



# Towards Pre-training an Effective Respiratory Audio Foundation Model

Daisuke Niizumi<sup>1</sup>, Daiki Takeuchi<sup>1</sup>, Masahiro Yasuda<sup>1</sup>, Binh Thien Nguyen<sup>1</sup>, Yasunori Ohishi<sup>1</sup>,  
Noboru Harada<sup>1</sup>

<sup>1</sup>NTT Communication Science Laboratories, NTT Corporation, Japan

daisuke.niizumi@ntt.com, d.takeuchi@ntt.com, masahiro.yasuda@ntt.com,  
binhthien.nguyen@ntt.com, yasunori.ohishi@ntt.com, harada.noboru@ntt.com

## Abstract

Recent advancements in foundation models have sparked interest in respiratory audio foundation models. However, the effectiveness of applying conventional pre-training schemes to datasets that are small-sized and lack diversity has not been sufficiently verified. This study aims to explore better pre-training practices for respiratory sounds by comparing numerous pre-trained audio models. Our investigation reveals that models pre-trained on AudioSet, a general audio dataset, are more effective than the models specifically pre-trained on respiratory sounds. Moreover, combining AudioSet and respiratory sound datasets for further pre-training enhances performance, and preserving the frequency-wise information when aggregating features is vital. Along with more insights found in the experiments, we establish a new state-of-the-art for the OPERA benchmark, contributing to advancing respiratory audio foundation models. Our code is available online.<sup>1</sup>

**Index Terms:** audio foundation model, pre-training, respiratory sound

## 1. Introduction

As a trend in non-invasive diagnostic methods using sound [1], respiratory audio foundation models have gained attention, driven by recent advancements in deep learning foundation models [2, 3, 4]. These models may serve as a fundamental block for health monitoring and disease diagnosis applications. Since training foundation models requires large-scale data, existing methods have curated their respiratory sound datasets, trained models using established pre-training schemes, and demonstrated their effectiveness through benchmarks [4, 5].

However, the datasets used for training these respiratory audio models are small and lack diversity compared to those for general audio models. While general audio datasets contain a wide variety of sounds on a large scale, respiratory audio datasets are more than ten times smaller than general audio ones [4, 5], and its sounds, such as coughs and breaths, are monotonic. In addition, the effectiveness of the conventional pre-training schemes using such a small and monotonic dataset is unclear.

Towards pre-training a respiratory audio foundation model that is practically effective, we empirically investigate what makes the pre-trained models advantageous for respiratory sound analysis. To do this, we evaluate numerous pre-trained audio models through a unified respiratory task benchmark and compare the results to explore better practices. The various models have taken different pre-training approaches; thus, comparing them on a unified benchmark can surface the factors of

<sup>1</sup><https://github.com/nttcs-lab/eval-audio-repr/tree/main/plugin/OPERA>

Table 1: OPERA benchmark tasks used in our study.

Dataset	ID	Task	Modality	Class samples
COUGHVID [6]	T5	Covid/Non-covid	Cough	547 / 5628
	T6	Female/Male	Cough	2468 / 4795
ICBHI [7]	T7	COPD <sup>†</sup> /Healthy	Lung sounds	793 / 35
Coswara [8]	T8	Smoker/Non-smoker	Cough	201 / 747
	T9	Female/Male	Cough	759 / 1737
KAUH [9]	T10	Obstructive/Healthy	Lung sounds	129 / 105
Resp.@TR [10]	T11	COPD <sup>†</sup> severity (0-4)	Lung sounds	72/60/84/84/204

<sup>†</sup>Chronic obstructive pulmonary disease, a common lung disease.

effective pre-training for respiratory sounds. Our research questions are the following.

- **RQ1** What are the better practices for pre-training effective respiratory audio foundation models?
- **RQ2** What data lead to learning effective features?
- **RQ3** Are the intermediate layer features effective?

The experiments reveal various insights, such as the advantage of combining a diverse large-scale dataset with respiratory datasets and the importance of preserving frequency-wise information in feature aggregation. In addition, our results renew the state-of-the-art (SOTA) performance on the OPERA benchmark with a large margin. Our evaluation code is available online for future advancement of respiratory audio foundation models.

## 2. Experimental Setup

We utilized a unified respiratory audio benchmark to evaluate models under the following experimental setup.

