Vietnamese Automatic Speech Recognition: A Revisit

Thi Vu, Linh The Nguyen, Dat Quoc Nguyen
Qualcomm AI Research[*]
{thivu, linhnt, datnq}@qti.qualcomm.com

Abstract

Automatic Speech Recognition (ASR) performance is heavily dependent on the availability of large-scale, high-quality datasets. For low-resource languages, existing open-source ASR datasets often suffer from insufficient quality and inconsistent annotation, hindering the development of robust models. To address these challenges, we propose a novel and generalizable data aggregation and preprocessing pipeline designed to construct high-quality ASR datasets from diverse, potentially noisy, open-source sources. Our pipeline incorporates rigorous processing steps to ensure data diversity, balance, and the inclusion of crucial features like word-level timestamps. We demonstrate the effectiveness of our methodology by applying it to Vietnamese, resulting in a unified, high-quality 500-hour dataset that provides a foundation for training and evaluating state-of-the-art Vietnamese ASR systems. Our project page is available at `https://github.com/qualcomm-ai-research/PhoASR`.

1 Introduction

Recent advances in Automatic Speech Recognition (ASR) have demonstrated that pretraining on massive, weakly supervised or self-supervised corpora produces models with improved robustness and cross-domain generalization (Baevski et al., 2020; Radford et al., 2023). However, when adapting these models to specific languages or downstream tasks, the benefits of scale alone begin to plateau. At this point, high-quality supervised datasets become essential for achieving further performance improvements (Radford et al., 2023).

Prior efforts to build supervised ASR datasets have typically followed one of two approaches: scripted read speech (Panayotov et al., 2015; Paul and Baker, 1992; Veaux et al., 2017; Luong and Vu, 2016) or spontaneous speech collection (Chen et al., 2021; Galvez et al., 2021; Tran et al., 2024). Read speech datasets are typically recorded in controlled environments, which ensures high audio quality and accurate transcriptions. While this approach offers clear advantages in terms of data quality, it comes with drawbacks: these datasets are expensive to produce and often lack the linguistic variability and prosodic richness characteristic of natural conversations. This limitation reduces their effectiveness for real-world ASR applications.

In contrast, datasets based on spontaneous speech offer several advantages: they are cheaper and more scalable to produce, and they better reflect real-world usage patterns due to their diverse sources, which often include publicly available content such as podcasts and YouTube videos. The creation pipeline for these datasets typically employs existing ASR models and forced alignment tools for automatic transcription, often supplemented by manual human review to ensure transcript quality (Dinh et al., 2024; Le-Duc, 2024).

For Vietnamese, existing open-source datasets exemplify these trade-offs. Some are based on controlled read speech but suffer from limited diversity (Luong and Vu, 2016), while others draw from spontaneous sources but struggle with inconsistent transcription quality (Dinh et al., 2024). Furthermore, these datasets often vary in sampling rate and format, with preprocessing steps that are poorly documented (Nhut et al., 2024). This inconsistency makes it challenging to combine datasets or reuse them effectively in unified training setups.

Timestamps are rarely included in these datasets, which limits their use for training models that require fine-grained alignment—such as those used in subtitle generation or audio editing. While post hoc alignment tools like Montreal Forced Aligner (McAuliffe et al., 2017) or other alignment models (Baevski et al., 2020) can be applied to ASR transcription outputs to provide times-

[*] Qualcomm Vietnam Company Limited. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Findings of the Association for Computational Linguistics: EACL 2026, pages 6557–6568
March 24-29, 2026 ©2026 Association for Computational Linguistics

tamps (Bain et al., 2023), this approach increases computational cost and creates formatting challenges. These alignment tools often require transcripts to be in standard written form. For example, numbers must be spelled out (e.g., “forty five” instead of “45”) to avoid out-of-vocabulary issues. However, ASR systems typically output digits (e.g., “45”) rather than words (“forty five”), creating a mismatch with alignment model requirements. This incompatibility is problematic because the spelled-out format may not match what end-users actually want.

Similarly, post-processing models for punctuation restoration operate only on text without access to the original audio, preventing them from correctly placing punctuation based on acoustic cues like pauses or intonation.

To address these limitations, we propose a generalizable pipeline for aggregating and preprocessing ASR data from diverse sources. Our pipeline leverages existing deep learning models and auxiliary tools to clean, normalize, and align raw audio-text pairs, producing word-level timestamps and datasets that are high-quality, balanced, and feature-rich. We demonstrate this pipeline’s effectiveness by applying it to Vietnamese, creating a 500-hour dataset suitable for fine-tuning and evaluating modern ASR systems. This pipeline can be adapted to different languages facing similar data quality challenges, making it an important contribution to advancing speech processing research.

2 Related Work

2.1 Vietnamese ASR Datasets

The landscape of Vietnamese ASR is shaped by a variety of datasets, each with its own strengths and limitations. These can be broadly categorized into read speech and spontaneous speech corpora. Read speech datasets include CMV-vi-14 (Ardila et al., 2020), VIVOS (Luong and Vu, 2016) and FPT Open Speech Dataset (FOSD) (Tran, 2020). Spontaneous speech datasets include VSV-1100 (Nhut et al., 2024), viVoice (Gia et al., 2024), BUD500 (Pham et al., 2024), VLSP 2020,[1] Vietnam-Celeb (Pham et al., 2023), ViMD (Dinh et al., 2024), LSVSC (Tran et al., 2024), VietMed-L (Le-Duc, 2024). In this section, we will briefly highlight some datasets from both categories.

