VLSP 2025 ASR-SER
From Data Exploration to Model Training: A Strategic Approach

Nhat-Minh Nguyen
Cake By VPBank, Viet Nam
minh.nguyen12@cake.vn

Ngoan Pham Van
Cake By VPBank, Viet Nam
ngoan.pham@cake.vn

Abstract

This paper presents a strategic, data-driven approach to the Vietnamese Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) tasks at the VLSP 2025 challenge. Our approach focuses on the strategic adaptation of large pre-trained models to address the complexities of Vietnamese speech, such as dialectal diversity and inconsistent data. Through a progressive, two-phase training schedule that blends extensive general domain data with focused in-domain adaptation, we optimized the OpenAI Whisper Small model for ASR. We used a hybrid strategy for SER, extracting strong acoustic features for a lightweight downstream classifier using the Emotion2Vec model. Enhanced by thorough data pre-processing and analysis, our system produced extremely competitive results, placing fourth with an accuracy of 79.5% for SER and a Word Error Rate (WER) of 19.12% for ASR. These findings demonstrate the effectiveness of a data-centric methodology in adapting foundational models for robust Vietnamese speech processing.

1 Introduction

Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) for Vietnamese remain challenging tasks due to limited labeled resources, strong dialectal and pronunciation variability, and inconsistent annotation conventions across public dataset. Robust systems for these tasks are increasingly important for real-world applications such as voice-driven assistants, call-center analytics, and affect-aware human–computer interaction in Vietnamese. In this work, we present a practical, data-driven approach that combines a modern end-to-end ASR backbone with a feature-based SER pipeline, together with careful Exploratory Data Analysis (EDA) and targeted preprocessing to mitigate noisy labels and domain mismatch.

We chose the Whisper Small (Radford et al., 2023) model as the ASR backbone. Whisper Small (Radford et al., 2023) is a pretrained end-to-end acoustic model that supports Vietnamese and the largest model of Whisper's series has demonstrated very low average WER on widely used English test-sets, suggesting strong modeling capacity that can be leveraged when adapted to Vietnamese data. Motivated by this potential, we fine-tuned the entire Whisper Small (Radford et al., 2023) model from its pre-trained initialization using Vietnamese speech dataset, so the model internalizes both generic acoustic patterns and language-specific characteristics.

For SER, we adopted Emotion2Vec_plus_large (Ma et al., 2023) as a feature extractor. The publicly released base of this model produces nine emotion classes with competitive accuracy on public benchmarks; rather than using the model's original classifier head, we used Emotion2Vec_plus_large (Ma et al., 2023) to extract robust audio embeddings and then train a compact neural classifier to predict the two contest-specific emotion labels. This design decouples representation learning (handled by a large pretrained encoder) from the low-parameter task-specific classifier, which is beneficial when labeled emotion data are scarce.

A major part of our methodology focused on data analysis and cleaning. We analyzed the organizer-provided public test domain and the available training sets (VLSP 2023 dataset plus public Vietnamese ASR dataset such as viVoice (Capleaf, 2024), Viet_Bud500 (Anh Pham, 2024), phoaudiobook (Vu et al., 2025)) and 28k_vietnamese_voice_augmented_of_VinBigData (natmin322, 2023). During EDA, we identified numerous problematic samples: incorrect or inconsistent transcriptions, non-standard tokenization (mathematical notations, mixed numeric symbols, inconsistent punctuation), and

other annotation artifacts that introduce label noise. To reduce this source of error, we applied a set of conservative preprocessing steps to remove or correct samples likely to confuse the models, prioritizing data quality over raw quantity.

Our training strategy was tailored to each task. For ASR, we perform full-model fine-tuning of Whisper Small (Radford et al., 2023) in two-phase: (1) a broad-phase training on large and diverse Vietnamese dataset to capture general acoustic–linguistic patterns, followed by (2) a domain-focused phase where the model is further adapted with data whose characteristics closely match the public test set. For SER, we trained the compact classifier on the organizer's labeled SER data and run extended training (multiple epochs) to stabilize convergence and improve generalization for Vietnamese.