### 2.1. Benchmark

We conducted experiments on the publicly available OPERA [4] benchmark and used seven tasks (T5 to T11) for which the data are publicly available. The seven tasks listed in Table 1 are health condition classification problems, where all are binary classification tasks except T11, which is a five-class classification. The benchmark takes a linear evaluation protocol, where model weights are frozen. A model encodes task data samples into features, and a linear layer is trained to classify the task using the features as input. The performance metric is AUROC (Area Under the Receiver Operating Characteristic), and the results are the statistics for five attempts.

### 2.2. Evaluated Models

We used audio foundation models with various backgrounds. For models trained specifically on respiratory sounds, we used OPERA-CT/GT [4]. For models trained on speech similar to respiratory sounds, we used speech self-supervised learning (SSL) models wav2vec 2.0 [14], HuBERT [15], and WavLM [16], which are all base models. For models trained

Table 2: Performance comparison among audio foundation models on OPERA benchmark.

Model	Task	T5 Covid	T6 Gender	T7 COPD	T8 Smoker	T9 Gender	T10 Obstructive	T11 COPD (5 cls)	Avg.
<i>Supervised learning models.</i>									
1. AST [11]		0.607 ±0.008*	0.739 ±0.002	<b>1.000 ±0.000*</b>	0.671 ±0.011	0.833 ±0.000	0.837 ±0.007*	0.652 ±0.020*	0.763*
2. PANNs Cnn14 [12]		0.533 ±0.006	0.566 ±0.003	0.606 ±0.037	0.533 ±0.014	0.549 ±0.007	0.447 ±0.057	0.527 ±0.022	0.537
3. HTS-AT [13]		0.577 ±0.002	0.615 ±0.001	0.765 ±0.013	0.590 ±0.010	0.699 ±0.002	0.778 ±0.014*	0.521 ±0.031	0.649
<i>Speech SSL models (with the best performing layer used for extracting features).</i>									
4. wav2vec2 <sub>Layer#7</sub> [14]		0.480 ±0.005	0.634 ±0.003	0.172 ±0.013	0.589 ±0.020	0.606 ±0.004	0.620 ±0.021	0.560 ±0.022	0.523
5. HuBERT <sub>Layer#7</sub> [15]		0.558 ±0.002	0.736 ±0.001	0.644 ±0.012	0.683 ±0.004	0.807 ±0.002	0.689 ±0.019	0.658 ±0.018*	0.682
6. WavLM <sub>Layer#6</sub> [16]		0.555 ±0.002	0.700 ±0.001	0.599 ±0.016	0.687 ±0.004*	0.771 ±0.001	0.703 ±0.020	0.624 ±0.011	0.663
<i>CLAP models.</i>									
7. LAION-CLAP [17]		0.549 ±0.001	0.660 ±0.001	0.674 ±0.062	0.531 ±0.003	0.714 ±0.002	0.776 ±0.015*	0.584 ±0.033	0.641
8. CLAP <sub>2022</sub> [18]		0.599 ±0.007*	0.665 ±0.001	0.933 ±0.005*	0.680 ±0.009	0.742 ±0.001	0.697 ±0.004	0.636 ±0.045*	0.707
9. CLAP <sub>2023</sub> [19]		0.602 ±0.007*	0.779 ±0.001	0.988 ±0.004*	0.687 ±0.007*	0.866 ±0.001	0.795 ±0.012*	0.606 ±0.037	0.760*
<i>General audio SSL models.</i>									
10. BYOL-A [20]		0.531 ±0.011	0.702 ±0.006	0.950 ±0.031*	0.581 ±0.035	0.807 ±0.011	0.710 ±0.047	0.566 ±0.029	0.693
11. ATST-Clip [21]		0.609 ±0.009*	0.793 ±0.001	0.977 ±0.007	0.679 ±0.010	0.850 ±0.002	0.828 ±0.012	0.606 ±0.029	0.763*
12. ATST-Frame [21]		<b>0.621 ±0.007*</b>	<b>0.801 ±0.001*</b>	0.998 ±0.001*	0.687 ±0.010*	<b>0.908 ±0.001*</b>	<b>0.843 ±0.006*</b>	0.657 ±0.003*	<b>0.788*</b>
13. AudioMAE [22]		0.554 ±0.004	0.628 ±0.001	0.886 ±0.017*	0.549 ±0.022	0.724 ±0.001	0.616 ±0.041	0.510 ±0.021	0.638
14. BEATs [23]		0.555 ±0.002	0.644 ±0.001	0.823 ±0.011	0.631 ±0.004	0.695 ±0.003	0.723 ±0.031*	0.623 ±0.015	0.670
15. MSM-MAE [24]		0.569 ±0.003	0.781 ±0.000	<b>1.000 ±0.000*</b>	<b>0.721 ±0.008*</b>	0.879 ±0.001*	0.746 ±0.009*	0.662 ±0.006*	0.765*
16. M2D [25]		0.595 ±0.008*	0.797 ±0.000*	<b>1.000 ±0.000*</b>	0.703 ±0.024*	0.905 ±0.001*	0.756 ±0.013*	<b>0.720 ±0.012*</b>	0.782*
<i>Ensemble SSL model (CED), large parameter model (Dasheng 1.2B), and audio-visual contrastive SSL model (OpenL3).</i>									
17. CED [26]		0.614 ±0.001*	0.782 ±0.001	0.997 ±0.001*	0.713 ±0.007*	0.873 ±0.001	0.833 ±0.019*	0.597 ±0.117	0.773*
18. Dasheng-1.2B [27]		0.582 ±0.005*	0.734 ±0.002	0.915 ±0.031*	0.662 ±0.016	0.772 ±0.002	0.700 ±0.072	0.660 ±0.021*	0.718
19. OpenL3[28]		0.608 ±0.011*	0.754 ±0.006	0.978 ±0.007*	0.695 ±0.018*	0.845 ±0.007	0.751 ±0.027*	0.639 ±0.025*	0.753*
<i>Respiratory audio SSL models.</i>									
20. OPERA-CT [4]		0.578 ±0.001	0.795 ±0.001	0.855 ±0.012	0.685 ±0.012	0.874 ±0.000	0.722 ±0.016	0.625 ±0.038	0.733
21. OPERA-GT [4]		0.552 ±0.003	0.735 ±0.000	0.741 ±0.011	0.650 ±0.005	0.825 ±0.001	0.703 ±0.016	0.606 ±0.015	0.687
<i>Reference from Table 3: M2D further pre-trained on AudioSet + Respiratory sound data. Bold results are better than the models above.</i>									
M2D+Resp		<b>0.627 ±0.009*</b>	<b>0.856 ±0.001*</b>	<b>1.000 ±0.001*</b>	<b>0.757 ±0.004*</b>	<b>0.954 ±0.001*</b>	0.794 ±0.016*	0.714 ±0.007*	<b>0.814*</b>

