УДК 004

СИНТЕТИЧЕСКИЕ ДАННЫЕ ДЛЯ ПРЕДОБУЧЕНИЯ МОДЕЛИ НЕЙРОННОЙ ДИАРИЗАЦИИ NSD-MS2S В КОНКУРСЕ CHIME8

Тимофеева Т.Н. (ИТМО)

Научный руководитель – кандидат технических наук, Романенко А.Н. (ИТМО)

Введение. Несмотря на быстрое развитие речевых технологий в последнее десятилетие, задача распознавания и диаризации спонтанной речи в реалистичных условиях в присутствии пересечения речи дикторов остается актуальной. Модель нейронной диаризации NSD-MS2S [1], обучаемая на длинных записях длительностью 15-90 минут, с хорошей точностью предсказывает метки дикторов в условиях пересекающейся речи. Однако из-за проблемы ограниченного количества реальных данных возникает задача генерации синтетических данных для предобучения модели. Традиционные методы генерации не в полной мере воспроизводят естественную разговорную динамику и акустическую изменчивость в синтетических данных [2,3]. Так, для получения реверберированных аудио импульсные характеристики комнаты (RIR) рассчитываются перебором параметров комнаты и расстояний между источником и диктором независимо, в результате, некоторые RIR могут соответствовать нереалистичным сценариям. К тому же, сигналы разных дикторов сворачиваются с одной и той же импульсной характеристикой комнаты (RIR), при этом не учитывается отличное взаимное расположение разных дикторов и приемников.

Основная часть. В данной работе для генерации данных предлагается отбирать реалистичные RIR-ы и моделировать перемещение дикторов за счет изменения RIR-ов на протяжении одной сессии.

На первом этапе подготовки синтетических данных было обучено пять классификаторов RIR-ов, каждый из которых насчитывал 100 тысяч классов [4]. Применение этих классификаторов к реальным данным позволило выделить импульсные характеристики комнат, наиболее точно соответствующие целевым условиям. Затем на основе отобранных RIR ов были сгенерированы 100 тысяч многоканальных RIR с двадцатью источниками и десятью приемниками на каждую комнату. Процесс генерации синтетических данных осуществлялся с использованием инструмента BUT EEND [3], который опирается на статистику пауз и перекрытий речи, извлеченную из реальных данных. Этот инструмент был модифицирован таким образом, чтобы сигналы каждого диктора сворачивались с их собственными многоканальными RIR. Разные сегменты речи одного диктора сворачивались с различными RIR, что позволяло имитировать его перемещение. Затем к записям добавлялись реальные шумы из CHiME8 датасетов с уровнем SNR в диапазоне от 5 до 20 дБ.

Выводы. В результате, полученный по предложенной методике 2500-часовой синтетический датасет был использован при обучении NSD-MS2S модели, что позволило повысить точность диаризации в целевых сценариях на 5-20 процентов.

Таким образом, можно заключить, что отбор реалистичных RIR-ов улучшает качество синтетических данных, повышая точность предобученной модели.

Список использованных источников:

1. Yang, Gaobin et al. Neural Speaker Diarization Using Memory-Aware Multi-Speaker Embedding with Sequence-to-Sequence Architecture // ICASSP, 2024, pp. 11626-11630.
2. Chen Z., Yoshioka T. et al. Continuous Speech Separation: Dataset and Analysis // ICASSP 2020.
3. Landini F., Lozano-Diez A., Diez M., and Burget L. From Simulated Mixtures to Simulated Conversations as Training Data for End-to-End Neural Diarization. // Interspeech 2022, pp. 5095–

5099.

4. Khokhlov Y., Prisyach T., Mitrofanov A., Dutov D., Agafonov I., Timofeeva T., Romanenko A., Korenevsky M. Classification of room impulse responses and its application for channel verification and diarization. // Interspeech, 2024.