Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Meta-Learning Approaches for Speaker-Dependent Voice Fatigue Models

Roseline Polle[1], *Agnes Norbury*[1], *Alexandra Livia Georgescu*[1,2], *Nicholas Cummins*[1,2], *Stefano Goria*[1]

[1]thymia, London, UK
[2]Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK
roseline@thymia.ai,stefano@thymia.ai

Abstract

Speaker-dependent modelling can substantially improve performance in speech-based health monitoring applications. While mixed-effect models are commonly used for such speaker adaptation, they require computationally expensive retraining for each new observation, making them impractical in a production environment. We reformulate this task as a meta-learning problem and explore three approaches of increasing complexity: ensemble-based distance models, prototypical networks, and transformer-based sequence models. Using pre-trained speech embeddings, we evaluate these methods on a large longitudinal dataset of shift workers (N=1,185, 10,286 recordings), predicting time since sleep from speech as a function of fatigue, a symptom commonly associated with ill-health. Our results demonstrate that all meta-learning approaches tested outperformed both cross-sectional and conventional mixed-effects models, with a transformer-based method achieving the strongest performance.

Index Terms: meta-learners, sequential modelling, speech, paralinguistic, health, repeated measures

1. Introduction

Fatigue is a common symptom in many physical and mental health conditions and is especially widespread in occupational settings known for high rates of health problems, such as shift work [1, 2]. Monitoring fatigue through readily available devices such as smartphones could enable earlier intervention and better outcomes through more targeted treatment adjustments and timely rest provision in safety-critical contexts [3, 4].

Recent work has shown that fatigue manifests detectably in speech, offering a scalable, low-burden monitoring approach that can be deployed in various real-world settings [5, 6]. Studies have demonstrated that speaker-specific modelling substantially improves detection performance [7], likely by better separating speaker characteristics (e.g., age, sex) from fatigue-related changes. However, current speaker adaptation methods such as mixed-effects modelling are impractical for production environments, as they require computationally expensive retraining for each new observation [8]. While solutions such as federated learning could distribute this computational load, it would still require complex server-client coordination and would face performance constraints on edge devices [9]. An ideal monitoring solution would require users to report their fatigue levels during an initial calibration period, after which the system could maintain accurate predictions from speech alone without retraining.

Meta-learning offers a promising framework to achieve this goal while remaining tractable for real-world deployment [10]. Under this 'learning to learn' paradigm, models learn to leverage a few examples, known as a support set, from initial user interactions to make predictions about new speech samples, without requiring further user input or model retraining [11]. Meta-learning has been demonstrated for healthcare applications such as rare disease diagnosis [12, 13], and for speech processing tasks including emotion classification [14].

In this work, we reformulate fatigue monitoring as a meta-learning problem and explore three approaches of increasing complexity: (i) a simple ensemble-based distance model that learns changes between speech samples, (ii) a prototypical network that learns a metric space for comparing speech samples [15], and (iii) A transformer-based sequence model that learns to predict fatigue from sequences of previous observations [16].

Using a large longitudinal dataset of shift workers [7], we compare our meta-learning approaches against two baselines: (i) a cross-sectional model that ignores speaker identity, and (ii) a linear mixed-effects model [17]. Results demonstrate that our meta-learning methods outperform both baselines after collecting just a few user-reported fatigue levels. The transformer-based approach performs strongest while offering practical advantages for deployment; it requires no speaker-specific retraining, uses a single model artefact, and handles both classification and regression naturally.

2. Data and Preprocessing

2.1. Dataset

Our analysis uses data from a large-scale longitudinal study of shift workers [7], comprising 1,185 participants monitored over a two-week period. Study procedures were reviewed by an independent research ethics expert (Dr. David Carpenter) working under the auspices of the Association of Research Managers and Administrators (ARMA, https://arma.ac.uk/, involved in formulating national standards for research ethics support and review) and received a favourable ethics opinion on 22nd November 2023. All participants gave written informed consent. Access to the dataset can be obtained through reasonable requests to the senior author.