\*Results better than OPERA-CT, the previous SOTA, pre-trained on respiratory sounds only.

on general audio from AudioSet, we utilized supervised learning models PANNs [12], AST [11], and HTS-AT [13], as well as SSL models Audio-MAE [22], MSM-MAE [24], BEATs [23], BYOL-A [20], ATST-Clip/Frame [21], and M2D [25]. Most of the models are base-sized transformers with about 90M parameters, except for PANNs and BYOL-A.

Additionally, we examined CED [26] (distills multiple Masked Autoencoders (MAE) [29]), Dasheng [27] (pre-trains MAE on large datasets and parameters), OpenL3 [28] (a multi-modal SSL connecting video and audio), and CLAP (contrastive language-audio pre-training) models [17, 18, 19]. These models employ diverse pre-training paradigms, datasets, network architectures, and output feature aggregation techniques.

### 2.3. Pre-training Dataset

To evaluate the impact of the pre-training dataset, we utilized respiratory sound databases, a speech corpus LibriSpeech [30], and a general audio dataset AudioSet [31]. For respiratory sounds, we employed COUGHVID [6], HF\_Lung [32], and ICBHI2017 [7], which form a subset of the data used in OPERA. COUGHVID consists of cough sounds recorded via microphones, and we used 7054 samples with a cough detection ratio of 0.95 or higher. HF\_Lung contains lung sounds recorded via a stethoscope and another device; we used 3839 samples of stethoscope recordings. ICBHI2017 also consists of lung sounds recorded via stethoscopes, comprising 539 samples. To balance the amount of cough and lung sound data, we augmented the ICBHI2017 sample list sixfold and combined it with the lists from COUGHVID and HF\_Lung, resulting in a total of 14,127 samples used as the respiratory sound dataset.