The contributions of this report are threefold:

- A reproducible pipeline that adapts a pretrained Whisper Small model (Radford et al., 2023) for Vietnamese ASR through a two-phase fine-tuning schedule that addresses domain mismatch.
- A hybrid SER approach that uses Emotion2Vec_plus_large (Ma et al., 2023) as a fixed feature extractor and a lightweight neural classifier to map rich embeddings to the contest's two emotion classes.
- A data-quality driven preprocessing and EDA methodology that documents common labeling issues in available Vietnamese speech dataset and shows how conservative cleaning improves the reliability of downstream training.

2 Related Work

2.1 ASR Models for Vietnamese Speech Recognition

OpenAI's Whisper (Radford et al., 2023) is an end-to-end ASR model based on a sequence-to-sequence Transformer (encoder–decoder) architecture (Vaswani et al., 2017). Audio is fed as 80-channel log-Mel spectrograms (in 30-second chunks) through a small convolutional front-end followed by stacked encoder Transformer blocks, and a decoder predicts text tokens. During training, special tokens are used to specify tasks: for example, tokens indicate the language, whether to transcribe or translate, and where timestamps should be placed.

Whisper was pre-trained on an unprecedented 680,000 hours of multilingual, multitask data (collected from the web) spanning 99 languages. This large and diverse dataset (including roughly 1.7K hours of Vietnamese speech) makes Whisper highly robust to accents, background noise, and domain-specific terminology, and enables it to perform multilingual speech transcription and speech-to-text translation without per-language fine-tuning. Indeed, Whisper (Radford et al., 2023) achieves near-human ASR accuracy on English benchmarks while still generalizing better out-of-distribution than specialized models. In multilingual benchmarks (e.g., Multilingual LibriSpeech (Pratap et al., 2020)), large Whisper models outperform or match prior state-of-the-art multilingual systems.

At the same time, practitioners have noted that for very low-resource languages like Vietnamese, further adaptation can improve performance. PhoWhisper (Le et al., 2024), released in 2024, achieved state-of-the-art results by fine-tuning Whisper in 844 hours of diverse Vietnamese speech, reaching WER 13.75% in VLSP 2020 Task-1 with the large version and 15.93% with the small version. Similarly, recent comparative work (Song et al., 2024) shows that while Whisper achieves a WER of around 17. 9% in Vietnamese without fine tuning, LoRA adaptation reduces this to 16.1%. Moreover, LLM-based ASR yields further improvements, with relative WER reductions of 35.7% over Whisper and 12.8% over Whisper-finetuned models on average across low-resource languages. These findings underscore both the robustness of Whisper as a foundation and the importance of domain-specific adaptation. Building on this line of work, we leverage Whisper as the backbone for Vietnamese ASR, applying targeted fine-tuning to better capture the phonetic and tonal characteristics of the language.

2.2 Speech Emotion Recognition Advances

Speech emotion recognition (SER) has evolved from traditional methods based on hand-made acoustic characteristics, such as MFCCs and filter banks, to deep learning models that learn representations directly from raw audio (El Ayadi et al., 2011). More recently, self-supervised learning (SSL) has emerged as a powerful paradigm, enabling models to capture generalizable

speech representations without requiring large labeled datasets.

Among SSL approaches, Emotion2Vec (Ma et al., 2023) represents a breakthrough in universal emotion representation. It adopts a teacher–student framework, where both networks share the same architecture (a CNN-based feature extractor followed by a Transformer encoder (Vaswani et al., 2017)). During pre-training, the student receives masked audio frame features and learns to reconstruct both frame-level content and an utterance-level embedding, while the teacher provides online soft targets. This dual-level self-supervision reflects the intuition that global and local cues are crucial for emotion. The resulting fixed-dimensional embeddings (e.g., 768-d) encode rich affective information.

Emotion2Vec has demonstrated consistent superiority over general-purpose SSL models such as WavLM and HuBERT. On benchmarks like IEMOCAP, it surpasses state-of-the-art emotion-specific systems, and in the recent EmoBox benchmark (Ma et al., 2024), which evaluated 10 pre-trained models across 32 emotion datasets in 14 languages, it consistently ranked at the top, underscoring its strong cross-lingual generalization. Furthermore, evaluations on 13 SER datasets covering 10 languages show its broad applicability, with additional success on related tasks such as song emotion recognition, conversational emotion prediction, and speech sentiment analysis.