Read speech datasets. Common Voice (Ardila et al., 2020) is a large-scale, multilingual crowd-sourced dataset initiative by Mozilla, aiming to provide free and publicly available voice data for speech technology development. Volunteers contribute by recording themselves reading sentences from a public-domain text corpus, and other users validate the recordings by listening to them. While it has grown to be one of the largest public voice datasets, covering over 100 languages, the data distribution is highly skewed. For many low-resource languages, the available data is very limited. For instance, the Vietnamese portion CMV-vi-14 of the dataset in the 14th version contains less than five hours of validated audio,[2] which is insufficient for training high-performance ASR systems from scratch. The text in the dataset does include punctuation and casing, but timestamps are unavailable.

VIVOS (Luong and Vu, 2016) is a Vietnamese speech corpus originally created for text-to-speech research. It consists of 15 hours of read speech from a small number of speakers recorded in a controlled environment. While the audio quality is high, the provided transcripts often lack punctuation and capitalization and do not include word-level timestamps. These limitations make it less suitable for training models that require rich text features or precise audio-text alignment without further processing.

Spontaneous speech datasets. BUD500 (Pham et al., 2024) is presented as a large corpus of spontaneous Vietnamese speech. However, limited information is publicly available regarding its data collection and annotation methodology, making it difficult to assess its quality comprehensively. Our analysis in Table 2 shows that a substantial portion of the dataset fails to pass our quality filters, suggesting inaccuracies in the provided transcripts. Additionally, the dataset lacks standardized punctuation, capitalization, and word-level timestamps, which limits its direct applicability for training robust ASR models.

VietMed-L (Le-Duc, 2024) is a 16-hour labeled subset of the VietMed dataset for Vietnamese medical ASR. The dataset was created through a multi-stage, computer-assisted annotation workflow. The process began with initial transcripts generated by YouTube’s automatic captioning service, which were then independently corrected by two native

[1] https://vlsp.org.vn/vlsp2020/eval/asr

[2] https://huggingface.co/datasets/mozilla-foundation/common_voice_14_0

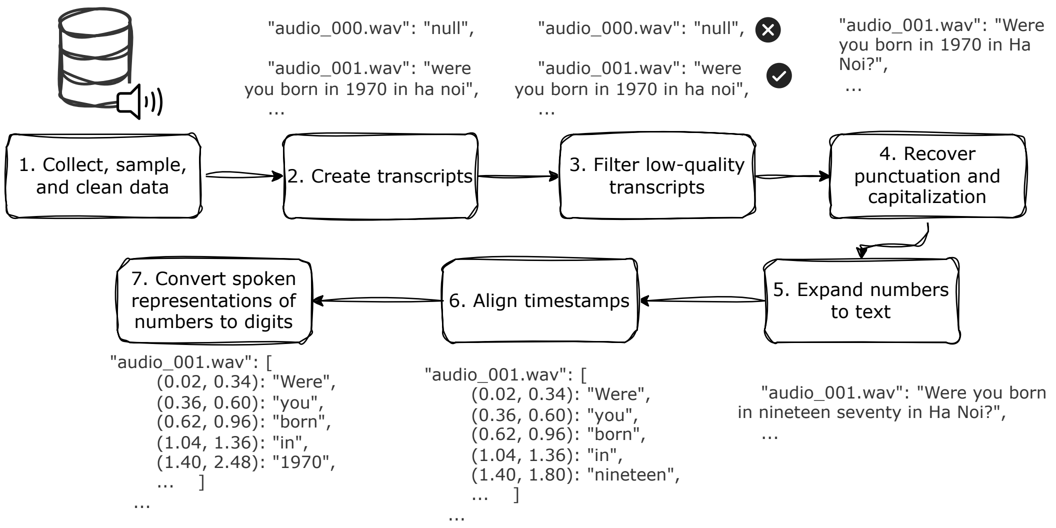
"audio_000.wav": "null",
"audio_001.wav": "were you born in 1970 in ha noi",
...
"audio_000.wav": "null",
"audio_001.wav": "were you born in 1970 in ha noi",
...
"audio_001.wav": "Were you born in 1970 in Ha Noi?",
...
1. Collect, sample, and clean data
2. Create transcripts
3. Filter low-quality transcripts
4. Recover punctuation and capitalization
5. Expand numbers to text
6. Align timestamps
7. Convert spoken representations of numbers to digits
"audio_001.wav": [
(0.02, 0.34): "Were",
(0.36, 0.60): "you",
(0.62, 0.96): "born",
(1.04, 1.36): "in",
(1.40, 2.48): "1970",
...]
...
"audio_001.wav": [
(0.02, 0.34): "Were",
(0.36, 0.60): "you",
(0.62, 0.96): "born",
(1.04, 1.36): "in",
(1.40, 1.80): "nineteen",
...]
...
"audio_001.wav": "Were you born in nineteen seventy in Ha Noi?",
...

Figure 1: Overview of our pipeline for building a high-quality ASR dataset from multiple sources. The figure uses an English example for illustration. We apply this pipeline to Vietnamese. But, it is adaptable to other languages.

Vietnamese speakers. The two corrected versions were compared, and segments with significant transcription differences were excluded from the dataset. For final quality assurance, a small validation portion was manually transcribed from scratch by three annotators with medical backgrounds and subsequently merged with the computer-assisted versions. Although the dataset transcripts benefit from both human and machine annotations, they lack punctuation, capitalization, and timestamps.

Note that other datasets in the read and spontaneous categories face similar challenges to those reviewed above: they lack punctuation, capitalization, and word-level timestamps.

2.2 Forced Alignment for Timestamp Generation

A key challenge in preparing ASR datasets is generating accurate word-level timestamps through a process known as **forced alignment**.

