УДК 004.932:004.934

Крикунов Данила Анатольевич, магистрант, Тюменский Индустриальный Университет, г. Тюмень

АЛГОРИТМИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ТЕМПОРАЛЬНОЙ СИНХРОНИЗАЦИИ ГЕТЕРОГЕННЫХ ДАННЫХ В СИСТЕМАХ РЕЧЕВОЙ АНАЛИТИКИ

Аннотация

В статье рассматриваются методы темпоральной синхронизации гетерогенных данных в системах речевой аналитики. Актуальность исследования обусловлена широким распространением технологий автоматического распознавания речи и диаризации спикеров, результаты которых используются совместно для построения расшифровок диалогов, анализа коммуникаций и формирования аналитических отчетов. Особое внимание уделяется проблеме рассогласования временных меток, возникающего при независимой работе модулей Automatic Speech Recognition и Speaker Diarization.

Предлагается подход к построению алгоритма temporal alignment, обеспечивающего сопоставление словарных временных меток ASR-модели с интервалами активности спикеров, полученными в результате диаризации. Рассматриваются методы overlap-based matching, boundary matching, nearest-segment assignment и оценки confidence score при назначении слов конкретным участникам диалога. Анализируются типовые источники ошибок, включая timestamp drift, overlapping speech, boundary fragmentation и асинхронность вычислительных процессов.

Показано, что применение алгоритмов темпорального выравнивания позволяет существенно повысить точность speaker attribution, уменьшить

количество ошибок распределения реплик между спикерами и повысить качество последующих этапов обработки данных, включая анализ эмоций, семантическую интерпретацию и формирование коммуникационных метрик. Практическая значимость исследования связана с использованием предложенных методов в системах речевой аналитики, контакт-центрах, сервисах анализа переговоров и интеллектуальных диалоговых платформах.

Annotation

The article discusses methods of temporal synchronization of heterogeneous data in speech analytics systems. The relevance of the study is determined by the widespread use of automatic speech recognition and speaker diarization technologies, whose outputs are jointly applied for dialogue transcription, communication analysis and analytical report generation. Particular attention is paid to the problem of timestamp mismatch arising from the independent operation of Automatic Speech Recognition and Speaker Diarization modules.

An approach to the construction of a temporal alignment algorithm is proposed, providing the matching of ASR word-level timestamps with speaker activity intervals obtained through diarization. Methods such as overlap-based matching, boundary matching, nearest-segment assignment and confidence score estimation for speaker attribution are examined. Typical error sources are analyzed, including timestamp drift, overlapping speech, boundary fragmentation and asynchronous inference processes.

The study demonstrates that the application of temporal alignment algorithms significantly improves speaker attribution accuracy, reduces speaker assignment errors and enhances the quality of downstream processing tasks, including emotion analysis, semantic interpretation and communication metric generation. The practical significance of the proposed approach is associated with its application in speech analytics systems, contact centers, negotiation analysis services and intelligent dialogue platforms.

Ключевые слова: темпоральная синхронизация, речевая аналитика, диаризация спикеров, автоматическое распознавание речи, временное

выравнивание данных, временные метки, распределение реплик по спикерам, обработка речевых сигналов

Keywords: temporal synchronization, speech analytics, speaker diarization, automatic speech recognition, temporal data alignment, timestamps, speaker attribution, speech signal processing

Развитие систем речевой аналитики привело к широкому распространению технологий автоматического распознавания речи и диаризации спикеров. Современные платформы анализа переговоров, контакт-центров и корпоративных коммуникаций используют одновременно несколько независимых алгоритмов обработки аудиосигнала, формирующих различные типы данных [1]. Наиболее распространенной архитектурой является сочетание Automatic Speech Recognition (ASR), выполняющего транскрибацию речи, и Speaker Diarization, определяющего принадлежность речевых сегментов конкретным участникам разговора.

В большинстве современных решений модуль распознавания речи и модуль диаризации функционируют независимо друг от друга. Система автоматического распознавания речи формирует текстовую расшифровку с временными метками слов и фраз, тогда как диаризация возвращает временные интервалы активности каждого спикера [2]. В результате возникают ситуации, когда границы речевых сегментов не совпадают с временными метками распознанных слов, что приводит к ошибкам speaker attribution и снижает качество последующего анализа.

Дополнительную сложность создает использование современных ASR-моделей семейства Whisper и faster-whisper. Данные модели демонстрируют высокую точность распознавания речи, однако исходные временные метки сегментов не всегда обеспечивают достаточную точность для последующего связывания слов со спикерами. Для решения данной проблемы применяется технология forced alignment, позволяющая получить word-level timestamps с

более высокой точностью позиционирования слов относительно аудиосигнала [3].

Параллельно с распознаванием речи выполняется процедура speaker diarization, основной задачей которой является определение того, кто и в какой момент времени говорит. Современные системы диаризации используют каскадную архитектуру, включающую Voice Activity Detection, выделение speaker embeddings, сегментацию речевого потока и последующую кластеризацию речевых фрагментов [4]. Одним из наиболее распространенных инструментов в данной области является библиотека pyannote.audio, обеспечивающая построение высокоточных моделей диаризации на основе глубоких нейронных сетей.

