Speech enhancement augmentation for robust speech recognition in noisy environments

II Международная Школа-семинар HMMOCS-II 2023: Гибридные методы моделирования и оптимизации в сложных системах

Рауф Насретдинов, Андрей Лепендин, Илья Ильяшенко
Алтайский государственный университет

Актуальность

Развитие современных систем распознавания речи (ASR) в значительной степени обусловлено прогрессом в области глубокого обучения. Использование нейросетевых моделей и подходов привело к существенному повышению точности до уровня, близкого к человеческому

Искажение речевого сигнала за счет фоновых шумов, помех в канале регистрации или передачи сигнала, влияние эффекта реверберации заметно снижает эффективность нейросетевых моделей распознавания речи

Подходы к решению этой проблемы:

- применение предварительной обработки для улучшения речи
- использование аугментации за счет искусственного “ухудшения” голосовых образцов

Они не всегда могут быть использованы одновременно, так как обученная на аугментированных данных модель распознавания речи может плохо работать на неискаженных или очищенных от шума примерах

Цели и задачи

Цель работы: разработка нового подхода к аугментации речевых данных при обучении систем распознавания речи, предназначенного для совместного применения нейросетевых моделей распознавания с алгоритмами / моделями улучшения качества речи

Задачи:

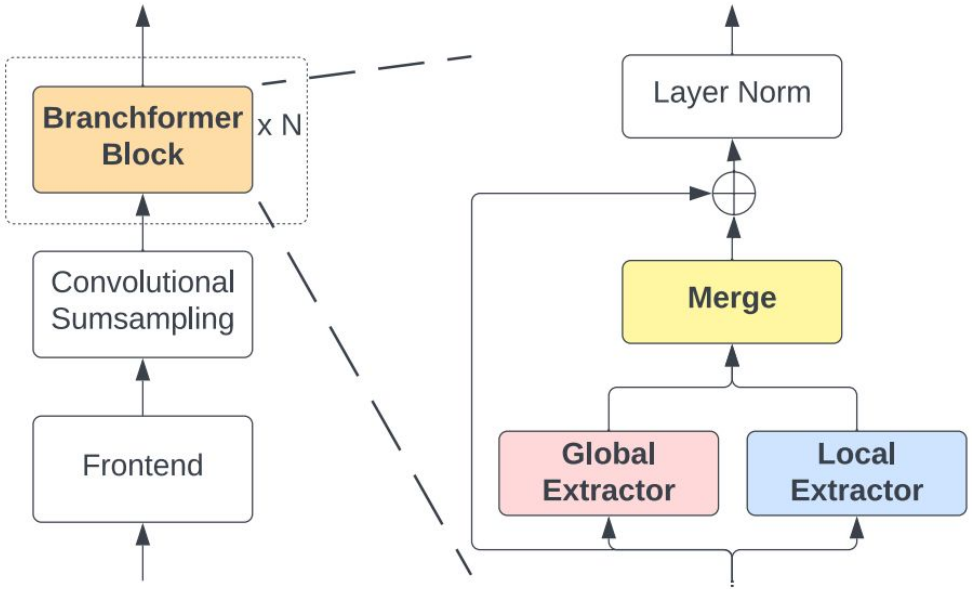
1) создание нескольких модифицированных выборок голосовых примеров для обучения системы ASR на основе корпуса LibriSpeech
2) проведение численных экспериментов на модельной системе ASR для апробации предложенного подхода с использованием 100-часовых выборок
3) анализ результатов экспериментов с качественной оценкой вклада каждого из аугменированных наборов в качество работы на тестовых выборках
4) проверка эффективности подхода на большой 960-часовой выборке

Модель распознавания речи

В данной работе использовалась сеть E-Branchformer, развивающая подход работы Conformer

За извлечение глобальных признаков сигнала отвечали слои самовнимания, а часть, извлекающая локальные признаки, состояла из блока на основе Convolutional Spatial Gating Unit

Отличие E-Branchformer от Conformer состояло в том, что блоки извлечения локальных и глобальных признаков работали параллельно