WhisperX (Bain et al., 2023) adopts a two-stage, model-based approach to forced alignment. In the first stage, a large, pre-trained ASR model (by default, OpenAI's Whisper) generates a highly accurate transcript of the audio. In the second stage, the system performs forced alignment using a separate model, typically `wav2vec2` (Baevski et al., 2020), that has been fine-tuned specifically for alignment tasks. This alignment model employs a Connectionist Temporal Classification (CTC) head, which outputs a probability distribution over vocabulary tokens (e.g., characters or phonemes) for each time step of the audio representation. Given the transcript from the first stage, the aligner identifies the most probable sequence of token spikes in the CTC output matrix that corresponds to the transcript. The time steps associated with the start and end of each word's tokens are then merged to produce the final word-level timestamps.

This methodology bypasses the need for a pronunciation dictionary, making it highly adaptable to various languages and domains. However, the process is computationally intensive due to the use of large models, and the alignment step often requires transcripts to be in standard written form. For instance, numbers may need to be converted to their full textual representation (e.g., 45 to forty-five), creating an intermediate processing step that may not be desired in the final output. Our pipeline incorporates this approach while managing these complexities to achieve high-quality, feature-rich results.

3 Our pipeline

Our dataset creation pipeline, shown in Figure 1, consists of the following steps:

1. **Collect, sample, and clean data:** This step comprises: (i) collecting a diverse set of open-source audio samples from publicly available repositories, (ii) filtering out audio samples containing invalid characters or non-standard punctuation, and (iii) normalizing the audio volume

and resampling each sample to a predefined target sampling rate to ensure consistency across the dataset.

2. **Create transcripts:** For audio datasets lacking associated transcripts, this step comprises: (i) applying two state-of-the-art pre-trained ASR models to independently generate transcripts, (ii) computing the word error rate (WER) by comparing the transcript output of one ASR model against that of the other, and (iii) retaining only those audio-transcript pairs for which the computed WER falls below a predefined threshold, thus ensuring transcript reliability.
3. **Filter low-quality transcripts:** For datasets that include pre-existing transcripts, this step involves: (i) applying an additional ASR model to generate a secondary transcript for each audio sample, (ii) comparing the secondary transcript with the provided transcript to compute the WER, (iii) discarding samples where the computed WER exceeds a predefined threshold, and (iv) using manually annotated transcripts without modification, where available.
4. **Recover punctuation and capitalization:** To enhance transcript readability, this step includes: (i) training and deploying a neural network model to restore punctuation and capitalization in the transcripts, (ii) normalizing both the model input and output by converting them to lowercase and removing punctuation, and (iii) retaining only those samples for which the normalized input and output match exactly, thereby ensuring the fidelity of the restored formatting.
5. **Expand numbers to text:** All numeric tokens in the transcripts are converted to their corresponding full-text (spoken) representations. For example, the numeral "30" is transformed into the word "thirty." This transformation is performed to ensure compatibility with a subsequent alignment model that requires textual input for accurate processing.
6. **Align timestamps:** This step involves applying an alignment model to the audio signal and its corresponding transcript. The alignment model generates temporal metadata by assigning timestamps to each syllable in the transcript, thereby enabling word-level synchronization between the audio and the textual transcript.
7. **Convert spoken representations of numbers to digit:** This step is to convert the spoken textual representations of numbers back into their digit form. In instances where a single numeric expression spans multiple words and associated timestamps (e.g., “one hundred twenty-three”), the corresponding timestamps are merged to represent the numeric entity as a unified temporal segment.

This pipeline ensures that the resulting dataset is clean, standardized, and enriched with word-level timestamps, making it well-suited for training and evaluating modern ASR systems.

4 Our PhoASR dataset

We demonstrate how the proposed pipeline is applied to Vietnamese to construct a 500-hour high-quality dataset PhoASR, as well as an extended version of its training set (PhoASR-3100h).

4.1 PhoASR Dataset Creation

Step 1. We download multiple datasets, including: VSV-1100 (Nhut et al., 2024), viVoice (Gia et al., 2024), BUD500 (Pham et al., 2024), VLSP 2020[3], Vietnam-Celeb (Pham et al., 2023), ViMD (Dinh et al., 2024), LSVSC (Tran et al., 2024), FPT Open Speech Dataset (FOSD) (Tran, 2020), VietMed-L (Le-Duc, 2024), VIVOS (Luong and Vu, 2016), CMV-vi-14 (Ardila et al., 2020), all of which are released under licenses that permit use for research purposes. Table 1 reveals considerable heterogeneity across the source datasets. Sampling rates vary widely from 8kHz to 48kHz, while the datasets exhibit diverse speech types ranging from read speech (CMV-vi-14, VIVOS) to spontaneous conversations (VSV-1100, VietMed-L) and mixed content. The domains are equally varied, covering general-purpose speech, medical conversations, broadcast news, and entertainment shows. However, some datasets lack proper documentation, with unknown domains and speech types. Text lengths and audio durations also show significant variation, with ViMD containing the longest average samples (19s) and BUD500 the shortest (3s). While this heterogeneity presents significant preprocessing challenges, it results in a dataset whose diversity is crucial for enhancing the generalizability of any model trained on it. Table 2 shows that the total duration of audio data is approximately 3342 hours, which includes 3171 hours of split training data.