AudioSet, widely used in general audio models, comprises various sounds, such as music, speech, and environmental sounds. It is worth mentioning that AudioSet also includes classes such as respiratory sounds, breathing, and coughing, similar to respiratory databases. We utilized 2,005,132 samples from the balanced/unbalanced train segments.

## 3. Empirical Analysis

We evaluated 21 audio foundation models on the OPERA benchmark (Table 2) and conducted ablation experiments on the M2D model (Table 3) and on data (Table 4). We also evaluated the layer-wise performance (Figures 1 and 2).

### 3.1. RQ1 What are the better practices for pre-training effective respiratory audio foundation models?

#### 3.1.1. Comparing benchmark results for all models

Table 2 shows the results for 21 audio foundation models and a reference model. We compare these models based on their pre-training settings and performance and especially focus on comparing models with the respiratory audio foundation models OPERA-CT/GT.

**Pre-training with a sufficiently large and diverse audio dataset is more advantageous than using only respiratory sound data.** The results shows that eight out of 19 general audio models outperform the respiratory audio model OPERA-CT (The asterisks \* in the table indicate a better performance than OPERA-CT). As the eight models that outperform OPERA-CT employ a variety of pre-training backgrounds while using AudioSet, the pre-training data may have a greater impact on performance than other factors (e.g., learning methods). Notably, while using the respiratory audio dataset should be advantageous for learning respiratory audio features, AudioSet also contains these sounds, as described in Section 2.3. Besides, the limited performance of OPERA-GT (which employs the SSL method MAE [29]) may indicate that the pre-training task of predicting masked parts becomes too simple with monotonous respiratory sound data, making it challenging to learn useful features. To summarize, the observations indicate that pre-training on AudioSet’s large-scale data with approximately 2M samples ( $\approx 5569$ h) is more effective than pre-training on the respiratory sound data in OPERA with 136K samples (404.1h).

Table 3: M2D ablations on OPERA benchmark.

Model	Task	T5 Covid	T6 Gender	T7 COPD	T8 Smoker	T9 Gender	T10 Obstructive	T11 COPD (5 cls)	Avg.
M2D [25]		0.595 ±0.008	0.797 ±0.000	1.000 ±0.000	0.703 ±0.024	0.905 ±0.001	0.756 ±0.013	0.720 ±0.012	0.782
<i>Summarizing output features by mean pooling instead of concatenating frequency-wise features.</i>									
(i) Mean pooling		0.593 ±0.003	0.728 ±0.002	0.955 ±0.004	0.642 ±0.015	0.754 ±0.002	0.728 ±0.030	0.582 ±0.068	0.712
<i>Training objective ablations.</i>									
(ii) M2D ftAS [25]		0.597 ±0.012	0.813 ±0.001	0.999 ±0.000	0.732 ±0.007	0.907 ±0.001	0.827 ±0.025	0.649 ±0.113	0.789
(iii) M2D-CLAP [33]		0.604 ±0.007	0.809 ±0.000	0.995 ±0.004	0.737 ±0.032	0.910 ±0.001	0.752 ±0.039	0.720 ±0.049	0.790
(iv) M2D-S [34]		0.547 ±0.007	0.660 ±0.001	0.881 ±0.031	0.591 ±0.027	0.727 ±0.001	0.725 ±0.019	0.554 ±0.049	0.669
<i>Feature time-frame resolution ablations with patch sizes of <math>16 \times 4</math> and <math>80 \times 2</math>.</i>									
(v) 40ms ( $16 \times 4$ )		0.577 ±0.002	0.787 ±0.001	1.000 ±0.000	0.706 ±0.007	0.880 ±0.001	0.738 ±0.018	0.630 ±0.025	0.760
(vi) 20ms ( $80 \times 2$ )		0.582 ±0.005	0.778 ±0.001	0.956 ±0.006	0.678 ±0.007	0.882 ±0.003	0.757 ±0.061	0.663 ±0.135	0.757

**Although speech shares the same pathway as breath sounds, speech SSLs underperform other models.** One reason could be that these models utilize training signals obtained through clustering features of a speech signal. While they contain phonemes and other linguistic information, they are likely to have fewer non-linguistic breath features, thus making the models learn less to represent respiratory features. Similarly, BEATs, which use clustered features as training signals for pre-training on AudioSet, exhibit limited performance, suggesting that respiratory sounds are not well-represented when learned from clustered features. Furthermore, speech datasets are generally clear speech recordings, likely not to include trainable respiratory sounds, potentially resulting in the lower performance of the speech SSLs.

