HPP-Voice: A Large-Scale Evaluation of Speech Embeddings for Multi-Phenotypic Classification

David Krongauz [1] Hido Pinto [1] Sarah Kohn [1] Yanir Marmor [1] Eran Segal [1]

Abstract

Human speech contains paralinguistic cues that reflect a speaker's physiological and neurological state, potentially enabling non-invasive detection of various medical phenotypes. We introduce the Human Phenotype Project Voice corpus (HPP-Voice): a dataset of 7,188 recordings in which Hebrew-speaking adults count for 30 seconds, with each speaker linked to up to 15 potentially voice-related phenotypes spanning respiratory, sleep, mental health, metabolic, immune, and neurological conditions. We present a systematic comparison of 14 modern speech embedding models, where modern speech embeddings from these 30-second counting tasks outperform MFCCs and demographics for downstream health condition classifications. We found that embedding learned from a speaker identification model can predict objectively measured moderate to severe sleep apnea in males with an AUC of 0.64 ± 0.03, while MFCC and demographic features led to AUCs of 0.56 ± 0.02 and 0.57 ± 0.02, respectively. Additionally, our results reveal gender-specific patterns in model effectiveness across different medical domains. For males, speaker identification and diarization models consistently outperformed speech foundation models for respiratory conditions (e.g., asthma: 0.61 ± 0.03 vs. 0.56 ± 0.02) and sleep-related conditions (insomnia: 0.65 ± 0.04 vs. 0.59 ± 0.05). For females, speaker diarization models performed best for smoking status (0.61 ± 0.02 vs 0.55 ± 0.02), while Hebrew-specific models performed best (0.59 ± 0.02 vs. 0.58 ± 0.02) in classifying anxiety compared to speech foundation models. Our findings provide evidence that a simple counting task can support large-scale, multi-phenotypic voice screening and highlight which embedding families generalize best to specific conditions, insights that can guide future vocal biomarker research and clinical deployment.

1. Introduction

Human speech is a richly layered signal. Beyond lexical content, it carries prosody, articulation, timing, respiration, and other paralinguistic cues that mirror a speaker's physiology and neurological state(Fant, 1971; Härmä et al., 2024). Recent deep learning work has capitalized on these cues for emotion recognition (Latif et al., 2021), speaker verification (Mittal & Dua, 2022), and detection of neurodegenerative (Tao et al., 2025) or pulmonary disorders (Sharma et al., 2020). However, most clinical studies remain narrow: they target a single disease (e.g., Parkinson's disease (Moro-Velazquez et al., 2021) and Alzheimer's disease (Luz et al., 2021)), use small cohorts, or focus on non-speech events such as coughs and breaths (Baur et al., 2024).

In this study, we leverage the extensive and uniquely comprehensive data from the Human Phenotype Project (HPP) (Shilo et al., 2021) to investigate associations between voice characteristics and health conditions. Furthermore, we systematically compare multiple voice feature extraction techniques, ranging from classical spectral features to advanced embeddings derived from state-of-the-art self-supervised foundation models for speech.

Why fluent counting? A brief "count-to-30" prompt is easy to administer remotely, highly reproducible, and intuitively understood by speakers of any language (Wardle et al., 2011; Schöbi et al., 2022). It also elicits sustained phonation and prosodic variability, which are not captured in cough-centric corpora such as Coswara (Sharma et al., 2020).

The HPP-Voice corpus. We introduce *HPP-Voice*, comprising **7,188 recordings** from 6,760 adults (3,211 males / 3,549 females, age 52 ± 10 years). Each utterance is paired with 15 health conditions spanning across 6 body systems: respiratory, metabolic, neurological, mental health, immune, and sleep (Shilo et al., 2021).

[1] Department of Computer Science and Applied Mathematics, Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. Correspondence to: David Krongauz <david.krongauz@weizmann.ac.il>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

Our contributions are:

1. **Large, clinically diverse corpus.** We release a novel dataset of fluent speech paired with verified diagnoses across six body systems.

2. **Comprehensive benchmark.** Fifteen state-of-the-art encoders (e.g., MFCC (Davis & Mermelstein, 1980), x-vector (Snyder et al., 2018), wav2vec 2.0 (Baevski et al., 2020), WavLM (Chen et al., 2022)) are evaluated under identical splits.

3. **Phenotype-specific insights.** We demonstrate which speech representations outperform for specific clinical domains, providing a comprehensive evaluation of their relative efficacy across different healthcare applications.

