Speech Foundation Models Generalize to Time Series Tasks from Wearable Sensor Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Both speech and sensor time series data encode information in both the time- and frequency- domains, like spectral powers and waveform shapelets. We show that speech foundation models learn representations that are domain-independent and achieve state-of-the-art performance on time series tasks from wearable sensors. Probes trained on features extracted from HuBERT and wav2vec 2.0 outperform those extracted from self-supervised models trained directly on modality specific datasets for mood classification, arrhythmia detection, and activity classification tasks. We find a particularly strong relevance of the convolutional feature encoders from speech models for wearable sensor tasks. The methods proposed here improve performance and robustness for data-scarce time series tasks, using simple probing methods. This work is a step towards generalized time series models for speech and sensor data, a topic for further exploration.

1 Introduction and Related Work

Time series models have been trained for applications spanning numerous domains—including health, activity recognition, gesture recognition, weather forecasting, and infrastructure modeling. Classical time series modeling methods include dynamic time warping, shapelet-based methods, convolution-based methods like ROCKET, and numerous other approaches Bagnall et al. [2017], Middlehurst et al. [2024]. Recent work has also explored deep learning-based methods and representation learning strategies to enable shared learning among tasks Xu et al. [2023, 2024], Abbaspourazad et al. [2023], Erturk et al. [2025]. In the realm of physiological and wearable sensor data, most approaches have focused on within-domain representation learning where there is a domain match between pre-training data and evaluation tasks. Prior works on generalized time series modeling across domains often focus on time series data like power consumption, traffic, and weather and often on forecasting tasks Woo et al. [2024], Liu et al. [2023], Jin et al. [2023].

Voice2Series Yang et al. [2021] explored re-programming speech processing models for time series tasks using task-specific target data along with a transformer-based speech model, and found strong performance across the evaluated UCR datasets—including some sensor-based tasks like ECG modeling. Voice2Series trained speech embedding models from scratch, re-trained layers within the model for each time series task, and then identified source-to-target label for inference in the new domains. While successful, this approach required both re-training and label mapping, as well as from-scratch training of speech foundation models. Between domain knowledge transfer has been successful in other areas—for instance, the Audio Spectrogram Transformer Gong et al. [2021] showed improved performance on speech tasks by training a ViT model on ImageNet data.

Speech and wearable data streams have related key signal properties: including frequency band powers, periodic structures, and shapelets in the time domain (see Figure 1). Many sensor domains have limited data availability—learning relevant structure from speech data can have high impact

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

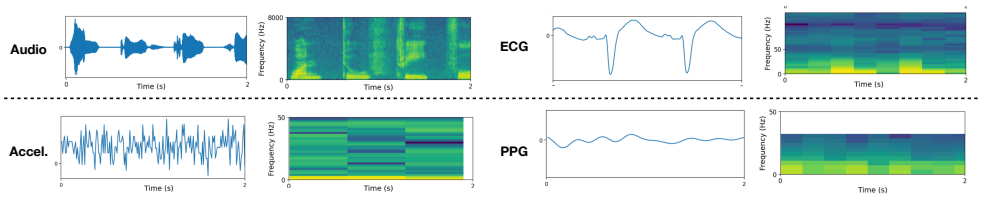

Figure 1: There are important time series characteristics in the time and frequency domain across modalities like frequency band powers, periodic structures, and shapelets in the time domain

in improving performance on data-scarce time series tasks. Additionally, using a single embedding model across time series modalities can improve computational efficiency in multi-modal systems by enabling the deployment of a single model with task-specific adapters across modalities.

Here, we explore the lightweight adaptation of pre-trained state-of-the-art foundation models like HuBERT (Large) and wav2vec 2.0 (Large) via probing and LoRA adapters. To our knowledge, we are the first to directly use pre-trained speech models as feature extractors in the time series domain. We envision this as a step towards more efficient cross-modality adaptation and generalized time series models that can leverage data-hungry architectures and learning from large-scale cross-modality datasets including speech and audio.

2 Methods and Results

We use pre-trained speech models as feature extractors for a variety of time series tasks. We train probes using on four tasks using other sensor modalities: activity classification from accelerometer data from a range of datasets and device placements, arrhythmia detection from ECG data, and stress classification using ECG. For each task, we train linear probes and MLPs. We report performance with an MLP probe at the first and last layer in each results table, along with per-layer performance across the transformer module for each task for both HuBERT and wav2vec 2.0. We also trained LoRA adapters for each transformer layer (results summarized in the Discussion). See the Appendix for additional details on each dataset and evaluation scheme.

