vietnamese dataset by collecting utterances from non-native speakers.

6. Acknowledgements

This work was supported by the NAVER Corporation within the framework of collaboration with the International Research Center for Artificial Intelligence (BKAI), School of Information and Communication Technology, Hanoi University of Science and Technology under Project NAVER.2022.DA08.

7. References

[1] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, pp. 95–108, 02 2000.

[2] W. Hu, Y. Qian, F. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, 01 2015.

[3] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," 10 2020.

[4] 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.

[5] L. Wai Kit, S. Zhang, and H. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," pp. 765–768, 09 2010.

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

[7] H. Meng, Y. Lo, L. Wang, and W. Lau, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," *2007 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2007, Proceedings*, pp. 437 – 442, 01 2008.

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

[9] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.

[10] Y. Feng, G. Fu, Q. Chen, and K. Chen, "Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," pp. 3492–3496, 2020.

[11] 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," 04 2021.

[12] 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," 10 2021.

[13] Y. Xie, X. Feng, B. Li, J. Zhang, and J. Yujia, "A mandarin l2 learning app with mispronunciation detection and feedback," 05 2021.

[14] R. Tong, N. Chen, B. Ma, and H. Li, "Context aware mispronunciation detection for mandarin pronunciation training," 09 2016, pp. 3112–3116.

[15] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," pp. 2494–2498, 2014.

[16] T. Nguyen, "Hmm-based vietnamese text-to-speech: Prosodic phrasing modeling, system design, corpus design and evaluation," PhD thesis, Universite Paris Sud 11 - Hanoi University of Science and Technology, 15 Rue Georges Clemenceau, 91400 Orsay, France, September 2015. [Online]. Available: /2015/Ngu15f

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," p. 369–376, 2006. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 06 2020.

[20] "Vlsp 2020 asr dataset," https://vlsp.org.vn/vlsp2020/eval/asr, 2020.

[21] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.