**УДК 004.9**

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ ОТКРЫТЫХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ОПРЕДЕЛЕНИЯ ВОЗРАСТА ПО РЕЧЕВОМУ СИГНАЛУ

**Иванова М.К.**[1]
**Научный руководитель – Зорькина А.А.**[1]
[1]Университет ИТМО
m.k.ivanova314@gmail.com

## Введение

Определение возраста человека по речевому сигналу – одна из актуальных задач в области обработки речи. Она решается в последние годы многими учеными. Существуют различные разработки на основе методов глубокого машинного обучения, достигающие достаточно высоких результатов. Однако практически все решения имеют низкую обобщаемость и показывают хорошие результаты лишь на определенных наборах данных. Перед учеными стоят следующие задачи: уменьшенить ошибку, улучшив точность предсказания, и улучшить обобщаемость результата.

## Основная часть

В области речевых технологий современными учеными предложен ряд моделей для автоматического определения возраста человека по речевому сигналу, демонстрирующих передовые результаты на текущем этапе развития. Среди них существуют решения, демонстрирующие открытый исходный код, что делает возможным воспроизведение экспериментальных результатов. В ходе данного исследования проведено тестирование двух моделей на основе глубоких нейронных сетей: многопрофильного бенчмарка VoxProfile[1], использующего для извлечения признаков фундаментальную модель WavLM[2] и SpeakerProfiling – модели “смесь экспертов”, описанной в “Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model”[3] на основе дообученной для определения возраста модели wav2vec[4]. Для верификации заявленных показателей точности, оценки способности модели обобщаться на новые данные и определения потенциала дальнейшего улучшения проводится тестирование всех трех моделей на разнородных наборах данных. В качестве таких наборов берутся AgeVoxCeleb[5] – крупный мультимодальный сбалансированный набор данных, включающий как аудио- так и видеоматериалы (в данном исследовании будут полезны аудиозаписи в формате .wav, содержащиеся в корпусе данных); TIMIT[6] – классический набор, широко использующийся при решении данной задачи; SeniorTalk[7] – китайский корпус данных, содержащий образцы речи пожилых людей; NNCES[8], включающий в себя записи детских голосов, а также VoiceBiometricAge, собранный в университете ИТМО.

## Выводы

В ходе исследования подтвердились текущие тенденции в области оценки возраста по голосу. Несмотря на активное развитие методов глубокого обучения, ключевая проблема остается нерешенной: системы демонстрируют низкую обобщающую способность на данные, которые кардинально отличаются от обучающей выборки. Для создания устойчивого решения требуется дообучение на сбалансированных выборках, охватывающих весь возрастной диапазон.

## Литература

1. Feng T. et al. Vox-Profile: A Speech Foundation Model Benchmark for Characterizing

Diverse Speaker and Speech Traits //arXiv preprint arXiv:2505.14648. – 2025.
2. Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pretraining for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6):1505–1518, 2022.
3. Gupta T., Truong T. D., Anh T. T., Chng E. S. Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model // Proceedings of Interspeech. – 2022. – С. 1978–1982.
4. Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations //Advances in neural information processing systems. – 2020. – Т. 33. – С. 12449-12460.
5. Tawara N. et al. Age-vox-celeb: Multi-modal corpus for facial and speech estimation //ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2021. – С. 6963-6967.
6. John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. NASA STI/Recon technical report n, 93:27403, 1993.
7. Chen Y. et al. SeniorTalk: A Chinese Conversation Dataset with Rich Annotations for Super-Aged Seniors //arXiv preprint arXiv:2503.16578. – 2025.
8. https://www.kaggle.com/datasets/kodaliradha20phd7093/nonnative-children-english-speech-nnces-corpus