Toward a Phonetic Approach for Multi-Dialect Speech Recognition in Vietnamese

Anonymous ACL submission

Abstract

Vietnamese automatic speech recognition (ASR) remains challenging due to systematic dialectal variation across Northern, Central, and Southern regions, where identical lexical items often exhibit substantially different pronunciations. Most existing approaches address this variability primarily at the word level, relying on vocabularies that implicitly assume dialect-invariant mappings between orthography and pronunciation, which is linguistically inappropriate for Vietnamese. In this work, we propose a dialect-aware phonetic framework that explicitly models Vietnamese phonological structure and dialectal variation at both the vocabulary and decoding levels. We introduce a phonetic vocabulary that decomposes each syllable into structured phonetic components and maps them to dialect-specific IPA representations. Building on this representation, we design a phonetic-structure decoder that jointly predicts these components, enabling consistent and interpretable modeling. Experiments on the ViMD dataset demonstrate that the proposed approach consistently outperforms or matches strong pretrained baselines across dialects, achieving a WER of 13.35%, a PER of 8.45%, and dialect identification accuracy exceeding 95%, while using fewer parameters and no external pretraining. We will release code and phonetic resources for experimental reproducibility upon the acceptance of this paper.

1 Introduction

Automatic Speech Recognition (ASR) for Vietnamese multi-dialect speech remains challenging due to the language's extensive dialectal variation. Speakers across the Northern, Central, and Southern regions exhibit systematic differences in pronunciation for the same lexical items, resulting in substantial acoustic-lexical mismatches (Phạm and McLeod, 2016; Nga et al., 2021; Dinh et al., 2024) (Figure 1). These mismatches often lead to degraded performance, as current ASR systems assume stable mappings between acoustic features and vocabulary units.

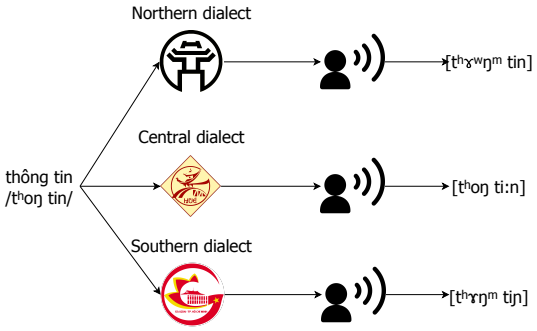

Figure 1: Examples of Vietnamese multi-dialect speech variations.

Prior work has primarily addressed dialectal variation from the acoustic perspective, through multi-task training (Elfeky et al., 2016; Dan et al., 2022), dialect-specific transfer learning (Ta et al., 2024; Suwanbandit et al., 2023), or multi-stage training combined with augmentation techniques that simulate pronunciation diversity (Park et al., 2019; Wang et al., 2023). While effective to some degree, these approaches implicitly treat dialect differences as variability to be normalized. Crucially, these methods do not address a core limitation at the vocabulary level: word-based and subword-based tokenizers presuppose a dialect-invariant correspondence between orthography and pronunciation. This assumption is particularly problematic for Vietnamese, where dialect-specific phenomena such as vowel shifts, coda neutralization, and tone contour variation are systematic and phonetically rule-based.

Vietnamese further motivates rethinking vocabulary design. As a monosyllabic, tonal language, each word corresponds to a structured combination of an initial consonant, a rhyme, and a tone. These components vary predictably across dialect families

and encode distinctions that are essential for both human perception and machine recognition (Phạm and McLeod, 2016). Conventional Vietnamese ASR systems that operate on orthographic words or subwords overlook this structure, limiting their ability to generalize across dialects and exacerbating data sparsity and alignment inconsistencies, particularly for low-resource dialects.

In this work, we propose a dialect-aware phonetic vocabulary representation for Vietnamese ASR, grounded in linguistic structure rather than surface orthography, explicitly encoding Vietnamese phonology and dialectal variation. Our method decomposes each syllable into linguistically meaningful components, then maps them to dialect-specific International Phonetic Alphabet (IPA) patterns, yielding a compact and interpretable token set that represents systematic cross-dialectal differences. Beyond vocabulary design, we introduce a multi-token decoder that jointly predicts the initial, rhyme, and tone of each syllable. Unlike conventional single-token auto-regressive decoders, our architecture leverages the internal structure of Vietnamese phonology to model cross-component dependencies, thereby improving phonetic consistency, reducing ambiguity, and stabilizing alignment during training. This paper makes three main contributions:

1. We develop the first dialect-aware phonetic vocabulary for Vietnamese ASR that integrates linguistic structure with dialect-specific pronunciation patterns.

2. We propose a structured multi-token decoding mechanism that describes syllable components jointly to enhance phonetic discrimination and decoding stability.

3. We demonstrate through extensive experiments that integrating dialectal phonetics at both the vocabulary and decoding levels substantially improves word recognition accuracy, phonetic consistency, and vocabulary coverage compared with standard methods and strong pretrained models.

2 Related Works

Research on dialectal ASR has primarily focused on mitigating performance degradation caused by pronunciation variations. A prominent direction centers on multi-task and multi-model architectures that explicitly incorporate dialect labels during training. Multi-task learning frameworks jointly optimize dialect classification and ASR to better align dialect-sensitive acoustic cues with recognition objectives (Dan et al., 2022). Complementarily, systems equipped with dialect identification modules route utterances to dialect-specific ASR models, reducing cross-dialect interference through specialized decoding paths (Elfeky et al., 2016). While these methods effectively utilize dialect labels, they typically require large-scale annotated multi-dialect corpora and remain limited to modeling differences primarily at the acoustic level.

Another line of research addresses data scarcity and adaptation challenges using transfer learning or multi-stage training strategies. Models pretrained on high-resource standard-accent corpora are fine-tuned on small dialectal datasets to improve robustness (Ta et al., 2024; Suwanbandit et al., 2023). More recent architectures, such as Aformer, adopt progressive, multi-stage optimization schedules to capture diverse acoustic patterns (Wang et al., 2023). Mixture-of-Experts (MoE) frameworks further enhance cross-dialect generalization by combining dialect-sensitive experts with general experts under dynamic routing strategies (Zhou et al., 2024). Despite their effectiveness, these approaches largely conceptualize dialectal variation as acoustic variability to be normalized or compensated, thereby underexploiting the deeper lexical and phonological structure.

For Vietnamese, research has increasingly focused on multi-dialect corpora (Tran et al., 2024; Dinh et al., 2024), allowing systematic analysis of pronunciation variation and syllable structure. Resources such as ViMD (Dinh et al., 2024) provide valuable data for examining these linguistic patterns and evaluating dialect-aware ASR models at the provincial level. Nevertheless, most Vietnamese ASR systems continue to operate on orthographic word or subword units (Nga et al., 2021). Such units assume a dialect-invariant mapping between text and pronunciation, a problematic assumption for Vietnamese, where systematic cross-dialect phonological shifts are pervasive. As a result, the lexical and phonological levels remain comparatively underexplored in Vietnamese ASR.

Overall, existing work predominantly addresses dialect variation through acoustic modeling and adaptation techniques. In contrast, vocabulary-level modeling that explicitly captures dialect-dependent phonological structure remains limited. This gap directly motivates our approach: we introduce a

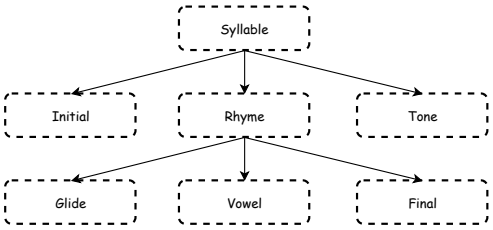

Figure 2: Vietnamese syllable structure.

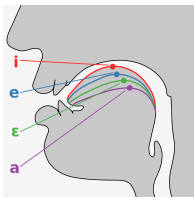

Figure 3: Highest tongue positions of front vowels.

(a) (b)

Figure 4: Mouth configurations during production of the voiced velar nasal [ŋ] (a) and the voiceless velar plosive [k] (b)

phonologically grounded, dialect-aware vocabulary representation and a structured decoding mechanism that jointly model syllable components, enabling Vietnamese ASR systems to more faithfully represent, generalize across, and ultimately benefit from systematic dialectal variation.

3 Preliminaries

3.1 Vietnamese Syllable Structure

Vietnamese syllables exhibit a highly regular structure consisting of three core components: initial, rhyme, and tone (Đoàn Thiện Thuật, 2016; Hạo, 1998). Phonetic studies further decompose the rhyme into glide, vowel, and final, where the vowel is mandatory, and the remaining components are optional (Đoàn Thiện Thuật, 2016; Hạo, 1998). Consequently, the rhyme serves as the central carrier of phonological information in Vietnamese syllables. Linguistic (Giáp, 2008, 2011; Hạo, 1998) and natural language processing studies (Hieu Nguyen et al., 2025) consistently characterize Vietnamese as a monosyllabic language, in which each word contains at most one syllable. Coupled with its phonemic orthography and strong grapheme–phoneme correspondence, this property enables a straightforward alignment between orthographic units and phonetic components. These characteristics form the basis of our proposed dialect-aware phonetic vocabulary, described in Section 4.1. In this work, phonemes are denoted using slashes, while dialect-specific phones are denoted using square brackets. Figure 1 illustrates a valid example of dialectal variation.

3.2 Phonetic Variation in Multi-Dialect Vietnamese at the Phone Level

Multi-dialect variation in Vietnamese primarily manifests at the phonetic level rather than the word level. For example, a salient characteristic of the Northern dialect is the interaction between front vowels /i, e, ɛ/ and velar consonants /k, ŋ/. In rhymes such as /iŋ, eŋ, ɛŋ, ik, ek, ɛk/, front vowels require a high front tongue position (Figure 3), while velar consonants require contact with the velum (Figure 4), resulting in articulatory incompatibility. This incompatibility leads to vowel centralization (/i, e, ɛ/ → [ɨ, ɘ, ɜ]) and consonant palatalization (/k, ŋ/ → [c, ɲ]), often accompanied by a transitional glide [j]. These patterns reflect systematic, phone-level variation rather than changes in lexical identity, providing strong motivation for phone-based modeling in Vietnamese multi-dialect ASR. A detailed dialect-specific analysis is in Appendix C.

4 Methodology

Our goal is to develop a vocabulary representation and decoding framework that explicitly encodes Vietnamese phonological structure and dialectal regularities. This section presents: (i) the construction of a dialect-aware phonetic vocabulary, (ii) a structured multi-dialect transcript representation, and (iii) a phonetic-structure decoder that jointly predicts internal syllable components.