Проблема возникает на этапе объединения результатов ASR и диаризации. Даже при обработке одного и того же аудиофайла обе системы могут формировать различные временные границы сегментов. Причинами подобных расхождений являются особенности алгоритмов сегментации, наличие шумов, наложение речи нескольких участников, различия в алгоритмах Voice Activity Detection и эффект timestamp drift [5]. В результате одно и то же слово может частично пересекаться сразу с несколькими сегментами диаризации либо не иметь явного пересечения ни с одним из них.

Для устранения подобных ошибок предлагается использовать алгоритм temporal alignment, основанный на анализе временного пересечения интервалов. На вход алгоритма поступают word-level timestamps, сформированные модулем ASR, и интервалы активности спикеров, полученные в результате диаризации. Для каждого слова определяется временной диапазон его произнесения, после чего вычисляется степень пересечения данного диапазона с каждым сегментом диаризации [6]. Слово назначается тому спикеру, для которого величина пересечения является максимальной. Подобный подход позволяет существенно повысить точность распределения слов между участниками диалога.

Особую важность имеет обработка ситуаций с наложением речи нескольких участников. В реальных коммуникациях перебивания и одновременная речь встречаются достаточно часто, особенно в переговорах, дискуссиях и клиентских обращениях. В подобных случаях стандартный алгоритм сопоставления временных интервалов может приводить к неоднозначному распределению слов между спикерами. Для решения данной проблемы используются дополнительные критерии, включая коэффициент пересечения, анализ ближайших сегментов и оценку confidence score назначения спикера [7]. Такой подход позволяет минимизировать количество ошибочных назначений даже при наличии overlapping speech.

После назначения слов спикерам выполняется группировка слов в реплики. Реплика представляет собой непрерывную последовательность слов одного участника разговора, ограниченную временными рамками его речевой активности. Формирование реплик позволяет перейти от анализа отдельных слов к анализу завершенных коммуникативных высказываний. Полученная структура используется для построения транскриптов, расчета аналитических метрик и последующего применения методов обработки естественного языка.

Качество темпоральной синхронизации оказывает непосредственное влияние на эффективность последующих этапов обработки данных. Ошибки speaker attribution приводят к искажению статистики участия спикеров, ухудшают результаты анализа эмоционального состояния и могут существенно снизить достоверность выводов, формируемых интеллектуальными аналитическими системами.

Перспективным направлением дальнейших исследований является развитие адаптивных методов temporal alignment, учитывающих не только временные пересечения сегментов, но и семантические, акустические и поведенческие признаки коммуникации. Использование комплексного подхода позволит повысить устойчивость алгоритмов синхронизации при работе с шумными данными, большим количеством участников и сложными сценариями взаимодействия.

Таким образом, темпоральная синхронизация гетерогенных данных является одним из ключевых этапов современных систем речевой аналитики. Использование алгоритмов временного выравнивания, основанных на анализе пересечений временных интервалов, позволяет существенно повысить точность распределения реплик между спикерами и обеспечить более высокое качество последующего анализа речевых коммуникаций.

Литература

1. Ронжин А.Л., Будков В.Ю. Анализ современных методов и систем диаризации дикторов // Труды V Российской мультиконференции по проблемам управления. 2012.
2. Карпов А.А. Методология оценивания работы систем автоматического распознавания речи // Труды СПИИРАН. 2012.
3. Титов Ф.М. Обзор задачи автоматического распознавания речи // КиберЛенинка. 2021.
4. Налчаджи К.В. Обзор актуальных открытых решений в области распознавания речи // КиберЛенинка. 2022.
5. Нутфуллин Б.М. Применение синусоидального моделирования речи к задаче диаризации звука // КиберЛенинка. 2021.
6. Bain M., Huh J., Han T., Zisserman A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio // Interspeech. 2023.
7. Bredin H., Laurent A. pyannote.audio: Neural Building Blocks for Speaker Diarization // ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. 2021.

Literature

1. Ronzhin A.L., Budkov V.Yu. Analysis of Modern Methods and Speaker Diarization Systems // Proceedings of the V Russian Multiconference on Control Problems. 2012.
2. Karpov A.A. Methodology for Evaluating Automatic Speech Recognition Systems // SPIIRAS Proceedings. 2012.

3. Titov F.M. Overview of the Automatic Speech Recognition Task // CyberLeninka. 2021.
4. Nalchadzhi K.V. Review of Current Open-Source Speech Recognition Solutions // CyberLeninka. 2022.
5. Nutfullin B.M. Application of Sinusoidal Speech Modeling to the Audio Diarization Task // CyberLeninka. 2021.
6. Bain M., Huh J., Han T., Zisserman A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio // Interspeech. 2023.
7. Bredin H., Laurent A. pyannote.audio: Neural Building Blocks for Speaker Diarization // ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. 2021.