Unified Acoustic Representation Learning for Vietnamese Speech Classification Tasks

Xuan-Truong HA
VinBigData, VinGroup
Ha Noi, Vietnam
hatruong3996@gmail.com

Van-Huy NGUYEN
VinBigData, VinGroup
Thai Nguyen University of Technology, Vietnam
Ha Noi, Vietnam
huynguyen@tnut.edu.vn

Tri-Nhan DO
VinBigData, VinGroup
Ha Noi, Vietnam
v.nhandt23@vinbigdata.org

Trung-Kien PHAN
VinBigData, VinGroup
Ha Noi, Vietnam
v.kienpt@vinbigdata.org

Dang-Khoa MAC
VinBigData, VinGroup
Ha Noi, Vietnam
v.khoamd@vinbigdata.org

***Abstract*—Analyzing multiple attributes of Vietnamese speech concurrently, such as gender, dialect, and emotion, is an important yet challenging task. Traditional methods often isolate each task, potentially overlooking correlations and leading to suboptimal performance. This paper proposes a unified architecture that integrates features derived from two different representations of the same input audio signal: raw waveforms and Mel spectrograms. The architecture employs a fixed, pre-trained Wav2Vec 2.0 encoder processing the waveform and a trainable Vision Transformer (ViT) encoder processing the spectrogram. These representations are fused using Feature-wise Linear Modulation (FiLM). This single multi-task model is designed to concurrently classify gender, dialect, and emotion. We detail a practical training strategy for a multi-dataset scenario using probabilistic sampling and a masked loss function, alongside a tailored hybrid evaluation protocol (fixed test set for ViMD, 5-fold cross-validation for VNEMOS). The proposed approach achieves competitive Macro F1-scores: 98.74% (gender), 91.81% (dialect), and 95.53% (emotion, 5-fold average). This research demonstrates the effectiveness of fusing different acoustic representations within a unified multi-task model for complex Vietnamese speech analysis.**

***Index Terms*—Vietnamese Speech, Multi-Task Learning, Speech Representation Fusion, Speech Emotion Recognition, Dialect Identification.**

I. Introduction

Human speech conveys rich information beyond linguistic content, including paralinguistic characteristics like speaker gender, dialect, and emotional state [1], [2]. The automatic extraction of this non-linguistic information—through tasks such as gender recognition, dialect identification, and Speech Emotion Recognition (SER)—is increasingly vital for applications ranging from human-computer interaction systems to large-scale audio data analysis.

However, developing systems capable of performing these diverse speech analysis tasks simultaneously presents significant hurdles. Conventional approaches typically build separate models for each task [3]–[5]. This strategy is resource-intensive and fails to exploit potential synergies and shared information across related tasks. Multi-Task Learning (MTL) [6] offers a compelling alternative, enabling a single model to learn multiple tasks concurrently. This often leads to improved generalization and performance through shared representations and implicit regularization.

These challenges are particularly pronounced for the Vietnamese language due to its complex tonal system, considerable variations between major dialects (Northern, Central, Southern), and the relative scarcity of labeled data compared to resource-rich languages [7], [8]. These factors necessitate robust and data-efficient methods. A unified architecture that effectively integrates information from different acoustic representations within a multi-task framework offers a promising direction.

This paper introduces such an architecture for the simultaneous classification of gender, dialect, and emotion in Vietnamese speech. The core of our approach involves processing the input audio signal through two parallel streams. The first stream utilizes a pre-trained Wav2Vec 2.0 model[1], kept **frozen**, to extract features directly from the raw audio waveform. The second stream employs a custom, **trainable** Vision Transformer (ViT) [10] to analyze Mel spectrogram images derived from the same audio. Features from these two representations are then integrated using Feature-wise Linear Modulation (FiLM) [11], allowing the spectrogram-based features to dynamically modulate the waveform-based features. The resulting fused representation is fed into separate classification heads for gender (Male/Female) and dialect (North/Central/South) using the ViMD dataset [8], and emotion (5 classes: anger, happiness, sadness, neutral, and fear) using the VNEMOS dataset [12].

