Exploring the Adaptability of Large Speech Models to Non-Verbal Vocalization Task

Juan José Márquez Villacís[1,†], Federico D'Asaro[1,2,*,†], Giuseppe Rizzo[1,2] and Andrea Bottino[2]

[1]LINKS Foundation – AI, Data & Space (ADS)
[2]Politecnico di Torino – Dipartimento di Automatica e Informatica (DAUIN)

Abstract

Large Speech Models (LSMs), pre-trained on extensive speech corpora, have recently emerged as powerful foundations in the audio processing field, demonstrating strong transfer capabilities to downstream tasks such as speaker identification and emotion recognition. However, while these models excel on speech-centric tasks, limited research has investigated their adaptability to Non-Verbal Vocalization (NVV) tasks, which involve vocal bursts like laughter, sighs, shrieks, and moans.

In this work, we examine how well LSMs, specifically Wav2Vec 2.0, HuBERT, WavLM, and Whisper, can be adapted to NVV tasks. We conduct experiments using both linear probing to evaluate the pre-trained knowledge relevant to NVVs, and Parameter-Efficient Fine-Tuning (PEFT) techniques, including LoRA, Adapters, and Prompt Tuning. Experimental results on NVV datasets—*ASVP-ESD, CNVVE, Non-Verbal Vocalization Dataset, ReCANVo, VIVAE*—indicate that Whisper-based models consistently achieve superior performance, which is further enhanced through the application of LoRA. Additionally, our layer-wise analysis reveals that applying PEFT specifically to layers with lower NVV information is key to effective model adaptation, providing valuable insights for optimizing fine-tuning strategies in future work. The repository associated with this work can be found here: https://github.com/links-ads/kk-nonverbal-vocal-class

Keywords

Non-Verbal Vocalization Large Speech Models Parameter Efficient Fine-Tuning

1. Introduction

Understanding and correctly identifying emotional cues in human vocalizations is essential for building conversational systems capable of engaging with people in an emotionally aware and natural manner [1, 2]. Emotional information in the human voice is transmitted mainly through two distinct pathways: *speech prosody*—which encompasses features such as intonation, rhythm, and vocal quality [3]—and non-verbal vocal sounds, commonly referred to as *vocal bursts* [4], which include expressions like laughter, sighs, screams, and moans. Importantly, these non-speech sounds serve as critical communicative tools, particularly for individuals with profound disabilities or speech limitations, since more than 96% of people with speech impairments are still able to produce non-verbal vocalizations [5].

While much research has focused on speech-related tasks such as speaker recognition, speaker diarization, and emotion recognition from prosody [6], the domain of Non-Verbal Vocalizations (NVV) has received comparatively little attention [7, 1]. Early approaches for NVV analysis often relied on Hidden Markov Models or Convolutional Neural Networks. However, the advent of Transformer architectures [8] has led to the development of Large Speech Models (LSMs), including Wav2Vec 2.0 [9], HuBERT [10], WavLM [11], and Whisper [12], which have demonstrated impressive transfer learning capabilities on speech-based tasks. Despite this success, the adaptability of these models to NVV tasks remains largely unexplored.

In this work, we systematically investigate how various LSMs perform as feature extractors for NVV recognition, aiming to understand the extent to which non-verbal knowledge is already embedded in their pre-trained representations. To further enhance their adaptation to NVV tasks, we apply Parameter-Efficient Fine-Tuning (PEFT) strategies [13], including Adapters [14], Prompt Tuning [15], and LoRA [16].

Our experimental results, conducted across five NVV datasets—*ASVP-ESD, CNVVE, Non-Verbal Vocalization Dataset, ReCANVo, VIVAE*—indicate that Whisper consistently outperforms Wav2Vec 2.0, HuBERT, and WavLM, especially when fine-tuned with PEFT techniques. Among these, LoRA achieves the best overall performance. Further analysis of the Transformer layers reveals that non-verbal information is primarily captured in the later layers of Whisper. Interestingly, we find that applying LoRA exclusively to earlier, less important lay-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
juan.marquez@linksfoundation.com (J.J. Márquez Villacís); federico.dasaro@polito.it (F. D'Asaro); giuseppe.rizzo@linksfoundation.com (G. Rizzo); andrea.bottino@polito.it (A. Bottino)
http://conceptbase.sourceforge.net/mjf/ (A. Bottino)
0009-0008-3098-5492 (J.J. Márquez Villacís); 0009-0003-8727-3393 (F. D'Asaro); 0000-0003-0083-813X (G. Rizzo); 0000-0002-8894-5089 (A. Bottino)

ers yields better adaptation compared to focusing on the layers already rich in non-verbal knowledge. This counterintuitive result suggests that adjusting the layers with initially limited task relevance is crucial, as these layers benefit most from targeted adaptation.