The Emotion2Vec_plus_large variant we adopted extends the base model with iterative fine-tuning on 40,000 hours of pseudo-labeled emotional speech. It supports nine emotion classes, angry, disgust, fear, happy, neutral, sad, surprised, other and unknown (Ma et al., 2023). This model has proven particularly effective in cross-lingual transfer, making it especially suitable for low-resource contexts like Vietnamese SER, where annotated emotional speech data remains scarce.

3 Methodology

3.1 Exploratory Data Analysis

We assembled several Vietnamese speech dataset to train our ASR model. All audio was standardized to a 16 kHz sampling rate. In total we used four datasets: the VLSP 2023 ASR dataset, phoaudiobook (Vu et al., 2025), Viet_Bud500 (Anh Pham, 2024), viVoice (Capleaf, 2024) and 28k_vietnamese_voice_augmented_of_VinBigData (natmin322, 2023). The dataset we trained the SER model is the VLSP 2023 dataset provided by the organizers. Below we summarize each dataset's source, size, and key characteristics.

VLSP 2023 ASR Dataset

This dataset (released by the VLSP 2023 challenge organizers) contains 56-hours Vietnamese speech from diverse media (e.g. TV series, entertainment videos, social media clips, narrated lectures, news broadcasts, etc.). Its domain and characteristic closely matches that of the public test set.

However, the provided transcripts are not fully clean: they include non-alphabetic symbols (digits, math symbols) and inconsistent spelling conventions (such as mixed "i"/"y"). In other words, VLSP 2023 is not a "clean" dataset, and its labels require extensive preprocessing to remove spurious characters and normalize the text.

phoaudiobook ASR Dataset

phoaudiobook (Vu et al., 2025) is a high-quality 941-hour Vietnamese speech dataset derived from professionally-recorded audiobooks. The audio comes from studio-quality readings with 735 distinct speakers, yielding very clear speech. Transcriptions have been thoroughly validated: they applied two ASR models (Whisper-large-v3 and PhoWhisper-large) to each segment and retained only those samples where their outputs exactly agreed. The result is a dataset with very accurate, normalized transcripts (numbers converted to words, consistent punctuation).

Because phoaudiobook originates from literary narration, the speaking style is primarily formal and controlled; it lacks conversational or emotional variation.

Viet_Bud500 ASR Dataset

The Viet_Bud500 (Anh Pham, 2024) dataset is a 500-hour Vietnamese ASR dataset designed for diversity. It covers a broad spectrum of topics, including podcasts, travel, literature, food, and more – and spans accents from northern, central, and southern Vietnam. The data were collected from freely available audio sources (online media) and uniformly sampled at 16 kHz.

Importantly, Viet_Bud500 transcripts were manually annotated and carefully checked, so the labeling quality is high (comparable to phoaudiobook). This dataset adds conversational and topical variety (e.g. informal speech, modern

content) that is missing from audiobooks, although it is still mostly single-speaker and relatively clean speech.

viVoice ASR Dataset

viVoice (Capleaf, 2024) is a very large (≈1016.97-hour) multi-speaker speech dataset assembled from YouTube videos. It consists of ∼888k utterances extracted from 186 YouTube channels, covering a wide range of topics and recording conditions. All audio clips have been cleaned (noise and music removed) and trimmed to sentence-length.

However, the transcripts in viVoice are less reliable. They are provided in raw form (no text normalization), so they often contain extraneous characters or numbers, and even entire phrases in English. In fact, a manual sample found an estimated 1.8% of utterances with transcription errors (insertions, deletions, or substitutions). In summary, viVoice greatly increases the volume and diversity of data, but its transcripts require careful post-processing (removing special symbols, normalizing words) before training.

VinBigData Augmented Voice Dataset 28k samples

This dataset was released on Huggingface as 28k_vietnamese_voice_augmented_of_VinBigData (natmin322, 2023) , consists of approximately 28,000 Vietnamese audio samples spanning multiple domains such as news reports, movies, and other broadcast media.

A distinctive feature of this dataset is the inclusion of singing voice recordings, which enriches the acoustic diversity compared to standard speech-only datasets. While the transcripts are generally accurate, some utterances contain occasional English words inserted into Vietnamese sentences, which introduces minor code-switching noise.

Overall, the dataset shares many domain characteristics with the VLSP 2023 ASR dataset and was therefore considered valuable for enhancing in-domain robustness of the ASR model.