During the study, participants completed speech recordings and self-report assessments twice daily on their personal devices, coinciding with the start and end of their work shifts. From the collected metrics, we focus on time since sleep (S) as our target variable, representing a key component of fatigue [18]. Among the various speech tasks completed by participants, we analysed the reading of fourteen different texts from the speech literature, including the Aesop fables, The North Wind and the Sun [19], The Boy who Cried Wolf [20], and the Rainbow Passage [21]. On each day, different texts (but the same for all participants, although translated into the relevant language, i.e. English or Spanish), were assigned before

10.21437/Interspeech.2025-2055

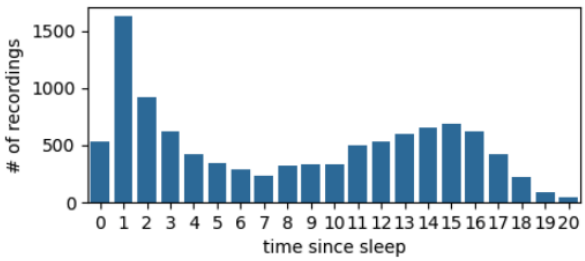

Figure 1: *Time since sleep samples distribution*

and after each work shift. These standardized reading tasks provide a controlled context where variations in speech primarily reflect paralinguistic features rather than content or speaking style [22].

2.2. Data Cleaning and Processing

The dataset underwent several preprocessing steps to ensure quality and consistency. For the target variable, time since sleep values were constrained to the range [0,24] hours, resulting in 10,286 valid observations across 1,185 speakers. The distribution of these values shows expected bimodality with peaks at 1 and 15 hours (Figure 1), reflecting the pre/post-shift recording schedule. To support regression and classification analysis, we additionally defined a binary 'fatigued' state where participants with time since sleep greater than or equal to 10 hours were labelled as fatigued. This threshold was selected based on prior research in shift work environments [7] and results in a reasonably balanced dataset with 45.4% of the samples labelled as fatigued. The regression task thus aims to predict the continuous time since sleep value, while the classification task predicts this binary fatigue state.

For the audio data, all recordings were standardized to WAV format with 16kHz sampling rate, 32-bit resolution, and mono-channel audio. From each recording, leading and trailing silence were removed, and then, to maintain a standardized approach, the first 20 seconds were selected for feature analysis. Speech embeddings were then generated using Trillsson-5 [23], a model specifically trained to capture paralinguistic information. This approach produces fixed-dimension feature vectors that encapsulate speaker state, while being robust to variations in recording conditions [24].

The final processed dataset characteristics are detailed in Table 1. The dataset's size and diversity, combined with its longitudinal nature, make it particularly suitable for evaluating speaker-dependent modelling approaches.

3. Methods

Given a speech embedding $x_j^i \in \mathbb{R}^d$ from speaker i at time j, our goal is to predict the associated fatigue level y_j^i. The target y_j^i can represent either time since sleep (regression) or a binary fatigued/non-fatigued state (classification). Each speaker i provides a sequence of ordered observations (x_j^i, y_j^i) with $j = 0, ..., T$. We aim to learn a model f that can adapt to speaker-specific patterns. At time t, the model makes predictions using both the current speech sample x_t^i and a support set $S_t^i = (x_j^i, y_j^i)$ with $j < t$ containing all previous observation pairs from that speaker. The support set grows as more observations become available, allowing the model to better adapt to speaker-specific patterns.

Table 1: *Dataset Characteristics*

(a) *Demographics*

Group	Participants	%	avg samples
Total	1,185	100.0	8.7
Gender			
Female/Male/Other	553/528/4	55.1/44.6/0.3	8.2/9.3/4.2
Age			
Under 40/ Over 40	936/250	78.9/21.1	8.4/9.6
Language			
GB/US/es/eng-other	701/268/170/46	59.2/22.6/14.3/3.9	8.6/7.1/12.1/6.0

(b) *Target Variable Characteristics*

Characteristic	Value	Distribution
Total Samples	10,286	(100%)
Time Since Sleep (S - hours)	8[6]	avg[std]
Fatigued ($S \geq 10$)	4,671	(45.4%)
Not fatigued ($S < 10$)	5,615	(54.6%)

We explore three meta-learning approaches of increasing complexity, comparing them against simpler benchmarks (mixed-effects and cross-sectional models that ignore speaker identity) on both regression and classification tasks.