A key contribution of this work is not only the architecture itself but also a practical training and evaluation pipeline designed for a multi-task model operating across multiple datasets with potentially different labeling schemes and evaluation protocols. We employ probabilistic sampling to construct

[1]the `nguyenvulebinh/wav2vec2-large-vi` model [9].

batches containing data from different sources and utilize a masked loss function to handle instances where labels for certain tasks are unavailable. A hybrid evaluation protocol is adopted, involving evaluation on the fixed ViMD test set and 5-fold cross-validation on VNEMOS.

Our main contributions are threefold: we propose a unified architecture fusing waveform and spectrogram representations (Frozen Wav2Vec 2.0 + Trainable ViT + FiLM) for multi-task Vietnamese speech analysis; we develop a practical methodology for training and evaluating such a model across multiple datasets with varying labels and protocols; and we demonstrate competitive performance across all three tasks (gender, dialect, emotion) within this single, unified model.

II. Related Work

Speech processing has seen significant advancements, especially in feature extraction. While handcrafted features like Mel-frequency cepstral coefficients (MFCCs) [13], [14] remain relevant, deep learning-based representations have become dominant. Self-supervised learning (SSL) models, such as Wav2Vec 2.0 [15], [16], HuBERT [17], and WavLM [18], learn powerful representations from large unlabeled audio datasets and can be effectively fine-tuned for downstream tasks. The availability of pre-trained SSL models for Vietnamese [9] is particularly advantageous, allowing their use as frozen feature extractors to conserve computational resources while leveraging strong acoustic features.

Exploiting complementary information derived from different representations of the same input signal is a common strategy. Combining features from raw audio waveforms, which capture fine temporal details, with those from visual representations like Mel spectrograms, which highlight frequency patterns over time, has proven effective in speech processing [19], [20]. Models originally designed for vision, including Convolutional Neural Networks (CNNs) [21] and Vision Transformers (ViTs) [10], have shown success in analyzing spectrograms for tasks like SER [22]–[24] and acoustic event detection. Fusion techniques are critical for integrating information from these different representations. Methods range from simple concatenation and attention mechanisms [25] to more sophisticated modulation approaches like FiLM [11], which enable flexible, conditional interactions between feature sets.

Specifically for Vietnamese, research has addressed individual tasks such as gender recognition [26], dialect identification [8], [27], [28], and SER [12], [29], [30]. However, work on unified systems that address these tasks concurrently by fusing different acoustic representations remains limited. Our work aims to contribute to this area by proposing and evaluating such a system tailored for Vietnamese speech.

III. Methodology: Model Architecture

The proposed architecture, illustrated in Figure 1, is an end-to-end system designed for multi-task learning by fusing features derived from the input audio's waveform and spectrogram representations. It comprises two parallel encoding streams processing these representations, a fusion mechanism, and separate task-specific prediction heads.

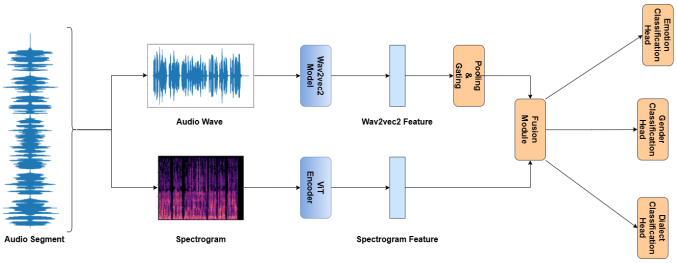

Fig. 1. Overview of the proposed system architecture. The waveform stream employs a frozen Wav2Vec 2.0, while the spectrogram stream uses a trainable ViT. Features derived from these representations are integrated using FiLM. The fused features pass through a shared layer before feeding into task-specific heads for Gender, Dialect, and Emotion classification.