4.1 Dialect-aware Phonetic Vocabulary

Vietnamese syllables follow a regular phonological template consisting of an *initial consonant*, *glide*, *vowel*, *coda*, and *tone*. Because dialectal variation systematically affects each of these components, our vocabulary construction begins with a

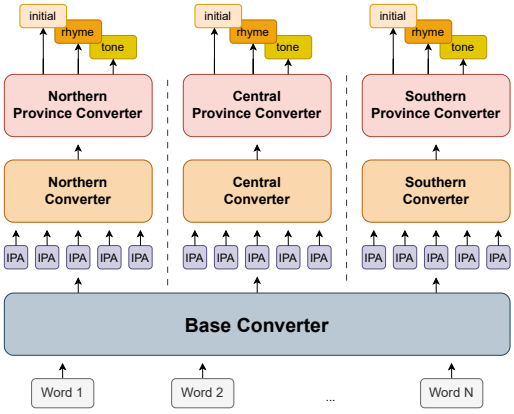

Figure 5: Pipeline for constructing a dialect-aware phonetic vocabulary for Vietnamese.

deterministic syllable parser that decomposes every orthographic word into these five constituent units. The parser enforces phonotactic constraints, including initial–vowel compatibility, coda admissibility, and some orthographic exceptions, ensuring that only phonotactically valid Vietnamese syllables enter the vocabulary. Each component is then mapped to its canonical IPA representation via a *Base Converter* encoding the standard Vietnamese phonology. These IPA arrays serve as a dialect-invariant reference layer that abstracts away from orthographic inconsistencies (Figure 5).

Dialect-specific regularities are flexibly integrated through a modular extension of the converter. A *Dialectal Converter* applies rule sets capturing systematic Northern, Central, and Southern shifts, such as initial mergers (e.g., /v/ → [j]), vowel raising/lowering, diphthong simplification, and coda neutralization. A further *Province Converter* refines these transformations to model a few province-level sociophonetic differences. This hierarchical design allows fine-grained dialect modeling without entanglement with surface orthography. The final representation for each word yields a compact, interpretable, and dialect-aware phonetic vocabulary suitable for Vietnamese ASR.

4.2 Multi-dialect Transcript Representation

Rhymes, comprising the medial, nucleus, and coda, constitute the core and consistently stable obligatory phonological unit of Vietnamese and encode the majority of dialect-induced variation in vowel quality and coda realization (Phạm and McLeod, 2016). To balance linguistic granularity with decoding efficiency, we adopt a structured three-part

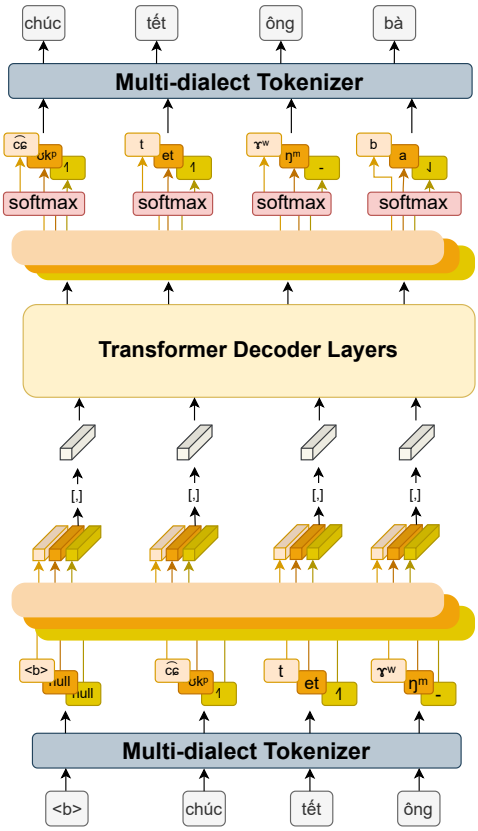

Figure 6: Vietnamese phonetic-structure decoder. The English text shown corresponds to Tet (Vietnamese New Year) wishes addressed to grandparents.

representation (initial, rhyme, tone) rather than using all five phonological components. The resulting vocabulary consists of three disjoint subsets: V_{initial} (27 categories), V_{rhyme} (∼200 categories), and V_{tone} (6 categories). During training, output is a sequence of structured triplets. At inference time, predicted phone sequences are deterministically mapped back to orthographic forms using a reverse lexicon, which relies on the high consistency between phonetics and orthography in Vietnamese.

4.3 Vietnamese Phonetic-structure Decoder

To leverage the internal structure of Vietnamese syllables, we introduce a *multi-token autoregressive decoder* that predicts the initial, rhyme, and tone in parallel at each timestep, while maintaining autoregression across syllables (Figure 6). **Component-Specific Embeddings**: At timestep t, the decoder receives three syllabic components $(\text{initial}_{t-1}, \text{rhyme}_{t-1}, \text{tone}_{t-1})$. Each component is embedded in its own learned subspace, preserving

the distinct phonological roles of each one. **Unification Layer**: The component embeddings are fused using a linear learned unification function, producing a unified representation that captures inter-component phonotactic dependencies.

$$h_t = f_{\text{unify}}(e_t^{\text{initial}}, e_t^{\text{rhyme}}, e_t^{\text{tone}}).$$

Attention over Acoustic Context: This step is similar to the standard decoding mechanism (Vaswani et al., 2017), where the unified representation attends to encoder outputs via multi-head attention, thereby integrating acoustic context with phonetic structure and enabling the decoder to capture dialect-sensitive cues and long-range temporal dependencies. **Parallel Multi-Token Prediction**: A shared projection layer produces a single probability vector that is partitioned into three distributions.

$$p(\text{initial}_t), \quad p(\text{rhyme}_t), \quad p(\text{tone}_t).$$

Joint prediction enforces internal consistency, avoids invalid syllables, and aligns naturally with the phonological structure of Vietnamese.

5 Experiments

5.1 Baselines, Datasets, and Metrics

Baselines. We compare the proposed dialect-aware phonetic vocabulary and structured decoding mechanism against a diverse set of strong baselines that reflect common design choices in Vietnamese ASR. *Zero-shot* baselines include publicly available large-scale Vietnamese ASR models, namely PhoWhisper (Le et al., 2024) and wav2vec 2.0 variants (Baevski et al., 2020) (wav2vec2-base-vietnamese, wav2vec2-base-vietnamese-160h, and wav2vec2-base-vietnamese-250h), which serve as off-the-shelf references. For fully supervised evaluations, we implement the Speech Transformer and Conformer (Dong et al., 2018; Gulati et al., 2020) architectures trained with standard *word-based* tokenization, where the output units correspond to orthographic Vietnamese words. To isolate the effect of vocabulary design, we construct *phone-based* counterparts that replace word tokens with phonetic units (initial, rhyme, tone) derived from the proposed dialect-aware vocabulary, while keeping the encoder–decoder architecture, parameter budget, and training configuration fixed. We additionally include a multi-task (MT) baseline that jointly predicts dialect labels and transcriptions, representing prior approaches that inject dialect information at the acoustic modeling level (Dan et al., 2022). Finally, our full phonetic method, *Vietnamese phonetic-structure decoder* models, combine the phonetic vocabulary with the proposed Vietnamese phonetic-structure multi-token decoder, allowing us to disentangle the contributions of phonetic representation and decoding structure.

Datasets. All experiments are conducted on the publicly available Multi-Dialect Vietnamese (ViMD) corpus (Dinh et al., 2024), which comprises speech from all Vietnamese provinces. We adopt the official train–validation–test splits, enforcing strict speaker disjointness while preserving the original dialectal distribution. Only minimal preprocessing is applied: utterances containing non-Vietnamese tokens are filtered out, and audio is kept in its original mono 16kHz waveform to retain natural pronunciation and recording conditions. As a result, ViMD constitutes a challenging and realistic benchmark for evaluating the robustness of ASR systems across Vietnamese dialects.

Metrics. We evaluate models from complementary lexical and phonetic perspectives. At the orthographic level, Word Error Rate (WER) is reported as the primary metric, together with dialect-wise results. Northern provinces use the Northern dialect, provinces from Nghe An–Ha Tinh to Thua Thien Hue use the Central dialect, and the remaining provinces use the Southern dialect. For phone-based systems, we assess phonetic accuracy using the Phone Error Rate (PER) over the predicted phonetic sequences. We also report component-wise error rates for syllable initials, rhymes, and tones to analyze the phonological modeling behavior. Together, these metrics capture both transcription accuracy, linguistic fidelity, and dialectal robustness of the proposed framework.

5.2 Configuration

For all supervised experiments, models were trained from scratch on the ViMD corpus without using any external language models on a single NVIDIA H100 GPU. All experiments were implemented using a unified pipeline based on the SpeechBrain end-to-end speech processing toolkit (Ravanelli et al., 2021). Speech signals were represented by 80-dimensional log-Mel filterbank (Fbank) features, with a window size of 25ms and a step size of 10ms. To improve robustness, SpecAugment (Park et al., 2019) was applied

Model	Setup	Params	WER↓	PER↓	WER_N	WER_C	WER_S	Time(s)↓
Published Zero-shot Baselines†								
PhoWhisper-base	–	74M	21.78	–	18.17	27.18	23.80	0.73
PhoWhisper-small	–	244M	17.11	–	15.25	18.50	18.38	1.21
wav2vec2-base-vi	–	95M	22.48	–	19.87	27.15	23.82	0.03
wav2vec2-base-vi-160h	–	95M	31.12	–	27.09	40.49	32.83	0.03
wav2vec2-base-vi-250h	–	95M	16.83	–	14.51	19.73	18.23	0.07
Published Fine-tuned Baselines†								
PhoWhisper-base	FT	74M	16.30	–	–	–	–	–
wav2vec2-base-vi	FT	95M	15.80	–	–	–	–	–
wav2vec2-base-vi-160h	FT	95M	17.49	–	–	–	–	–
wav2vec2-base-vi-250h	FT	95M	13.56	–	–	–	–	–
Our Implementations								
Transformer	–	29M	16.66	–	14.68	18.64	17.95	0.37
Conformer	–	31M	17.56	–	16.05	19.92	18.39	0.29
Transformer	P	26M	15.13	12.52	13.55	**15.99**	16.27	1.25
Conformer	P	28M	13.96	11.29	12.01	17.20	14.99	0.93
MT-Transformer	P+M	26M	15.93	12.99	14.22	17.49	17.05	1.58
MT-Conformer	P+M	28M	16.75	13.58	14.71	19.56	17.93	1.21
Our-Transformer	**P+V**	26M	**13.35**	**8.45**	**11.37**	16.06	**14.51**	0.33
Our-Conformer	P+V	28M	13.80	8.91	11.78	17.07	14.89	0.28

Table 1: Performance comparison on Vietnamese multi-dialect ASR. Metrics include word error rate (WER, %) and phone error rate (PER, %). WER_N, WER_C, and WER_S denote Northern, Central, and Southern dialects. Setup: phonetic supervision (P), multi-task learning (M), Vietnamese phonetic-structure decoder (V). † Results reported from (Dinh et al., 2024).

during training, including both time masking and frequency masking with fixed mask widths.