VLSP 2023 SER Dataset

The dataset we trained the SER model is the VLSP 2023 dataset provided by the organizers. We used the entire organizer's data for training without any preprocessing or filtering.

3.1.1 Data Pre-processing

We applied the following pre-processing steps to improve data quality and align training data with the characteristics of the public test set:

- **Empty or invalid transcripts:** removed samples with empty transcripts or transcripts not containing Latin alphabetic characters; discarded transcripts containing digits. For samples with transcripts containing dates, months, years, or monetary values, we also removed them without converting them to readable text. Because we could not determine their specific reading, for example "nghìn" or "ngàn".
- **Normalization:** converted all transcripts to lowercase and removed all punctuation.
- **Non-Vietnamese symbols:** excluded samples containing non-Latin characters (e.g., mathematical operators, special symbols).
- **Tone marks filtering:** removed samples whose transcripts contained no Vietnamese tone-marked characters.
- **Duration filtering:** discarded samples longer than 15 seconds or with transcripts containing more than 85 words, consistent with the distribution of the public test set.

3.2 Model Training

3.2.1 ASR Model

The model we used for the ASR task was Whisper Small (Radford et al., 2023). We started with the pretrained model and then fine-tuned the entire network on the Vietnamese datasets prepared in Section 3.1. The training process was divided into two-phase: in the first phase, we trained the model on a large-scale general Vietnamese ASR dataset; in the second phase, we fine-tuned it on Vietnamese ASR data with characteristics and domains similar to the public test set, aiming to optimize the model's performance on the test data.

Training Phase 1 - General Training

First training: The Whisper Small model was trained on the combination of the first 600,000 samples from viVoice and the first 600,000 samples from the phoaudiobook datasets. Both datasets provide large-scale and diverse speech samples. Training was conducted for 1 epoch with a batch size of 16.

Second training: The model from the first training was further trained on Viet_Bud500 and 28k Vietnamese Voice Augmented of VinBigData.

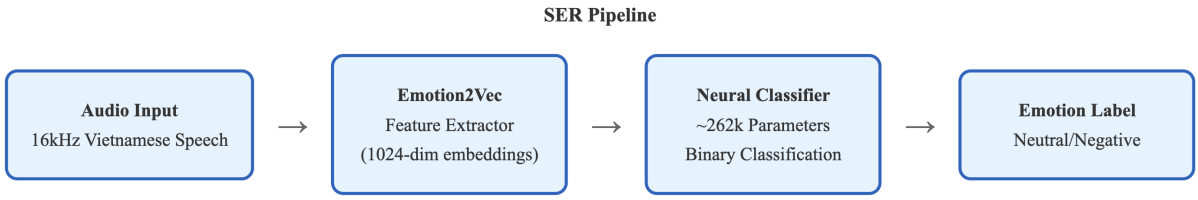

Figure 1: Our Speech Emotion Recognition Model

These dataset share acoustic and domain properties with the VLSP 2023 dataset and the public test set (e.g., news, films, conversational media, and broadcast speech). Training was performed for 1 epoch with a batch size of 16.

Dividing the training into different phases allowed us to more clearly observe the model's improvements and verify the effectiveness of our training strategy.

Training Phase 2 - In-domain Training

Following the general adaptation, the model was further fine-tuned using the VLSP 2023 ASR dataset to capture the distributional characteristics of the public test set. Training was conducted for 2 epochs, with a batch size of 16 in the first epoch and 64 in the second epoch.

This progressive training schedule ensured that the model first internalized broad Vietnamese phonetic and lexical knowledge, then specialized toward the target domains, and finally adapted closely to the characteristics of the VLSP 2023 test distribution.