A. Waveform Encoding Stream

This stream processes the raw audio signal $\boldsymbol{x}_{\text{audio}} \in \mathbb{R}^{B \times T}$, where B is the batch size and T is the sequence length.

1) Fixed Wav2Vec 2.0 Encoder: We utilize a pre-trained Vietnamese Wav2Vec 2.0 model [9] as a fixed feature extractor. It takes $\boldsymbol{x}_{\text{audio}}$ as input and produces a sequence of hidden representations $\boldsymbol{h}_{\text{audio}} \in \mathbb{R}^{B \times L \times D_{\text{w2v}}}$, where L is the representation sequence length and D_{w2v} is the hidden dimension (1024 for large). All weights of this Wav2Vec 2.0 model remain **frozen** during training.

2) Pooling & Gating Module: To obtain a fixed-size vector from the variable-length sequence $\boldsymbol{h}_{\text{audio}}$, we employ a combination of pooling and gating. Mean pooling, max pooling, and attention-weighted pooling are computed across the temporal dimension (L) yielding $\boldsymbol{A}_{\text{mean}}$, $\boldsymbol{A}_{\text{max}}$, and $\boldsymbol{A}_{\text{att}}$. These pooled features are combined using a small gating network:

$$\boldsymbol{A}_{\text{pooled}} = g_{\text{mean}}\boldsymbol{A}_{\text{mean}} + g_{\text{max}}\boldsymbol{A}_{\text{max}} + g_{\text{att}}\boldsymbol{A}_{\text{att}} \quad (1)$$

where g_k are learned scalar gate weights. The resulting vector $\boldsymbol{A}_{\text{pooled}}$ is projected down to a shared dimension $D_{\text{shared}} = 512$, followed by dropout and Layer Normalization, producing the final waveform-based feature $\boldsymbol{A} \in \mathbb{R}^{B \times D_{\text{shared}}}$.

$$\boldsymbol{A} = \text{LayerNorm}(\text{Dropout}(\text{Linear}_{D_{\text{w2v}} \to D_{\text{shared}}}(\boldsymbol{A}_{\text{pooled}}))) \quad (2)$$

B. Spectrogram Encoding Stream

This stream processes Mel spectrogram images $\boldsymbol{x}_{\text{vision}} \in \mathbb{R}^{B \times 3 \times 224 \times 224}$, which are derived from the same original audio signal $\boldsymbol{x}_{\text{audio}}$. The single-channel spectrogram is replicated across three channels to match the standard ViT input format.

1) Trainable ViT Encoder: We employ a standard ViT-Base encoder [10], the weights of which are **trainable**. The encoding process begins by dividing the 224×224 spectrogram image into $N = 196$ non-overlapping 16×16 patches. Each patch is linearly projected into an embedding of dimension $D_{\text{vit}} = 768$. A learnable [CLS] token embedding $\boldsymbol{x}_{\text{class}}$ is prepended to the sequence of patch embeddings, and positional embeddings $\boldsymbol{E}_{\text{pos}}$ are added. This sequence then passes through $L_{\text{vit}} = 12$ Transformer Encoder blocks,

each comprising Multi-Head Self-Attention (MHSA), a Feed-Forward Network (FFN), Layer Normalization, and residual connections. Finally, the state of the [CLS] token from the last encoder block, $\boldsymbol{z}^0_{L_{\text{vit}}}$, is Layer Normalized and linearly projected to the shared dimension $D_{\text{shared}} = 512$, yielding the spectrogram-based feature $\boldsymbol{V} \in \mathbb{R}^{B \times D_{\text{shared}}}$.

$$\boldsymbol{V} = \text{Linear}_{D_{\text{vit}} \to D_{\text{shared}}}(\text{LayerNorm}(\boldsymbol{z}^0_{L_{\text{vit}}})) \tag{3}$$

C. Feature Fusion

This stage integrates the waveform-derived feature $\boldsymbol{A}$ and the spectrogram-derived feature $\boldsymbol{V}$.

