УДК 004.8

КОМПЛЕКСНАЯ СИСТЕМА ВЕРИФИКАЦИИ ДИКТОРА ДЛЯ СЛОЖНЫХ КРОСС-КАНАЛЬНЫХ И КРОСС-ЯЗЫКОВЫХ УСЛОВИЙ

Малых С.Ю. (ИТМО)

Научный руководитель – кандидат технических наук Новоселов С.А.
(ИТМО)

Введение. Задача автоматической верификации диктора по голосу (Automatic Speaker Verification, ASV) - одна из главных задач машинного обучения, связанного с аудио. Данная задача широко используется в системах безопасности, биометрической аутентификации и голосовых помощниках. В верификации диктора алгоритму предоставляются аудиозапись с речью и утверждение о том, что это голос конкретной личности. Алгоритм должен проверить это утверждение, сравнив голос с известным образцом голоса этого человека.

Демонстрируя передовые результаты в области микрофонного канала (протоколы VoxCeleb), современные системы верификации диктора редко тестируются в сложных акустических условиях, таких как телефонный канал или микрофон дальнего поля. В данной работе исследуется устойчивость моделей к сложным кросс-канальным и кросс-языковым условиям. Путем объединения моделей создается единая комплексная система. В качестве бенчмарка используется датасет, предоставленный в рамках международного конкурса NIST Speaker Recognition Evaluation (SRE) 2024.

Основная часть. Модели верификации диктора исследовались в двух вариантах условий: фиксированные (с ограничением в допустимых данных) и открытые (без ограничений в данных).

Для фиксированных условий исследовались нейросетевые модели, основанные на сверточной архитектуре: ResNet, ECAPA-TDNN [1], CAM++, NeXt-TDNN и др. Количество обучаемых параметров фиксировалось на уровне 15-45 миллионов во избежание переобучения. Во время расчета градиентов использовались сегменты с большой длительностью (16-20 секунд), что значительно ускоряло сходимость. Использовались дополнительные способы постобработки эмбеддингов и результатов сравнений: классовые эмбеддинги (выходы классификационного слоя), канальная нормализация, доменная адаптация через центрирование.

Для открытых условий рассматривались большие (250+ миллионов обучаемых параметров) нейросети с архитектурой трансформер, предобученные на больших объемах неразмеченных данных: wav2vec XLS-R 1B [2], w2v-BERT 2.0 v2, XEUS. Количество эпох обучения фиксировалось на уровне 10-15 во избежание переобучения. Данные для обучения задаче верификации диктора аккуратно подбирались путем проведения большого количества экспериментов. Использовался алгоритм диаризации для обработки записей, в которых присутствовало несколько дикторов.

Для тестирования моделей применялся датасет NIST SRE 2024, в котором были представлены сложные случаи сравнений: мультиязычность, кросс-язычность, кросс-канальность, шумы и реверберация. Для создания системы, устойчивой к подобным условиям, была использована комплексная структура. Система состояла из нескольких модулей: предобработки, диаризации, экстрактора эмбеддингов, постобработки эмбеддингов и результатов сравнений. Было продемонстрировано, что подобная структура позволяет улучшить результаты на 3.71 п.п. в фиксированных условиях и на 0.87 п.п. в открытых условиях в терминах метрики Equal Error Rate (EER).

Выводы. Разработана комплексная система верификации диктора, устойчивая к эффектам кросс-канала, кросс-языка и доменного смещения. Применение данной системы улучшило значение метрики EER на 11.4 п.п. относительно open-source моделей.

Список использованных источников:

1. Desplanques B., Thienpondt J., Demuynck K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification // In Proceedings Interspeech 2020 – 2020. – С. 3830-3834.
2. Babu A. et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale // In Proceedings Interspeech 2022 – 2022. – С. 2278-2282.