Detail Training Configs

- batch_size = 16
- learning_rate = 1e-5
- lr_scheduler_type = "linear"
- optim = adamw_torch
- fp16 = True
- generation_max_length = 225

Predictions Post-processing

To further enhance the accuracy of the ASR system, we applied a rule-based post-processing module to the model's predictions. The procedure consisted of several steps:

- **Normalization:** All transcripts were converted to lowercase and stripped of punctuation marks.
- **Error-driven mapping:** We reviewed the model's predictions on the public test set and designed optimized mapping rules in addition to those provided by the organizers (only expanded the Vietnamese phonetic variants of the existing English words and did not introduce any new words independently). For example:
  "cô viết" → "covid"
- **Refinement of organizer-provided mappings:** Some mappings supplied by the organizers were not sufficiently general to handle foreign word transcriptions, especially those involving ambiguous consonants such as "s" and "x". We introduced supplementary mappings to address these cases, e.g.:
  "olympic" → "ô lim pích", "ô lim pít", "ô lim píc"
  "triton" → "trai tần", "trai tờn"
- **Reliability constraint:** To avoid unintended corrections, a mapping was only applied when the phonetic transcription matched a predefined key in at least two or more words.
- **Final normalization:** A conversion of "y" to "i" was performed according to the official rules provided by the organizers.

This post-processing step ensured more robust handling of frequent phonetic ambiguities and domain-specific transcriptions, resulting in improved final recognition accuracy.

3.2.2 SER Model

We leveraged the Emotion2Vec_plus_large model (Ma et al., 2023) as a feature extractor for binary speech emotion recognition (SER) task. The task was formulated as a two-class classification problem: *neutral* vs *negative* (where *negative* aggregates emotions such as angry, sad, fearful, and disgusted), as shown in Figure 1. The training procedure is summarized as follows:

- **Pretrained model:** Emotion2Vec_plus_large (Ma et al., 2023), used for extracting utterance-level embeddings.
- **Binary classification head:**
  Input: 1024-dimensional embeddings.
  Architecture: $\text{Dropout}\ (0.1) \rightarrow \text{Linear}\ (1024 \rightarrow 256) \rightarrow \text{ReLU} \rightarrow \text{Dropout}\ (0.1) \rightarrow \text{Linear}\ (256 \rightarrow 2)$
- **Dataset:** VLSP 2023 SER dataset
- **Data split:** We randomly sampled 5000 samples from the VLSP 2023 SER dataset for validation set, the remaining samples are used to train model.
- **Training configuration:**
  Batch size: 64
  Optimizer: AdamW with learning rate 1×10^{-5}, weight decay 0.01
  Loss function: Weighted cross-entropy to mitigate class imbalance
  Learning rate scheduler: ReduceLROnPlateau (patience = 3, factor = 0.5)
  Epochs: 50
- **Training strategy:**
  Utterance embeddings extracted from Emotion2Vec_plus_large were fed into the binary classifier.
  The classifier is trained end-to-end while the pre-trained model part is kept intact (only responsible for extracting input features for the classifier).
  Validation metrics (accuracy) were computed at the end of each epoch.
  The best model was selected based on validation accuracy.
- **Output:**
  The trained classifier weights and configuration were saved for downstream prediction.

This training setup enabled efficient adaptation of a large pretrained speech representation model to the VLSP 2023 SER task while addressing label imbalance and phonetic variability across the dataset.

4 Experiments

4.1 ASR Task

Detail Inference Configs

- batch_size = 1
- temperature = 0.0
- repetition_penalty = 1.0
- num_beams = 5
- no_repeat_ngram_size = 5

Table 1: Table comparing the error rate between the pretrained model and our fine-tuned model for the ASR task on the public and private test sets. Evaluation metric: Word Error Rate (WER), Unit: %.

	Public Test	**Private Test**
Pre-trained Model	35.92	35.81
Ours - Phase 1 (first training)	26.21	26.01
Ours - Phase 1 (second training)	21.42	21.35
Ours - Phase 2 (pre-mapping)	20.09	19.31
Ours - Phase 2 (post-processed)	**19.87**	**19.12**

The experimental results demonstrate a clear performance improvement when progressively fine-tuning the pretrained ASR model. Specifically:

- The pretrained model exhibited relatively high error rates, with WERs of 35.92% (public test) and 35.81% (private test).
- After first training in Phase 1 (general fine-tuning), WER was significantly reduced by nearly 10 percentage points on both test sets, showing that domain adaptation plays a crucial role in improving recognition accuracy.
- In second training of Phase 1, further general-domain training continued to yield improvements, achieving WERs of 21.42% and 21.35%, confirming the effectiveness of iterative fine-tuning for consolidating the model's generalization ability.
- Finally, by fine-tuning the model on the VLSP 2023 dataset in Phase 2, which closely matches the target domain, we achieved the best performance, with WER of 20.09% (public test) and 19.31% (private test). After applying the post-processing that we have

defined, the final WER results we got are 19.87% for public test and 19.12% for private test. This highlights the importance of domain-specific training data in bridging the gap between general speech recognition capabilities and task-specific requirements.