1) FiLM Layer: We adopt Feature-wise Linear Modulation (FiLM) [11] for fusion. FiLM allows the spectrogram feature $\boldsymbol{V}$ to modulate the waveform feature $\boldsymbol{A}$ via an element-wise affine transformation. First, the waveform feature is normalized: $\boldsymbol{A}_{\text{norm}} = \text{LayerNorm}(\boldsymbol{A})$. Then, modulation parameters (scale $\boldsymbol{\gamma}$ and shift $\boldsymbol{\beta}$) are generated from the spectrogram feature $\boldsymbol{V}$ using separate linear layers:

$$\boldsymbol{\gamma} = \text{Linear}_{\gamma}(\boldsymbol{V}) \in \mathbb{R}^{B \times D_{\text{shared}}} \tag{4}$$

$$\boldsymbol{\beta} = \text{Linear}_{\beta}(\boldsymbol{V}) \in \mathbb{R}^{B \times D_{\text{shared}}} \tag{5}$$

The FiLM operation is applied as $\boldsymbol{F}_{\text{pre}} = \boldsymbol{\gamma} \odot \boldsymbol{A}_{\text{norm}} + \boldsymbol{\beta}$, where $\odot$ denotes element-wise multiplication. The resulting modulated feature is normalized (e.g., using BatchNorm1d) to produce the final fused feature vector $\boldsymbol{F} = \text{BatchNorm1d}(\boldsymbol{F}_{\text{pre}}) \in \mathbb{R}^{B \times D_{\text{shared}}}$.

D. Multi-Task Prediction Heads

The fused feature vector $\boldsymbol{F}$ serves as a shared input representation for the downstream classification tasks.

1) Shared Transformation Layer: Optionally, $\boldsymbol{F}$ can be passed through a shared non-linear transformation block (e.g., Linear -> LayerNorm -> ReLU -> Dropout) to obtain a further refined representation $\boldsymbol{T} \in \mathbb{R}^{B \times D_{\text{out}}}$ (e.g., $D_{\text{out}} = 512$).

$$\boldsymbol{T} = \text{SharedBlock}(\boldsymbol{F}) \tag{6}$$

2) Task-Specific Classification Heads: From the shared representation $\boldsymbol{T}$, three separate classification heads (typically small MLPs) generate logits for each task: gender ($\boldsymbol{z}_{\text{gender}} \in \mathbb{R}^{B \times 2}$), dialect ($\boldsymbol{z}_{\text{dialect}} \in \mathbb{R}^{B \times 3}$), and emotion ($\boldsymbol{z}_{\text{emotion}} \in \mathbb{R}^{B \times 5}$).

$$\boldsymbol{z}_{\text{gender}} = \text{Head}_{\text{gender}}(\boldsymbol{T}) \tag{7}$$

$$\boldsymbol{z}_{\text{dialect}} = \text{Head}_{\text{dialect}}(\boldsymbol{T}) \tag{8}$$

$$\boldsymbol{z}_{\text{emotion}} = \text{Head}_{\text{emotion}}(\boldsymbol{T}) \tag{9}$$

These logits are used to compute the respective task losses, typically via Cross-Entropy.

IV. Experimental Setup

This section details the datasets, data preparation, training strategy, evaluation protocol, and implementation specifics.

A. Datasets

We utilize two publicly available Vietnamese speech datasets:

1) **ViMD** [8]: Provides natural conversational speech labeled for gender (Male/Female) and dialect (North/Central/South). We strictly adhere to its predefined training and testing splits for evaluating gender and dialect classification.
2) **VNEMOS** [12]: Contains emotional speech (acted and natural/elicited) annotated with 5 emotion classes (Anger, Happiness, Sadness, Neutral, Anxiety). Due to its structure, we employ speaker-independent 5-fold cross-validation for evaluating emotion recognition.