3.1. Training and Evaluation

All models were evaluated using a 5-fold cross-validation scheme with speaker stratified 70%/10%/20% train/validation/test splits; except for the mixed-effects baseline in which we replicated the evaluation procedure in [7]. To account for potential sequence ordering effects, we created five random sequence orderings (as per section 4.2) for each speaker within each cross-validation split, resulting in 25 total evaluation iterations. We report AUC, balanced accuracy, precision, and recall for classification tasks. We compute Pearson correlation coefficients and root mean squared error for regression tasks. All reported metrics are averaged across the 25 iterations, with standard deviations indicating performance stability across different sequence orderings and data splits. For classification models, decision thresholds were optimised on the validation set to equalise precision and recall, ensuring balanced performance across classes without test set leakage.

3.2. Benchmark Models

Cross-sectional Model: Our simplest baseline ignores speaker identity and treats each observation independently, learning a mapping $f_{cs} : x \rightarrow y$ directly from speech embeddings to fatigue levels. For regression, we use ridge regression [25], with a parameter search for the regularisation parameter α in $(0.1, 1, 10, 100, 1000)$. For classification, we use logistic regression, with parameter search for the inverse regularisation strength C in $(0.001, 0.01, 0.1, 1, 10)$. Parameters are tuned on the validation set.

Mixed-effects Model: We implement a mixed-effects approach [17], following [7] to incorporate speaker-specific patterns (with fixed effects for age, sex, and language, and random effects for participants). For each speaker i and time point t, we train a separate ridge regression model using all previous observations $(x_j^i, y_j^i)_{j<t}$ from that speaker. Unlike our meta-learning models, which can be trained once and then adapted to new speakers, this approach requires full model retraining with each new observation, making it considerably less practical in deployment. The evaluation methodology for this base-

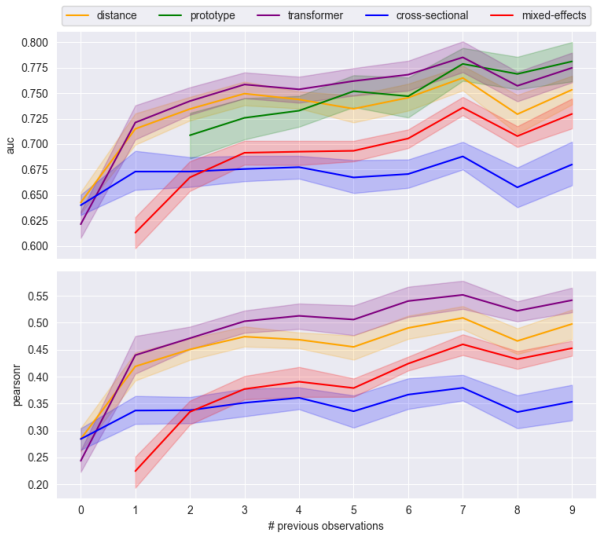

Figure 2: *Performance comparison across participants for different numbers of previous observations (with 95% CI). Top panel: classification - AUC. Bottom panel: regression - Pearson correlation.*

line differs from our other models: following [7]'s approach on the same dataset, we use speakers' past observations as training data and their current observations as test data. As this model serves as a benchmark for comparison with our proposed methods, we deem this methodological difference acceptable.

3.3. Meta-learning Approaches

Distance-based Model: Our first meta-learning approach focuses on learning changes in fatigue levels rather than absolute values. By analyzing fatigue-related variations in speech within each individual, this framework inherently accommodates speaker-specific differences. Instead of predicting fatigue directly, we train a ridge regression model ($\alpha = 1000$) to predict differences:

$$\Delta y = f_d(x_j - x_t) \tag{1}$$

For a new observation x_t^i, we predict:

$$y_t^i = \begin{cases} f_{cs}(x_t^i) & \text{if } t = 0 \\ \frac{1}{|S_t^i|} \sum_{j \in S_t^i} (y_j^i + f_d(x_t^i - x_j^i)) & \text{otherwise} \end{cases} \tag{2}$$

where S_t^i is the set of indices of previous observations for speaker i. Note that this approach requires two models: the cross-sectional model f_{cs} for the initial prediction and the distance model f_d for subsequent predictions based on changes.

Prototypical Network: We implement a metric-learning approach following [15]. We learn a projection function $g : \mathbb{R}^{1024} \to \mathbb{R}^{64}$ (implemented as three linear layers with ReLU activations and hidden dimension of 32) that maps speech embeddings to a space where similar fatigue states cluster together.