The main contributions of this work are:

- We evaluate the adaptability of Large Speech Models (Wav2Vec 2.0, HuBERT, WavLM, and Whisper) to Non-Verbal Vocalization tasks using both linear probing and Parameter-Efficient Fine-Tuning techniques on five NVV datasets.
- We demonstrate that Whisper achieves the strongest performance across all datasets, and that LoRA is the most effective PEFT method when compared to Adapters and Prompt Tuning.
- Through layer-wise importance analysis, we observe that non-verbal information is predominantly encoded in the later layers of Whisper. Surprisingly, we find that adapting less important layers is more beneficial for task-specific performance than focusing solely on the most informative layers.

2. Related Work

2.1. Non Verbal Vocalization

Early approaches to recognizing Non-Verbal Vocalizations (NVVs) primarily relied on Hidden Markov Models (HMMs), which analyzed vocal signals based on acoustic features such as intensity, pitch, and vowel articulation patterns [17, 18]. Despite their initial success, these models were limited by their dependence on linear modeling, susceptibility to noise interference, and challenges in handling large or complex datasets.

To address these limitations, subsequent research transitioned towards employing convolutional neural networks (CNNs) that process time-frequency representations like Mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) [7]. Recent progress has been driven by the adoption of Transformer-based frameworks capable of learning from massive audio datasets. Drawing inspiration from large-scale speech models such as Wav2Vec 2.0 and Whisper, these state-of-the-art systems have enabled the classification of up to 67 distinct types of vocal expressions [1].

Following this research direction, Koudounas et al. [19] proposed a new foundation model trained on 125 hours of non-verbal vocalization data, demonstrating significantly improved performance on downstream classification tasks.

2.2. Large Speech Models

Recent advancements in natural language processing (NLP) and computer vision (CV) have leveraged vast amounts of unlabeled data using Self-Supervised Learning [20, 21] and Weakly Supervised Learning [22]. Inspired by techniques such as masked language modeling in NLP and image modeling in CV, Wav2Vec 2.0 [9] introduced a Large Speech Model (LSM) trained through masked speech modeling on large-scale audio datasets, including the LibriSpeech corpus [23] and LibriVox [24].

Following Wav2Vec 2.0, subsequent LSMs such as HuBERT [10] and WavLM [11] further advanced self-supervised pretraining approaches. In parallel, Whisper [12] was introduced, trained with large-scale weak supervision from paired audio and transcription data using an encoder-decoder transformer architecture.

These large speech models have demonstrated strong capabilities in learning rich and robust speech representations from large datasets, leading to significant improvements in various tasks, including language modeling, audio classification, and speech-to-text transcription.

2.3. Parameter Efficient Finetuning

Large-scale models demonstrate strong adaptability across a wide range of downstream tasks, but this often comes at a significant computational cost. To address this, Parameter-Efficient Fine-Tuning (PEFT) techniques have emerged, aiming to introduce minimal task-specific parameters while keeping the majority of the pretrained model unchanged. This approach preserves the model's generalization ability and reduces the number of parameters that require modification.

As outlined by Han et al. [13], PEFT methods can be broadly categorized into two types: *Additive PEFT* and *Reparameterized PEFT*. Additive PEFT methods include techniques such as Adapters [14] and Prompt Tuning [15], which introduce additional learnable components at either the activation level or through prompt-based conditioning without altering the core model parameters. Reparameterized PEFT approaches, such as LoRA [16], apply low-rank adaptations to the weight matrices, effectively transforming the model's parameter space while maintaining the original architecture and inference speed.

These parameter-efficient strategies have shown strong results in English Speech Emotion Recognition tasks [25, 26, 27], with LoRA in particular demonstrating notable performance. In this work, we investigate the application of Adapters, Prompt Tuning, and LoRA for adapting Large Speech Models to the classification of Non-Verbal Vocalizations.