B. Data Preparation and Augmentation

Audio preprocessing involves converting all files to mono, resampling to 16kHz, and normalizing amplitude values to the range [-1, 1]. From this preprocessed audio, Mel spectrograms are computed using a 25ms window, 10ms hop size, and 128 Mel frequency bins. These log Mel spectrograms are resized to 224×224 pixels and channel-replicated to create RGB-like images suitable for ViT input. During training only, data augmentation is applied. For the raw audio, this includes random pitch shifting, speed perturbation, and the addition of background noise from the MUSAN dataset [31] at varying Signal-to-Noise Ratios (SNRs). For the Mel spectrograms, SpecAugment, involving frequency and time masking, is utilized.

C. Multi-Dataset Training Strategy

To handle the heterogeneity between datasets and tasks within a single model, we adopt a specific training strategy. Each training batch is constructed via probabilistic sampling, drawing data points equally ($p = 0.5$) from the ViMD training set and the current VNEMOS training folds (4 out of 5 folds in the active cross-validation split). The total loss is calculated as a weighted sum of Cross-Entropy losses (with label smoothing $\epsilon = 0.1$) for each task. Crucially, a masked loss approach is used: the loss for a specific task is computed only for samples within the batch that possess a valid label for that task. This is implemented using binary masks $m_{T,i}$ (1 if sample i has a label for task T, 0 otherwise):

$$\begin{aligned}\mathcal{L}_{\text{total}} = \frac{1}{N_{valid}} \sum_{i=1}^{B} [& m_{G,i} \lambda_G \mathcal{L}_{CE}(\boldsymbol{z}_{\text{gender},i}, y_{\text{gender},i}) \\ & + m_{D,i} \lambda_D \mathcal{L}_{CE}(\boldsymbol{z}_{\text{dialect},i}, y_{\text{dialect},i}) \\ & + m_{E,i} \lambda_E \mathcal{L}_{CE}(\boldsymbol{z}_{\text{emotion},i}, y_{\text{emotion},i})]\end{aligned} \tag{10}$$

Here, N_{valid} represents the total number of valid task labels in the batch ($\sum_i m_{G,i} + m_{D,i} + m_{E,i}$), and task weights are set equally ($\lambda_G = \lambda_D = \lambda_E = 1$). For additional regularization and to prevent over-reliance on a single representation stream, we apply *representation dropout* ($p = 0.1$) before the FiLM fusion layer, randomly zeroing out either the entire waveform-derived feature $\boldsymbol{A}$ or the entire spectrogram-derived feature $\boldsymbol{V}$ for a given sample.

D. Evaluation Protocol

We employ a hybrid evaluation protocol combining fixed-set evaluation and cross-validation:

1) Five independent training runs are performed.
2) Within each run k (where $k = 1...5$):
   - VNEMOS is split into 5 speaker-independent folds. Fold k serves as the validation/test set for emotion in this run, while the remaining 4 folds are used for training the emotion task.
   - The model is trained using the ViMD training set (for gender/dialect) and the 4 active VNEMOS training folds (for emotion), following the multi-dataset strategy described above.
   - Post-training, the model's performance is evaluated on the fixed ViMD test set (for gender/dialect) and the held-out VNEMOS fold k (for emotion).
3) Final performance metrics are reported as follows:
   - For gender and dialect: Mean $\pm$ standard deviation across the 5 runs evaluated on the fixed ViMD test set.
   - For emotion: Mean $\pm$ standard deviation across the 5 held-out VNEMOS folds (representing the 5-fold cross-validation result).

The primary evaluation metric is the **Macro F1-score**. We also report Unweighted Average Recall (UAR) for the emotion task, as it is commonly used in SER literature.