**SSL models learned only from audio excels across benchmark tasks.** Models employing a wide range of learning methods have demonstrated their effectiveness. However, AST, supervised learning, CLAP<sub>2023</sub>, contrastive learning with audio captions, and OpenL3, contrastive learning with videos, tend to perform slightly lower in gender classification tasks (T6 and T9). In contrast, models with masked prediction-based learning only from audio (ATST-Frame utilizing data augmentation, M2D enhancing MAE for training signal and prediction task, and CED distilling many MAE models) achieve consistently strong performance across tasks. Notably, ATST-Frame and M2D outperform OPERA in all tasks.

**Increasing network parameters or the amount of data used does not necessarily improve performance.** Dasheng, pre-trained on a large-scale dataset of 1.2B parameters and 97M samples, achieves performance comparable to BYOL-A, which has 5M parameters pre-trained on a dataset with 2M samples. In the case of the CLAP model, CLAP<sub>2023</sub>, pre-trained on 4.6M samples, shows significant performance improvement, especially in gender classification, compared to CLAP<sub>2022</sub>, pre-trained on 128K samples. On the other hand, the top-performing models, ATST and M2D trained on the 2M-sample AudioSet dataset using about 90M parameters.

Finally, the reference result from the M2D ablation study, M2D+Resp, demonstrates a new SOTA performance on the OPERA benchmark by further pre-training M2D using respiratory sound data. The M2D+Resp shows an average result of 0.814, outperforming the former SOTA of the OPERA-CT’s 0.733 with a large margin. The improvement highlights the potential for further advancements by tailoring the top-performing model specifically to respiratory sound analysis.

### 3.1.2. Ablations using M2D

We conducted ablation experiments using the top-performing model, M2D, to determine the impact on the feature aggregation, training objectives, and feature time-frame resolution, and

Table 3 shows the results.

**Preserving frequency-wise information in feature aggregation is crucial.** Models using vision transformers divide spectrograms into  $16 \times 16$  patches and encode each patch into features (e.g., AST, BEATs, M2D, and OPERA-GT). While most approaches aggregate the features for an input audio clip by averaging the patch features, M2D and MSM-MAE concatenate frequency-wise features for a time frame, preserving information for each frequency in the aggregated features. The results show that (i) Mean pooling, which simply averages the M2D output features for patches, yields an average task performance of 0.712, a significant degradation from the original M2D’s 0.782. In particular, tasks T6 and T9, which involve gender classification, become more challenging, suggesting that frequency components are critical for identifying gender from cough sounds. Similarly, the performance degradation in T11, which involves COPD (chronic obstructive pulmonary disease, a common lung disease) severity classification, highlights the importance of frequency components for estimating severity. In the case of (vi) M2D with a patch size of  $80 \times 2$ , where each patch represents the entire frequency range of a single time frame, the configuration shows less degradation than (i) Mean pooling. Similarly, ATST, which preserves the entire frequency in a patch, also shows strong performance in Table 2. These observations indicate that, as reported in the previous study for general audio tasks [35], preserving frequency-wise information is crucial and effective for the respiratory benchmark.

**Learning representations from the acoustic patterns of AudioSet samples is more important than learning them from its semantics or labels.** In the training objective ablations, both (ii) M2D ftAS, an M2D further fine-tuned with AudioSet labels, and (iii) M2D-CLAP, which jointly learns from M2D and CLAP using captions, learn representations from the data distributions of labels or semantics as well as from the acoustic patterns via M2D training. However, their average performance is close to M2D’s 0.782. Additionally, (iv) M2D-S, trained with speech feature clusters of LibriSpeech to learn speech patterns, shows a significant performance drop. These observations suggest that learning acoustic patterns from the AudioSet samples is more effective for the respiratory sound benchmark.

Furthermore, adjusting the patch size (v) and (vi) to narrow the temporal resolution of features does not contribute to performance improvement.

### 3.2. RQ2 What data lead to learning effective features?

We discuss the impact of the datasets on performance based on the results for M2D pre-trained on various datasets in Table 4. We used the same setting with M2D and replaced the pre-training dataset. For further pre-training experiments, we pre-trained the AudioSet-pre-trained M2D for 50 epochs on each

Table 4: Data ablations of M2D on OPERA benchmark. The respiratory pre-training data contains the data from tasks with  $\in \text{Resp}$ .