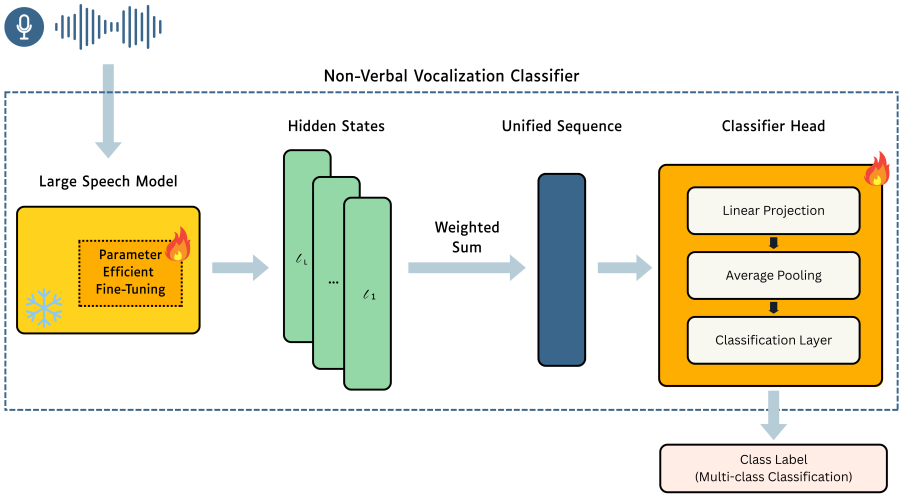

Figure 1: Overview of our Non-Verbal Vocalization Classifier, which consists of a Large Speech Model fine-tuned using the Parameter Efficient approach, followed by a classification head.

3. Non-Verbal Vocalization Classifier

In this section, we describe the architecture of the Non-Verbal Vocalization classifier illustrated in Figure 1. The model is composed of a Large Speech Model serving as the backbone $\mathcal{B}$, with a classifier $\mathcal{C}$ stacked on top. Additionally, we describe the integration of PEFT techniques, which can be selectively applied to the Transformer layers of the LSM to enhance adaptability while minimizing the number of trainable parameters.

3.1. Large Speech Models

Wav2Vec 2.0 Wav2Vec 2.0 demonstrated, for the first time, that it is possible to learn powerful speech representations directly from raw audio without requiring labels. The architecture consists of a multi-layer 1D convolutional feature encoder, which takes raw audio input X and produces latent representations $Z = \{z_1, \ldots, z_T\}$, where T denotes the number of frames, each corresponding to 25 ms of audio. These latent representations Z are then passed through a Transformer network to obtain contextualized representations $C = \{c_1, \ldots, c_T\}$. Additionally, the output of the feature encoder is discretized using product quantization in the latent space [28]. This discretization enables the application of masked speech modeling, the core innovation of Wav2Vec 2.0's self-supervised learning strategy. The model is trained to solve a contrastive task, where it must correctly identify the true quantized latent representation of a masked time step from a set of distractor candidates.

HuBERT HuBERT [10] introduced the use of an acoustic unit discovery system, such as k-means clustering applied to MFCC features, to generate frame-level targets for both masked and unmasked tokens. By adjusting the number of clusters (k), the system produces targets of varying granularity, ranging from broad vowel categories to more fine-grained senones. Similar to Wav2Vec 2.0, the HuBERT architecture employs a 1D convolutional feature encoder with seven layers, using a frame size of 20 ms, followed by a series of Transformer blocks for contextual representation learning.

WavLM The WavLM framework [11] further extends the pretraining approach introduced by Wav2Vec 2.0 by integrating both masked speech prediction and speech denoising into the pretraining process. Specifically, WavLM introduces masked speech denoising, where portions of the input are artificially corrupted with simulated noise or overlapping speech. The model is then tasked with predicting the pseudo-labels of the original clean speech in the masked regions, similar to the approach used in HuBERT. This strategy enhances the model's robustness in complex acoustic environments.

Like previous models, WavLM employs a 1D convolutional feature encoder followed by a Transformer encoder. The Transformer in WavLM is augmented with gated relative position bias, which improves the modeling of interactions between speech segments and enhances the model's ability to capture long-range dependencies.

Whisper Unlike previous models, Whisper adopts a weakly supervised learning paradigm that relies on paired audio and transcription data. Specifically, it predicts raw text transcripts directly from audio without requiring significant text standardization. Whisper employs an encoder-decoder Transformer architecture, consisting of an encoder E and a decoder D, which processes Mel spectrograms instead of raw waveforms as used in earlier models. Formally, given an input audio signal X, the model first applies two 1D convolutional layers with GELU activation as a feature encoder, followed by Transformer blocks to produce contextualized internal representations. These representations are then used by the BERT-like decoder D to generate the output text.