E. Implementation Details

Experiments were conducted using PyTorch, Hugging Face Transformers, and torchaudio libraries on NVIDIA RTX 3090 GPUs (or equivalent). We used a batch size of 32 and the AdamW optimizer [32] with a learning rate of 3×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and weight decay of 0.01. The learning rate schedule included a linear warmup for the first 10% of training steps, followed by a linear decay to zero. Models were trained for a maximum of 50 epochs, employing early stopping based on validation performance with a patience of 5-10 epochs.

F. Baseline Models for Comparison

We compare our proposed fusion model against two main types of baselines:

1) **Original Reported Results:** Performance metrics cited in the original papers for ViMD [8] and VNEMOS [12]. Direct comparison should be viewed cautiously due to potential differences in evaluation setups or metrics. We select representative results where available for context.
2) **W2V2 Finetuned (Single-Task):** The same pre-trained Wav2Vec 2.0 model used in our fusion architecture [9], but fine-tuned separately for each individual task (gender, dialect, emotion). These models are trained and evaluated using our exact hybrid protocol (fixed test for ViMD tasks, 5-fold CV for VNEMOS emotion) to serve as a strong single-representation, single-task baseline.

G. Computational Cost and Inference Time

To assess the practical viability of our proposed architecture, we analyzed its computational demands during both training and inference. Key aspects of our design, particularly the use of a frozen Wav2Vec 2.0 encoder, were intentionally chosen to balance high performance with manageable computational costs.

1) Model Complexity and Training Cost: The total number of parameters in our model is approximately 405 million. However, a crucial design choice was to **freeze the 317M parameters** of the Wav2Vec 2.0 Large backbone. Consequently, only about 88M parameters—primarily from the Vision Transformer (ViT-Base, $\sim$86M) and the smaller fusion/classification head components ($\sim$2M)—were updated during training. This strategy significantly reduces the GPU memory required for storing gradients and optimizer states, making the training process feasible on a single consumer-grade GPU like the NVIDIA RTX 3090. Table I provides a breakdown of the model's parameters. Each of the five cross-validation runs, training for up to 50 epochs with early stopping, required approximately 4-5 hours of computation.

TABLE I
APPROXIMATE PARAMETER DISTRIBUTION IN THE PROPOSED MODEL.

Component	Total Parameters	Trainable Parameters
Wav2Vec 2.0 Large (Frozen)	$\sim$317 M	0
Vision Transformer (ViT-Base)	$\sim$86 M	$\sim$86 M
Fusion, Pooling, & Heads	$\sim$2 M	$\sim$2 M
Total	**$\sim$405 M**	**$\sim$88 M**

2) Inference Speed: Inference efficiency is critical for real-world deployment. On an NVIDIA RTX 3090 GPU, the model achieves an inference speed of approximately 5 audio samples per second when processing audio sequences of 1-20 seconds in length.

V. RESULTS AND DISCUSSION

This section presents the main experimental results and discusses the findings. Table II summarizes the Macro F1-Score performance of our proposed unified fusion model compared to the baselines.

Our proposed unified model achieves strong results across all tasks: an average Macro F1-score of 98.74% for gender classification and 91.81% for dialect identification on the ViMD test set, and an average Macro F1-score of 95.53% for emotion recognition over the 5-fold cross-validation on VNEMOS.

Comparing these results with the baselines highlights the effectiveness of our approach. While acknowledging potential differences in evaluation setups for the originally reported baselines, our model achieves highly competitive performance levels. More significantly, when compared against the strong Wav2Vec 2.0 baseline fine-tuned on single tasks using our identical evaluation protocol, the proposed fusion model demonstrates clear advantages. The improvement is particularly substantial for emotion recognition (+14.7% absolute

TABLE II
TASK CLASSIFICATION PERFORMANCE (MACRO F1-SCORE %). GENDER/DIALECT SCORES ARE MEAN ± STD DEV OVER 5 RUNS ON THE VIMD TEST SET. EMOTION SCORE IS MEAN ± STD DEV FROM 5-FOLD CV ON VNEMOS.