2. Dataset: HPP-Voice

HPP-Voice is the voice arm of the *Human Phenotype Project* (HPP), a longitudinal, multi-omics registry designed to unravel chronic-disease mechanisms at the population scale.

Cohort description The HPP cohort currently comprises 11,460 adults. Between December 2019 and December 2024, 6,760 unique speakers were invited for voice sampling and completed the protocol, yielding **7,188** recordings. Among the 6,760 speakers, 3,549 were women (52%) aged 52.4±10.7 years and 3,211 were men (48%) aged 52.6±9.9 years.

Recording protocol Speech was captured in a controlled laboratory environment using 32-bit 384 kHz sampling, then down-sampled to 16 kHz mono. Each participant performed a single 30-second counting task at a comfortable pace.

Medical labels Medical and phenotype labels were derived from the HPP questionnaire and medical records. Sleep apnea was defined as present if the averaged Apnea-Hypopnea Index (AHI) measured across three nights of monitoring exceeded 15 events per hour (Kohn et al., 2025). Cases were also included if the subject self-reported the condition. All other health conditions were based on self-report. For analysis, we grouped the conditions into six body systems: respiratory (asthma, current smoker, past smoker, chronic sinusitis), metabolic (anemia, hyperthyroidism, hypothyroidism), neurological (migraine, headache), mental health (depression, anxiety), immune (COVID-19, allergy), and sleep-related (sleep apnea, insomnia). Notable condition prevalence differences by gender include sleep apnea (14.89

3. Method

Our pipeline converts every HPP-Voice recording into a single fixed-length vector, feeds these vectors into downstream classifiers, and compares performance across a diverse set of embedding models.

3.1. Representation families

To compare how different types of acoustic representations capture health-relevant voice information, we systematically evaluate 14 different speech embeddings across five categories, each capturing unique aspects of the speech signal (see Table 1). In addition, we compute and evaluate the performance of Mel-Frequency Cepstral Coefficients (MFCCs) as a classical, non-deep learning baseline.

3.2. Experimental setup

We conducted gender-specific stratification based on confounding analysis, as preliminary testing revealed that all embedding models could predict participant gender with very high performance (AUCs 0.92-0.98). After applying quality control and removing repeated visits, we retained 2,150 recordings from unique female participants and 1,993 recordings from unique male participants.

For classifier training and evaluation, we employed a LightGBM classifier (Ke et al., 2017) trained using 4-fold cross-validation, with hyperparameter optimization conducted via Optuna (Akiba et al., 2019). The entire process was repeated across 20 random seeds to ensure robustness. Age was included as an additional input feature to adjust for its potential confounding effect.

4. Results

We begin our analysis with sleep apnea, as this condition was curated using objective physiological recordings. Our analysis focused on sleep apnea prediction in males, where this disorder is more prevalent (14.89% in our cohort). As shown in Figure 1, speaker identification models demonstrated superior performance. The x-vector model achieved the highest performance with an AUC of 0.64 ± 0.03, significantly outperforming both MFCC features (AUC = 0.56 ± 0.02) and baseline demographic features (AUC = 0.57 ± 0.02). Among speech foundation models, WavLM-Large exhibited strong performance, while models trained for emotion recognition and language-specific models did not significantly improve over baseline.

Beyond sleep apnea, we extended our analysis to multiple self-reported medical conditions, as shown in Figure 2. The analysis reveals pronounced gender-specific patterns in model effectiveness. For males, SI and SD models consistently outperformed speech foundation models for respira-

Table 1. Speech-embedding models evaluated in this study. Dataset acronyms: LS (LibriSpeech) (Panayotov et al., 2015), LL (Libri-Light) (Kahn et al., 2020), LVG (LibriLight+VoxPopuli (Wang et al., 2021)+Gigaspeech (Chen et al., 2021)). PT (Pre-training), FT (Fine-tuning).