In this work, we utilize the Whisper model solely as a feature extractor by using the encoder E as backbone $\mathcal{B}$ and discarding the decoder D.

3.2. PEFT Methods

Adapter Adapters introduce small, trainable modules within Transformer layers to enable efficient fine-tuning. Each adapter consists of a down-projection matrix $W_{\text{down}} \in \mathbb{R}^{r \times d}$, a non-linear activation $\sigma(\cdot)$, and an up-projection matrix $W_{\text{up}} \in \mathbb{R}^{d \times r}$, where d is the hidden size and r is the bottleneck dimension.

Given input h_{in}, the adapter output with residual connection is:

$$\text{Adapter}(c) = W_{\text{up}}\, \sigma\left(W_{\text{down}}\, c\right) + c \tag{1}$$

Prompt Tuning Unlike adapters, embedding prompts introduce learnable prompt vectors that are prepended to the input sequence at each Transformer layer. Formally, the input sequence to layer l is:

$$X^{(l)} = \left[p_1^{(l)}, \ldots, p_{N_P}^{(l)}, c_1^{(l)}, \ldots, c_{N_C}^{(l)}\right] \tag{2}$$

where $p_i^{(l)}$ are the continuous prompt tokens and $c_i^{(l)}$ are the original input tokens. Here, N_P denotes the number of continuous prompt tokens, and N_C is the length of the original input. This approach allows task-specific information to be injected directly into the model without modifying its internal weights.

LoRA LoRA enhances each Transformer layer by applying a low-rank decomposition to the pretrained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, enabling parameter-efficient fine-tuning without altering the original model weights. It adds two additional trainable matrices: $W_{\text{down}} \in \mathbb{R}^{r \times k}$ and $W_{\text{up}} \in \mathbb{R}^{d \times r}$, where r is the rank, typically much smaller than $\min(d, k)$.

Given an input h_{in}, the original output $W_0 h_{\text{in}}$ is updated with a task-specific adjustment:

$$h_{\text{out}} = W_0 h_{\text{in}} + \frac{\alpha}{r} W_{\text{up}} W_{\text{down}} h_{\text{in}} \tag{3}$$

where α is a scaling coefficient that balances the adaptation impact. At initialization, W_{up} is set to zero and W_{down} is randomly initialized, ensuring that the model initially behaves as the pretrained base without modification. This strategy allows LoRA to inject task-specific knowledge while preserving the original model's structure and maintaining fast inference.

3.3. Classifier Head

To perform non-verbal event classification, we append a classifier $\mathcal{C}$ to the backbone $\mathcal{B}$ of the Large Speech Model. From the Transformer encoder, we obtain hidden representations across all layers denoted by $\{h_t^l\}$, where $l = 1, \ldots, L$ indexes the layers and $t = 1, \ldots, T$ indexes the sequence frames.

We aggregate these multi-layer representations into a unified sequence $\{h_t^*\}_{t=1}^T$ by applying a learnable weighted sum across layers. This aggregation is formalized by the function $\mathcal{S} : \mathbb{R}^{L \times T \times d} \to \mathbb{R}^{T \times d}$, defined as:

$$h_t^* = \sum_{l=1}^{L} w_l \cdot h_t^l, \quad \forall t \in \{1, \ldots, T\} \tag{4}$$

where each weight w_l satisfies $w_l \geq 0$ and the weights are normalized such that $\sum_{l=1}^{L} w_l = 1$.

The resulting sequence $\{h_t^*\}$ is first projected using a frame-wise linear transformation $\mathcal{L}_1 : \mathbb{R}^d \to \mathbb{R}^m$. Following standard practices in speech emotion recognition [26], we apply temporal aggregation via average pooling $\mathcal{P}$ over the T frames to produce a single vector summarizing the input audio. This pooled representation is then fed into a classification layer $\mathcal{O} : \mathbb{R}^m \to \mathbb{R}^k$, which outputs the logits corresponding to the target classes.