Method	Gender	Dialect	Emotion
Reported VNEMOS Baseline [12]*	–	–	∼87 UAR
Reported ViMD Baseline [8]*	-	91.47	–
Baselines (Our Protocol):			
W2V2-Large-Vi Finetuned (Single-Task) [9]	98.51	91.21	$80.83 \pm 0.65^{\ddagger}$
Proposed Multi-Task Model:			
Proposed Method	**98.74**	**91.81**	$\mathbf{95.53 \pm 0.49}^{\ddagger}$

*Original baseline results cited for reference.

‡Emotion score is Macro F1-score average over 5-fold CV on VNEMOS (5 classes).

Macro F1), but also noticeable for dialect (+0.60% absolute Macro F1) and gender (+0.23% absolute Macro F1). This underscores the benefits derived from fusing features from both the raw waveform and the Mel spectrogram, combined with the representational power gained through the multi-task learning paradigm. The model successfully learns complementary information from the different acoustic representations and leverages the shared learning process inherent in MTL.

Table III provides a more detailed breakdown of the emotion recognition performance per class (averaged over the 5-fold CV) for the proposed fusion model.

TABLE III
DETAILED PER-EMOTION CLASSIFICATION PERFORMANCE (PROPOSED FUSION MODEL - VNEMOS 5 CLASSES - 5-FOLD CV AVERAGE %).

Emotion	Precision	Recall (UAR)	F1-Score
Anger	96.12	94.35	95.23
Happiness	95.05	95.11	95.08
Sadness	97.88	97.02	97.45
Neutral	97.15	99.18	98.15
Fear	92.15	91.99	92.07
Macro Average	**95.67**	**95.53**	**95.60**

The per-class emotion results show relatively strong and balanced performance. "Neutral" achieves the highest F1-score (98.15%), a common finding as neutral speech often forms a distinct acoustic baseline. "Sadness" and "Anger" also exhibit high recognition rates. "Fear" shows a slightly lower F1-score (92.07%), potentially reflecting the inherent difficulty in distinguishing its subtle acoustic cues or possible overlap with other affective states. Overall, the model demonstrates effective discrimination between the five target emotion classes within the VNEMOS dataset.

VI. CONCLUSION

This paper introduced a unified architecture for simultaneous gender, dialect, and emotion classification in Vietnamese speech. By integrating features from raw waveforms via a frozen pre-trained Wav2Vec 2.0 encoder and from Mel spectrograms via a trainable Vision Transformer, fused using Feature-wise Linear Modulation (FiLM), we constructed a robust multi-task model. We also presented a practical multi-dataset training strategy using probabilistic sampling and masked loss, evaluated with a tailored hybrid protocol. Our results demonstrate that this single model achieves competitive performance across all three tasks, significantly outperforming strong single-representation, single-task baselines, particularly for emotion recognition.

The findings highlight the benefits of fusing different acoustic representations derived from the same audio input and leveraging multi-task learning for complex speech analysis challenges. The proposed architecture is effective and computationally mindful due to the frozen Wav2Vec 2.0 backbone. Furthermore, the multi-dataset methodology offers a practical approach for handling diverse data sources.

VII. FUTURE WORK

Future research could explore several promising avenues. Investigating more sophisticated fusion mechanisms beyond FiLM, as well as advanced techniques for optimizing multi-task and multi-dataset training such as dynamic loss weighting or gradient balancing, could yield further improvements. Incorporating linguistic information, potentially extracted via an Automatic Speech Recognition (ASR) component, could provide valuable contextual cues. Extending the framework to encompass other relevant paralinguistic tasks like speaker identification or age estimation is another logical step. Additionally, adapting and applying the proposed architecture to other languages facing similar challenges (e.g., tonal languages, those with significant dialectal diversity) would be valuable. Finally, exploring the integration of Large Language Models (LLMs) to potentially leverage semantic context or enhance zero/few-shot generalization presents an exciting direction. We believe this work provides a solid foundation for building more comprehensive, efficient, and accurate systems for analyzing the rich information embedded within Vietnamese speech.