Pre-training: the data used <sup>†</sup>	Task	T5 $\in \text{Resp}$ Covid	T6 $\in \text{Resp}$ Gender	T7 $\in \text{Resp}$ COPD	T8 Smoker	T9 Gender	T10 Obstructive	T11 COPD (5 cls)	Avg.
Scratch: AS only (M2D) [25]		0.595 $\pm$ 0.008	0.797 $\pm$ 0.000	<b>1.000</b> $\pm$ 0.000	0.703 $\pm$ 0.024	0.905 $\pm$ 0.001	0.756 $\pm$ 0.013	0.720 $\pm$ 0.012	0.782
Fur: Resp only		0.612 $\pm$ 0.004	0.832 $\pm$ 0.000	0.990 $\pm$ 0.008	0.717 $\pm$ 0.021	0.922 $\pm$ 0.001	0.773 $\pm$ 0.016	0.635 $\pm$ 0.022	0.783
Fur: AS+Resp 100K		0.608 $\pm$ 0.010	0.847 $\pm$ 0.000	0.999 $\pm$ 0.000	0.761 $\pm$ 0.011	0.948 $\pm$ 0.001	0.752 $\pm$ 0.014	0.705 $\pm$ 0.019	0.803
Fur: AS+Resp 200K		0.617 $\pm$ 0.011	0.848 $\pm$ 0.001	0.999 $\pm$ 0.002	0.758 $\pm$ 0.004	0.948 $\pm$ 0.001	0.792 $\pm$ 0.048	0.689 $\pm$ 0.066	0.807
Fur: AS+Resp 300K		0.610 $\pm$ 0.009	0.851 $\pm$ 0.001	0.998 $\pm$ 0.004	<b>0.768</b> $\pm$ 0.004	0.951 $\pm$ 0.001	0.773 $\pm$ 0.030	<b>0.724</b> $\pm$ 0.026	0.811
Fur: AS+Resp 400K (M2D+Resp)		0.627 $\pm$ 0.009	<b>0.856</b> $\pm$ 0.001	<b>1.000</b> $\pm$ 0.001	0.757 $\pm$ 0.004	<b>0.954</b> $\pm$ 0.001	<b>0.794</b> $\pm$ 0.016	0.714 $\pm$ 0.007	<b>0.814</b>
Fur: AS+Resp 500K		0.632 $\pm$ 0.006	<b>0.856</b> $\pm$ 0.001	0.995 $\pm$ 0.007	0.758 $\pm$ 0.002	0.953 $\pm$ 0.001	0.760 $\pm$ 0.021	0.710 $\pm$ 0.009	0.809
Scratch: AS+Resp 400K		<b>0.644</b> $\pm$ 0.005	0.854 $\pm$ 0.000	0.981 $\pm$ 0.010	0.745 $\pm$ 0.003	0.942 $\pm$ 0.000	0.755 $\pm$ 0.010	0.591 $\pm$ 0.010	0.787
Scratch: LibriSpeech		0.582 $\pm$ 0.002	0.725 $\pm$ 0.001	0.833 $\pm$ 0.008	0.662 $\pm$ 0.006	0.747 $\pm$ 0.016	0.747 $\pm$ 0.036	0.595 $\pm$ 0.035	0.699

<sup>†</sup>Scratch: pre-training from scratch; Fur: further pre-training on M2D; AS: AudioSet; Resp: the set of respiratory data and size. The size of the respiratory sound data (Resp) is 14K, and the versions of 100, 200, 300, 400, and 500K were created by augmenting the data list by 7, 15, 22, 29, and 36 times, respectively, to increase the proportion relative to AudioSet with 2M samples.

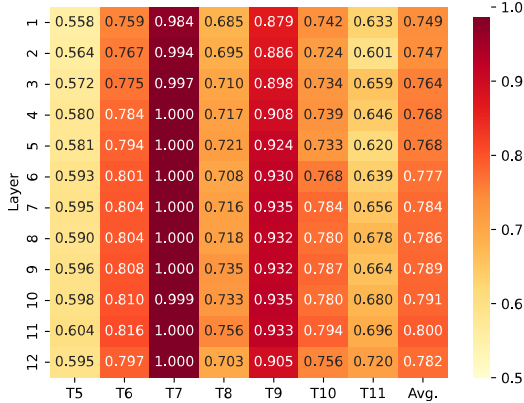


Figure 1: M2D performance by layers.