Family	Model	Params	Training Data
Speech foundation	wav2vec2-Base (Baevski et al., 2020)	95M	PT: LS 960 hr
	wav2vec2-Large (Baevski et al., 2020)	317M	PT: LL 60k hr
	WavLM-Base (Chen et al., 2022)	95M	PT: LS 960 hr
	WavLM-Large (Chen et al., 2022)	317M	PT: LVG 94k hr
	XLSR-53 (Conneau et al., 2021)	300M	PT: 53-language mixed corpus 56k hr
Hebrew-specific	XLSR Hebrew-PT	300M	PT: XLSR-53 corpus + ivrit.ai (Marmor et al., 2023)
	XLSR Hebrew-FT	300M	PT: same as above; FT: Hebrew ASR on Common Voice (Ardila et al., 2020)
Speaker diarization	WavLM-SD (Chen et al., 2022)	317M	PT: LVG 94k hr; FT: SD on LibriMix
	pyannote (Bredin et al., 2019)	4.3M	PT: VoxCeleb (Nagrani et al., 2020)
Speaker identification	pyannote-FT	4.3M	PT: VoxCeleb; FT: SI on HPP-Voice
	x-vector (Snyder et al., 2018)	4.2M	PT: VoxCeleb
	EffNet (Tan & Le, 2020)	6M	PT: HPP-Voice
Emotion	wav2vec2-SER (Ravanelli et al., 2021)	95M	PT: LS 960 hr; FT: IEMOCAP (Busso et al., 2008)
	WavLM-SED (Wang et al., 2023)	317M	PT: LVG 94k hr; FT: IEMOCAP, RAVDESS (Livingstone & Russo, 2018)

tory conditions (e.g., asthma: 0.61 ± 0.03 vs. 0.56 ± 0.02) and sleep-related conditions (insomnia: 0.65 ± 0.04 vs. 0.59 ± 0.05). In contrast, among females, SD models performed best for smoking status (0.61 ± 0.02 vs. 0.55 ± 0.02), while Hebrew-specific models achieved superior performance (0.59 ± 0.02 vs. 0.58 ± 0.02) in classifying anxiety compared to speech foundation models.

5. Discussion and Conclusion

This study introduces HPP-Voice and systematically benchmarks 14 speech embedding models for multi-condition classification. Using only a 30-second counting task per subject, we demonstrate that modern speech representations, particularly those trained for speaker identification and diarization, can detect clinically relevant signals for various conditions, accounting for age and gender as potential confounders. The optimal embedding family varied by both medical domain and gender, indicating the presence of condition-specific and population-specific acoustic markers.

Several hypotheses may explain why counting, despite its constrained linguistic content, encodes such health-relevant information. The dynamics of breathing pauses during fluent counting may reflect respiratory health and sleep-disordered breathing patterns (Sharma et al., 2020). Prosodic features such as pitch variability, rhythm, and speech rate are known to reflect psychological and neurological states (Luz et al., 2021; Latif et al., 2021). Articulatory stability and vocal tract control may be modulated by metabolic or cognitive impairments (Tracey et al., 2023).

Our findings underscore the clinical potential of voice as a scalable, low-effort health monitoring modality. A short counting task, easily administered in remote settings, may support non-invasive pre-screening for multiple health conditions. By identifying which embedding families generalize best to specific conditions, this work offers practical guidance for future development of voice-based diagnostic systems.

Acknowledgments

E.S. is supported by the Crown Human Genome Center, by D. L. Schwarz, J. N. Halpern, and L. Steinberg, and by grants funded by the European Research Council and the Israel Science Foundation. The authors extend sincere appreciation to Y. Sieradzki, whose invaluable consultation and insights were instrumental in shaping this work.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework, July 2019. URL `http://arxiv.org/abs/1907.10902`. arXiv:1907.10902 [cs].

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European

Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/.

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

Baur, S., Nabulsi, Z., Weng, W.-H., Garrison, J., Blankemeier, L., Fishman, S., Chen, C., Kakarmath, S., Maimbolwa, M., Sanjase, N., et al. Hear–health acoustic representations. *arXiv preprint arXiv:2403.02522*, 2024.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. pyannote.audio: neural building blocks for speaker diarization, November 2019. URL http://arxiv.org/abs/1911.01255. arXiv:1911.01255 [eess].

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *Interspeech 2021*, 2021.

Davis, S. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

Fant, G. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Number 2. Walter de Gruyter, 1971.

Härmä, A., Brinker, B. d., Grossekathofer, U., Ouweltjes, O., Nallanthighal, S., Abrol, S., and Sharma, V. Survey on biomarkers in human vocalizations. *arXiv preprint arXiv:2407.17505*, 2024.

Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., 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)*, pp. 7669–7673. IEEE, 2020.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Kohn, S., Diament, A., Godneva, A., Dhir, R., Weinberger, A., Reisner, Y., Rossman, H., and Segal, E. Phenome-wide associations of sleep characteristics in the Human Phenotype Project. *Nature Medicine*, 31(3):1026–1037, March 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03481-x. URL https://www.nature.com/articles/s41591-024-03481-x. Publisher: Nature Publishing Group.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., and Schuller, B. Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 14(2):1634–1654, 2021.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Luz, S., Haider, F., Fuente, S. D. L., Fromm, D., and MacWhinney, B. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Interspeech 2021*, pp. 3780–3784. ISCA, August 2021. doi: 10.21437/Interspeech.2021-1220. URL https://www.isca-archive.org/interspeech_2021/luz21_interspeech.html.

Marmor, Y., Misgav, K., and Lifshitz, Y. ivrit.ai: A Comprehensive Dataset of Hebrew Speech for AI Research and Development, July 2023. URL http://arxiv.org/abs/2307.08720. arXiv:2307.08720.

Mittal, A. and Dua, M. Automatic speaker verification systems and spoof detection techniques: review and analysis. *International Journal of Speech Technology*, 25(1): 105–134, 2022.

Moro-Velazquez, L., Gomez-Garcia, J. A., Arias-Londoño, J. D., Dehak, N., and Godino-Llorente, J. I. Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. *Biomedical Signal Processing and Control*, 66: 102418, 2021.

Nagrani, A., Chung, J. S., Xie, W., and Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. SpeechBrain: A General-Purpose Speech Toolkit, June 2021. URL http://arxiv.org/abs/2106.04624. arXiv:2106.04624 [eess].

Schöbi, D., Zhang, Y.-P., Kehl, J., Aissani, M., Pfister, O., Strahm, M., van Haelst, P., and Zhou, Q. Evaluation of speech and pause alterations in patients with acute and chronic heart failure. *Journal of the American Heart Association*, 11(21):e027023, 2022.

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., Ghosh, P. K., Ganapathy, S., et al. Coswara–a database of breathing, cough, and voice sounds for covid-19 diagnosis. *arXiv preprint arXiv:2005.10548*, 2020.

Shilo, S., Bar, N., Keshet, A., Talmor-Barkan, Y., Rossman, H., Godneva, A., Aviv, Y., Edlitz, Y., Reicher, L., Kolobkov, D., Wolf, B. C., Lotan-Pompan, M., Levi, K., Cohen, O., Saranga, H., Weinberger, A., and Segal, E. 10 K: a large-scale prospective longitudinal study in Israel. *European Journal of Epidemiology*, 36(11): 1187–1194, November 2021. ISSN 1573-7284. doi: 10.1007/s10654-021-00753-5.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, Calgary, AB, April 2018. IEEE. ISBN 978-1-5386-4658-8. doi: 10.1109/ICASSP.2018.8461375. URL https://ieeexplore.ieee.org/document/8461375/.

Tan, M. and Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. URL http://arxiv.org/abs/1905.11946. arXiv:1905.11946 [cs].

Tao, F., Mirheidari, B., Pahar, M., Young, S., Xiao, Y., Elghazaly, H., Peters, F., Illingworth, C., Braun, D., O'Malley, R., et al. Early dementia detection using multiple spontaneous speech prompts: The process challenge. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2. IEEE, 2025.

Tracey, B., Volfson, D., Glass, J., Haulcy, R., Kostrzebski, M., Adams, J., Kangarloo, T., Brodtmann, A., Dorsey, E. R., and Vogel, A. Towards interpretable speech biomarkers: exploring MFCCs. *Scientific Reports*, 13(1): 22787, December 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-49352-2. URL https://www.nature.com/articles/s41598-023-49352-2. Publisher: Nature Publishing Group.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021.

Wang, Y., Ravanelli, M., and Yacoubi, A. Speech Emotion Diarization: Which Emotion Appears When?, October 2023. URL http://arxiv.org/abs/2306.12991. arXiv:2306.12991 [cs].

Wardle, M., Cederbaum, K., and de Wit, H. Quantifying talk: developing reliable measures of verbal productivity. *Behavior research methods*, 43:168–178, 2011.