REFERENCES

[1] C. G. Clopper, B. Conrey, and D. B. Pisoni, "Effects of talker gender on dialect categorization," *Journal of Language and Social Psychology*, vol. 24, no. 2, pp. 182–206, 2005.

[2] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614–36, 1996.

[3] Q. Ouyang, "Speech emotion detection based on MFCC and CNN-LSTM architecture," *Applied and Computational Engineering*, vol. 5, no. 1, pp. 243–249, May 2023.

[4] M. Gupta and S. Chandra, "Speech emotion recognition using MFCC and wide residual network," in *Proceedings of the Thirteenth International Conference on Contemporary Computing*, ser. IC3-2021. New York, NY, USA: Association for Computing Machinery, 2021, pp. 320–327.

[5] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," 2021.

[6] S. Ruder, "An overview of multi-task learning in deep neural networks," *ArXiv*, vol. abs/1706.05098, 2017.

[7] v. Tuan, C. d'Alessandro, and S. Rosset, "A phonetic study of vietnamese tones: acoustic and electroglottographic measurements," 09 2002.

[8] N. V. Dinh, T. C. Dang, L. Thanh Nguyen, and K. V. Nguyen, "Multi-dialect vietnamese: Task, dataset, baseline models and challenges," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 7476–7498.

[9] T. B. Nguyen, "Vietnamese end-to-end speech recognition using wav2vec 2.0," 09 2021.
[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020.
[11] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," in *International Conference on Learning Representations*, 2021.
[12] N. Quang Anh, M.-H. Ha, Q. C. Nguyen, T. H. N. Thi, Q. Vu, D. Minh-Duc, D.-C. Nguyen, and T. K. Dinh, "Vnemos: Vietnamese speech emotion inference using deep neural networks," in *9th International Conference on Integrated Circuits, Design, and Verification (ICDV)*, 2024, pp. 97–101.
[13] V. T. Tiwari, "Mfcc and its applications in speaker recognition," *Int. J. Emerg. Technol.*, vol. 1, 01 2010.
[14] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.
[15] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 2019.
[16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.
[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.
[18] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
[19] L. Wang, P. Luc, A. Recasens, J.-B. Alayrac, and A. van den Oord, "Multimodal self-supervised learning of general audio representations," 2021.
[20] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7367–7371.
[21] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015.
[22] H. Aouani and Y. B. Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251–260, 2020, knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020.
[23] R. Begazo, A. Aguilera, I. Dongo, and Y. Cardinale, "A combined cnn architecture for speech emotion recognition," *Sensors*, vol. 24, no. 17, 2024.
[24] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," *Scientific Reports*, vol. 14, 06 2024.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
[26] H. Bui, "Vietnamese voice classification based on deep learning approach," *International Journal of Machine Learning and Networked Collaborative Engineering*, vol. 04, pp. 171–180, 01 2020.
[27] P. N. Hung, T. V. Loan, and N. H. Quang, "Automatic identification of vietnamese dialects," *Journal of Computer Science and Cybernetics*, vol. 32, no. 1, pp. 19–30, Jul. 2016.
[28] A. Schweitzer and T. Vu, "Cross-gender and cross-dialect tone recognition for vietnamese," 09 2016, pp. 1064–1068.
[29] Q.-A. N. D., M.-H. Ha, T. K. Dinh, M.-D. Pham, and N. N. Van, "Emotional vietnamese speech-based depression diagnosis using dynamic attention mechanism," 2024.
[30] T. D. T. Le, T. V. Loan, and N. H. Quang, "Gmm for emotion recognition of vietnamese," *Journal of Computer Science and Cybernetics*, vol. 33, no. 3, pp. 229–246, Mar. 2018.
[31] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015.
[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.