For encoder–decoder architectures, Transformer employed 12 encoder and 6 decoder layers. Conformer consisted of 8 encoders and 4 decoders, each with an attention dimension of 256, 4 self-attention heads, and a feed-forward dimension of 2048. All models were optimized using the Adam optimizer (Kingma and Ba, 2015) with the Noam learning-rate schedule. Warm-up steps (Gotmare et al., 2019) were set to 40k for the Transformer and 20k for the Conformer, with initial learning rates of 10^{-3} and 4×10^{-4}, respectively. Label smoothing (Szegedy et al., 2016) and dropout were both set to 0.1 for regularization. Training used a joint CTC–Attention objective (Kim et al., 2017), with the CTC loss weight set to 0.30 (Transformer) and 0.15 (Conformer). This joint objective promotes monotonic alignment through CTC while preserving the modeling flexibility of attention-based decoding, resulting in more stable training.

5.3 Main Results

Table 1 presents the main ASR results on the ViMD benchmark, comparing our proposed dialect-aware phonetic framework against strong zero-shot, fine-tuned, and supervised baselines. Overall, the results demonstrate that explicitly modeling Vietnamese phonological structure and dialectal variation at both the vocabulary and decoding levels yields consistent improvements across model architectures and dialect regions.

Large pretrained models, such as PhoWhisper and wav2vec 2.0, achieve competitive performance in both zero-shot and fine-tuned settings by leveraging large-scale pretraining. However, even after fine-tuning, these models remain constrained by word- or subword-based vocabularies that implicitly assume dialect-invariant pronunciations. In contrast, our models, trained from scratch on ViMD, achieve comparable or lower WERs with substantially fewer parameters, highlighting that linguistic inductive bias, specifically phonetic and dialect-aware representations, can partially compensate for the absence of massive pretraining when dialectal variation is systematic and structured.

Effect of the phonetic vocabulary design. Replacing word-level outputs with the proposed phonetic units yields clear improvements over standard Transformer and Conformer baselines. For example, Transformer-P reduces the WER from 16.66% to 15.13%, while Conformer-P achieves a larger reduction, from 17.56% to 13.96%. These results

Design	Transformer				Conformer			
	PER↓	Initial	Rhyme	Tone	**PER↓**	Initial	Rhyme	Tone
Phonetic baseline	12.52	12.29	14.53	10.52	11.29	11.07	13.33	9.26
+ Multitask	12.99	12.66	15.15	10.90	13.58	13.01	15.94	11.47
+ Phonetic-structure decoder (our)	**8.45**	**8.36**	**10.07**	**6.88**	8.91	8.82	10.63	7.24

Table 2: Phonetic error analysis of phone-based ASR models. Metrics include overall phone error rate (PER, %) and component-wise PERs for initial, rhyme, and tone units.

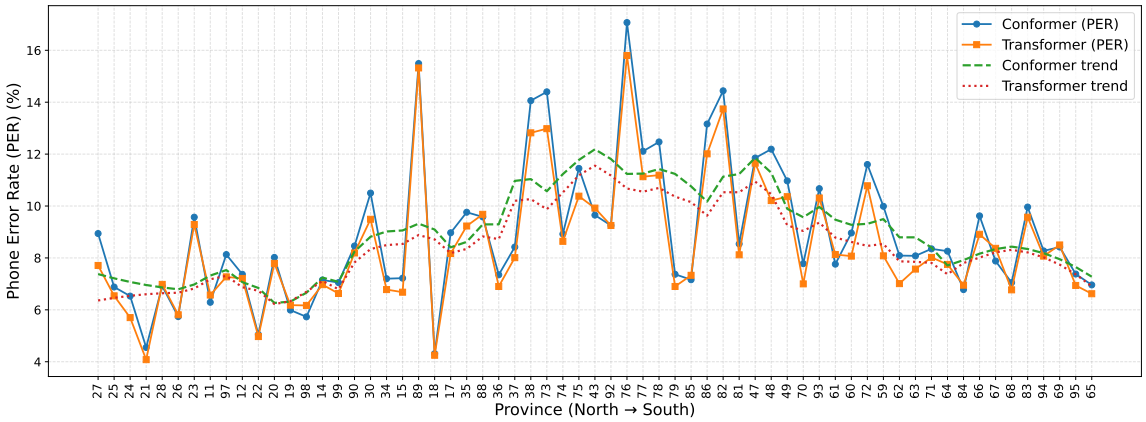

Figure 7: Phone error rate (PER, %) by province for Transformer and Conformer models using the proposed Vietnamese phonetics method.

suggest that explicit phone-level modeling reduces acoustic-lexical mismatch and improves generalization across dialects, with more pronounced gains for Conformer architectures.

Contribution of the Vietnamese phonetic-structure decoder. The largest gains are obtained with the full phonetic method, which further reduces WER to 13.35% (Transformer) and 13.80% (Conformer) without increasing model capacity or introducing additional supervision. By jointly predicting initial, rhyme, and tone, the decoder enforces internal phonological consistency and reduces invalid syllable hypotheses, leading to a more compact token representation and more stable decoding.

Dialect-specific analysis shows consistent improvements across Northern, Central, and Southern regions. Central dialect provinces, which exhibit greater internal phonetic variability, also benefit from the proposed approach. Figure 7 further corroborates this trend at the provincial level, where phone error rates are generally lower and more stable across regions.

5.4 Results Analysis

Phonetic Structure and Dialect Modeling Table 2 provides deeper insight into the main results. While phonetic baselines already improve over word-based systems, they still exhibit relatively high initial, rhyme, and tone error rates, reflecting ambiguity when these components are predicted independently without structural constraints. The phonetic-structure decoder substantially reduces overall PER, from 12.52% to 8.45% for Transformer and from 11.29% to 8.91% for Conformer, with consistent improvements across all three phonetic components.

By comparison, multi-task learning with dialect classification improves dialect identification accuracy (Table 4), but this does not consistently translate into better ASR performance. In some cases, PER can even degrade relative to phonetic-only baselines, suggesting that injecting dialect information solely at the acoustic level is insufficient and may introduce competing optimization objectives, particularly when training on a modest-sized dataset such as ViMD. In contrast, our method integrates dialectal knowledge directly into the vocabulary and decoding process, resulting in more consistent and stable performance gains.

Word Level vs. Phonetic Level Modeling Table 3 compares lexical diversity and frequency bias between word-level and phonetic-level ASR models. Phonetic-level models correctly recognize a

Modeling Unit	Architecture	Unique Correct Words↑	Pearson r	Spearman ρ
Word-level	Transformer	1,696	0.79	0.76
	Conformer	1,718	0.77	0.74
Phonetic-level	Transformer	**1,950**	0.57	0.45
	Conformer	**1,943**	0.58	0.47

Table 3: Lexical diversity and frequency bias analysis for word-level and phonetic-level ASR models. Metrics include the number of unique word types correctly recognized at least once in the test set, as well as Pearson/Spearman correlations between word frequency in the training data and per-word recall.

Model	Accuracy	Macro-F1
Published baselines[†]		
PhoWhisper-base	-	86.97
Whisper-base	-	85.59
wav2vec2-base-vi	-	91.02
wav2vec2-large-vi	-	91.47
wav2vec2-xlsr-300m	-	89.01
wav2vec2-large-xlsr-35	-	87.36
Our implementations		
Our-Transformer	**95.75**	**96.11**
Our-Consformer	95.47	95.80

Table 4: Classification performance (%) of multi-task phonetic models. Metrics include macro F1 and accuracy. [†]Results reported from (Dinh et al., 2024).

substantially larger number of unique word types in the test set (from 1696 to 1950 for Transformer and from 1718 to 1943 for Conformer), indicating improved generalization to infrequent lexical items. Moreover, both Pearson and Spearman correlations between training-set word frequency and test-time recall are markedly lower for phonetic-level models, suggesting reduced reliance on frequency-driven memorization. These results suggest that phonetic-level modeling promotes structural generalization by capturing systematic phonological patterns across dialects, rather than exploiting surface-level lexical statistics.

Dialect Classification Performance Table 4 reports dialect classification results for the proposed multi-task phonetic models. Both Transformer and Conformer architectures achieve high overall accuracy (95.8% and 95.5%) and macro-F1 scores (96.1% and 95.8%), consistently matching fine-tuned PhoWhisper-base and wav2vec2-base-vi baselines reported in (Dinh et al., 2024). These results suggest that the learned phonetic representations encode dialect-discriminative information, despite dialect identification being treated as an auxiliary task. This finding provides empirical evidence that explicitly modeling Vietnamese phonological structure facilitates robust dialect characterization within a unified ASR and dialect classification framework through multi-task learning.

6 Conclusion

In this work, we address a key limitation in Vietnamese multi-dialect ASR: the assumption of a dialect-invariant correspondence between orthography and pronunciation. Motivated by linguistic evidence that Vietnamese dialectal variation is systematic and primarily realized at the phone level, we propose a dialect-aware phonetic framework that integrates phonological structure into both vocabulary design and decoding. We introduce a phonetic vocabulary grounded in Vietnamese syllable structure, representing each syllable as an initial, rhyme, and tone mapped to canonical and dialect-specific IPA forms, together with a phonetic-structure decoder that jointly predicts these components while explicitly modeling their internal dependencies. Experimental results demonstrate that incorporating phonological structure and dialectal regularities provides an effective inductive bias for multi-dialect ASR and complements large-scale pretraining for languages with structured and systematic phonology.

Limitations

Despite its effectiveness, the proposed approach has several limitations. The structured decoder jointly predicts initials, rhymes, and tones at the syllable level; it does not explicitly model cross-syllabic or cross-word phonological and prosodic phenomena, which remain open research directions. Moreover, while the framework is linguistically grounded, extending it to other languages would require language-specific phonological analysis and rule design, and the current implementation is limited to Vietnamese. Future work includes combining rule-based and data-driven phonetic modeling, integrating large pretrained acoustic models, and extending the framework to other low-resource, multi-dialect languages.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hoàng Thị Châu. 2002. *Phương ngữ học tiếng Việt*. Nhà xuất bản Đại học Quốc gia Hà Nội.