The overall classifier $\mathcal{C}$ can be concisely expressed as:

$$\mathcal{C}\left(\{h_t^*\}_{t=1}^T\right) = \mathcal{O}\left(\mathcal{P}\left(\mathcal{L}_1\left(\{h_t^*\}_{t=1}^T\right)\right)\right) \tag{5}$$

4. Experiments

4.1. Datasets

ASVP-ESD The ASVP-ESD (Audio, Speech and Vision Processing Lab Emotional Sound Database) [29] comprises 12,625 emotion-related audio samples, including both speech and non-speech vocalizations. These samples were collected from movies, YouTube channels, and various other online sources. Each recording is annotated with one of 12 emotion categories, plus an additional "breath" label. All audio files are mono-channel and sampled at 16 kHz.

Table 1
Linear probing results of Large Speech Models are reported for the ASVP-ESD, CNVVE, Non-Verbal Vocalization Dataset, ReCANVo, and VIVAE datasets, using Accuracy and Macro F1 as evaluation metrics. For each dataset, the best results are highlighted in gray.

Model	ASVP ESD		CNVVE		Nonverbal		ReCanVo		ViVAE	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Whisper Tiny	54.75	38.81	80.43	81.01	45.21	44.09	44.50	35.95	36.81	31.75
Whisper Base	59.17	45.06	84.78	84.98	57.53	57.21	45.74	37.31	38.04	36.86
Whisper Small	61.98	46.32	73.91	72.97	57.53	56.49	45.58	35.77	38.04	33.58
HuBERT Base	52.48	35.13	60.87	57.04	47.95	47.81	40.62	30.20	34.97	30.01
WavLM Base Plus	45.25	28.94	45.65	39.14	36.99	37.88	32.09	21.26	20.86	10.07
Wav2Vec2 Base	51.94	34.3	56.52	53.18	47.95	45.24	39.22	32.77	30.06	20.71

CNVVE The Dataset and Benchmark for Classifying Non-verbal Voice Expressions (CNVVE) [7] consists of 950 audio recordings from 42 participants. Each recording is labeled with one of six non-verbal voice expression categories. The audio samples are mono-channel and sampled at 16 kHz.

Non-verbal Vocalization Dataset The Non-verbal Vocalization Dataset[1] includes crowdsourced audio recordings of non-verbal vocalizations categorized into 16 distinct labels. All recordings are sampled at 16 kHz, with 16-bit resolution and mono-channel format.

ReCANVo The Real-World Communicative and Affective Nonverbal Vocalizations (ReCANVo) dataset [30] contains over 7,000 vocalizations produced by minimally speaking individuals aged between 6 and 25 years. Each vocalization is annotated with one of six communicative or affective labels.

VIVAE The Variably Intense Vocalizations of Affect and Emotion (VIVAE) dataset [31] comprises 1,085 audio recordings from 11 speakers. The recordings are sampled at 42 kHz with 16-bit resolution and are annotated with six emotion labels. These labels capture both positive and negative affective states, as well as emotional intensity.

4.2. Metrics

For the experimental evaluation, we report both *Accuracy* and *Macro F1* score. Since the datasets are imbalanced, the macro F1 score offers a more reliable assessment of the model's performance across all classes.

[1] https://www.openslr.org/99/

4.3. Experimental Details

All experiments were conducted using a consistent setup across datasets. Each dataset was split into training, validation, and test sets, with 80% of the audio samples used for training, 10% for validation, and the remaining 10% for testing.

The Large Speech Models evaluated in this study include: Whisper Tiny[2], Whisper Base[3], Whisper Small[4], HuBERT Base[5], WavLM Base Plus[6], and Wav2Vec2 Base[7].

Training was performed for 50 epochs with the following hyperparameters: an initial learning rate of $1e{-}4$, weight decay of 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e{-}8$ for the Adam optimizer. A batch size of 16 was used along with a gradient accumulation step of 2.

All experiments were executed on a single NVIDIA A100 GPU.

4.4. Results

4.4.1. Linear Probing on Large Speech Models

To compare the Large Speech Models introduced in Section 3.1, we adopt a linear probing setup where the backbone $\mathcal{B}$ is kept frozen, and only the classifier $\mathcal{C}$ is trained. In this configuration, each model—Wav2Vec 2.0, HuBERT, WavLM, and Whisper—is used purely as a feature extractor for the Non-Verbal Vocalization task. This approach allows us to evaluate the extent to which task-relevant representations are already captured in the pre-trained models.