For the classification task, we compute class prototypes from the support set:

$$c_k^i = \frac{1}{|S_{t,k}^i|} \sum_{j \in S_{t,k}^i} g(x_j^i) \tag{3}$$

where $S_{t,k}^i$ contains previous observations from speaker i and class k. Prototypes are therefore created for each speaker and fatigue state. Classification probabilities are then derived from Euclidean distances to these prototypes. While effective for classification, this approach does not naturally extend to regression, which is why we exclude it from our regression analysis.

Transformer-based Model: Our final approach treats speaker adaptation as a sequence modelling problem, leveraging transformers' ability to learn from in-context examples [26, 16]. For each speaker, we construct sequences:

$$[x_0^i, y_0^i, x_1^i, y_1^i, ..., x_t^i] \tag{4}$$

Using a GPT2 architecture (15 layers, 8 heads), we frame fatigue prediction as next-token prediction, computing loss only at target positions (the ys). To form the sequences, targets are augmented to match feature dimensionality by sampling each dimension from $\mathcal{N}(y_j^i, 0.1^2)$. We use a mean squared error loss for each dimension of the augmented target, which is then averaged to produce the final loss.

Table 2: *Performance averaged for points with 6 and 7 previous observations. Dist, Proto, Tr, CS, ME respectively stand for Distance, Prototype, Transformer, Cross-sectional and Mixed-effects methods. bacc, rec, prec, rho stand for balanced accuracy, recall, precision and pearson correlation. Standard deviation is in bracket.*

	Method				
	Dist	**Proto**	**Tr**	**CS**	**ME**
Classification					
auc	.755[.034]	.763[.049]	.777[.040]	.679[.036]	.721[.020]
bacc	.684[.041]	.694[.044]	.706[.034]	.627[.032]	.657[.018]
rec	.671[.060]	.666[.071]	.690[.078]	.606[.057]	.630[.029]
prec	.640[.058]	.664[.064]	.671[.056]	.582[.052]	.618[.027]
Regression					
rho	.500[.056]	n/a	.546[.069]	.373[.069]	.442[.028]
RMSE	5.32[.27]	n/a	5.04[.31]	5.59[.28]	5.38[.09]

4. Results

4.1. Model Performance

We observed that, overall, prediction accuracy increases with the number of available observations per speaker (Figure 2). The cross-sectional baseline, having no speaker adaptation, maintains constant performance (AUC $\sim$ 0.68, correlation $\sim$ 0.37). The prototype model and mixed-effects approaches require at least one or two previous observations to make predictions, as they need examples to build speaker-specific representations.

The meta-learning approaches demonstrate notable improvements as more speaker observations become available, surpassing both the cross-sectional and mixed-effects baselines. The transformer model performs best, reaching an AUC of 0.78 and a correlation of 0.55 with a minimum of 6 prior observations (Table 2). Compared to conventional approaches, this represents improvements of 14% in AUC and 46% in correlation versus cross-sectional modeling, and 8% and 23% respectively versus mixed-effects. The transformer approach also offers practical advantages. Unlike the prototype model, it handles both classification and regression. Unlike the distance model, it doesn't require ensembling predictions from all previous observations. And critically, unlike both alternatives, it provides 'cold start' predictions for speakers with no previous observations, making it more flexible in deployment.

Table 3: *Null model average AUC on different sequence types. The standard deviation is in brackets.*

	Tested on		
Trained on	**Periodic**	**Balanced**	**Random**
Periodic	.922 [.090]	.440 [.046]	.498 [048]
Balanced	.664 [.186]	.751 [.123]	.506 [.037]
Random	.519 [.047]	.515 [.036]	.502 [.029]

4.2. Sensitivity Analysis

Sequence Effect Control: Given twice-daily recordings (pre/post shift), participants' fatigue patterns could show periodicity that models might exploit without learning from speech content. To test possible impacts of this effect, we created a control experiment using a null model that maintains our transformer architecture but replaces speech features with dataset-wide averages. Testing on periodic (alternating high/low fatigue), balanced (fixed-length), and random sequences, we found the model achieves high performance (AUC=0.922) only when both trained and tested on periodic sequences (Table 3). Performance drops to chance level with random ordering. Based on these findings, we use random sequence ordering in all previous experiments to ensure performance gains reflect learning from speech content rather than temporal patterns.