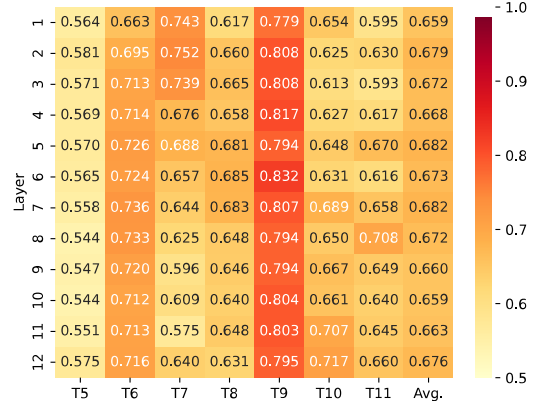


Figure 2: HuBERT performance by layers.

dataset. Note that the respiratory sound dataset used for the pre-training contains the training data from tasks T5 to T7.

**Further pre-training on data combining AudioSet and respiratory sound datasets improves performance.** In the experiments, we pre-trained M2D from scratch or further pre-trained it on various datasets. Among various attempts, further pre-trainings of AS+Resp 100 to 500K consistently demonstrated average performance improvement. Further pre-training on the combination of AudioSet and respiratory sound for 400K samples (M2D+Resp) shows the best performance.

When further pre-trained using only respiratory sounds (*Fur: Resp only*), the average performance becomes comparable to that of the original M2D. In this experiment, we employed the M2D-X [25] framework using AudioSet as background noise to achieve successful pre-training on a small-sized dataset, which showed the best performance in the preliminary study. The results suggest that additional training with only respiratory sounds does not further enhance the effectiveness of the representations gained from AudioSet.

Pre-training from scratch on the AS+Resp 400K dataset (*Scratch: AS+Resp 400K*) improves the performance for the cough sound tasks (T5, T6, T8, and T9) while degrading the performance for other lung sound tasks. Furthermore, training solely with the speech corpus LibriSpeech in M2D also underperforms even with no feature clustering as done in M2D-S.

To summarize the observations, the most effective approach that updates the SOTA on the OPERA benchmark is to further pre-train the AudioSet pre-trained M2D on the combination of AudioSet and respiratory sound data. While this combination can refine the learned effective representations for respiratory sounds, we leave the background reason for future investigation.

### 3.3. RQ3 Are the intermediate layer features effective?

We evaluated each layer’s output and investigated their performance on the benchmark as in the previous study [35]. We tested M2D from top-performing models and the best-performing speech SSL, HuBERT. We specifically added a speech SSL because layer-wise feature performance is crucial in the speech domain. Figures 1 and 2 shows the results.

HuBERT performs better in its earlier layers for some tasks, while the performance gain is limited. In particular, the early layers of HuBERT perform better on task T7, while the middle layers consistently outperform the later layers on other tasks except T10. Similar to speech domain practice, utilizing features from the middle layers may lead to better performance.

In contrast, the deeper layers of M2D consistently show better performance across all tasks. This observation confirms that the typical use of the last-layer features of general audio models is also effective for respiratory sound tasks. Furthermore, the penultimate layer demonstrates the highest average performance, suggesting that leveraging this layer could be most effective for respiratory sound tasks.

## 4. Conclusion

We investigated the better practices for pre-training an effective respiratory audio foundation model by comparing numerous audio models under the unified respiratory benchmark OPERA. Experiments provided various insights, such as the effectiveness of pre-training on general large-scale audio as well as respiratory sound datasets and the significance of preserving frequency-wise information in feature aggregation. We renewed SOTA performance on the OPERA benchmark by a large margin. Our code is available online to contribute to the progress of respiratory audio foundation models.