Activity Classification We evaluated activity classification on four 3-axis accelerometer data datasets spanning three sensor positions and two window sizes. In both evaluations, we included a benchmark using a Random Forest and engineered features, as in Yuan et al. [2024]. We included results and benchmarks from two window sizes along with performance comparisons from Xu et al. [2024], Yuan et al. [2024], Haresamudram et al. [2022]: (1) following Yuan et al. [2024], a leave one subject out subject out evaluation scheme with 10 second windows at 100 Hz with the PAMAP2 wrist data Reiss

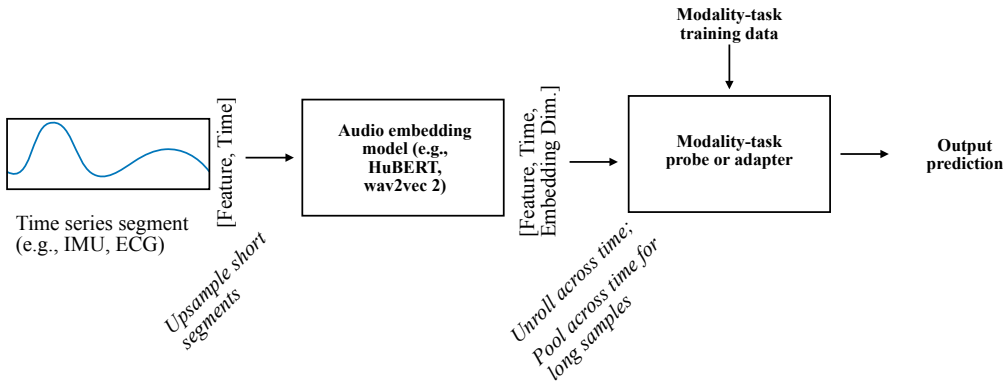

Figure 2: Speech foundation models as feature extractors for other modalities. Time series data is fed as inputs into audio embedding models, with short segments upsampled. Task specific probes are trained on the extracted features, and used to generate predictions across time series tasks.

		PAMAP2 (Wrist)	Opportunity (Wrist)	HHAR (Wrist)	Motionsense (Waist)	PAMAP2 (Leg)
		Yuan et al. [2024] Eval		Haresamudram et al. [2022] Eval		
		F1 (macro avg.)				
	Feat eng. + RF	66.6 ± 9.3	33.4 ± 8.8	61.2 ± 10.2	77.7 ± 3.5	**67.0 ± 10.1**
Pretrain accel	MLP (Yuan'24)	72.5 ± 5.4	57.0 ± 7.8	–	–	–
	SimCLR + Linear (Hare'22)	–	–	55.9 ± 1.8	83.9 ± 1.8	50.8 ± 3.0
	SimCLR + MLP (Hare'22)	–	–	58.6 ± 2.2	85.6 ± 2.5	60.2 ± 2.3
	RelCon + MLP	**85.4 ± 3.5**	**69.1 ± 8.3**	57.6 ± 3.2	80.4 ± 0.7	54.0 ± 0.8
Pretrain speech	HuBERT + Linear	71.3 ± 7.4	49.3 ± 8.2	65.6 ± 8.6	80.0 ± 3.7	54.5 ± 4.3
	HuBERT + MLP	73.6 ± 7.0	50.3 ± 3.9	69.0 ± 11.1	93.1 ± 2.5	60.5 ± 5.9
	wav2vec 2.0 + Linear	67.1 ± 7.0	47.3 ± 3.9	70.2 ± 3.2	89.1 ± 3.4	52.1 ± 4.7
	wav2vec 2.0 + MLP	68.5 ± 6.0	50.8 ± 3.4	**74.5 ± 3.7**	**93.4 ± 2.3**	54.5 ± 6.2

Table 1: **Activity classification results**. Probes with pre-trained speech models compared to probes with pre-trained IMU models from Yuan et al. [2024] "(Yuan'24)" and Haresamudram et al. [2022] "(Hare'22)" and a baseline Random forest model with engineered features.

and Stricker [2012] (8 classes, n=2,869) and the Opportunity dataset (4 classes, n=3,842) Roggen et al. [2010], and (2) following Haresamudram et al. [2022], a 5-fold cross-validation evaluation on 2 second windows of 100 Hz 3-axis accelerometer data: HHAR Reyes-Ortiz et al. [2015] wrist data (6 classes, n=3,370), Motionsense waist data (6 classes, n=14,121), Malekzadeh et al. [2018], and PAMAP2 leg data (12 classes, n=9,709). The 2 second windows were upsampled by a factor of 2 to have sufficient sample lengths to use with the pre-trained speech models. The embeddings extracted from 10 second windows were pooled across the time dimension before being passed to the probe, in order to reduce dimensionality. Table 1 shows results from five activity classification evaluations: two with 10 second windows following the evaluation in Yuan et al. [2024] and three with 2 second windows following the evaluation in Haresamudram et al. [2022].