To ensure dataset diversity and prevent dominance by particular speakers or dialects, we sam-

[3] https://vlsp.org.vn/vlsp2020/eval/asr

Dataset	Sample Rate (Hz)	Speech Type	Domain	Text Length	Duration	License
VSV-1100	16000	Spontaneous	General	3–43 | 14	2–8 | 4	Apache 2.0
viVoice	24000	Spontaneous	General	1–60 | 16	1–19 | 4	CC-BY-NC-SA 4.0
BUD500	16000	Spontaneous	General	6–32 | 10	1–8 | 3	CC-BY-NC-SA 4.0
VLSP2020	16000	Spontaneous	Unknown	1–132 | 17	1–30 | 5	Custom
Vietnam-Celeb	16000	Spontaneous	Interviews, Shows	W/o transcripts	1–30 | 8	Custom
ViMD	44100	Spontaneous	Broadcast News	6–132 | 62	2–30 | 19	CC-BY-NC-ND 4.0
LSVSC	16000	Spontaneous	General	1–69 | 26	1–13 | 7	CC-BY 4.0
FOSD	48000	Read	Unknown	1–59 | 12	3–6 | 4	Custom
VietMed-L	8000	Spontaneous	Medical	5–39 | 24	2–10 | 7	CC-BY-4.0
VIVOS	16000	Read	Unknown	2–30 | 13	1–18 | 5	CC-BY-NC-SA-4.0
CMV-vi-14	32000	Read	General	1–14 | 8	2–10 | 5	CC0

Table 1: Characteristics of the source datasets. "Text Length" and "Duration" are accounted for "words" and "seconds", respectively, with "min-max | mean" values. All datasets permit research use.

ple only portions of the downloaded datasets. For datasets with available speaker information, namely Vietnam-Celeb (Pham et al., 2023) and ViMD (Dinh et al., 2024), we limit each speaker to a maximum of 10 minutes. For the viVoice dataset (Gia et al., 2024), where audio is collected from YouTube, we treat the channel name as the speaker identifier and sample at most 30 minutes per channel. For VSV-1100 (Nhut et al., 2024) and BUD500 (Pham et al., 2024), which lack speaker information, we train a classification model to predict the province of each audio sample and limit sampling to 50 minutes per province. The classification model is trained using the ViMD (Dinh et al., 2024) dataset, which includes province information. We use the soxan codebase [4] to finetune the wav2vec2-base-vi (Nguyen, 2021) model, achieving 43.48% province prediction accuracy on the ViMD test set. Even with low classification accuracy, the model assigns different labels to different voices and groups similar voices together, ensuring speaker diversity in the sampling process.

To ensure a consistent dataset, we perform cleaning and normalization. The data is first filtered by removing samples longer than 30 seconds and those with special characters in their transcripts, ensuring only standard punctuation (, . ! ?) and alphabet characters remain. All audio is then normalized in volume and resampled to 16 kHz using the sox library.[5] This initial processing phase results in an 809-hour dataset.

Step 2. This step addresses datasets that lack pre-existing transcripts. Among the 11 datasets collected, Vietnam-Celeb (Pham et al., 2023) does not contain transcripts as it was constructed for speaker verification tasks. We use two state-of-the-art Vietnamese ASR models—PhoWhisper-large (Le et al., 2024) and ChunkFormer-large-vie (Le et al., 2025)—to predict the transcripts, then calculate the word error rate (WER) between their output predictions. Only audio samples with the corresponding WER lower than 5% are retained.

Step 3. For datasets with pre-existing transcripts, we run PhoWhisper-large and compute the WER between the predicted transcript and the ground truth transcript provided by each dataset. Only samples with WER lower than 5% are retained.

Step 4. We fine-tune the bartpho-syllable-base model (Tran et al., 2022) for capitalization and punctuation recovery. The training data is taken from the "news-corpus",[6] which contains main content texts from Vietnamese news articles. We select a subset of 5M samples without special characters and apply four transformation strategies to generate training variations: 85% of samples are lower-cased and stripped of punctuation, 5% have only punctuation removed, 5% are only lower-cased, and the final 5% are left unchanged. We then fine-tune the model for 20 epochs and use the fine-tuned one to predict punctuation and capitalization for the audio transcripts. To ensure prediction quality, we first lowercase both the fine-tuned model's transcript input and prediction output, then remove punctuation from both lowercased variants. We compare these processed versions of the input and output, retaining only samples where

[4] https://github.com/m3hrdadfi/soxan

[5] https://sourceforge.net/projects/sox/

[6] https://github.com/binhvq/news-corpus

Dataset	Original	Sampled	PhoASR	Train.	Valid.	Test	PhoASR-3100h
VSV-1100	1144.53	48.85	25.69	25.69	N/A	N/A	1144.52
viVoice	1016.97	89.41	33.42	33.42	N/A	N/A	997.07
BUD500	461.90	45.24	20.61	20.55	0.03	0.03	456.58
VLSP2020	261.83	261.83	205.87	196.48	2.06	7.33	258.91
Vietnam-Celeb	187.37	98.95	37.37	37.37	N/A	N/A	62.58
ViMD	102.56	97.57	61.60	53.13	6.54	1.93	98.10
LSVSC	100.66	100.66	73.47	61.94	7.68	3.85	98.05
FOSD	30.18	30.18	21.97	20.15	1.08	0.74	30.12
VietMed-L	15.93	15.93	4.20	4.08	N/A	0.12	15.92
VIVOS	15.66	15.66	14.31	12.88	0.91	0.52	15.66
CMV-vi-14	4.79	4.79	4.15	2.91	0.39	0.85	4.79
Total	3342.38	809.07	502.67	468.60	18.70	15.37	3100.52

Table 2: Dataset sizes (in hours) at different stages of the pipeline and the final split distribution. "Train." and "Valid." refer to the Training and Validation splits of PhoASR, respectively. N/A indicates that the validation or test split is not available for a given dataset. "PhoASR-3100h" denotes the extended training set, which combines minimally processed data with the high-quality data from the PhoASR training split (see details in Section 4.2).

they are identical, ensuring the model has not added, removed, or altered any words during the capitalization and punctuation recovery process.