Zhengjia Dan, Yue Zhao, Xiaojun Bi, Licheng Wu, and Qiang Ji. 2022. Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition. *Entropy*, 24(10):1429.

Alexandre de Rhodes. 1651. *Dictionarium Annamiticum Lusitanum et Latinum*.

Nguyen Dinh, Thanh Dang, Luan Thanh Nguyen, and Kiet Van Nguyen. 2024. Multi-dialect vietnamese: Task, dataset, baseline models and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7476–7498. Association for Computational Linguistics.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 5884–5888. IEEE.

Mohamed Elfeky, Meysam Bastani, Xavier Velez, Pedro J. Moreno, and Austin Waters. 2016. Towards acoustic model unification across dialects. In *2016 IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA, December 13-16, 2016*, pages 624–628. IEEE.

Nguyễn Thiện Giáp. 2008. *Từ vựng học tiếng Việt*. Nhà xuất bản Giáo dục Việt Nam.

Nguyễn Thiện Giáp. 2011. *Vấn đề "từ" trong tiếng Việt*. Nhà xuất bản Giáo dục Việt Nam.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 5036–5040. ISCA.

Nghia Hieu Nguyen, Dat Tien Nguyen, and Ngan Luu-Thuy Nguyen. 2025. Vietnamese words are not constructed from syllables: Rethinking the role of word segmentation in natural language processing for vietnamese texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):24069–24077.

Cao Xuân Hạo. 1998. *Tiếng Việt mấy vấn đề ngữ âm - ngữ pháp - ngữ nghĩa*. Nhà xuất bản Giáo dục Việt Nam.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 4835–4839. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. 2024. Phowhisper: Automatic speech recognition for vietnamese. In *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, May 11, 2024*. OpenReview.net.

Cao Hong Nga, Chung-Ting Li, Yung-Hui Li, and Jia-Ching Wang. 2021. A survey of vietnamese automatic speech recognition. In *2021 9th International Conference on Orange Technology (ICOT)*, pages 1–4.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 2613–2617. ISCA.

Ben Phạm and Sharynne McLeod. 2016. Consonants, vowels and tones across vietnamese dialects. *International Journal of Speech-Language Pathology*, 18(2):122–134.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. Speechbrain: A general-purpose speech toolkit. abs/2106.04624.

Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, and Ekapol Chuangsuwanich. 2023. Thai dialect corpus and transfer-based curriculum learning investigation for dialect automatic speech recognition. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 4069–4073. ISCA.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Bao Thang Ta, Nhat Minh Le, and Van Hai Do. 2024. Transfer learning methods for low-resource speech accent recognition: A case study on vietnamese language. *Eng. Appl. Artif. Intell.*, 132:107895.

Linh Thi Thuc Tran, Han-Gyu Kim, Hoang Minh La, and Su Van Pham. 2024. Automatic speech recognition of vietnamese for a new large-scale corpus. *Electronics*, 13(5).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xuefei Wang, Yanhua Long, Yijie Li, and Haoran Wei. 2023. Multi-pass training and cross-information fusion for low-resource end-to-end accented speech recognition. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 2923–2927. ISCA.

Jie Zhou, Shengxiang Gao, Zhengtao Yu, Ling Dong, and Wenjun Wang. 2024. Dialectmoe: An end-to-end multi-dialect speech recognition model with mixture-of-experts. In *Chinese Computational Linguistics - 23rd China National Conference, CCL 2024, Taiyuan, China, July 25-28, 2024, Proceedings*, volume 14761 of *Lecture Notes in Computer Science*, pages 243–258. Springer.

Đoàn Thiện Thuật. 2016. *Ngữ âm tiếng Việt*. Nhà xuất bản Đại học Quốc gia Hà Nội.

A Appendix: Experimental Details

A.1 Dataset Details

Split	Duration (h)	#Utterances	#Unique Words
Train	69.58	13,061	3,859
Dev	8.65	1,631	2,381
Test	9.45	1,787	2,441

Table 5: Statistics of the ViMD dataset across training, development, and test splits.

Table 5 presents the detailed statistics of the ViMD dataset used in our experiments, including the duration, number of utterances, and number of unique words for each split (training, development, and test).

A.2 Metrics

We evaluate model performance using complementary metrics that capture orthographic transcription accuracy, phonetic fidelity, dialectal robustness, and lexical generalization behavior.

Word Error Rate (WER). End-to-end transcription quality is primarily measured using *Word Error Rate* (WER), defined as

$$\mathrm{WER} = \frac{S + D + I}{N}, \tag{1}$$

where S, D, and I denote the numbers of word-level substitutions, deletions, and insertions obtained via minimum-edit-distance alignment, and N is the number of words in the reference transcription. In addition to overall WER, we report dialect-wise WER for Northern, Central, and Southern Vietnamese, following the regional grouping defined in the ViMD benchmark.

Phone Error Rate (PER). For phonetic-level systems, we report *Phone Error Rate* (PER), computed analogously at the phone level:

$$\mathrm{PER} = \frac{S_p + D_p + I_p}{N_p}, \tag{2}$$

where S_p, D_p, and I_p are phone-level substitutions, deletions, and insertions, and N_p is the number of reference phones. PER provides a linguistically grounded assessment of pronunciation modeling, which is particularly relevant for Vietnamese due to systematic dialectal variation at the phone level.

Component-wise Phonetic Error Rates. To analyze phonological modeling behavior with respect to Vietnamese syllable structure, we further compute component-wise error rates for syllable *initials*, *rhymes*, and *tones*. For a component $c \in \{\text{initial}, \text{rhyme}, \text{tone}\}$, the error rate is defined as

$$\mathrm{ER}_c = \frac{S_c + D_c + I_c}{N_c}, \tag{3}$$

where S_c, D_c, I_c, and N_c are computed over the aligned component sequence.

Dialect Classification Metrics. For multi-task models that jointly perform ASR and dialect identification, we report *classification accuracy* and *macro-averaged F1-score*. Given a set of dialect classes C, macro-F1 is computed as

$$\text{Macro-F1} = \frac{1}{|C|} \sum_{c \in C} \mathrm{F1}_c, \tag{4}$$

ensuring equal weighting across dialects regardless of class imbalance.

Lexical Diversity and Frequency Bias. To assess lexical generalization beyond aggregate error rates, we report the number of *unique word types* correctly recognized at least once in the test set. We further analyze frequency bias by measuring the relationship between training-set word frequency and per-word recall.

For a word type w, its training-set frequency is defined as

$$f_{\text{train}}(w) = \sum_{u \in \mathcal{D}_{\text{train}}} \sum_{i=1}^{|u|} \mathbf{1}\left(u_i = w\right), \qquad (5)$$

where $\mathcal{D}_{\text{train}}$ denotes the training corpus and $\mathbf{1}(\cdot)$ is the indicator function. To mitigate the heavy-tailed frequency distribution, we apply a logarithmic transformation:

$$\tilde{f}_{\text{train}}(w) = \log\left(1 + f_{\text{train}}(w)\right). \qquad (6)$$

Per-word recall is defined as

$$\text{recall}(w) = \frac{\sum_{(x,y)\in\mathcal{D}_{\text{test}}} \sum_{i=1}^{|\hat{y}|} \mathbf{1}\left(\hat{y}_i = w\right)}{\sum_{(x,y)\in\mathcal{D}_{\text{test}}} \sum_{i=1}^{|y|} \mathbf{1}\left(y_i = w\right)}, \qquad (7)$$

where (x, y) is a test utterance–reference pair and $\hat{y}$ is the predicted word sequence after optimal alignment.

The relationship between training frequency and recognition performance is quantified using Pearson and Spearman correlation coefficients:

$$r = \text{Pearson}\left(\tilde{f}_{\text{train}}(w), \text{recall}(w)\right), \qquad (8)$$

$$\rho = \text{Spearman}\left(\tilde{f}_{\text{train}}(w), \text{recall}(w)\right), \qquad (9)$$

computed over all word types appearing at least once in the test set. Lower correlation values indicate reduced dependence on frequency-driven memorization and stronger structural generalization.

A.3 Additional Dialect Classification Results

Table 6 shows that the Central dialect achieves the highest classification accuracy among the three dialects for both Transformer and Conformer models, as reflected by the strong diagonal entries and minimal cross-dialect confusions. In contrast, most classification errors arise from confusion between the Northern and Southern dialects, which are mutually misclassified more frequently than either is with the Central dialect. This pattern is consistent across architectures and suggests that Central Vietnamese exhibits more distinctive acoustic-phonetic characteristics, while Northern and Southern dialects share partially overlapping phonological properties that make them harder to separate. Overall, the confusion matrices provide fine-grained evidence that complements the aggregate performance metrics in Table 4.

B Appendix: Phonetics and Orthography in Vietnamese

Traditionally, the Vietnamese had the Nom alphabet as the main orthography system. However, the Nom alphabet was developed based on the ancient Han alphabet. This means we have to be fluent in the ancient Han to have the ability to use the Nom alphabet for reading and writing. Later on, (de Rhodes, 1651) used the Latin alphabet to describe the speech sound of this language. This orthography system is simple and effective for describing almost all phonetic phenomena in Vietnamese. With its advantages that counterbalance its disadvantages, this Latin alphabet is gradually improved over time and finally becomes the national alphabet system of modern Vietnamese.

Although having the Nom alphabet or the Latin alphabet, these orthography systems all reflect two consistent characteristics of Vietnamese:

1. Vietnamese is a monosyllabic language.
2. The correspondence between graphemes and phonemes in Vietnamese is consistent.

That is, in this language, we do not have the linking pronunciation as in English, and every phoneme has persistent writing forms. In Vietnamese, each syllable has three components: initials, rhymes, and tones. Rhyme has smaller components, which are glide, vowel, and final. We provide the list of all phonemes according to the syllabic structure of Vietnamese:

- 22 phonemes as the initials:
  - Plosive consonants: /b, t, tʰ, k/.
  - Fricative consonants: /f, d, ɣ, z, j, s, ʂ, c͡ɕ, t͡ʂ, ŋ, x, v/.
  - Nasal consonants: /n, m, ŋ, ɲ/.
  - Vibrant consonants: /r, l/.
- 01 phonemes as the glide: /u̯/.

True	Predicted Dialect			True	Predicted Dialect		
	Northern	Central	Southern		Northern	Central	Southern
	Conformer				**Transformer**		
Northern	711	0	13	Northern	714	0	10
Central	2	148	5	Central	0	152	3
Southern	58	3	847	Southern	57	6	845

Table 6: Dialect confusion matrices for multi-task Transformer and Conformer models on the ViMD test set. Rows denote ground-truth dialects and columns denote predicted dialects.

