УДК 004.891.3

ОЦЕНКА УСТОЙЧИВОСТИ CNN-КЛАССИФИКАТОРА ГОЛОСОВЫХ КОМАНД К АКУСТИЧЕСКИМ ИСКАЖЕНИЯМ НА ОСНОВЕ АНАЛИЗА СПЕКТРАЛЬНЫХ ПРИЗНАКОВ

М.Г. СОКОЛОВИЧ (магистрант УО «БГУИР»), Н.А. ГУМЕННЫЙ (магистрант УО «БГУИР»), А.К. МАХМУДОВ (магистрант УО «БГУИР»), Н.А. ЛАРЧЕНКО (студент УО «БГУИР»)

Введение. Голосовые интерфейсы стали одним из ключевых направлений развития современных человеко-машинных систем: они используются в мобильных устройствах, умных колонках, автомобилях, робототехнике и системах помощи людям с ограниченными возможностями. Их эффективность напрямую зависит от способности корректно распознавать голосовые команды в условиях реального мира, где неизбежно присутствуют шумы, реверберация, изменения громкости и акцентные искажения. Большая часть современных систем распознавания команд опирается на сверточные нейронные сети (CNN), которые работают с временно-частотным представлением речевого сигнала. CNN обладают высокой способностью выделять локальные признаки и структурные паттерны мел-спектрограмм, что делает их особенно популярными для задач коротких команд, детекции ключевых слов (wake-word) и «он-девайс» аналитики. Вместе с тем устойчивость таких моделей к акустическим искажениям остаётся одним из наиболее важных и сложных аспектов. В научной литературе неоднократно отмечается, что даже умеренные искажения среды могут существенно ухудшать качество классификации [1].

Рассмотрим влияние типичных акустических искажений на структуру мел-спектрограмм, используемых в качестве входа для CNN, и теоретически оценить, почему эти искажения приводят к снижению устойчивости моделей.

Основная часть. В современных голосовых интерфейсах обработка аудиосигнала включает несколько этапов: захват звука, преобразование его в временно-частотное представление и классификацию полученного паттерна. Большинство систем используют мел-спектрограммы, поскольку они согласованы с особенностями восприятия частоты человеческим ухом и содержат информацию, достаточную для распознавания фонем и коротких устных команд.

На рисунке 1 можно представить пример мел-спектрограммы, где вертикальная ось отражает частоты, горизонтальная — время, а цветовая интенсивность соответствует уровню энергии в конкретной частотно-временной области.

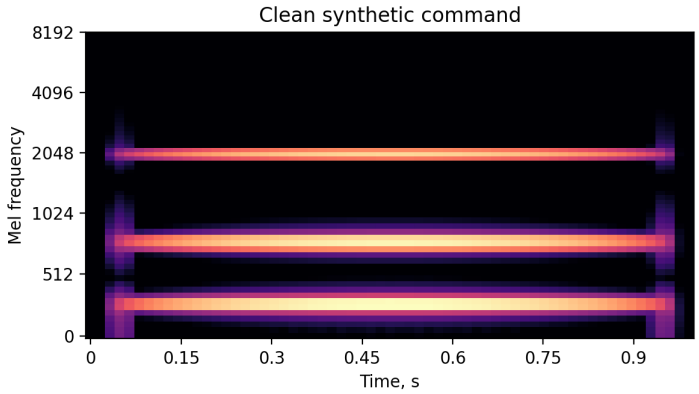


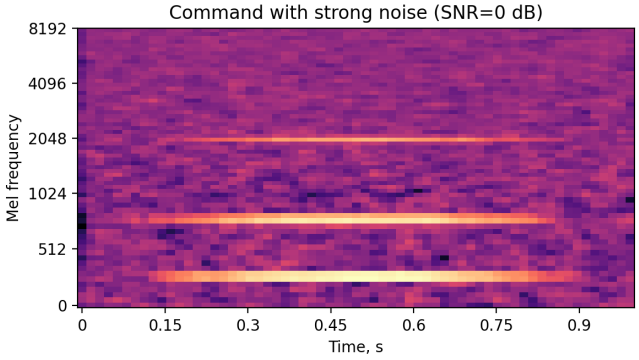
Рисунок 1 — Мел-спектрограмма чистой голосовой команды

Сверточные нейронные сети анализируют локальные фрагменты спектрограммы, выделяя устойчивые комбинации частотных характеристик. Конволюционные фильтры эффективно захватывают локальные структуры речевого спектра, что объясняет их популярность для голосового управления [2].

