УДК 004.934

ЭКСПРЕСС-ОЦЕНКА КАЧЕСТВА ЗАПИСИ РЕЧИ В УЧЕБНОМ КЛАССЕ

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

Введение

Задача оценки разборчивости речевых сигналов (РРС) является важной в различных областях: телекоммуникациях, криминалистике, при разработке систем синтеза речи, архитектурной акустике и многих других областях. Одним из методов, которые признанны в научной среде и хорошо показали свою эффективность является STOI (англ. Short-Time Objective Intelligibility) [1]. Однако, он эталонный, то есть основан на сравнении эталонного и тестируемого сигналов, что очевидно накладывает на возможность применения этого метода ограничения, связанные с необходимостью наличия эталонного сигнала при оценке РР. В связи с этим актуальной задачей является разработка безэталоных методов, для работы которых, эталонный сигнал не нужен. По этой причине, в последнюю декаду лет было предложено несколько решений в основу которых лёг data driven подход (подход на основе машинного обучения), например NISA [2], LCQA [3]. Эти способы оценки РР параметрические [2, 3], а значит их точность зависит от качества и полноты признаков, которые конструируются вручную. Учитывая это, перспективным направлением безэталонной оценки РРС является использование нейронных сетей. Целью данной работы является разработка и оценка метода безэталонной оценки разборчивости речи на основе глубокой сверточной нейронной сети, предсказывающей значение индекса STOI непосредственно по искаженному сигналу.

Основная часть

Предлагается решение, основанное на применении глубокой свёрточной нейронной сети для предсказания индекса разборчивости STOI. В отличие от методов NISA [2] и LCQA [3], где используется ручное извлечение признаков для предсказании РР, разработанная нейронная сеть обучается непосредственно на искажённых аудиосигналах, что даёт ей возможность выделять самостоятельно из них наиболее информативнее признаки, связанные с разборчивостью речи, и предсказывать РР непосредвенно по искажённому сигналу. Архитектура нейронной сети состоит из каскада свёрточных блоков с увеличивающимся числом фильтров, за которыми следуют полносвязные слои с регуляризацией.

Для обучения и тестирования нейронной сети был собран корпус данных, в основу которого лёг англоязычный датасет CMU-MOSEI [4]. К аудиозаписям были добавлены различные искажения: аддитивный шум с отношениями С/Ш от -10 до +20 дБ и реверберация, с временем реверберации RT_{60} от 0,25 до 1,8 с. Для каждой полученной пары сигналов (эталон – искажённая запись) была вычислена мера STOI [1]. Такой подход к конструированию датасета обеспечил широкий диапазон акустических условий для обучения модели. Для обучения модели использовался оптимизатор Adam, в качестве функции потерь использовалась среднеквадратичная ошибка (MSE). Тестирование проводилось на модельных и натурных данных.

Тестирование на модельных данных показало высокую эффективность предложенного метода: коэффициент детерминации $R^2 = 0{,}90$, среднеквадратическая

ошибка $MSE = 0{,}10$ и средняя абсолютная ошибка $MAE = 0{,}06$. Коэффициент ранговой корреляции Спирмена 0,93 свидетельствует о сильной монотонной связи между предсказаниями и эталонными значениями. Анализ распределения ошибок показал отсутствие систематического смещения.

Тестирование на натурных данных проводилось на фонограммах, записанных в учебном классе. В качестве эталонного сигнала использовалась фонограмма, воспроизводимая акустической колонкой. Оценивалась PP на аудиозаписях, записанных на расстоянии в 1 м, 2 м, 4 м и 8 м от колонки. Средние значения индекса STOI, предсказанные нейронной сетью, составили: 0,7583 (1 м), 0,7499 (4 м), 0,7431 (8 м) и 0,7358 (2 м). Значения, полученные с помощью эталонного алгоритма, составили в среднем от 0,4544 до 0,5161, что свидетельствует о необходимости дальнейшего совершенствования модели.

Выводы

В работе представлен метод безэталонной оценки показателя разборчивости речи STOI на основе глубокой сверточной нейронной сети. На модельных данных экспериментально подтверждена высокая точность модели ($R^2 = 0{,}90$). Дальнейшая работа будет направлены на расширение корпуса данных, совершенствование архитектуры нейронной сети и для анализа фонограмм в криминалистике и других областях.

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. – pp. 2125-2136.
2. Sharma D. et al. A data-driven non-intrusive measure of speech quality and intelligibility //Speech Communication. – 2016. – Т. 80. – pp. 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. – pp. 2236-2246.