Overall, the step-by-step training strategy from general fine-tuning (Phase 1) to domain-specific adaptation (Phase 2) proved effective in enhancing model accuracy. The results indicate that carefully combining diverse datasets for general training with in-domain fine-tuning on VLSP 2023 dataset leads to a robust ASR model, significantly outperforming the pretrained baseline.

4.2 SER Task

In our experiments with the original Emotion2Vec_plus_large (Ma et al., 2023) model, we mapped the 9 classes into 2 classes, namely neutral and negative, as follows:

- Neutral, Happy, Unknown, Surprised, Other $\rightarrow$ Neutral
- Angry, Sad, Fearful, Disgusted $\rightarrow$ Negative

Table 2: Table comparing the accuracy between the pretrained model and our fine-tuned model for the SER task on the public and private test sets. Evaluation metric: Accuracy, Unit: %.

	Public Test	**Private Test**
Pre-trained Model	71.08	73.45
Our model	**79.18**	**79.5**

The pretrained model achieved accuracies of 71.08% (public test) and 73.45% (private test). After fine-tuning on the target datasets, our model significantly improved performance to 79.18% (public test) and 79.5% (private test).

The improvements of +8.1% (public) and +6.05% (private) demonstrate the effectiveness of fine-tuning. These results confirm that fine-tuning enhances the model's ability to capture emotion-related features, leading to more robust performance across both public and private evaluations.

5 Conclusion

This technical report presents a practical, data-driven approach to tackling the VLSP 2025 ASR and SER challenges. We achieved positive results and a 4th place ranking by combining careful data analysis, targeted preprocessing, and a strategic, multi-phase training methodology.

For the Automatic Speech Recognition (ASR) task, we fine-tuned the Whisper Small model in a two-phase process. The first phase involved broad training on large Vietnamese datasets to capture general acoustic and linguistic patterns, while the second phase adapted the model to the specific domain of the public test set. This progressive training schedule, along with meticulous preprocessing and rule-based post-processing, significantly reduced the Word Error Rate (WER) from 35.81% to 19.12% on the private test set. This outcome highlights the crucial role of domain-specific adaptation.

For the Speech Emotion Recognition (SER) task, we employed a hybrid approach using Emotion2Vec_plus_large as a fixed feature extractor and a lightweight neural classifier. This design is particularly beneficial when labeled emotion data is scarce, as it decouples representation learning from the task-specific classifier. Fine-tuning on the VLSP 2023 dataset improved our model's accuracy to 79.5% on the private test set, a notable increase of 6.05% over the pre-trained model. This demonstrates the effectiveness of fine-tuning for emotion-related features.

In summary, the methodologies presented, including a data-quality-driven approach, a two-phase fine-tuning schedule, and a hybrid model architecture, proved effective in developing robust ASR and SER systems for Vietnamese, leading to significant performance improvements over the baseline models.

References

Linh Nguyen Thanh Duy Cao Phuc Phan Duong A. Nguyen Anh Pham, Khanh Linh Tran. 2024. Bud500: A comprehensive vietnamese asr dataset.

Capleaf. 2024. vivoice: Enabling vietnamese multi-speaker speech synthesis. https://huggingface.co/datasets/capleaf/viVoice.

Moataz M.H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. 2024. Phowhisper: Automatic speech recognition for vietnamese. *arXiv preprint arXiv:2406.02555*.

Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. 2024. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Proc. INTERSPEECH*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

natmin322. 2023. 28k vietnamese voice augmented of vigbigdata. https://huggingface.co/datasets/natmin322/28k_vietnamese_voice_augmented_of_VigBigData.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Zheshu Song, Ziyang Ma, Yifan Yang, Jianheng Zhuo, and Xie Chen. 2024. A comparative study of llm-based asr and whisper in low resource and code switching scenario. *arXiv preprint arXiv:2412.00721*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thi Vu, Linh The Nguyen, and Dat Quoc Nguyen. 2025. Zero-shot text-to-speech for vietnamese. *arXiv preprint arXiv:2506.01322*.