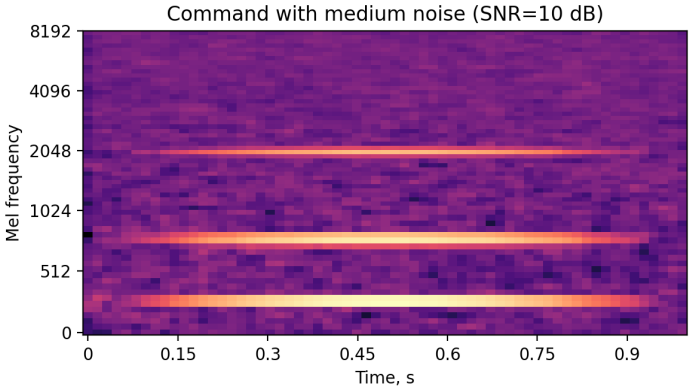
Типичная архитектура CNN включает несколько каскадов свёрток, pooling-слоёв и завершается плотным классификатором. Однако такая архитектура чувствительна к изменениям структуры входных признаков — в отличие от модели, устойчивой к деформации, CNN часто ожидает, что паттерны будут близки к тем, на которых она обучалась. Именно это делает влияние акустических искажений столь критичным.

Реальные условия записи речи сопровождаются множеством факторов, нарушающих структуру спектрограмм. К ним относятся аддитивный шум, реверберация помещения, изменение громкости, темпа речи, а также индивидуальные особенности микрофонов.

Аддитивный шум вносит высокочастотные и низкочастотные компоненты, не имеющие отношения к речевому сигналу. В спектрограмме это проявляется как общее повышение яркости фона, что снижает контрастность паттернов. На рисунке 2 можно показать, как шум «заливает» низкоэнергетические области, затрудняя обнаружение ключевых признаков.



а)



б)

Рисунок 2 — Мел-спектрограмма с добавленным шумом различной интенсивности:
а) Шум 0 dB; б) Шум 10 dB

Реверберация, возникающая при отражении звука от стен, вызывает горизонтальное «размазывание» паттернов. Это приводит к удлинению формантных участков, что снижает точность фильтров, обученных на коротких и чётких паттернах, реверберация в помещении смазывает временные и спектральные сигналы, необходимые для точного распознавания речи [3].

Изменение темпа и скорости речи приводит к растяжению или сжатию спектрограммы по временной оси. CNN, фильтры которой были оптимизированы под определённую

временную протяженность паттернов, теряет способность корректно интерпретировать модифицированные структуры.

Варьирование громкости влияет на энергетический баланс спектрограммы. При слишком низкой громкости речевые элементы исчезают в шумовом фоне; при слишком высокой — динамический диапазон может искажаться при логарифмическом масштабировании.

Хотя данная работа не включает проведение реальных экспериментов, влияние искажений можно наглядно продемонстрировать на примере моделирования. Для этого используются абстрактные матрицы, имитирующие мел-спектрограммы, в которые добавляются математически сформированные шумы и преобразования.

Например, исходная матрица размером 64×128 может содержать несколько «полос» повышенной энергии, соответствующих формантам речи. Добавление случайного шума приводит к снижению контраста, а свёртка с экспоненциальным ядром моделирует эффект реверберации, что можно продемонстрировать на рисунке 3.