- 15 phonemes as the vowels:
  - Diphthongs: /ie, uo, ɯə/.
  - Monophthongs: /a, ă, ɜ̆, i, ɛ, e, u, ɯ, o, ɔ, ɔː, ə/.
- 10 phonemes as the finals:
  - Nasal consonants: /n, t/.
  - Labial consonants: /m, p/.
  - Velar consonants: /ŋ, k/.
  - Palatal consonants: /ɲ, c/.
  - Semivowels: /u̯, i̯/.
- 6 phonemes denote tones:
  - Flat tone is denoted by nothing.
  - Low falling tone: /˨˩/.
  - Mid raising tone: /˧˥/.
  - Mid falling tone: /˧˩/.
  - Mid glottalized-falling tone: /˧ʔ˩/.
  - Mid glottalized-raising tone: /˧ʔ˥/.

The writing form of these phonemes is consistent regardless of grammar. In particular:

- 6 tones are denoted by a mark above or below the graphemes of vowels:
  - Flat tone is denoted by nothing (a).
  - Low falling tone /˨˩/ is denoted by a grave accent (à).
  - Mid raising tone /˧˥/ is denoted by an acute accent (á).
  - Mid falling tone /˧˩/ is denoted by a hook above (ả).
  - Mid glottalized-falling tone /˧ʔ˥/ is denoted by a tilde above (ã).
  - Mid glottalized-raising tone /˧ʔ˩/ is denoted by a dot below (ạ).

  In the following texts, we provide writing forms of vowels regarding the mentioned phonemes. These examples might include the tone mark above or below the graphemes. Readers can discard these marks to see the true writing form of the vowels in Vietnamese. For instance, the dot below **iê** /iə/ in word **kiệm** /kiem˧ʔ˥/ denotes the mid glottalized-raising tone /˧ʔ˥/.
- 22 initials have 26 writing forms:
  - b /b/. Eg: **b**a mẹ, **b**ánh kẹo, **b**uôn **b**án.
  - t /t/. Eg: **t**âm **t**ư, **t**ịnh **t**iến, **t**ính cách.
  - th /t^{h}/. Eg: **th**ách **th**ức, **th**ành **th**ạo.
  - k, c, or q /k/. Eg: **c**ách mạng, **q**uan hệ, hiện **k**im.
  - ph /f/. Eg: **ph**ụ huynh, **ph**ong cách, **ph**ân định.
  - đ /d/. Eg: **đ**ưa **đ**ón, **đ**ậm **đ**à.
  - gh or g /ɣ/. Eg: **g**a tàu, **g**ánh hát, **g**anh **gh**ét.
  - gi /z/. Eg: **gi**ếng nước, **gi**ống loài.
  - đ /d/. Eg: **đ**ứng **đ**ắn, **đ**êm tối, **đ**èn **đ**uốc.
  - d /j/. Eg: **d**a **d**ẻ, tiêu **d**ùng, **d**ụng cụ.
  - x /s/. Eg: sản **x**uất, **x**uất **x**ứ, **x**e cộ.
  - s /ʂ/. Eg: xác **s**uất, **s**o **s**ánh, **s**ao chép.
  - ch /c͡ɕ/. Eg: **ch**ứa **ch**an, **ch**e **ch**ở/.
  - tr /t͡ʂ/. Eg: **tr**anh chấp, tiệt **tr**ùng.
  - ng or ngh /ŋ/. Eg: mong **ng**óng, tình **ngh**ĩa.
  - nh /ɲ/. Eg: **nh**à cửa, nỗi **nh**ớ, **nh**ung lụa.
  - l /l/. Eg: **l**ấm **l**em, **l**ung **l**inh, **l**ối về.
  - r /r/. Eg: **r**ậm **r**ạp, **r**ón **r**én, **r**ực **r**ỡ.
  - kh /x/. Eg: **kh**ó **kh**ăn, **kh**ởi sắc, **kh**ấm **kh**á.
  - v /v/. Eg: **v**ui **v**ẻ, **v**ương **v**ấn, **v**ẫy **v**ùng.
  - m /m/. Eg: **m**ong **m**ỏi, **m**ay **m**ắn, **m**ênh **m**ông.
  - n /n/. Eg: đất **n**ước, **n**úi **n**on, **n**ông cạn.
- The glide /u̯/ has two writing forms: u or o. Eg: q**u**ê nhà, h**o**a cỏ, kh**u**yến khích.

- 03 diphthongs have 8 writing forms:
  - iê, yê, ia, or ya /iɛ/. Eg: kh**iế**m thị, **yê**n ắng, ch**ia** sẻ, khu**ya** khoắt.
  - uô or ua /uo/. Eg: kh**uô**n khổ, m**ua** bán.
  - ươ or ưa /ɯə/. Eg: kh**ươ**u giác, dây d**ưa**.
- 12 monophthongs have 13 writing forms:
  - a /a/. Eg: b**a** mẹ, tr**a**nh vẽ, ng**ã** b**a**, m**ả**i miết, làng ch**à**i.
  - ă or a /ă/. Eg: ánh **ắ**ng, n**ắ**m tay, n**ă**m tháng, **áy** n**áy**, chạy nhảy.
  - â /ə̆/. Eg: n**â**ng niu, **ấ**n định, ng**ân** vang.
  - i or y /i/. Eg: th**i** cử, tr**ĩ**u nặng, b**ĩ**u môi.
  - ê /e/. Eg: k**ế**t quả, thể hi**ện**, mân m**ê**.
  - e /ɛ/. Eg: mùa h**è**, x**e** cộ, t**é** ngã.
  - u /u/. Eg: th**u** mua, m**ủ**m mỉm, l**u**ng lay, tr**u**ng thành.
  - ư /ɯ/. Eg: tr**ư**ng cầu, xây dự**ng**, **ư**ng ý.
  - o /ɔ/. Eg: chăm s**ó**c, m**o**ng ng**ó**ng, tr**o**ng veo.
  - oo /ɔː/. Eg: x**oo**ng chảo.
  - ô /o/. Eg: tr**ố**ng đồng, **ố**ng hút, c**ố** nhân.
  - ơ /ə/. Eg: m**ơ** mộng, c**ơ** nhỡ, chơi bời.
- 10 final consonants have 12 writing forms:
  - i or y /i̯/. Eg: làng chà**i**, mỏ**i** mệt, chạ**y** đua, bay nhả**y**.
  - m /m/. Eg: ê**m** ấ**m**, nhiệ**m** màu, mâ**m** cỗ.
  - n /n/. Eg: na**n** giải, no**n** nớt, tả**n** mạ**n**.
  - ng /ŋ/. Eg: sa**ng** trọ**ng**, trố**ng** trải, su**ng** túc.
  - nh /ɲ/. Eg: nha**nh** nhẹn, bê**nh** vực, bi**nh** quyền.
  - p /p/. Eg: phậ**p** phồng, thấ**p** thỏm, thá**p** tùng, gượng é**p**, ức hiế**p**, tẩm ướ**p**.
  - t /t/. Eg: lấn á**t**, bá**t** đĩa, kế**t** quả, hớ**t** hả.
  - c /k/. Eg: cú**c** áo, chợ**c** chờ, bố**c** vá**c**.
  - ch /c/. Eg: cá**ch** thức, chí**ch** ngừa.
  - u or o /u̯/. Eg: tra**u** chuốt, chị**u** đựng, xoong chả**o**, xiêu vẹ**o**.

It is important to note that the writing form of phonemes in Vietnamese is consistent in all usage cases, regardless of grammar (tense, aspect, mood). This feature defines Vietnamese as an isolating language whose morphology is an aspect of phonetics rather than grammar, as in inflectional languages. To this end, given a Vietnamese word,

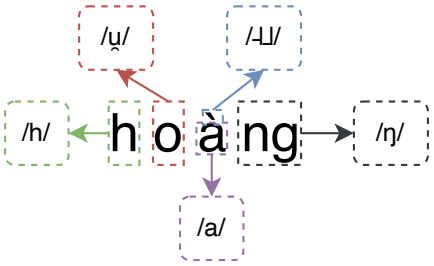

Figure 8: Example for the consistency between graphemes and phonemes in Vietnamese.

native speakers can find no difficulty in pronouncing it. On the other hand, Vietnamese natives can write down any Vietnamese word by listening to its speech sound without knowing how to spell it. For instance, given the word **hoàng**, its phonetic representation is /huaŋ˨˩/. We can determine the grapheme of the initial /h/ is **h**, of the glide /u̯/ is **o**, of the vowel /a/ is **a**, of the /ŋ/ is **ng**, and tone is denoted by the grave accent above **à** (Figure 8).

However, although the conversion from grapheme to phonemes is straightforward (that is, many graphemes correspond to unique phonemes), the inversion is not all ways many-to-one mapping. There are some phonemes having one-to-many mapping with graphemes, such as the initial /ɤ/ can be written as **g** or **gh**, or the diphthong /ie/ can be written as **iê, yê, ia** or **ya**. Actually, a particular writing form of such phonemes is determined consistently via the neighbor phonemes. We detail here rules for determining writing forms of phonemes having multiple respective graphemes:

- The initial /k/ is written as:
  - **k** if followed by **i** /i/, **ê** /e/, **e** /ɛ/, **iê** /ie/. Eg: **k**ịp thời, **ki**ểu cách, **k**èm cặp.
  - **q** if followed by **u** /u̯/ as the glide. Eg: **qu**ê hương, **qu**à cáp.
  - **c** otherwise. Eg: **c**ứng **c**ỏi, **c**ởi mở, hạt **c**át, rau **c**ủ.
- The initial /ɤ/ is written as:
  - **gh** if followed by **i** /i/, **ê** /e/, **e** /ɛ/, **iê** /ie/. Eg: bàn **gh**ế, **gh**i chép, **gh**e tàu.
  - **g** otherwise. Eg: thanh gươm, gồng gánh, gọi điện.
- The initial /ŋ/ is written as:

  - **ngh** if followed by textbfi /i/, **ê** /e/, **e** /ɛ/, **iê** /ie/. Eg: **nghe** ngóng, **nghiê**m nghị, **nghệ** sĩ.
  - **ng** otherwise. Eg: **ngà**nh nghề, **ngỗ** nghịch, **ngọt ngà**o.

- The diphthong /ie/ is written as:
  - **iê** if the rhyme has a final consonant and no glide. Eg: k**iến** thức, tiết k**iệm**.
  - **yê** if the rhyme has a final consonant and the glide written as **u**. Eg: kh**uyên** bảo, **uyển** ch**uyển**, câu ch**uyện**.
  - **ya** if the rhyme has no final consonant and the glide written as **u**. Eg: đêm kh**uya**.
  - **ia** if the rhyme has no final consonant and no glide. Eg: b**ìa** sách, ch**ia** sẻ.