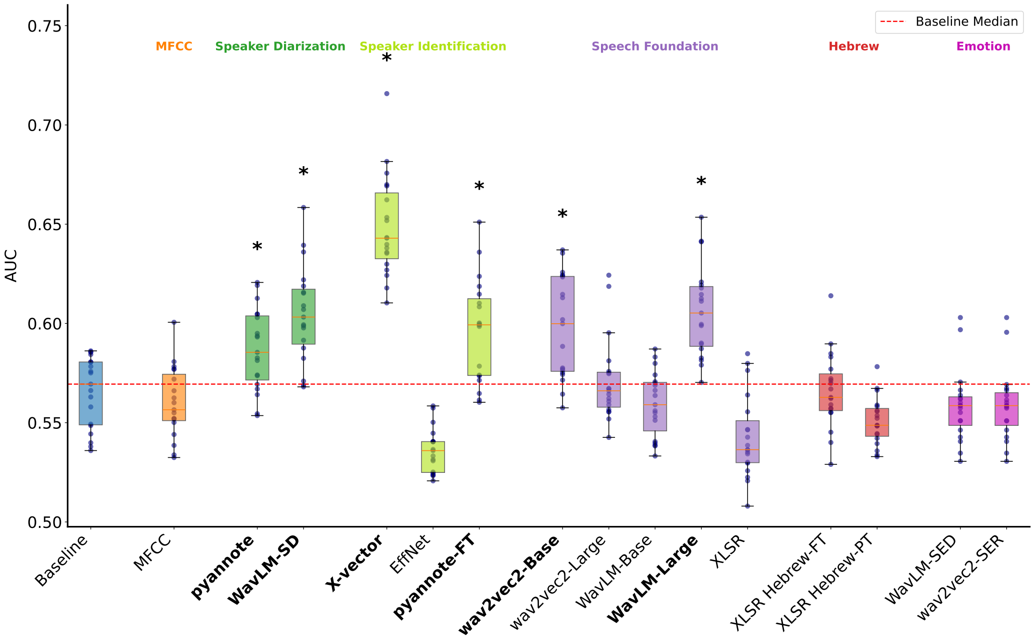

Figure 1. **Sleep apnea detection performance in males.** Performance comparison of speech representation models for sleep apnea detection in males. Models are grouped by speech processing domain (color-coded at top). Boxplots show AUC distribution across 20 random seeds. Asterisks indicate statistically significant improvement over baseline.

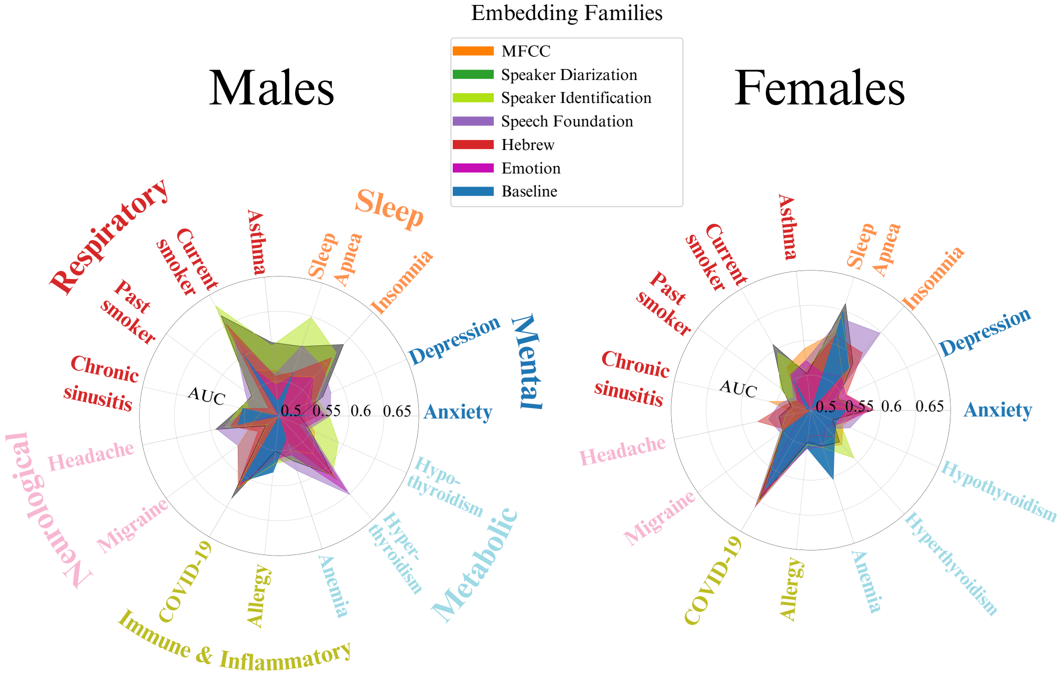

Figure 2. **Gender-specific model performance across medical domains.** Radar plots showing average AUC scores of the best-performing model from each speech representation family across medical conditions grouped by domain. The left panel shows male subjects; the right panel shows female subjects.