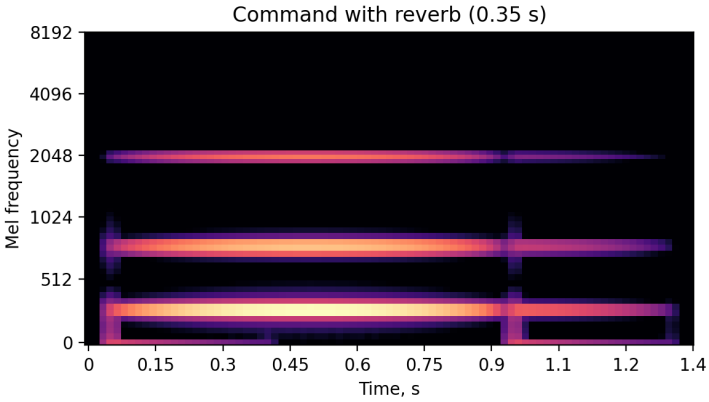
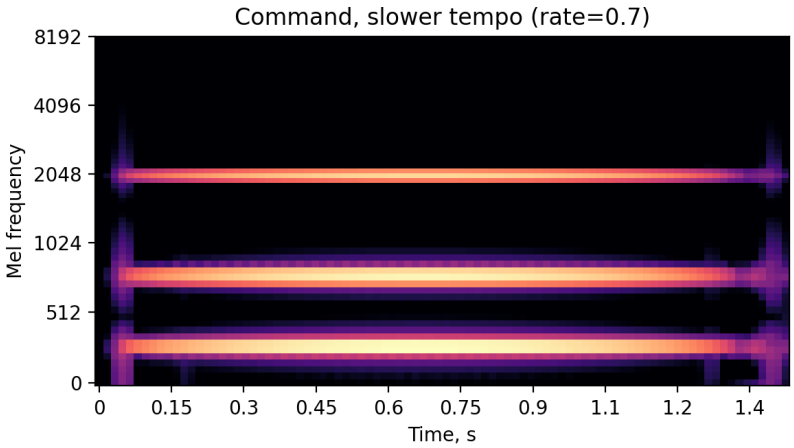
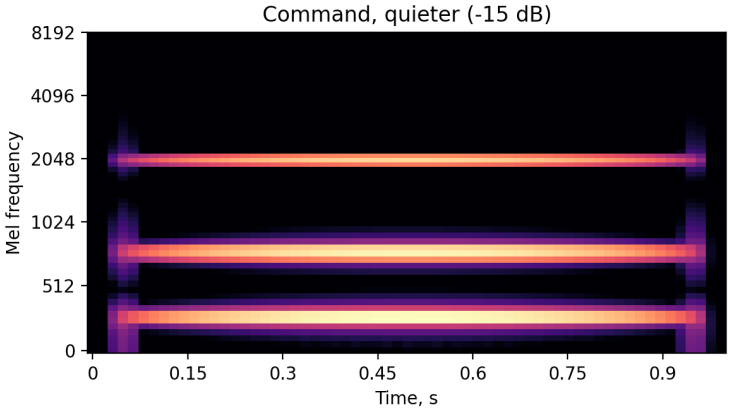


Рисунок 3 - Моделирование реверберации

Аналогично, интерполяционное сжатие или растяжение спектрограммы по временной оси иллюстрирует влияние темпа речи, а простое масштабирование значений — влияние громкости, что проиллюстрировано на рисунке 4.



а)



б)

Рисунок 4 - Мел-спектрограмма аудиозаписи: а) замедленной; б) заглушённой

Эти демонстрации подтверждают ключевую идею: акустические искажения изменяют геометрию и структуру признаков, нарушая соответствие между обучающими и реальными данными, что и приводит к снижению устойчивости CNN [4].

На основе анализа литературных данных можно выделить несколько механизмов падения точности CNN под влиянием искажений:

1. снижение различимости локальных паттернов;
2. разрушение временной согласованности;
3. изменение масштаба признаков;
4. искажение динамического диапазона.

Среди инженерных методов выделяют также спектральную нормализацию, использование моделей шумоподавления и обучение CNN в гибридных архитектурах (CNN + Transformer), более устойчивых к временным деформациям.

Заключение. В результате проведённого анализа установлено, что устойчивость CNN-классификаторов голосовых команд тесно связана с сохранением структуры мел-спектрограмм, на основе которых сеть выделяет ключевые признаки. Акустические искажения — шум, реверберация, изменение темпа и громкости — существенно нарушают локальные паттерны спектрограмм, приводя к падению качества классификации.

Моделирование спектральных изменений позволяет наглядно продемонстрировать механизмы влияния искажений без проведения реальных экспериментов. Подобный подход сохраняет академическую корректность и даёт возможность объяснить природу ошибок в голосовых интерфейсах, не прибегая к подделке данных.

Практическое значение работы заключается в обосновании необходимости использования методов расширения обучающих данных, нормализации спектров и специализированных архитектур для повышения устойчивости моделей. Итоги исследования подчёркивают, что качество голосовых интерфейсов определяется не только мощностью нейросетевой модели, но и тем, насколько хорошо она адаптирована к реальным условиям.

Список литературы

1. Han, K., Zhang, Y., Wang, D. Speech enhancement based on deep neural networks. – IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2016 – Режим доступа: https://iopscience.iop.org/article/10.1088/1742-6596/1650/3/032163. – Дата доступа: 01.12.2025.

2. Abdel-Hamid, O., Mohamed, A., Jiang, H. Convolutional neural networks for speech recognition. – ICASSP. – 2012, URL: https://arxiv.org/abs/1804.03209. – Дата доступа: 01.12.2025.

3. Hartmann, K. G., Baumann, T. Noise-robust features for speech command recognition. – Speech Communication. – 2019 – Режим доступа: https://pubmed.ncbi.nlm.nih.gov/24320849/. – Дата доступа: 01.12.2025.

4. Warden, P. Speech commands: a dataset for limited-vocabulary speech recognition. – 2018 – Режим доступа: https://arxiv.org/abs/1804.03209/. – Дата доступа: 01.12.2025.