Fairness Analysis: To assess potential biases, we first analysed the dataset using confound-aware metrics [27], finding balanced representation across demographic groups (confound-aware AUC $\sim$ 0.51 for sex, age, and language). However, examining model performance across these groups reveals some concerning patterns:

- Performance is consistently lower for female speakers across all methods, with this disparity most pronounced in the mixed-effects approach.
- Speakers under 40 show lower performance in most methods except cross-sectional, with the prototype model showing particularly strong age-related bias.
- While British and Spanish speakers show similar performance, US English speakers have markedly lower accuracy across all methods.

Figure 3 shows these disparities in detail, averaging performance across observations 0-7 to capture early-stage adaptation. While all methods show some demographic biases, the prototype network appears to amplify these differences more than other approaches as evidenced by the wider spread in AUC across demographic groups for 'Proto'. This suggests that while the transformer achieves the best overall performance, there remains important work to be done in ensuring equitable performance across demographic groups.

5. Conclusions and Future Work

In this work, we reformulate speaker-dependent fatigue monitoring as a meta-learning problem to eliminate the need for constant model retraining. We demonstrate that meta-learning approaches can effectively create personalized fatigue detection models for individual users while maintaining practical feasibility for real-world applications. Our transformer-based approach achieves strong performance (Pearson correlation of 0.55, AUC of 0.78) for individuals with at least 6 previous labelled speech recordings. This represents substantial improvements over traditional approaches (46% over cross-sectional,

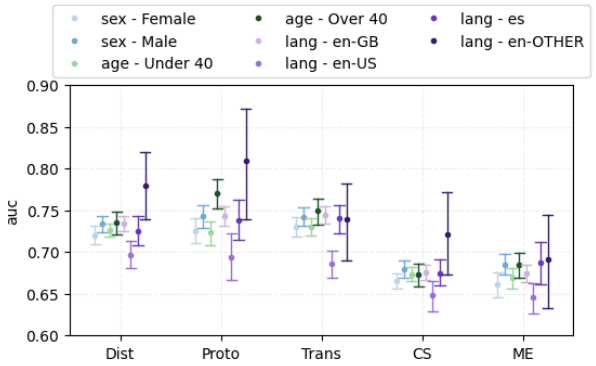

Figure 3: *AUC for points with up to 7 previous observations for different demographics (with 95% CI)*

23% over mixed-effects) while offering key practical advantages: 1) no need for continuous retraining, 2) ability to handle both regression and classification, and 3) availability of predictions even without previous speaker observations.

Our findings have important implications for real-world deployment. The requirement of only a few initial interactions with users to achieve personalisation makes it suitable for health monitoring applications. Users would need to report their fatigue levels for just a few days before receiving fully automated, personalised predictions from speech alone – a considerable advancement in making fatigue monitoring both accurate and practical.

Some limitations of our approach should be noted. While our models show promising overall performance – particularly given the variability in devices and recording settings – we observe performance variations across demographic groups, with differences in accuracy for female speakers and those under 40. Our investigation is also limited to the time since sleep variable, which is only one component of fatigue. Additionally, our current analysis uses fixed-length speech samples from standardised reading tasks. The same dataset includes other speech activities such as image descriptions and question-answering tasks, offering opportunities to explore more naturalistic speech patterns. Future work could investigate how our approach generalises across these different speech tasks, potentially enabling more flexible real-world applications that don't rely on specific text reading.

Importantly, beyond fatigue (operationalised as time since sleep), this approach could be adapted to other mental states or health symptoms where personalisation is crucial. Particularly interesting would be applications to conditions with slower temporal dynamics, such as depression or chronic pain, where collecting diverse samples may be more challenging, but the need for personalised monitoring is equally important. Research could also focus on making these models more robust across demographic groups through targeted data collection or architectural modifications.

Our work demonstrates that meta-learning is feasible for personalised health monitoring from speech while highlighting important considerations for moving such systems from research to real-world deployment. The performances achieved with just a few calibration samples per user suggest this approach could impact how we monitor fatigue and potentially other health conditions in real-world settings.

6. References

[1] J. Menting, C. J. Tack, G. Bleijenberg, R. Donders, H. A. Droogleever Fortuyn, J. Fransen, M. M. Goedendorp, J. S. Kalkman, R. Strik-Albers, N. van Alfen *et al.*, "Is fatigue a disease-specific or generic symptom in chronic medical conditions?" *Health Psychology*, vol. 37, no. 6, pp. 530–543, 2018.