Step 5. Since our end goal is to build a dataset with timestamps, we need to run the transcripts through an alignment model. However, available alignment models require text to be in standard spoken form, which is not the case for our filtered transcripts. Therefore, we need to convert numbers in the transcripts to their text form. To accomplish this, we use a pre-trained model for text normalization (Vu et al., 2025).[7] We compare the model's input and output to obtain the number-to-text mapping and replace the numbers in the transcripts with their word forms.

Step 6. With the transcripts in standard spoken form, we can now run the alignment model. We use whisperx (Bain et al., 2023) as the alignment framework, coupled with a wav2vec2-based model fine-tuned on Vietnamese (Duy Khanh, 2022), to generate word-level timestamps for the transcripts. To be compatible with the Whisper model, timestamp values obtained from whisperx are quantized into 20 ms (0.02 s) intervals—for example, <|0.00|>, <|0.02|>, ..., <|30.00|>—which serve as textual timestamp tokens during training.

Step 7. Converting the spoken textual representations of numbers back to the standard numerical form can be done by taking the number-to-text mapping from Step 5 and switching the texts with their corresponding numbers. We then take the starting timestamp of the text's first word and the ending timestamp of the text's last word as the starting and ending timestamps of the numerical form.

Discussion. The final result is a 502.67-hour high quality dataset (PhoASR). Table 2 shows the size of the datasets at the start, after sampling, and after intensive filtering and refining through the pipeline. We can observe substantial data reduction through our pipeline, with about 502.67 / 809.07 $\simeq$ 62% of the sampled data being retained after all processing steps. This substantial reduction highlights the prevalence of low-quality samples in the original datasets. Notably, medical domain data (VietMed-L) showed the highest reduction rate, retaining only 4.20 / 15.93 $\simeq$ 26% of its sampled size, despite the authors' claim of a manual verification process. In contrast, read speech datasets like VIVOS and CMV-vi maintained over 90% of their data, indicating their superior initial quality. This is expected, as in VIVOS and CMV-vi, the audio is recorded with people reading from prepared scripts, while in the rest of the datasets, the transcripts are generated by ASR models and thus are more prone to errors.

To investigate the balance of the dataset, we also add region information to each sample. Vietnamese has three main regional accents: Northern, Central, and Southern. We first train a dialect classifier using samples from raw datasets that contain region metadata: Vietnam-Celeb, ViMD, and LSVSC (389 hours in total). The wav2vec2-base-vi(Nguyen, 2021) model is fine-tuned on this data for 10 epochs and achieves 90%

[7] https://huggingface.co/thivux/PhoTextNormalization

Region	Training	Validation	Test	Total
North	266.17	11.48	8.52	286.17
South	138.09	4.01	4.37	146.47
Central	64.34	3.21	2.48	70.03
Total	468.60	18.70	15.37	502.67

Table 3: Our dataset distribution by region and split. From this point onward, **'64h' and '469h' refer to 64.34 hours and 468.6 hours of audio**, respectively.

accuracy. This classifier is then used to predict the regional accent for all samples in our dataset. The results in Table 3 show that Northern accent is the most dominant, followed by the Central accent, while the Southern accent is the least represented.

To assess the impact of our pipeline, we manually evaluated the quality of the original test set versus our refined PhoASR test set (containing 15.37 hours of audio). We sampled 20 examples from each component dataset, listened to the audio, created ground-truth transcripts, and compared them with the corresponding transcripts in the test sets. This manual check reveals that the average WER for the original test set is 2.73%, with VietMed-L having the highest WER at 14.89%, while the refined test has almost perfect transcripts, with an average WER of only **0.23**%. Therefore, we use the refined test set for evaluation in our experiments.

4.2 PhoASR-3100h Dataset Creation

In addition to our high-quality PhoASR dataset, we construct a larger mixed dataset to investigate the benefits of combining rigorous processing with increased scale. This mixed dataset is created by applying a minimal processing—defined as initial cleaning (Step 1) followed by punctuation and capitalization recovery (Step 4)—to the 3171 hours of raw training data and then merging the result with our PhoASR training set of 469h data.

In particular, recall that a portion of the Vietnam-Celeb audios already have transcripts generated in Step 2 of our PhoASR-specific pipeline. For the remaining Vietnam-Celeb audios that lack transcripts, we use PhoWhisper-large to generate them. We then apply the aforementioned minimal processing to the entire 3171-hour training data. For any audio files that also appear in our PhoASR training set, we replace their original or generated transcripts with the corresponding high-quality timestamped versions from the PhoASR training set. This process yields the PhoASR-3100h training set, which combines the precision of our rigorous pipeline with the lexical diversity of larger-scale data.

Table 2 presents the statistics of the PhoASR-3100h dataset. It represents nearly the entire training corpus (3100 out of 3171 hours), with a small amount of data excluded during the cleaning step.

5 Experiments

5.1 Setup

Text Accuracy: To measure the performance of text prediction, we use the Word Error Rate (WER) metric. WER is a standard metric for ASR that calculates the number of substitutions, deletions, and insertions between the predicted and reference texts, normalized by the total number of words in the reference text. A lower WER indicates better performance. We report two variants of WER (Hugging Face, 2024):

- **Orthographic WER (O-WER)**: This is calculated using the raw, unnormalized text, preserving the original capitalization and punctuation. It offers a strict measure of the model's ability to produce well-formatted output.
- **Normalized WER (N-WER)**: Before calculating WER, both the predicted and reference texts are normalized by converting them to lowercase and removing all punctuation. This approach emphasizes core lexical accuracy, disregarding formatting differences.

