УДК 004.85

ПОДХОДЫ ФЕДЕРАТИВНОГО ОБУЧЕНИЯ К ДИАРИЗАЦИИ И РАСПОЗНАВАНИЮ РЕЧИ В УСЛОВИЯХ ПЕРЕСЕКАЮЩИХСЯ ГОЛОСОВ

Байрамова Х.Б. (ИТМО)
Научный руководитель – Авдюшина А. Е. (ИТМО)

Введение. Современные технологии обработки речи активно развиваются, находя применение в широком спектре задач: от голосовых помощников до аналитики разговоров. Одной из ключевых задач в данной области является диаризация речи, заключающаяся в сегментации аудиозаписи на фрагменты, соответствующие разным говорящим. Однако в реальных условиях разговоры часто сопровождаются наложением голосов, что существенно усложняет задачу автоматической обработки. Традиционные подходы к диаризации и распознаванию речи опираются на централизованное обучение, при котором данные загружаются на сервер для обучения единой модели. Такой подход создает риски нарушения конфиденциальности и не учитывает индивидуальные особенности пользователей. Федеративное обучение (Federated Learning, FL) предлагает альтернативный метод, позволяя моделям обучаться локально на устройствах пользователей без передачи приватных данных. Федеративное обучение уже применяется в ряде задач обработки речи, включая диаризацию и распознавание речи. Однако проблема наложения голосов в контексте FL остается слабо изученной. В данной работе рассматривается возможность использования FL в задаче диаризации и распознавания речи в сценариях с перекрывающейся речью.

Основная часть. Методы диаризации традиционно основаны на кластеризационных алгоритмах, где из аудиозаписи извлекаются эмбеддинги говорящих, которые затем группируются. Ранние подходы использовали статистические методы, такие как BIC [1] и i-vectors [2], но они не обеспечивали высокой точности. С развитием глубокого обучения стали использоваться нейросетевые эмбеддинги (d-vectors, x-vectors) [3], что позволило улучшить качество диаризации. Однако такие методы не оптимизированы напрямую для минимизации ошибок диаризации и плохо работают в условиях наложения речи.

Для решения этих проблем были предложены end-to-end методы диаризации, которые исключают необходимость кластеризации и позволяют обучать нейросеть на задаче разметки говорящих непосредственно [4]. Такие подходы, включая их усовершенствованные версии [5], позволяют лучше обрабатывать записи с переменным числом говорящих, но по-прежнему требуют большого объема аннотированных данных и сложно адаптируются к индивидуальным особенностям пользователей.

Альтернативные методы диаризации речи в условиях наложения голосов используют различные способы представления и обработки говорящих в многоговорящей среде [6]. Они позволяют учитывать одновременно активных говорящих, но остаются зависимыми от централизованного обучения и не всегда сохраняют конфиденциальность данных.

Федеративное обучение уже применяется в задачах обработки речи, обеспечивая локальное обучение моделей без передачи данных пользователей [7]. Однако существующие FL-решения сосредоточены на сегментации речи и не учитывают специфику диаризации в условиях наложения голосов [8].

В данной работе рассматривается возможность объединения современных подходов к диаризации речи и методов распределенного обучения. Целью работы является разработка метода, обеспечивающего баланс между точностью диаризации, возможностью обработки перекрывающейся речи и сохранением конфиденциальности данных.

Выводы. В результате работы рассматриваются современные методы диаризации речи и их ограничения в условиях наложения голосов. Так же предлагается подход, сочетающий преимущества существующих решений, направленный на повышение точности диаризации и распознавания речи при сохранении конфиденциальности данных.

Список использованных источников:

1. Gish H., Schmidt M. R. Text-independent speaker identification // IEEE Signal Processing Magazine. – 1991. – Т. 11. – № 4. – С. 18–32.
2. Dehak N. et al. Front-end factor analysis for speaker verification // IEEE Transactions on Audio, Speech, and Language Processing. – 2011. – Т. 19. – № 4. – С. 788–798.
3. Snyder D. et al. X-vectors: Robust DNN embeddings for speaker recognition // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2018. – С. 5329–5333.
4. Fujita Y. et al. End-to-end neural speaker diarization with permutation-free objectives // IEEE Spoken Language Technology Workshop (SLT). – 2019. – С. 630–637.
5. Horiguchi S. et al. End-to-end neural diarization with encoder-decoder attractors // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2020. – Т. 28. – С. 1–15.
6. Zhihao D. et al. Speaker embedding-aware neural diarization using power-set encoding // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2022. – С. 7092–7096.
7. Bhuyan M. K. et al. Federated learning-based speaker diarization in IoT networks // IEEE Internet of Things Journal. – 2024. – Т. 11. – № 1. – С. 225–237.
8. Chen W. et al. Personalized federated learning for speaker recognition // IEEE Transactions on Neural Networks and Learning Systems. – 2023. – Т. 34. – № 6. – С. 3278–3290.