- The diphthong /uo/ is written as:
  - **uô** if the rhyme has a final consonant. Eg: nỗi b**uồn**, m**uối** biển, m**uộn** màn, ch**uồn** ch**uồn**.
  - **ua** if the rhyme has no final consonant. Eg: ch**ùa** chiền, nhảy m**úa**.

- The diphthong /ɯə/ is written as:
  - **ươ** if the rhyme has a final consonant. Eg: bia r**ượu**, h**ưởng** thụ, chiêm ng**ưỡng**.
  - **ua** if the rhyme has no final consonant. Eg: ch**ùa** chiền, nhảy m**úa**.

- The monophthong /ă/ is written as:
  - **a** if is is followed by character **y**. Eg: m**áy** bay, c**ay** nồng, t**ay** chân.
  - **ă** otherwise. Eg: b**ắt** tay, b**ằng** lòng, may m**ắn**.

- The monophthong /i/ is written as:
  - **i** if the rhyme has a final consonant. Eg: l**íu** r**ít**.
  - **y** if the rhyme has no consonant. Eg: k**ỷ** luật, l**ý** do.

- The final /i̯/ is written as:
  - **i** if the rhyme has the front vowel /a/, the central vowel /ɯ, ɯə/ or the back vowels /u, o, ɔ, uo/ as the vowel. Eg: m**ải** mê, m**ui** thuyền, h**ỏi** han, m**ồi** chài, g**ửi** gắm, l**ười** biếng, n**uôi** nấng.
  - **y** if the rhyme has /ă, ɤ̆/ as the vowel. Eg: b**ay** lượn, ch**ạy** nh**ảy**, c**ấy** c**ày**.

Following these orthographic rules in Vietnamese, there is no ambiguity in converting phonemes to graphemes and vice versa. Analysis in this Section together with the analysis in the following Appendix C serve as the fundamentals for our Dialect-aware Tokenization algorithm, which is described comprehensively in Appendix D.

C Appendix: Multi-dialect Speech in Vietnamese

Most studies in multi-dialect speech in Vietnamese (Hạo, 1998; Đoàn Thiện Thuật, 2016; Châu, 2002) agree that Vietnamese has three dialects of speech, which are the Northern dialect, the Central dialect, and the Southern dialect. These dialects were studied and grouped mostly depending on how the residents deal with the rhymes. However, when diving into each particular dialect, there are more complicated phonetic phenomena, and more variations are explored. Section 3.2 above briefly gives an overview of the main differences among the three main dialects in Vietnam. In this section, we provide the reader with a more detailed discussion and analysis of each dialect. These discussions form the fundamentals for our multi-dialect tokenization method, which is described in the following sections.

C.1 Northern Dialect

According to (Hạo, 1998), Northern dialect speech can be found in provinces from Thanh Hóa to the northern borders of Vietnam. In particular, the Northern dialect is distributed in Thanh Hóa, Hà Giang, Cao Bằng, Bắc Kạn, Lạng Sơn, Tuyên Quang, Thái Nguyên, Phú Thọ, Bắc Giang, Quảng Ninh, Lào Cai, Lai Châu, Yên Bái, Điện Biên, Sơn La, Hòa Bình, Hà Nội, Hà Nam, Bắc Ninh, Hải Dương, Hải Phòng, Hưng Yên, Nam Định, Ninh Bình, Thái Bình, and Vĩnh Phúc provinces.

In initials, Northern dialect does not give distinct pronunciation on the following phonemes (Hạo, 1998; Châu, 2002) (the writing forms are provided in the parentheses):

- /s/ (x) and /ʂ/ (s) are pronounced as [s].
- /c͡ɕ/ (ch) and /t͡ʂ/ (tr) are pronounced as [c͡ɕ].
- /z/ (gi), /r/ (r), and /j/ (d) are pronounced as [z].

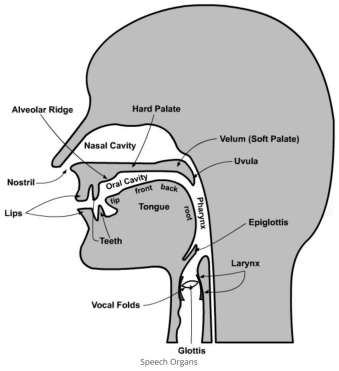

Figure 9: The speech organs.

- In Hà Nội, Vĩnh Phúc, Bắc Ninh, Hà Đông, and Hưng Yên, there is a mispronunciation between /l/ (l) and /n/ (n).

In rhymes, this dialect has the general following features (Hạo, 1998; Châu, 2002):

- The glide /u̯/ does not follows /u, o, ɔ/.
- The semivowel /j/ does not follow the front vowels /i, e, ɛ/ except for the cases where the final consonants are velar.
- The back diphthongs /ɯ, ɯə/ when being followed by the glide /u̯/ they are transmitted to the central diphthongs [ɨ, ɨə]. That is the reason why the following words **con *hươu*, ly *rượu*, cấp *cứu*, âm *mưu*** are pronounced approximately the same as **con *hiêu*, ly *riệu*, cấp *kíu*, âm *miu***, respectively.

In addition, this dialect exhibits two phonetic features on rhymes, including front vowels /i, e, ɛ/ or rounded back vowels /u, o, ɔ/ followed by velar consonants /ŋ, k/. The front vowels require the tongue to be raised and forward to the teeth in the mouth, while the velar consonants require the tongue to be raised and backward to the velum. This results in the placement of the tongue in the middle of the mouth, not forward to the teeth or backward to the velum (Figure 9). From that on, the velar consonants are palatalized to become [ɲ, c] while the front vowels are transmitted to the central vowels [ɨ, ə, ɜ]. The gliding between the central vowels and the palatalized consonants introduces a glide phone [j] in the middle. To this end, we have the following variations for the rhymes having front vowels /i, e, ɛ/ followed by velar consonants /ŋ, k/ (the writing forms are provided in the parentheses) (Hạo, 1998; Châu, 2002):

- /iŋ/ → [ɨʲɲ] (inh).
- /eŋ/ → [əʲɲ] (ênh).
- /ɛŋ/ → [ɜʲɲ] (anh).
- /ik/ → [ɨʲc] (ich).
- /ek/ → [əʲc] (êch).
- /ɛk/ → [ɜʲc] (ach).

In the cases of the rounded back vowels /u, o, ɔ/, the velar finals are labialized because of the effectiveness of labial vowels. These vowels are shortened according to the existence of the velar consonants, hence they lost the labial factor at the beginning. To this end, we have the following variations for the rhymes having rounded back vowels /u, o, ɔ/ followed by velar consonants /ŋ, k/ (the writing forms are provided in the parentheses) (Hạo, 1998; Châu, 2002):

- /uŋ/ → [ᵚᵘʊŋᵐ] (ung).
- /oŋ/ → [ɤʷŋᵐ] (ông).
- /ɔŋ/ → [ɑʷŋᵐ] (ong).
- /uk/ → [ᵚᵘʊkᵖ] (uc).
- /ok/ → [ɤʷkᵖ] (ôc).
- /ɔk/ → [ɑʷkᵖ] (oc).

C.2 Central Dialect

The Central dialect is distributed in more limited regions than the other two dialects. This dialect can be found in Nghệ An, Hà Tĩnh, Quảng Bình, Quảng Trị, and Huế provinces. Moreover, this dialect has a particular minor dialect for each province, which makes the Central dialect the most varied one in Vietnam. All minor variations of the Central dialect share the same features in initials and tones, while the rhymes are the only factor that makes them different from each other.

In general, this dialect has five tones:

- In Quảng Bình, Quảng Trị, and Huế, there is no distinction between tone /˧˩/ and /˧ʔ˥/.
- In Nghệ Tĩnh, there is no distinction between tone /˧ʔ˥/ and /˧ʔ˩/.

However, this dialect shares the same features relevant to initials in all provinces:

- /j, z/ → [z].
- /t͡ʂ/ → [ʈ].
- /c͡ɕ/ → [c].

The most significant factors that separate the minor dialects of the Central dialect are rhymes having front vowels /i, e, ɛ/ followed by the velar consonants /ŋ, k/. Rhymes having the rounded back vowels /u, o, ɔ/ followed by the velar consonants in the Central dialect share the same color as those in the Northern dialect.

C.2.1 Minor Dialect in Nghệ An and Hà Tĩnh

Rhymes in Nghệ An and Hà Tĩnh reflect the same rules as the Northern dialect for those having front vowels followed by velar consonants. In these rhymes, the velar consonants /ŋ, k/ are palatalized to become [ɲ, c], but the front vowels keep their original form. To this end, we have:

- /iŋ/ → [iɲ].
- /eŋ/ → [eɲ].
- /ɛŋ/ → [ɛɲ].
- /ik/ → [ic].
- /ek/ → [ec].
- /ɛk/ → [ɛc].

C.2.2 Minor Dialect in Quảng Trị

However, the rhymes mentioned in the above Section C.2.1 are pronounced slightly differently from those in Quảng Trị. In this province, both front vowels and velar consonants keep their original form:

- /iŋ/ → [iŋ].
- /eŋ/ → [eŋ].
- /ɛŋ/ → [ɛŋ].
- /ik/ → [ik].
- /ek/ → [ek].
- /ɛk/ → [ɛk].

C.2.3 Minor Dialect in Quảng Bình

In Quảng Bình, for rhymes having the nasal consonants /n, t/ following the front vowels /i, e, ɛ/, these vowels are pronounced longer than usual:

- /in/ → [iːn].
- /en/ → [eːn].
- /ɛn/ → [ɛːn].
- /it/ → [iːt].
- /et/ → [eːt].
- /ɛt/ → [ɛːt].

However, these vowels (iː, eː, ɛː) become shorter (i, e, ɛ) when being followed by the velar consonants /ŋ, k/. From that on, the distinction between these pairs /in/ - /iŋ/, /en/ - /eŋ/, /ɛn/ - /ɛŋ/, /it/ - /ik/, /et/ - /ek/, /ɛt/ - /ɛk/ is not necessarily dependent on the final but on the length of the vowel. To this end, this dialect replaces the velar consonants /ŋ, k/ by the nasal consonants /n, t/. These result in the following variations:

- /iŋ/ → [in].
- /eŋ/ → [en].
- /ɛŋ/ → [ɛn].
- /ik/ → [it].
- /ek/ → [et].
- /ɛk/ → [ɛt].