## 5. References

- [1] J. Cook, M. Umar, F. Khalili, and A. Taebi, "Body acoustics for the non-invasive diagnosis of medical conditions," *Bioengineering*, vol. 9, no. 4, 2022.
- [2] S. Baur, Z. Nabulsi, W.-H. Weng, J. Garrison, L. Blankemeier, S. Fishman, C. Chen, S. Kakarmath, M. Maimbolwa, N. Sanjase, B. Shuma, Y. Matias, G. S. Corrado, S. Patel, S. Shetty, S. Prabhakara, M. Muiyoyeta, and D. Ardila, "HeAR – health acoustic representations," *arXiv preprint arXiv:2403.02522*, 2024.
- [3] G. Mathew, D. Barbosa, J. Prince, and S. a. Venkatraman, "Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes," *npj Cardiovascular Health*, 2024.
- [4] Y. Zhang, T. Xia, J. Han, Y. Wu, G. Rizos, Y. Liu, M. Mosuily, J. Chauhan, and C. Mascolo, "Towards open respiratory acoustic foundation models: Pretraining and benchmarking," in *NeurIPS*, 2024.
- [5] P. Wang, Z. Zhao, L. Zhao, M. He, X. Sun, Y. Zhang, K. Sun, Y. Wang, and Y. Wang, "Auscultabase: A foundational step towards ai-powered body sound diagnostics," *arXiv preprint arXiv:2411.07547*, 2024.
- [6] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 156, 2021.
- [7] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, N. Maglaveras, R. P. Paiva, I. Chouvarda, and P. de Carvalho, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological Measurement*, vol. 40, p. 035001, Mar 2019.
- [8] D. Bhattacharya *et al.*, "Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection," *Scientific Data*, vol. 10, no. 397, 2023.
- [9] M. Fraiwan, L. Fraiwan, B. Khassawneh, and A. Ibnian, "A dataset of lung sounds recorded from the chest wall using an electronic stethoscope," *Data in Brief*, vol. 35, p. 106913, 2021.
- [10] G. Altan, Y. Kutlu, Y. Garbi, A. O. Pekmezci, and S. Nural, "Multimedia respiratory database (respiratorydatabase@tr): Auscultation sounds and chest x-rays," *Natural and Engineering Sciences*, vol. 2, no. 3, p. 59–72, 2017.
- [11] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Interspeech*, 2021, pp. 571–575.
- [12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020.
- [13] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP*, 2022, pp. 646–650.
- [14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.
- [15] 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," *IEEE/ACM Trans. Audio, Speech, Language Process.*, p. 3451–3460, 2021.
- [16] 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 J. Sel. Top. Signal Process.*, vol. 16, no. 6, p. 1505–1518, 2022.
- [17] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *ICASSP*, 2023.
- [18] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," in *ICASSP*, 2023.
- [19] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *ICASSP*, 2024, pp. 336–340.
- [20] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Exploring Pre-trained General-purpose Audio Representations," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, p. 137–151, 2023.
- [21] X. Li, N. Shao, and X. Li, "Self-Supervised Audio Teacher-Student Transformer for Both Clip-Level and Frame-Level Tasks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 1336–1351, 2024.
- [22] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in *NeurIPS*, 2022.
- [23] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATS: Audio Pre-Training with Acoustic Tokenizers," in *ICML*, 2023.
- [24] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Spectrogram Modeling using Masked Autoencoders for Learning General-purpose Audio Representation," in *HEAR: Holistic Evaluation of Audio Representations (NeurIPS 2021 Competition)*, vol. 166, 2022, pp. 1–24.
- [25] —, "Masked Modeling Duo: Towards a Universal Audio Pre-Training Framework," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 2391–2406, 2024.
- [26] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "CED: Consistent ensemble distillation for audio tagging," in *ICASSP*, 2024.
- [27] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Interspeech*, 2024, pp. 547–551.
- [28] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen and learn more: Design choices for deep audio embeddings," in *ICASSP*, Brighton, UK, May 2019, pp. 3852–3856.
- [29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, June 2022, pp. 16 000–16 009.
- [30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.
- [31] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.
- [32] F.-S. Hsu, S.-R. Huang, C.-W. Huang, Y.-R. Cheng, C.-C. Chen, J. Hsiao, C.-W. Chen, and F. Lai, "A progressively expanded database for automated lung sound analysis: An update," *Applied Sciences*, vol. 12, no. 15, 2022.
- [33] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto, "M2D-CLAP: Masked Modeling Duo Meets CLAP for Learning General-purpose Audio-Language Representation," in *Interspeech*, 2024, pp. 57–61.
- [34] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Modeling Duo for Speech: Specializing General-Purpose Audio Representation to Speech using Denoising Distillation," in *Interspeech*, 2023, pp. 1294–1298.
- [35] —, "Composing General Audio Representation by Fusing Multilayer Features of a Pre-trained Model," in *EUSIPCO*, 2022, pp. 200–204.