Arrhythmia Detection (ECG) We conducted binary arrhythmia classification using the MIT-BIH dataset Moody [1983], using 10 second windows sampled at 250 Hz. The extracted embeddings were pooled across the time dimension before being passed to the probe, in order to reduce dimensionality. We compare to previous results on this dataset reported in Xu et al. [2023], including a baseline supervised model and self-supervised models trained on ECG data. Results are reported in Table 2.

Mood Classification (PPG) We conducted four-class mood classification (baseline, stress, amusement, and meditation) using PPG data from WESAD Schmidt et al. [2018], in one minute windows sample at 64 Hz. Results are reported in Table 2, along with comparative results with identical evaluation from Xu et al. [2023] including a baseline supervised model and self-supervised models trained on PPG data.

3 Discussion

Probes trained on pre-trained speech foundation models had the best or competitive performance across tasks when using MLP probes trained on embeddings extracted from early layers. Pre-trained

		Arrythmia Detection (ECG) MIT-BIH		Mood Classification (PPG) WESAD	
		AUC	Accuracy	AUC	Accuracy
	NN (Xu'23)	0.93	78.1	0.62	41.4
Pre-train ECG/PPG	SimCLR (Xu'23)	0.83	69.9	0.62	34.5
	REBAR (Xu'23)	0.92	81.5	0.70	41.4
Pretrain speech	HuBERT + Linear	0.89	87.0	0.79	57.3
	HuBERT + MLP	0.95	94.0	**0.82**	**77.5**
	wav2vec 2.0 + Linear	0.96	**96.7**	0.72	52.8
	wav2vec 2.0 + MLP	**0.97**	96.1	0.80	70.8

Table 2: **Arrhythmia detection results and mood classification results**. Probes with pre-trained speech models compared to probes with pre-trained ECG models and pre-trained PPG models from Xu et al. [2023].

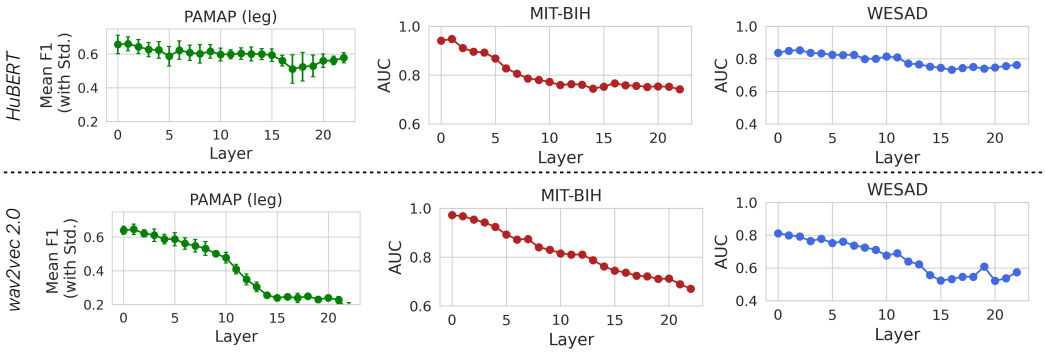

Figure 3: Performance by transformer layer with MLP probes for each task: activity classification (results shown with the PAMAP2 leg data), arrhythmia detection, and mood classification. Early layer performance is better across modalities, particularly for wav2vec 2.0

speech representations outperformed both baselines and self-supervised models trained directly on sensor data for most tasks—activity classification with two second windows, mood classification, and arrhythmia detection. Probes trained with pre-trained speech representations had competitive performance for activity classification with 10 second windows, though lower scores than the RelCon model trained on accelerometer data.

Earlier layer representations from the transformer module were consistently better than later layer representations, especially for wav2vec 2.0 (Figure 3). This suggests that the convolutional feature extractor layers preceding the transformer module learned by the speech encoders are particularly relevant across domains. Figure 4 shows sample convolution filters from HuBERT, which selected for visualization because they had high L2 norms and distinct properties. The filters capture periodic and spiked shapelets, and include filters like bandpass filters and high-pass filters. Additional work will further explore the interpretability of these filters across-domains. LoRA generally improved final layer by about 5-20%, though the best performing models tended to come from earlier layers. The presented experiments did not include adapter training combined with early layer representation extraction, which could potentially further improve performance and will be explored in future work.

4 Conclusions and Future Work

The presented analyses suggest that learned tokenization may be particularly impactful in the wearable sensor domain. Learning effective tokenizations along with masked representation learning, as in HuBERT and wav2vec 2.0, is enabled by large, well-curated speech datasets. Datasets in wearable sensor domains tend to be small and are often task-specific. Our results show that cross-domain learning, including speech and audio data, improves performance on data scarce tasks with data from wearable sensors.