Timestamp Accuracy: To evaluate the accuracy of word-level timestamps, we use the following metrics:

- **F_1-score**: This is the primary metric for evaluating timing accuracy, derived from True Positives (TP), False Positives (FP), and False Negatives (FN). A predicted word is considered a TP if it matches a reference word in content and their temporal overlap falls within a predefined collar. An FP refers to a predicted word with no corresponding reference, while an FN is a reference word that the model fails to predict.
- **mean Intersection over Union (mIoU)**: This metric assesses localization accuracy. For each predicted word that matches a reference word, the Intersection over Union (IoU) of their timestamps is computed. If no match is found, a score of 0 is assigned. The final metric is the average IoU across all matched words, with higher values indicating better performance.

	Dataset	North	Central	South	Overall
O-WER	64h-north	19.89	26.16	20.85	21.08
	64h-central	21.66	25.40	20.68	22.16
	64h-south	21.57	26.16	19.17	21.72
	192h	**15.88**	**19.86**	**15.81**	**16.53**
N-WER	64h-north	16.32	22.46	17.22	17.54
	64h-central	17.62	22.10	16.99	18.27
	64h-south	17.69	22.60	15.77	18.02
	192h	**12.18**	**16.56**	**12.36**	**12.97**

Table 4: WER scores (%) for regional subsets when fine-tuning whisper-small with 40 training epochs.

Ratio	O-WER ↓	N-WER ↓	F1 ↑	IoU ↑
0%	**19.21**	**15.79**	75.83	49.26
25%	22.66	19.23	76.42	49.65
50%	20.40	16.84	76.72	51.25
75%	22.16	18.38	79.45	54.84
100%	21.88	17.59	**80.61**	**57.21**

Table 5: Performance results when fine-tuning whisper-small with different timestamp ratios (%) on 100h training subset for 40 training epochs.

Implementation: Details of the implementation are provided in Appendix A.

5.2 Impact of Regional Accents

We investigate the impact of regional accents on ASR performance. To ensure a balanced comparison, we sample 64h of training data–the total available from the Central region–from each of the Northern and Southern regions, resulting in a combined dataset of 64 × 3 = 192 hours (192h). We fine-tune the whisper-small (Radford et al., 2023) model for 40 epochs on the 192-hour dataset, as well as on each of the 64-hour regional subsets.

We then evaluate each model on three regional subsets of the test set. As shown in Table 4, the model trained on the combined 192-hour dataset achieves the best overall performance, highlighting the benefits of training on diverse accents. In contrast, models trained on a single accent perform well on their matched region but are less effective on others. Furthermore, the Central accent is consistently the most difficult for models trained on other accents, which may indicate greater linguistic and phonetic divergence of that dialect from the others.

5.3 Timestamp Ratio

In this experiment, we fine-tune whisper-small to generate both timestamp and transcript tokens. Our goal is to determine the optimal proportion of timestamped data in the training set. To this end, we fine-tune the model on the 100h subset for 40 epochs using different timestamp ratios: 0%, 25%, 50%, 75%, and 100%. Here, the ratio refers to the proportion of timestamped data included in each training epoch.

Table 5 shows trade-offs between transcription and alignment performance across different timestamp ratios. While the 0% ratio (no timestamps) achieves the best transcription accuracy (lowest O-WER and N-WER), the 100% timestamp ratio delivers the highest timestamp alignment scores but at the cost of reduced transcription quality. The 50% timestamp ratio provides the optimal balance, maintaining competitive transcription and timestamp alignment performances. This suggests a clear trade-off in the multi-task learning setup: incorporating timestamp prediction improves alignment capabilities but can impair transcription accuracy when overemphasized.

5.4 Different Models

We compare Whisper with another leading ASR model, wav2vec2 (Baevski et al., 2020), by fine-tuning whisper-small and wav2vec2-xls-r-300m on our 469h training set for 40 epochs. In this setup, whisper-small is trained to generate both timestamp and transcript tokens using a 50% timestamp ratio, whereas wav2vec2-xls-r-300m is trained to generate transcript tokens only. This results in two fine-tuned models, PhoASR-whisper-small-469h and wav2vec2 (469h), respectively.

Table 6 shows that PhoASR-whisper-small-469h consistently outperforms wav2vec2 across all evaluation metrics. In particular, PhoASR-whisper-small-469h achieves an O-WER of 12.46% and an N-WER of 8.69%, which are substantially better than those of wav2vec2 (51.15% O-WER and 14.10% N-WER). Furthermore, fine-tuning yields a massive improvement over the pre-trained whisper-small baseline, reducing O-WER from 70.16% to 12.46% and N-WER from 64.07% to 8.69%, confirming the necessity of domain adaptation for Vietnamese.

A comparison between PhoWhisper-small, which was fine-tuned from whisper-small on a larger 844h dataset, and our model PhoASR-whisper-small-469h highlights the superiority of our high-quality dataset. Our model

Model	O-WER ↓	N-WER ↓	F1 ↑	IoU ↑
whisper-small (Pre-trained)	70.16	64.07	-	-
ChunkFormer-large-vi (25K)	32.43	**6.89**	-	-
PhoWhisper-small (844h)	33.90	8.97	-	-
wav2vec2 (469h)	51.15	14.10	-	-
PhoASR-whisper-small-469h	12.46	8.69	76.40	55.62
PhoASR-whisper-small-3100h	**11.70**	8.20	**83.68**	**57.39**

Table 6: Obtained scores for different models: "PhoASR-whisper-small-469h" trained for 40 epochs; "PhoASR-whisper-small-3100h" trained for 15 epochs. See Appendix B for results on each component dataset. ChunkFormer was trained on an internal non-public dataset of *25K* audio hours for *200* epochs. PhoWhisper was trained on a dataset of *844* audio hours, in which *586* hours are private data.

achieves a better N-WER (8.69% vs. 8.97%) and a notably lower O-WER (12.46% vs. 33.90%). This result underscores the importance of data quality over quantity; despite being trained on nearly half the data, our model produces more accurate and better-formatted transcripts, demonstrating the effectiveness of our data processing pipeline.