[2] J. P. Brown, D. Martin, Z. Nagaria, A. C. Verceles, S. L. Jobe, and E. M. Wickwire, "Mental health consequences of shift work: an updated review," *Current Psychiatry Reports*, vol. 22, no. 2, pp. 1–7, 2020.

[3] N. R. Adão Martins, S. Annaheim, C. M. Spengler, and R. M. Rossi, "Fatigue monitoring through wearables: A state-of-the-art review," *Frontiers in Physiology*, vol. 12, p. 790292, 2021.

[4] S. Karvekar, M. Abdollahi, and E. Rashedi, "Smartphone-based human fatigue level detection using machine learning approaches," *Ergonomics*, vol. 64, no. 5, pp. 600–612, 2021.

[5] M. Huckvale, A. Beke, and M. Ikushima, "Prediction of sleepiness ratings from voice by man and machine," in *Interspeech 2020*, 2020, pp. 4571–4575.

[6] S. Vekkot, S. Teja Chavali, C. Tej Kandavalli, R. Sai Abhishek Podila, D. Gupta, M. Zakariah, and Y. Ajami Alotaibi, "Continuous speech-based fatigue detection and transition state prediction for air traffic controllers," *IEEE Access*, vol. 13, pp. 3298–3319, 2025.

[7] A. Norbury, A. L. Georgescu, E. Molimpakis, S. Goria, and N. Cummins, "Predicting different dimensions of fatigue from speech data: a longitudinal study in shift workers," https://doi.org/10.1101/2024.06.17.24308769, 2024.

[8] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[9] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[10] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.

[11] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018, pp. 3490–3497.

[12] A. Rafiei, R. Moore, S. Jahromi, F. Hajati, and R. Kamaleswaran, "Meta-learning in healthcare: A survey," *SN Computer Science*, vol. 5, no. 6, p. 791, 2024.

[13] K. Singh and D. Malhotra, "Meta-health: learning-to-learn (meta-learning) as a next generation of deep learning exploring healthcare challenges and solutions for rare disorders: a systematic analysis," *Archives of Computational Methods in Engineering*, vol. 30, no. 7, pp. 4081–4112, 2023.

[14] S. Chopra, P. Mathur, R. Sawhney, and R. R. Shah, "Meta-learning for low-resource speech emotion recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, 2021, pp. 6259–6263.

[15] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA: Curran Associates Inc., 2017, p. 4080–4090.

[16] S. Yadlowsky, L. Doshi, and N. Tripuraneni, "Pretraining data mixtures enable narrow model selection capabilities in transformer models," https://arxiv.org/abs/2311.00871, 2023.

[17] A. L. Oberg and D. W. Mahoney, "Linear mixed effects models," *Methods in Molecular Biology*, vol. 404, pp. 213–234, 2007.

[18] J. A. Caldwell, J. L. Caldwell, L. A. Thompson, and H. R. Lieberman, "Fatigue and its management in the workplace," *Neuroscience & Biobehavioral Reviews*, vol. 96, pp. 272–289, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763418305220

[19] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, UK: Cambridge University Press, 1999.

[20] D. Deterding, "The north wind versus a wolf: Short texts for the description and measurement of english pronunciation," *Journal of the International Phonetic Association*, vol. 36, pp. 187 – 196, 12 2006.

[21] G. Fairbanks, *Voice and Articulation Drillbook*. Harper, 1960. [Online]. Available: https://books.google.co.uk/books?id=qN1ZAAAAMAAJ

[22] A. C. Lammert, J. Melot, D. E. Sturim, D. J. Hannon, R. DeLaura, J. R. Williamson, G. Ciccarelli, and T. F. Quatieri, "Analysis of phonetic balance in standard english passages," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 4, pp. 917–930, 2020. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/2020_JSLHR-19-00001

[23] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," in *Interspeech 2022*. Incheon, Korea: ISCA, 2022, pp. 356–360.

[24] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal paralinguistic speech representations using self-supervised conformers," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3169–3173.

[25] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[27] S. Goria, R. Polle, S. Fara, and N. Cummins, "Revealing confounding biases: A novel benchmarking approach for aggregate-level performance metrics in health assessments," in *Interspeech 2024*, 2024, pp. 1440–1444.