C.2.4 Minor Dialect in Huế

In this minor dialect, for rhymes having front vowels /i, e, ɛ/ followed by velar consonants /ŋ, k/, these vowels are transmitted to the central vowels [ɨ, ɘ, ɜ]. However, as the distinction between these pairs /iŋ/ - /in/, /ik - it/, /eŋ - en/, /ek - et/, /ɛŋ - ɛn/, /ɛk - ɛt/ largely depends on the difference of the vowels, hence having the velar consonants for these rhymes /iŋ, ik, eŋ, ek, ɛŋ, ɛk/ is not necessary. To this end, these velar consonants are replaced by the respective nasal consonants /n, t/. From then on, we have:

- /iŋ/ → [ɨn].
- /ik/ → [ɨt].
- /eŋ/ → [ɘn].

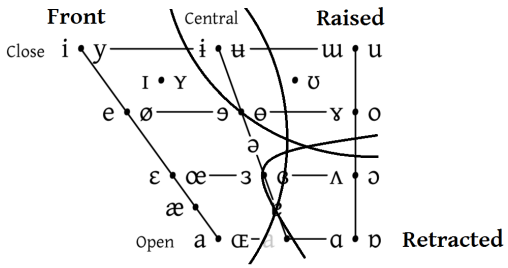

Figure 10: Vowel diagram.

- /ek/ → [ɘt].
- /ɛŋ/ → [ɜn].
- /ɛk/ → [ɜt].

The front vowels /i, e, ɛ/ preceding the nasal consonants /n, t/become longer. That is:

- /in/ → [iːn].
- /it/ → [iːt].
- /en/ → [eːn].
- /et/ → [eːt].
- /ɛn/ → [ɛːn].
- /ɛt/ → [ɛːt].

From these analyses, we can see that the Huế dialect exhibits the same behavior as the dialect in Quảng Bình. However, the way of pronouncing rhymes having the front vowels followed by the velar consonants is different in these two minor dialects. These contribute to the huge diversity of dialects in Vietnamese, which results in a significant challenge in modeling multi-dialect speech ASR in this language.

C.3 Southern Dialect

(Hạo, 1998; Châu, 2002) determined that the southern dialect can be found from the Đà Nẵng province to the southern border of Vietnam. In particular, the Southern dialect is distributed in Đà Nẵng, Quảng Nam, Quảng Ngãi, Kom Tum, Bình Định, Gia Lai, Phú Yên, Đắk Lắk, Khánh Hòa, Lâm Đồng, Ninh Thuận, Bình Thuận, Bình Phước, Đồng Nai, Bình Dương, Bà Rịa - Vũng Tàu, Hồ Chí Minh, Long An, Đồng Tháp, Tiền Giang, Bến Tre, Vĩnh Long, An Giang, Trà Vinh, Cần Thơ, Sóc Trăng, Kiên Giang, Bạc Liêu, and Cà Mau.

This dialect has five tones rather than six tones as the Northern dialect: there is no distinction between tone /˧˥/ and /˧ʔ˥/. In initials, Southern dialect does not give any difference between the following phonemes (Châu, 2002; Hạo, 1998):

- /z/ (gi), /j/ (d), and /v/ are all pronounced as [j].
- /k^{w}/ (qu) is pronounced as [w].
- This dialect also exhibits the same phonetic rule as the Northern dialect, where the velar consonants /ŋ, k/ follow the front vowels /i, e, ɛ/. These velar consonants are palatalized to become [ɲ, c], but the front vowels /i, e, ɛ/ keep their original.
- In provinces Quảng Nam, Quảng Ngãi, Kon Tum, Gia Lai, Đắk Lắk, Bình Định, Phú Yên, Khánh Hòa, Ninh Thuận, Bình Thuận, and Lâm Đồng: /t͡ʂ/ (tr) is pronounced as [ʈ]. While in other provinces, /t͡ʂ/ (tr) and /c͡ɕ/ (ch) is pronounced as [c], /ʂ/ (s) and /s/ (x) are all pronounced as [s].

In rhymes, when these diphthongs /iɛ, ɯɘ/ followed by the semivowel /i̯/ or labial consonants /m, p, u̯/, they are converted to the monophthongs in the same row of the vowel diagram (Figure 10). In particular, in this dialect we have:

- /iep/ → [ip].
- /ɯɘp/ → [ɯp].
- /iem/ → [im].
- /ɯɘm/ → [ɯm].
- /ieu̯/ → [iu̯].
- /ɯɘu̯/ → [ɯu̯].

This is the reason why in Southern dialect, the following words ***tiếp*** **tục, quả** ***mướp*****, đánh** ***chiếm*****,** ***ươm*** **mầm, tình** ***yêu*****, con** ***hươu*** are approximately pronounced as ***típ*** **tục, quả** ***múp*****, đánh** ***chím*****,** ***ưm*** **mầm, tình** ***iu*****, con** ***hưu***, respectively.

In addition to the general features of Southern dialect, two minor dialects share the same characteristics of Southern dialect but have their own particular features. The first one is the dialect distributed in provinces Quảng Nam and Quảng Ngãi, and the second one is the dialect found in the Mekong Delta.

C.3.1 Minor Dialect of Southern Dialect in Quảng Nam and Quảng Ngãi

In provinces Quảng Nam and Quảng Ngãi, the residents additionally pronounce the vowels differently from the Southern dialect. In rhymes without final consonants, the close monophthongs /i, u, ɯ/ are converted into diphthongs whose the first components are the respective more-open vowels (the row below of the vowel diagram, Figure 10), while the second components are the respective glides. In particular, we have:

- /i/ → [ɪj].
- /u/ → [ʊw].
- /ɯ/ → [ɤɯ].

In rhymes having /i̯/ or /u̯/ as the final consonants, the diphthongs /uo, ɯə, ie/ are transmitted into the monophthongs on the same row of the vowel diagram, but these monophthongs are longer to match the length of the original diphthongs:

- /uo/ → [uː].
- /ɯə/ → [ɯː].
- /ie/ → [iː].

Moreover, the following two rhymes are pronounced totally differently:

- /oi/ → [uə].
- /ai/ → [ae].

This is the reason why people in these two provinces pronounce the following words ***nói*** and ***hai*** approximately the same as ***núa*** and ***he***, respectively.

Furthermore, in rhymes having labial finals /m, p/ and the rounded semivowel /u̯/, the vowels in this dialect are pronounced as follows:

- /a/ → [ɔ].
- /ɔ/ → [o].
- /o/ → [ɤ].
- /u/ → [ʊ].
- /ă, ə̆/ → [a].
- In rhymes having /m/ as the final, /uo/ → [ʊ].

In rhymes having the velar consonants as the final, while the front vowels /i, e, ɛ/ do not vary, the following vowel changed as follows:

- /a/ → [ɑː].
- /ɔ/ → [ɑːᵐ].
- /o/ → [ʌ].
- /ə̆/ → [ɑ].
- /uo/ → [uː].
- /ie/ → [iː].
- /ɯə/ → [ɯː].

C.3.2 Minor Dialect of Southern Dialect in Mekong Delta

Mekong Delta includes Bình Dương, Bình Phước, Đồng Nai, Bà Rịa - Vũng Tàu, Tây Ninh, Hồ Chí Minh, Long An, Tiền Giang, Hậu Giang, Long An, Bến Tre, Đồng Tháp, Vĩnh Long, Trà Vinh, Cần Thơ, Sóc Trăng, An Giang, Kiên Giang, Bạc Liêu, and Cà Mau. These provinces have a unique dialect that shares many features with the Southern dialect, with their own distinctions.

In particular, the front vowels /i, e/ followed by velar consonants /ŋ, k/ do not transmitted to the respective central vowels, but the velar consonants are replaced by the nasal consonants /ŋ, k/ → /n, t/, which is similar to the Central dialect without changing in length of the front vowels. That is:

- /iŋ/ → [in].
- /ik/ → [it].

In addition, the front vowel /e/ followed by the velar consonants /ŋ, k/ is converted to the respective central vowel because of the velar consonants, but as the velar consonants /ŋ, k/ are replaced by the nasal consonants [n, t], which results in this vowel is longer. From then on, we have:

- /eŋ/ → [əːn].
- /ek/ → [əːt].

However, the alveolar vowels /n, t/ following vowel /a/, and the central vowels /ə, ɛ/ become the velar consonants /ŋ, k/. This is the reason why in this dialect, these words **buôn** ***bán***, ***út*** **ỏi**, ***ân cần***, ***chen lấn*** are pronounced approximately the same as **buôn** ***báng***, ***ích*** **ỏi**, ***âng cầng***, ***cheng lấng***, respectively.

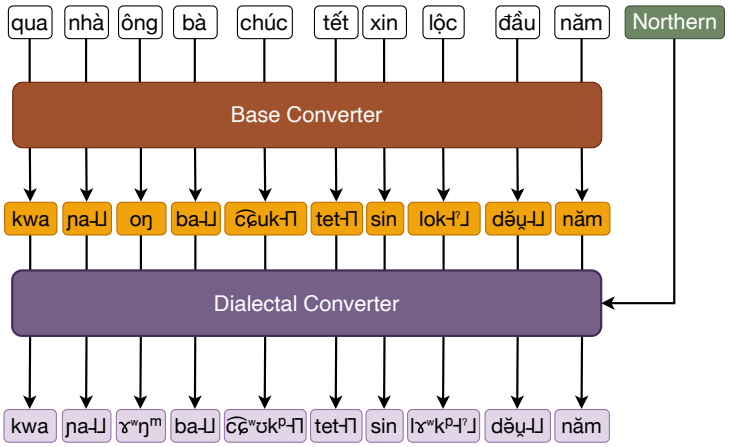

Figure 11: Dialect-aware tokenization algorithm.

On the other hand, the rounded back ones /u, o, ɔ)/ have variations depending on the following consonants. If these vowels are followed by the nasal consonants /n, t/, they keep their rounded features while their length is longer /u, o, ɔ/ → [uː, oː, ɔː]. In the case these vowels are followed by velar consonants /ŋ, k/, they lose the rounded characteristics as in the Northern dialect with a little bit difference, while these velar consonants are labialized to become [ŋ^{m}, k^{p}]. Moreover, the nasal consonants [n, t] followed by these vowels are replaced by the velar-labialized consonants [ŋ^{m}, k^{p}], respectively. In particular, we have:

- /un, uŋ/ → [^{ɯ}ʊŋ^{m}].
- /ut, uk/ → [^{ɯ}ʊk^{p}].
- /on, oŋ/ → [ɤŋ^{m}].
- /ɔt, ɔk/ → [ʌk^{p}].