Table 1 reports the performance of each model across all datasets, using Accuracy and Macro F1 as evaluation metrics. Results indicate that Wav2Vec 2.0, HuBERT, and

[2] https://huggingface.co/openai/whisper-tiny
[3] https://huggingface.co/openai/whisper-base
[4] https://huggingface.co/openai/whisper-small
[5] https://huggingface.co/facebook/hubert-base-ls960
[6] https://huggingface.co/microsoft/wavlm-base-plus
[7] https://huggingface.co/facebook/wav2vec2-base

Table 2
Comparison of PEFT strategies (LoRA, Adapter, Prompt Tuning) applied to Whisper models. The "Frozen" setting refers to linear probing, where the backbone remains fixed during training. For each model and dataset, the best-performing PEFT method is highlighted in gray.

		ASVP ESD		CNVVE		Nonverbal		ReCanVo		ViVAE	
Model	**Method**	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Whisper Tiny	Frozen	54.75	38.81	80.43	81.01	45.21	44.09	44.50	35.95	36.81	31.75
	LoRA	65.19	54.77	96.74	96.77	56.16	55.64	58.60	52.89	44.79	42.48
	Adapter	58.90	43.64	88.04	87.90	52.05	48.75	50.08	40.22	35.58	33.16
	Prompt Tuning	57.83	41.49	66.30	65.80	52.05	42.56	54.88	46.64	33.74	29.52
Whisper Base	Frozen	59.17	45.06	84.78	84.98	57.53	57.21	45.74	37.31	38.04	36.86
	LoRA	69.21	58.96	97.83	97.87	73.97	74.30	59.84	53.25	47.85	47.43
	Adapter	64.39	55.54	90.22	90.43	75.34	75.48	54.26	50.67	39.88	39.38
	Prompt Tuning	64.12	49.77	77.17	77.62	39.73	34.62	53.18	43.42	36.20	33.12
Whisper Small	Frozen	61.98	46.32	73.91	72.97	57.53	56.49	45.58	35.77	38.04	33.58
	LoRA	72.16	64.17	100.00	100.00	68.49	66.79	58.29	53.72	52.76	52.70
	Adapter	72.16	63.69	85.87	85.94	78.08	78.48	56.90	54.63	40.49	39.20
	Prompt Tuning	70.28	60.97	90.22	90.48	61.64	60.82	56.74	49.44	46.01	44.60

WavLM consistently underperform compared to Whisper, which achieves superior results across all datasets and model sizes (Tiny, Base, and Small).

Notably, the Whisper Base model delivers the best overall performance except on the ASVP-ESD dataset, where Whisper Small slightly outperforms it with a Macro F1 score of 46.32 compared to 45.06 achieved by Whisper Base.

4.4.2. Effect of Parameter-Efficient Fine-Tuning

For evaluating Parameter-Efficient Fine-Tuning (PEFT) techniques, we focus on Whisper models, which demonstrated the strongest performance in the previous section. Table 2 presents the results across different fine-tuning strategies applied to Whisper: Frozen Backbone, LoRA, Adapters, and Prompt Tuning.

Consistent with prior findings in audio classification tasks [26], LoRA emerges as the most effective PEFT method across various datasets and model sizes. However, an exception is observed in the Non-Verbal Vocalization dataset, where Adapters achieve superior performance for both the Whisper Base and Small models.

LoRA's strength lies in its ability to efficiently introduce minimal task-specific parameters while selectively modeling the non-verbal specific update ΔW, allowing it to effectively integrate pre-trained knowledge with new task-specific information.

4.4.3. Analysis of Transformer Layers

This subsection examines the contribution of each Transformer encoder layer within the Whisper backbone to the Non-Verbal Vocalization task. We concentrate on the Whisper model, given its superior performance as shown in Table 1.

For this analysis, we leverage the learned linear probing weights $w_1, \ldots, w_L$ corresponding to the L Transformer layers of the Whisper model. Figure 2 presents the average layer weights across all five datasets used in this study. We observe a consistent trend where deeper layers receive higher weights, indicating that features critical to non-verbal vocalizations are primarily encoded in the later layers. This observation is consistent with previous findings in Speech Emotion Recognition (SER) [32].

More specifically, the layers with the greatest influence vary by Whisper variant: layers 4 and 5 for Whisper Tiny, layers 5, 6, and 7 for Whisper Base, and layers 8 through 13 for Whisper Small.