5.5 Scaling with PhoASR-3100h

While Section 5.4 demonstrates that rigorous processing allows smaller datasets to outperform larger noisy ones, massive scale remains crucial for covering long-tail vocabulary and diverse acoustic conditions. To combine the precision of our high-quality data with the diversity of large-scale pre-training, we evaluate the scaling potential by training `whisper-small` on our PhoASR-3100h mixed dataset (described in Section 4.2) for 15 epochs, resulting in `PhoASR-whisper-small-3100h`.

Table 6 presents the evaluation results. As expected, adding more training data gives a boost to the performance. `PhoASR-whisper-small-3100h` improves across all metrics compared to `PhoASR-whisper-small-469h`. For example, it reduces the O-WER from 12.46% to 11.70% and improves timestamp prediction score F1 from 76.40% to 83.68%.

Compared to other baseline models, `PhoASR-whisper-small-3100h` demonstrates a strong balance between lexical accuracy and orthographic correctness. For instance, `ChunkFormer-large-vi` achieves a lower N-WER of 6.89%, benefiting from its massive 25,000-hour training set. However, its O-WER of 32.43% is notably higher than `PhoASR-whisper-small-3100h`'s 11.70%, indicating poor performance on punctuation and capitalization. In practical applications where transcripts are expected to be immediately usable, such as in automatic subtitling or meeting transcription, O-WER is a more critical metric than N-WER. A lower O-WER signifies that the output is well-formatted and requires minimal to no post-editing, making `PhoASR-whisper-small-3100h` a more suitable choice for such use cases.

6 Pipeline Component Analysis

We analyze the contribution of specific pipeline stages to the final model performance:

Filtering & Cleaning (Steps 1-3): Our rigorous filtering allows a smaller, cleaner dataset to outperform a larger, noisier one. As shown in Table 6, `PhoASR-whisper-small-469h` achieves a lower N-WER (8.69%) compared to `PhoWhisper-small` (8.97%), despite being trained on nearly half the amount of data (469h vs. 844h). This confirms that quality filtering is more effective than raw data quantity for lexical accuracy.

Punctuation & Capitalization (Step 4): The impact of this step is evident in the O-WER metric. `PhoWhisper-small` exhibits a high O-WER of 33.90%, while `PhoASR-whisper-small-469h` achieves 12.46%. This significant gap highlights the necessity of explicit restoration (Step 4) for producing ready-to-use transcripts with correct formatting.

Timestamp Alignment (Steps 5-7): The results in Section 5.3 (Table 5) serve as an ablation for the alignment steps. Increasing the timestamp ratio from 0% to 100% directly correlates with improved alignment performance (F1 and mIoU), confirming that our forced alignment process effectively imparts fine-grained temporal understanding to the model.

7 Conclusion

We present a pipeline for creating high-quality ASR datasets from noisy, open-source audio. Our method produces clean transcripts with reliable timestamps, punctuation, and capitalization, which eliminates the need for separate post-processing tools. To demonstrate its effectiveness, we construct a 500-hour Vietnamese corpus and demonstrate its potential as a strong foundation for both fine-tuning and benchmarking ASR models. While this pipeline was applied to Vietnamese, it is adaptable and practical for enhancing speech recognition across a broader range of languages.

Limitations

While our pipeline substantially improves data quality, the filtering process itself might introduce bias. By selectively retaining samples that are cleanly processed by our toolchain, we may unintentionally favor certain acoustic environments, speaking styles, or accents that upstream models handle more effectively. As a result, the final dataset—though high in quality—may be less representative of the full diversity of real-world speech. Our experiments with the PhoASR-3100h dataset suggest that combining high-quality filtered data with larger, minimally processed corpora may be the most effective strategy for training robust and generalizable ASR models.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of LREC*, pages 4211–4215.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of NeurIPS*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *Proceedings of INTERSPEECH*, pages 4489–4493.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proceedings of INTERSPEECH*, pages 3670–3674.

Nguyen Van Dinh, Thanh Chi Dang, Luan Thanh Nguyen, and Kiet Van Nguyen. 2024. Multi-Dialect Vietnamese: Task, Dataset, Baseline Models and Challenges. In *Proceedings of EMNLP*, pages 7476–7498.

Le Duy Khanh. 2022. Finetune Wav2vec 2.0 For Vietnamese Speech Recognition.

Daniel Galvez, Greg Diamos, Juan Manuel Ciro Torres, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage. In *Proceedings of NeurIPS Datasets and Benchmarks Track*.

Thinh Le Phuoc Gia, Tuan Pham Minh, Hung Nguyen Quoc, Trung Nguyen Quoc, and Vinh Truong Hoang. 2024. viVoice: Enabling Vietnamese Multi-Speaker Speech Synthesis.

Hugging Face. 2024. Unit 5: Evaluation metrics for asr. `https://huggingface.co/learn/audio-course/en/chapter5/evaluation`. Accessed: 2024-07-29.

Khanh Le, Tuan Vu Ho, Dung Tran, and Duc Thanh Chau. 2025. ChunkFormer: Masked Chunking Conformer For Long-Form Speech Transcription. In *Proceedings of ICASSP*.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. 2024. PhoWhisper: Automatic Speech Recognition for Vietnamese. In *Proceedings of the ICLR 2024 Tiny Papers track*.

Khai Le-Duc. 2024. VietMed: A Dataset and Benchmark for Automatic Speech Recognition of Vietnamese in the Medical Domain. In *Proceedings of LREC-COLING*, pages 17365–17370.

