

**АУГМЕНТАЦИЯ ОБУЧАЮЩЕГО МНОЖЕСТВА СИСТЕМЫ РАСПОЗНАВАНИЯ
РЕЧИ С ПОМОЩЬЮ МОДЕЛИ КЛАССИФИКАЦИИ ИМПУЛЬСНЫХ
ХАРАКТЕРИСТИК ПОМЕЩЕНИЯ**

Митрофанов А.А. (ИТМО)

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

Введение. Автоматическое распознавание речи (ASR) в реальных условиях сталкивается с проблемой снижения качества при работе с аудиозаписями, сделанными в акустических условиях, отличных от обучающих данных. Это явление, известное как проблема несоответствия доменов, проявляется при наличии реверберации, фонового шума или при записи через различные типы микрофонов. Традиционные методы аугментации данных, использующие случайный выбор импульсных характеристик помещения (RIR), не всегда эффективны и могут даже ухудшать качество работы системы, особенно когда характеристики целевой среды заранее неизвестны или сильно варьируются [1].

Основная часть. В данной работе предлагается метод аугментации обучающего множества для системы распознавания речи, основанный на использовании модели классификации импульсных характеристик помещения. Метод был разработан и протестирован в рамках международного конкурса по распознаванию речи CHiME-8 Challenge [2], аудиофайлы в котором представляют собой записи многодикторных разговоров в различных акустических условиях.

Для подготовки аугментированного набора данных были выполнены следующие шаги:

1. С начала был сгенерирован набор искусственных RIR с использованием специализированного генератора. Было обучено пять различных RIR-классификаторов, каждый на разных наборах из 100 тысяч искусственно сгенерированных RIR. Эти классификаторы обучались на данных Librispeech, аугментированных шумами, извлеченными из данных CHiME-8.

2. Все записи из обучающих данных CHiME-8 были классифицированы для определения, какой искусственный RIR наиболее похож на RIR каждого канала записи. Это было сделано для каждого классификатора, и 20% RIR с наивысшей вероятностью для каждого подмножества данных CHiME-8 были отобраны, что в итоге дало 20 тысяч отобранных RIR на каждый классификатор или 100 тысяч RIR в общей сложности.

3. Параметры помещений отобранных RIR были использованы для генерации набора многоканальных RIR с 20 источниками и 10 приемниками для каждого помещения. Для размещения источников и микрофонов в помещении был разработан специальный алгоритм, обеспечивающий реалистичное расположение с соблюдением ряда условий: источники не должны располагаться слишком близко друг к другу, к микрофонам и к границам помещения.

Выводы. Предложенный метод аугментации обучающей выборки с автоматическим выбором акустических характеристик среды с помощью нейросетевой модели классификации акустических условий демонстрирует значительное повышение качества распознавания речи на внедоменных данных. Это позволяет создавать более универсальные системы ASR, способные эффективно функционировать в разнообразных акустических условиях без необходимости переобучения или сложной адаптации под каждый конкретный сценарий использования.

Список использованных источников:

1. Cornell S. et al. The chime-8 dasr challenge for generalizable and array agnostic distant automatic speech recognition and diarization //arXiv preprint arXiv:2407.16447. – 2024.
2. Mitrofanov A. et al. STCON System for the CHiME-8 Challenge //arXiv preprint arXiv:2410.13411. – 2024.