УДК 004.8

AAM-SA-ASR: КОМПАКТНАЯ АРХИТЕКТУРА ДЛЯ ТРАНСКРИПЦИИ С АТРИБУЦИЕЙ ДИКТОРА

Аникин А. А.[1], Новоселов С. А.[1]
Научный руководитель – канд. техн. наук Новоселов С. А.[1]
[1]Университет ИТМО
anikin.aleksn@gmail.com

Введение

Задача Speaker-Attributed Automatic Speech Recognition (SA-ASR) заключается в автоматическом распознавании речи с одновременным определением, «кто что сказал» в многоголосной записи. В отличие от классического ASR, где требуется только транскрипция речи, SA-ASR предполагает также корректную атрибуцию фрагментов высказываний конкретным дикторам, что особенно важно для анализа совещаний, телефонных разговоров и групповых дискуссий.

Дополнительной проблемой является дефицит размеченных многоголосных данных для обучения SA-ASR-систем. В ряде исследований предлагается использовать синтетическую генерацию диалогов для расширения обучающей выборки [1], однако унифицированного подхода к формированию таких данных не существует.

Таким образом, актуальной научной проблемой является разработка компактной и вычислительно эффективной архитектуры SA-ASR, обеспечивающей сопоставимое с современными решениями качество распознавания и атрибуции дикторов при сниженных требованиях к вычислительным ресурсам и объёму обучающих данных.

Основная часть

В работе предлагается компактная архитектура AAM-SA-ASR, основанная на совместной модели распознавания речи и извлечения спикерских эмбеддингов MSA-ASR [2]. В отличие от полноценных сквозных моделей, предлагаемый подход использует предварительно обученную и «замороженную» модель распознавания речи Whisper [3] в качестве базового модуля ASR. Это позволяет избежать затратного переобучения акустической модели и существенно снизить вычислительную сложность.

Для решения задачи атрибуции дикторов используется облегчённый трансформерный декодер спикеров, который работает параллельно с декодером ASR и формирует последовательность спикерских эмбеддингов, синхронизированных с распознанными токенами. Ключевой особенностью архитектуры является механизм кросс-внимания между акустическими представлениями ASR и модулем спикера, что обеспечивает согласованность текстовой и спикерской информации без существенного увеличения числа параметров модели.

Для преодоления проблемы недостатка многоголосных данных предложен оригинальный алгоритм онлайн-генерации синтетических диалогов. Алгоритм формирует многоспикерские записи путём случайной выборки односпикерских аудиофайлов, переразметки границ речевой активности, перемешивания сегментов и их конкатенации с сохранением естественной структуры пауз. Генерация выполняется динамически в процессе обучения, что исключает необходимость хранения больших объёмов синтетических данных и повышает вариативность обучающей выборки. Данный подход развивает идеи синтетического формирования разговоров, предложенные в и делает их более экономичными с точки зрения хранения и вычислений.

Для повышения дискриминативной способности спикерских эмбеддингов применён метод Additive Angular Margin Softmax (AAM-Softmax) с использованием Margin Knowledge Distillation (MKD). В качестве «учителя» используется модель верификации дикторов на

основе Wav2Vec 2.0 [4], а её классификационная голова фиксируется при обучении «студента». Такой подход позволяет перенести знания из задачи автоматической верификации дикторов (ASV) в задачу SA-ASR и улучшить разделимость эмбеддингов различных спикеров без значительного увеличения модели.

Экспериментальные результаты на наборе LibriCSS демонстрируют, что предложенная модель обеспечивает сопоставимое качество с современными совместными решениями, такими как MSA-ASR, при существенном снижении числа параметров и ускорении работы на CPU. При этом достигается компромисс между точностью и вычислительной эффективностью, что делает модель применимой в реальных системах с ограниченными ресурсами.

Выводы

В результате проведённого исследования разработана компактная архитектура SA-ASR, обеспечивающая эффективное совместное распознавание речи и атрибуцию дикторов. Предложенный подход сочетает использование предварительно обученной ASR-модели, облегчённого спикерского декодера и методов AAM-Softmax с дистилляцией знаний, что позволяет приблизиться к уровню современных решений при значительном снижении вычислительных затрат. Практическое применение результатов возможно в системах протоколирования совещаний, мониторинга контакт-центров, судебной и медийной аналитики, где требуется автоматическая расшифровка с указанием говорящих.

Литература

1. Nguyen T. B., Waibel A. Synthetic conversations improve multi-talker ASR //ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2024. – С. 10461-10465.
2. Nguyen T. B., Waibel A. Msa-asr: Efficient multilingual speaker attribution with frozen asr models //ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2025. – С. 1-5.
3. Radford A. et al. Robust speech recognition via large-scale weak supervision //International conference on machine learning. – PMLR, 2023. – С. 28492-28518.
4. Malykh S. et al. STCON NIST SRE24 System: Composite Speaker Recognition Solution for Challenging Scenarios //Proc. Interspeech 2025. – 2025. – С. 3983-3987.