4.4.4. Optimizing PEFT via Layer Importance

In Section 4.4.2, we applied PEFT techniques uniformly across all Whisper layers, without considering their relative importance. However, as observed in the previous section, different layers contribute unevenly to the Non-Verbal Vocalization task. Therefore, in this subsection, we investigate whether the effectiveness of PEFT depends on layer importance, and if focusing on specific layers can further reduce adaptation parameters.

Table 3 presents different strategies for applying LoRA to Whisper models, as LoRA showed the best performance in most cases. For each model, *LoRA* refers to applying the technique to all Transformer layers, *LoRA[-]* applies LoRA only to the *less important* layers, and *LoRA[+]* applies it exclusively to the *important* layers, as determined in Section 4.4.3.

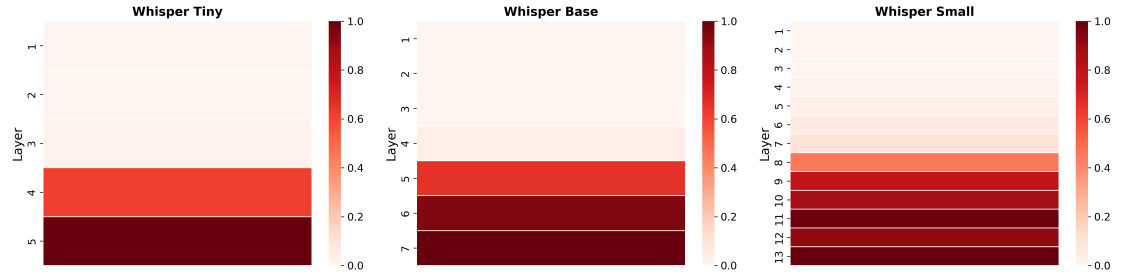

Figure 2: Layer importance scores normalized to the range $[0, 1]$ for Whisper Tiny, Base, and Small models. The importance values are averaged across all Non-Verbal Vocalization datasets. Darker shades correspond to higher importance.

Table 3
Effect of applying LoRA to different Transformer layers according to their importance for the Non-Verbal Vocalization task. LoRA[-] denotes applying LoRA exclusively to less important layers, while LoRA[+] applies it only to important layers. The best performance for each model and dataset is highlighted in gray, and the second best is underlined.

		ASVP ESD		CNVVE		Nonverbal		ReCanVo		ViVAE	
Model	**Method**	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Whisper Tiny	LoRA	65.19	54.77	96.74	96.77	56.16	55.64	58.60	52.89	44.71	42.12
	LoRA [-]	64.93	52.56	95.65	95.66	63.01	62.67	55.04	49.34	44.79	42.48
	LoRA [+]	57.97	41.39	83.7	83.72	41.1	40.33	45.27	37.45	39.88	38.12
Whisper Base	LoRA	69.21	58.96	97.83	97.87	73.97	74.30	59.84	53.25	47.85	47.43
	LoRA [-]	68.81	56.43	97.83	97.81	73.97	74.41	60.0	56.62	44.79	44.6
	LoRA [+]	62.25	48.08	84.78	85.18	58.9	59.27	52.25	44.14	44.17	42.43
Whisper Small	LoRA	72.16	64.17	100.00	100.00	68.49	66.79	58.29	53.72	52.76	52.70
	LoRA [-]	73.90	64.43	93.48	93.51	68.49	65.92	55.04	49.34	52.56	52.45
	LoRA [+]	68.67	56.83	93.48	93.55	69.86	68.70	45.27	37.45	45.60	46.21

Overall, we find that full LoRA adaptation typically yields the best results, followed by LoRA[-]. This suggests that adapting the less important layers has a greater positive impact than focusing solely on the important layers, for which performance is often significantly lower. Although this may seem counterintuitive, we hypothesize that adaptation is more necessary where the network retains less prior knowledge relevant to the task. Important layers already encode useful features, thus requiring less adjustment, while ignoring the less important layers limits the model's adaptability.

Hence, we propose that focusing on the less important layers is more beneficial than concentrating exclusively on the important ones. This insight offers valuable guidance for future work aimed at improving PEFT techniques by targeting the parts of the network that need the most adaptation.

5. Conclusion

In this work, we investigated the adaptability of Large Speech Models (LSMs) to Non-Verbal Vocalization (NVV) tasks using both linear probing and Parameter-Efficient Fine-Tuning (PEFT) techniques. Our experimental results demonstrate that Whisper models consistently outperform Wav2Vec 2.0, HuBERT, and WavLM across multiple NVV datasets.

