УДК 004.934

ЭКСПРЕСС-ОЦЕНКА КАЧЕСТВА ЗАПИСИ РЕЧИ В УЧЕБНОМ КЛАССЕ

Давыдов Д. А.[1],
Научный руководитель – канд. техн. наук Столбов М. Б.[1]
[1]Университет ИТМО
340193@niuitmo.ru

Введение

Объективная оценка разборчивости речевых сигналов (РРС) является важной задачей в телекоммуникациях, криминалистике, архитектурной акустике, при разработке систем синтеза речи и ряде других речевых технологий. Одним из эффективных методов объективной оценки РРС является метод STOI (англ. Short-Time Objective Intelligibility) [1]. Однако этот метод основан на сравнении тестируемого и эталонного сигнала, что ограничивает область его практического использования. Перспективным направлением развития подходов к оценке РРС является разработка безэталонных методов. В рамках этого направления в последнее десятилетие предложен ряд методов на основе машинного обучения, например NISA [2], LCQA [3]. Точность этих методов зависит от качества ручного конструирования признаков речевых сигналов. Другим перспективным направлением безэталонной оценки РРС является использование нейронных сетей. Целью данной работы является разработка и валидация метода безэталонной оценки разборчивости речи на основе глубокой сверточной нейронной сети, предсказывающей значение индекса STOI непосредственно по искаженному сигналу.

Основная часть

Суть предлагаемого решения заключается в применении глубокой сверточной нейронной сети для сквозного предсказания индекса разборчивости STOI. В отличие от классических методов, использующих ручное извлечение признаков, разработанная модель обучается непосредственно на аудиосигналах, что позволяет ей автоматически выявлять наиболее информативные иерархические представления, связанные с разборчивостью речи. Архитектура сети состоит из каскада сверточных блоков с увеличивающимся числом фильтров, за которыми следуют полносвязные слои с регуляризацией для решения задачи регрессии.

Для обучения и тестирования модели был сформирован репрезентативный корпус данных на основе англоязычного датасета CMU-MOSEI [4]. Высококачественные записи подвергались контролируемым искажениям: аддитивному белому гауссовскому шуму с уровнем от -10 до +20 дБ и реверберации с временем RT60 от 0,25 до 1,8 с. Для каждой из 15407 полученных фонограмм с помощью эталонного алгоритма было рассчитано целевое значение STOI. Такой подход обеспечил широкий охват акустических условий и наличие точной целевой метрики для обучения модели.

Обучение модели проводилось с использованием оптимизатора Adam и функции потерь MSE. Результаты тестирования на независимой выборке показали высокую эффективность предложенного метода. Модель достигла коэффициента детерминации R^2 = 0,90, среднеквадратичной ошибки RMSE = 0,10 и средней абсолютной ошибки MAE = 0,06. Коэффициент ранговой корреляции Спирмена составил 0,93, что свидетельствует о сильной монотонной связи между предсказаниями и эталонными значениями. Анализ распределения ошибок подтвердил отсутствие систематического смещения, а основная часть ошибок сосредоточена в интервале, приемлемом для практического применения.

Разработанный метод был применен для оценки РРС на фонограммах, записанных в учебном классе. В качестве опорного использовался речевой сигнал, воспроизводимый

акустической колонкой. Оценивалась разборчивость речи, записанной на дистанции 1, 2, 4 и 8 м от аудиоколонки. Средние значения индекса STOI, предсказанные разработанной моделью для этих условий, составили: 0.7583 (1 м), 0.7499 (4 м), 0.7431 (8 м) и 0.7358 (2 м) соответственно. Для сравнения, значения, полученные с помощью эталонного алгоритма, составили в среднем от 0.4544 до 0.5161, что подчеркивает чувствительность модели к изменениям акустических условий.

Выводы

В работе представлен метод безэталонной оценки разборчивости речи на основе глубокой сверточной нейронной сети. Основные результаты: разработана архитектура для сквозного предсказания STOI, создан репрезентативный корпус данных с контролируемыми искажениями, экспериментально подтверждена высокая точность (R^2 = 0,90). Разработанная модель может быть использована для экспресс-оценки разборчивости в учебных аудиториях, системах мониторинга качества голосовой связи и при предварительном анализе фонограмм в криминалистике. Дальнейшие исследования будут направлены на расширение типов искажений (нестационарные шумы, артефакты кодеков) и адаптацию модели к работе в реальном масштабе времени.

Литература

1. Taal C. H. et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech //IEEE Transactions on audio, speech, and language processing. – 2011. – Т. 19. – №. 7. – С. 2125-2136.
2. Sharma D. et al. A data-driven non-intrusive measure of speech quality and intelligibility //Speech Communication. – 2016. – Т. 80. – С. 84-94.
3. Grancharov V. et al/ Low complexity, nonintrusive speech quality assessment //IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 6, pp. 1948–1956, Nov. 2006.
4. Zadeh A. A. B. et al. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2018. – С. 2236-2246.