Hieu-Thi Luong and Hai-Quan Vu. 2016. A non-expert Kaldi recipe for Vietnamese Speech Recognition System. In *Proceedings of WLSI/OIAF4HLT*, pages 51–55.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of INTERSPEECH*, pages 498–502.

Vu Le Binh Nguyen. 2021. Wav2Vec2 Base Vietnamese Model. `https://huggingface.co/nguyenvulebinh/wav2vec2-base-vi`.

Pham Quang Nhut, Duong Pham Hoang Anh, and Nguyen Vinh Tiep. 2024. VSV-1100: Vietnamese social voice dataset.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of ICASSP*, pages 5206–5210.

Douglas B. Paul and Janet M. Baker. 1992. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of HLT*, pages 357–362.

Anh Pham, Khanh Linh Tran, Linh Nguyen, Thanh Duy Cao, Phuc Phan, and Duong A. Nguyen. 2024. Bud500: A Comprehensive Vietnamese ASR Dataset.

Viet Thanh Pham, Xuan Thai Hoa Nguyen, Vu Hoang, and Thi Thu Trang Nguyen. 2023. Vietnam-Celeb: A large-scale dataset for Vietnamese speaker recognition. In *Proceedings of INTERSPEECH*, pages 1918–1922.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*.

Duc Chung Tran. 2020. FPT Open Speech Dataset (FOSD) - Vietnamese.

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):977.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of INTERSPEECH*.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2017. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Thi Vu, Linh The Nguyen, and Dat Quoc Nguyen. 2025. Zero-Shot Text-to-Speech for Vietnamese. In *Proceedings of ACL*.

A Implementation details

For our experiments, we fine-tuned two publicly available pre-trained models: `whisper-small`[8] and `wav2vec2-xls-r-300m`.[9]

The fine-tuning settings for `whisper-small` are kept consistent across all experiments. We use a peak learning rate of 1.25×10^{-5} with an AdamW optimizer and a linear learning rate scheduler. For `wav2vec2-xls-r-300m`, we use a peak learning rate of 1×10^{-4} and train it with the Connectionist Temporal Classification (CTC) loss function. To improve generalization, we apply SpecAugment with a time masking probability of 0.75, a feature masking probability of 0.25, and a feature mask length of 64.

All models are trained for 40 epochs, with the first 5,000 steps used for warm-up, except for `whisper-small` trained on the PhoASR-3100h dataset for 15 epochs due to resource constraints. All experiments are conducted on a system with 4 NVIDIA A100 40GB GPUs. We use a per-device batch size of 4 and 4 gradient accumulation steps, which result in an effective global batch size of 64. The best-performing checkpoint for each model is selected based on the lowest N-WER achieved on the PhoASR validation set. We then employ the selected checkpoint to report final performance results on the PhoASR test set.

B Detailed Evaluation Results

This section contains the detailed evaluation results of different models on each dataset of our test set.

Table 7 shows the performance on our test set. `PhoASR-whisper-small-3100h` consistently achieves the best O-WER, F1 and mIoU, demonstrating the benefits of a larger training corpus. Notably, the `VLSP2020` dataset proves to be the most challenging for all models. Furthermore, while `wav2vec2` achieves competitive N-WER scores, its O-WER is considerably large, suggesting a weaker performance in predicting correct capitalization and punctuation compared to the `whisper`-based models.

[8] https://huggingface.co/openai/whisper-small

[9] https://huggingface.co/facebook/wav2vec2-xls-r-300m

Metric	Model (Training Hours)	BUD500	VLSP2020	ViMD	LSVSC	FOSD	VietMed-L	VIVOS	CMV-vi-14	Overall
O-WER	ChunkFormer-large-vi (25K)	18.22	38.25	13.13	14.16	18.9	14.63	15.38	22.44	32.43
	PhoWhisper-small (844h)	19.98	39.77	13.95	15.32	21.0	16.58	16.54	23.87	33.90
	wav2vec2 (469h)	21.76	60.89	17.44	19.45	28.79	22.23	33.45	31.04	51.15
	PhoASR-whisper-small-469h	8.27	14.25	8.75	6.72	7.73	**14.41**	5.92	9.18	12.46
	PhoASR-whisper-small-3100h	**4.22**	**13.64**	**8.08**	**6.4**	**5.53**	15.46	**3.06**	**7.02**	**11.7**
N-WER	ChunkFormer-large-vi (25K)	3.03	**8.71**	4.46	**1.19**	**0.79**	**5.95**	**1.37**	**2.97**	**6.89**
	PhoWhisper-small (844h)	6.59	10.95	5.55	2.64	3.36	7.9	2.69	4.81	8.97
	wav2vec2 (469h)	2.42	17.51	7.34	2.84	4.51	13.93	4.61	7.51	14.10
	PhoASR-whisper-small-469h	3.97	10.38	4.64	2.64	4.58	9.51	4.33	6.34	8.69
	PhoASR-whisper-small-3100h	**0.92**	10.09	**3.99**	2.49	2.35	10.2	1.92	4.35	8.2
F1	PhoASR-whisper-small-469h	77.31	73.55	73.77	83.10	75.69	74.08	76.25	67.09	76.40
	PhoASR-whisper-small-3100h	**85.71**	**82.27**	**79.46**	**89.65**	**81.05**	**77.60**	**78.70**	**72.36**	**83.68**
mIoU	PhoASR-whisper-small-469h	**48.90**	58.06	45.51	51.32	49.63	47.90	48.47	46.30	55.62
	PhoASR-whisper-small-3100h	48.03	**58.64**	**50.85**	**57.05**	**54.54**	**48.85**	**52.47**	**48.69**	**57.39**

Table 7: O-WER (%), N-WER (%), F1 (%), IoU (%) for different models on each dataset within our refined test set.