D Appendix: Dialect-Aware Tokenization Algorithm

We have analyzed in detail the phonetic and orthographic features of Vietnamese in Section B and comprehensively describe the multi-dialect variations of this language in Section C. From these analysis, we develop an Dialect-aware Tokenization algorithm (DiaToken). DiaToken receives a line of text which is the transcription of the given multi-dialect voice and the province ID, then processes and returns the sequence of IPA (International Phonetic Alphabet) characters describing the dialect speech sound of the given audio correspond to the dialect. DiaToken has two phases (Figure 11): (1) Retrieving phonemes from graphemes given the text of the transcript (the Base Converter module) and (2) Converting phonemes to IPA characters respective to dialectal phones (the Dialectual Converter module).

D.1 Base Converter Module

Algorithm 1: The algorithm for converting text to phonemes.

Data: Transcript of the audio $w = (w_1, w_2, ..., w_n)$.
Result: A sequence of syllables $p = (p_1, p_2, ..., p_n)$ of the given input transcript $w = (w_1, w_2, ..., w_n)$. Each phoneme $p_i = (p_i^{init}, p_i^{glide}, p_i^{vowel}, p_i^{final}, p_i^{tone})$ is a triplet of IPA for the initial, rhyme, and tone.

```
phonemes ← an empty list [];
for W in w do
    p_tone, W ← get_tone(W);
    p_initial, W ← get_initial(W);
    p_glide, W ← get_glide(W);
    p_vowel, W ← get_vowel(W);
    p_final ← get_final(W);
    phonemes ← Append
      (p_init, p_rhyme, p_tone);
end
return phonemes;
```

We describe in Alg. 1 a overview of algorithm of the Base Converter module for converting orthographic form of the transcript to the sequence of phonemes. In this algorithm, we repre-

Algorithm 2: Algorithm for determining initial of a word.

Function *get_initial(*W*)*:
  Input: A word W in Vietnamese
  Output: The initial i of W
  $onsets \leftarrow$ [ngh, tr, th, ph, nh, ng, kh, gi, gh, ch, q, đ, x, v, t, s, r, n, m, l, k, h, g, d, c, b];
  $i \leftarrow$ None;
  for *onset in onsets* **do**
    if W *starts with onset* **then**
      /* Words starting with "qu" are kept the onset for later process. */
      if $onset \neq k$ **then**
        Remove $onset$ from W;
      end
      $i \leftarrow$ IPA character of $onset$ according to Appendix B;
    end
    break;
  end
  return i, W;
end

Algorithm 4: Algorithm for determining vowel of a word.

Function *get_vowel(*W*)*:
  Input: A word W in Vietnamese
  Output: The vowel v of W
  $nuclei \leftarrow$ [oo, ươ, ưa, uô, ua, iê, yê, ia, ya, e, ê, u, ư, ô, i, y, ơ, â, a, o, ă];
  for *nucleus in nuclei* **do**
    if W *starts with nucleus* **then**
      $v \leftarrow$ IPA character of $nucleus$ according to Appendix B;
      Remove $nucleus$ from W;
      return v, W;
    end
  end
  return *None,* W;
end

sent each word as a vector of five syllabic components: initial, glide, vowel, final, and tone. From that on, these five phonemes of each word will be mapped to the dialectal phone according to the given province of the audio.

Algorithm 3: Algorithm for determining final of a word.

Function *get_final(*W*)*:
  Input: A word W in Vietnamese
  Output: The final f of W
  $codas \leftarrow$ [ng, nh, ch, u, n, o, p, c, m, y, i, t];
  if W *in codas* **then**
    return W;
  end
  return *None*;
end

Algorithm 5: Algorithm for determining tone of a word.

Function *get_tone(*W*)*:
  Input: A word W in Vietnamese
  Output: The tone t of W
end
$t \leftarrow$ tone of word according to Appendix B;
Remove tone from W;
return t, W;

Moreover, as described in Alg. 5, Alg. 2, Alg. 6, Alg. 4, and Alg. 3, the computational complexity of the Base Convert (Figure 11) is linear $\mathcal{O}(n)$ while using a fix sized of vocabulary of phonemes but tokenizing unlimited number of Vietnamese words.

Algorithm 6: Algorithm for determining glide of a word.

Function *get_glide(*W*)*:
  Input: A word W in Vietnamese
  Output: The glide g of W
  if W *starts with* ***qu*** **then**
    Remove **qu** from W;
    return $\underset{\frown}{u}$, W;
  end
  for *case in [oa, oă, oe]* **do**
    if W *starts with case* **then**
      Remove **o** from W;
      return $\underset{\frown}{u}$, W;
    end
  end
  for *case in [uê, uy, uơ, ua, uâ, uya]* **do**
    if W *starts with case* **then**
      Remove **u** from W;
      return $\underset{\frown}{u}$, W
    end
  end
  return *None, W*;
end

D.2 Dialectal Converter Module

Having the phoneme representation for the given transcript, the DiaToken continues to retrieve the dialectal representation of the given transcript according to the name of the province. Vietnam has 63 provinces and their speech sound are grouped into three large dialects as described in Appendix C. DiaToken receives the name of the province then determines the respective dialect and finally convert every phonemes to the corresponding phones. The overall algorithm is described in Alg. 7.

From Alg. 8, Alg. 9, and Alg. 10, the computational complexity is linear $\mathcal{O}(n)$, which means the complexity of the Dialectal Converter Module is $\mathcal{O}(n)$. Finally, the overall complexity of the DiaToken algorithm is $\mathcal{O}(n)$ which is efficient for tokenizing transcript in the context of multi-dialect.

Algorithm 7: The algorithm for converting phonemes to dialectal phones

Data:
- The name of the province **ProID**.
- A sequence $p = (p_1, p_2, ..., p_n)$ representing phonemic syllables of the transcript $w = (w_1, w_2, ..., w_n)$. Each $p_i = (p_i^{init}, p_i^{glide}, p_i^{vowel}, p_i^{final}, p_i^{tone})$ is a vector of five syllabic components of word w_i.

Result: A sequence $ph = (ph_1, ph_2, ..., ph_n)$ representing dialectal syllables of the transcript $w = (w_1, w_2, ..., w_n)$. Each $ph_i = (ph_i^{init}, ph_i^{rhyme}, ph_i^{tone})$ is a vector of three dialectal phonetic components of word w_i.

$northern \leftarrow$ set of provinces having the Northern dialect according to Appendix C;
$middle \leftarrow$ set of provinces having the Central dialect according to Appendix C;
$southern \leftarrow$ set of provinces having the Southern dialect according to Appendix C;
if ***ProID*** $\in northern$ **then**
  $pth \leftarrow \text{get\_northern}\ (p)$;
end
if ***ProID*** $\in middle$ **then**
  $pth \leftarrow \text{get\_middle}\ (p)$;
end
if ***ProID*** $\in southern$ **then**
  $pth \leftarrow \text{get\_southern}\ (p)$;
end
return ph;

Algorithm 8: Algorithm for determining Northern dialect of a syllable.

Function *get_nothern(s)*:
  Input: A sequence of phonemic syllables $s = (s_1, s_2, ..., s_n)$
  Output: A sequence of Northern phonetic syllables $s = (s_1, s_2, ..., s_n)$
  for s_i *in* s **do**
    $s_i^{init} \leftarrow s_i$;
    $s_i^{init} \leftarrow$ IPA character for Northern dialect of the initial s_i^{init} according to Appendix C.1;
    $s_i^{glide} \leftarrow$ IPA character for Northern dialect of the glide s_i^{glide} according to Appendix C.1;
    $s_i^{vowel} \leftarrow$ IPA character for Northern dialect of the vowel s_i^{vowel} according to Appendix C.1;
    $s_i^{final} \leftarrow$ IPA character for Northern dialect of the final s_i^{final} according to Appendix C.1;
    $s_i^{tone} \leftarrow$ IPA character for Northern dialect of the tone s_i^{tone} according to Appendix C.1;
    $s_i^{rhyme} \leftarrow s_i^{glide} \oplus s_i^{vowel} \oplus s_i^{final}$; /* $\oplus$ represents the concatenation operator */
    $s_i \leftarrow (s_i^{init}, s_i^{rhyme}, s_i^{tone})$;
  end
end

Algorithm 9: Algorithm for determining Central dialect of a syllable.

Function *get_nothern(s)*:
  Input: A sequence of phonemic syllables $s = (s_1, s_2, ..., s_n)$
  Output: A sequence of Central phonetic syllables $s = (s_1, s_2, ..., s_n)$
  for s_i *in* s **do**
    $s_i^{init} \leftarrow s_i$;
    $s_i^{init} \leftarrow$ IPA character for Central dialect of the initial s_i^{init} according to Appendix C.2;
    $s_i^{glide} \leftarrow$ IPA character for Central dialect of the glide s_i^{glide} according to Appendix C.2;
    $s_i^{vowel} \leftarrow$ IPA character for Central dialect of the vowel s_i^{vowel} according to Appendix C.2;
    $s_i^{final} \leftarrow$ IPA character for Central dialect of the final s_i^{final} according to Appendix C.2;
    $s_i^{tone} \leftarrow$ IPA character for Central dialect of the tone s_i^{tone} according to Appendix C.2;
    $s_i^{rhyme} \leftarrow s_i^{glide} \oplus s_i^{vowel} \oplus s_i^{final}$; /* $\oplus$ represents the concatenation operator */
    $s_i \leftarrow (s_i^{init}, s_i^{rhyme}, s_i^{tone})$;
  end
end

Algorithm 10: Algorithm for determining Southern dialect of a syllable.

Function *get_nothern(s)*:
  Input: A sequence of phonemic syllables $s = (s_1, s_2, ..., s_n)$
  Output: A sequence of Southern phonetic syllables $s = (s_1, s_2, ..., s_n)$
  for s_i *in* s **do**
    $s_i^{init} \leftarrow s_i$;
    $s_i^{init} \leftarrow$ IPA character for Southern dialect of the initial s_i^{init} according to Appendix C.3;
    $s_i^{glide} \leftarrow$ IPA character for Southern dialect of the glide s_i^{glide} according to Appendix C.3;
    $s_i^{vowel} \leftarrow$ IPA character for Southern dialect of the vowel s_i^{vowel} according to Appendix C.3;
    $s_i^{final} \leftarrow$ IPA character for Southern dialect of the final s_i^{final} according to Appendix C.3;
    $s_i^{tone} \leftarrow$ IPA character for Southern dialect of the tone s_i^{tone} according to Appendix C.3;
    $s_i^{rhyme} \leftarrow s_i^{glide} \oplus s_i^{vowel} \oplus s_i^{final}$; /* $\oplus$ represents the concatenation operator */
    $s_i \leftarrow (s_i^{init}, s_i^{rhyme}, s_i^{tone})$;
  end
end