Furthermore, we observe that applying PEFT methods significantly improves performance, with LoRA emerging as the most effective strategy compared to Adapters and Prompt Tuning. Through a detailed analysis of the Transformer layer weights in Whisper models, we find that non-verbal information is predominantly captured in the later layers.

Interestingly, we discover that fine-tuning only these later layers yields limited gains compared to adapting the layers that initially contain less non-verbal knowledge. We hypothesize that this is because the layers with less task-relevant information require a larger degree of adaptation to bridge the knowledge gap. This observation suggests a valuable pathway for optimizing PEFT methods by selectively targeting particular transformer layers based on the knowledge they embed, potentially minimizing the need for additional task-specific parame-

ters even further.

References

[1] P. Tzirakis, A. Baird, J. Brooks, C. Gagne, L. Kim, M. Opara, C. Gregory, J. Metrick, G. Boseck, V. Tiruvadi, et al., Large-scale nonverbal vocalization detection using transformers, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[2] T. Feng, S. Narayanan, Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 12116–12120.

[3] E. Liebenthal, D. A. Silbersweig, E. Stern, The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception, Frontiers in neuroscience 10 (2016) 506.

[4] A. Cowen, D. Sauter, J. L. Tracy, D. Keltner, Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression, Psychological Science in the Public Interest 20 (2019) 69–90.

[5] J. McCormack, S. McLeod, L. J. Harrison, L. McAllister, The impact of speech impairment in early childhood: Investigating parents' and speech-language pathologists' perspectives using the icf-cy, Journal of communication disorders 43 (2010) 378–396.

[6] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

[7] R. Hedeshy, R. Menges, S. Staab, Cnvve: Dataset and benchmark for classifying non-verbal voice (2023).

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[9] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM transactions on audio, speech, and language processing 29 (2021) 3451–3460.

[11] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.

[13] Z. Han, C. Gao, J. Liu, S. Q. Zhang, et al., Parameter-efficient fine-tuning for large models: A comprehensive survey, arXiv preprint arXiv:2403.14608 (2024).

[14] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: International conference on machine learning, PMLR, 2019, pp. 2790–2799.

[15] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, arXiv preprint arXiv:2104.08691 (2021).

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[17] J. Bilmes, X. Li, J. Malkin, K. Kilanski, R. Wright, K. Kirchhoff, A. Subramanya, S. Harada, J. Landay, P. Dowden, et al., The vocal joystick: A voice-based human-computer interface for individuals with motor impairments, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 995–1002.

[18] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, R. Palmer, A speech-controlled environmental control system for people with severe dysarthria, Medical Engineering & Physics 29 (2007) 586–593.

[19] A. Koudounas, M. La Quatra, S. M. Siniscalchi, E. Baralis, voc2vec: A foundation model for non-verbal vocalization, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin,

J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.
[23] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: an asr corpus based on public domain audio books, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 5206–5210.
[24] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, et al., Libri-light: A benchmark for asr with limited or no supervision, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7669–7673.
[25] L. Pepino, P. Riera, L. Ferrer, Emotion recognition from speech using wav2vec 2.0 embeddings, arXiv preprint arXiv:2104.03502 (2021).
[26] T. Feng, S. Narayanan, Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models, in: 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2023, pp. 1–8.
[27] T. Feng, R. Hebbar, S. Narayanan, Trust-ser: On the trustworthiness of fine-tuning pre-trained speech embeddings for speech emotion recognition, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 11201–11205.
[28] H. Jegou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, IEEE transactions on pattern analysis and machine intelligence 33 (2010) 117–128.
[29] D. Landry, Q. He, H. Yan, Y. Li, Asvp-esd: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances, Global Scientific Journals 8 (2020) 1793–1798.
[30] K. T. Johnson, J. Narain, T. Quatieri, P. Maes, R. W. Picard, Recanvo: A database of real-world communicative and affective nonverbal vocalizations, Scientific Data 10 (2023) 523.
[31] N. Holz, P. Larrouy-Maestri, D. Poeppel, The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception., Emotion 22 (2022) 213.
[32] F. D'Asaro, J. J. M. Villacís, G. Rizzo, A. Bottino, Using large speech models for feature extraction in cross-lingual speech emotion recognition, in: Titolo volume non avvalorato, Accademia University Press, 2024.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) and Grammarly in order to: Drafting content, Text translation, Paraphrase and reword, and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.