While speech and the investigated sensor domains (accelerometer data, ECG, and PPG) share enough commonalities to enable zero-shot transfer, the investigated sensor data streams were sampled at lower frequencies (100 Hz, 250 Hz, and 64 Hz respectively) than speech data. The pre-trained foundation models were designed for speech input at 16,000 Hz. Training strategies for foundation models that better enable learning over both the low and high frequency space could help improve performance across modalities and will also be explored in future work.

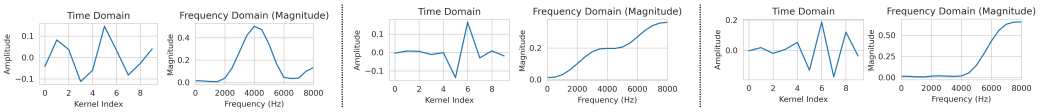

Figure 4: Visualization of selection of convolutional filters from HuBERT, from the first convolutional layer in the model. The filters capture periodic and spiked shapelets, and include filters like bandpass filters and high-pass filters.

References

Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017.

Matthew Middlehurst, Patrick Schäfer, and Anthony Bagnall. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38(4):1958–2031, 2024.

Maxwell A Xu, Alexander Moreno, Hui Wei, Benjamin M Marlin, and James M Rehg. Rebar: Retrieval-based reconstruction for time-series contrastive learning. *arXiv preprint arXiv:2311.00519*, 2023.

Maxwell A Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik J Raghuram, James M Rehg, and Shirley Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data. *arXiv preprint arXiv:2411.18822*, 2024.

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. *arXiv preprint arXiv:2507.00191*, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. 2024.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. Voice2series: Reprogramming acoustic models for time series classification. In *International conference on machine learning*, pages 11808–11819. PMLR, 2021.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ digital medicine*, 7(1):91, 2024.

Harish Haresamudram, Irfan Essa, and Thomas Plötz. Assessing the state of self-supervised human activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–47, 2022.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pages 233–240. IEEE, 2010.

Jorge Reyes-Ortiz, Davide Anguita, Luca Oneto, and Xavier Parra. Smartphone-based recognition of human activities and postural transitions. *UCI Machine Learning Repository*, 2015.

Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, W-P2DS'18, pages 2:1–2:6, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5654-1. doi: 10.1145/3195258.3195260. URL `http://doi.acm.org/10.1145/3195258.3195260`.

George Moody. A new method for detecting atrial fibrillation using rr intervals. *Proc. Comput. Cardiol.*, 10:227–230, 1983.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

5 Appendix

PAMAP2 (Wrist) We extracted 10 second segments at 100 Hz from PAMAP2 Reiss and Stricker [2012], and used a leave one subject out evaluation scheme following the procedure in Yuan et al. [2024]. The evaluation included eight classes: lying, sitting, standing, walking, ascending stairs, descending stairs, vacuum cleaning, and ironing. The experiments included n=2,860 samples from eight participants.

Opportunity (Wrist) We extracted 10 second segments at 100 Hz from Roggen et al. [2010], and used a leave one subject out evaluation scheme following the procedure in Yuan et al. [2024]. The evaluation included four classes: sitting, standing, walking, lying down. The experiments included n=3,842 samples from four participants.

HHAR (Wrist) We extracted 2 second segments at 100 Hz from Reyes-Ortiz et al. [2015], and used a 5-fold cross validation evaluation scheme using the procedure in Haresamudram et al. [2022]. The evaluation included six classes: stairs down, stairs up, walk, bike, stand, sit. The experiments included n=3,370 samples.

Motionsense (Waist) We extracted 2 second segments at 100 Hz from Malekzadeh et al. [2018], and used a 5-fold cross validation evaluation scheme using the procedure in Haresamudram et al. [2022]. The evaluation included six classes: stairs down, stairs up, walk, jog, stand, sit. The experiments included n=14,121 samples.

PAMAP2 (Leg) We extracted 2 second segments at 100 Hz from Reiss and Stricker [2012], and used a 5-fold cross validation evaluation scheme using the procedure in Haresamudram et al. [2022]. The evaluation included twelve classes: rope jumping, lying, sitting, standing, walking, running, cycling, Nordic walking, ascending stairs, descending stairs, vacuum cleaning, ironing. The experiments included n=9,709 samples.

MIT-BIH We extracted 10 second segments at 250 Hz from Moody [1983]. We followed the pre-processing and evaluation process as described in Xu et al. [2023] to allow for comparison with prior results from self-supervised models trained in-domain.

WESAD We extracted 60 second segments at 64 Hz from Schmidt et al. [2018]. We followed the pre-processing and evaluation process as described in Xu et al. [2023] to allow for comparison with prior results from self-supervised models trained in-domain.