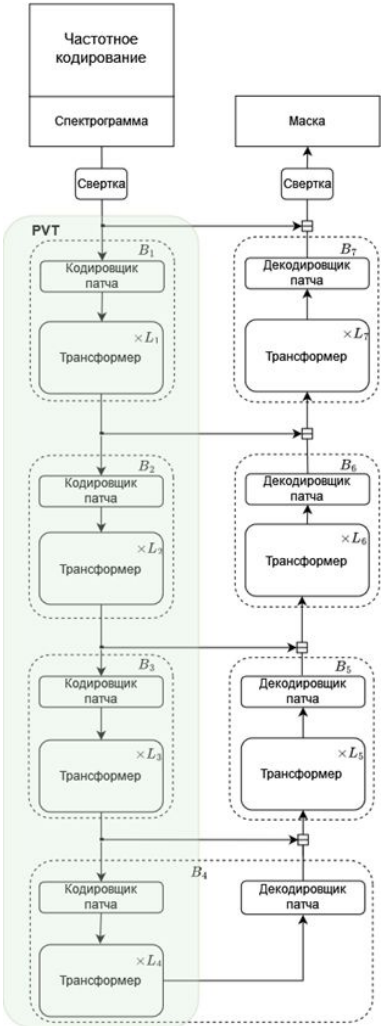


K. Kim, F. Wu, Y. Peng, *E-Branchformer: Branchformer with Enhanced Merging for Speech Recognition*, in Proceedings of the IEEE SLT Workshop (2023)

Модель шумоочистки

Для построения дополнительных обучающих и тестовых наборов данных в работе использовалась оригинальная модель улучшения качества речи Modified Encoder-Decoder Pyramid Transformer

Основана на архитектуре визуального трансформера, расширенного до пирамидальной структуры типа U-Net сети



R. Nasretdinov, I. Ilyashenko, J. Filin, A. Lependin, *Hierarchical Encoder-Decoder Neural Network with Self-Attention for Single-Channel Speech Denoising,* CCIS, **1733**, Springer, Cham (2023)

Используемые наборы данных

Корпус речевых образцов Librispeech

Записи аудиокниг на английском языке из корпуса LibriVox.

Выделялись подмножества:
- продолжительностью 100 ч («clean_100»)
- продолжительностью 360 ч («clean_360»)
- продолжительностью 500 ч («other_500»)

Использовались подмножества «clean_100» и объединение всех подмножеств (длительностью 960 ч)

Набор данных для обучения методов улучшения качества DNS Challenge 2021

Использовалось подмножество шумов из этого набора. Примеры в этом наборе отбирались из нескольких источников таким образом, что на каждый класс шумов приходилось не менее 500 записей, тем самым решалась проблема дисбаланса классов. Выборка включала в себя всего 150 классов шумов, общим числом более 65 000 образцов, каждый продолжительностью не менее 10 секунд.

Используемые метрики качества

1. Для оценки качества распознавания речи использовалась метрика **Word Error Rate (WER) -** точность распознавания слов в тестовых примерах

2. Для оценки качества аудиозаписей использовалась метрика **DNSMOS**, представляющая собой предсказанное предварительно обученной нейронной сетью среднее значение человеческой экспертной оценки. Шкала измерения DNSMOS - от 1 (неприемлемое качество) до 5 (отличное качество)

Модифицированные голосовые примеры

Обозначение выборки	Краткое описание примененного преобразования	Длительность, ч	DNSMOS на 100 ч
Clean	Без преобразований. Оригинальные образцы из набора данных LibriSpeech	100 / 960	4.02
Clean_sp	Применение аугментации изменения скорости речи с коэффициентами 0.9, 1.0, 1.1 к выборке Clean	300 / 2880	4.02
Enhanced	Применение модели улучшения качества к чистым примерам выборки Clean	100 / 960	4.06
Noisy	Добавление шума к подвыборке Clean с использованием шумовых образцов из DNS Challenge 2021	100 / 960	2.72
Denoised	Применение модели улучшения качества к образцам из выборки Noisy	100 / 960	3.59

Детали реализации и обучения

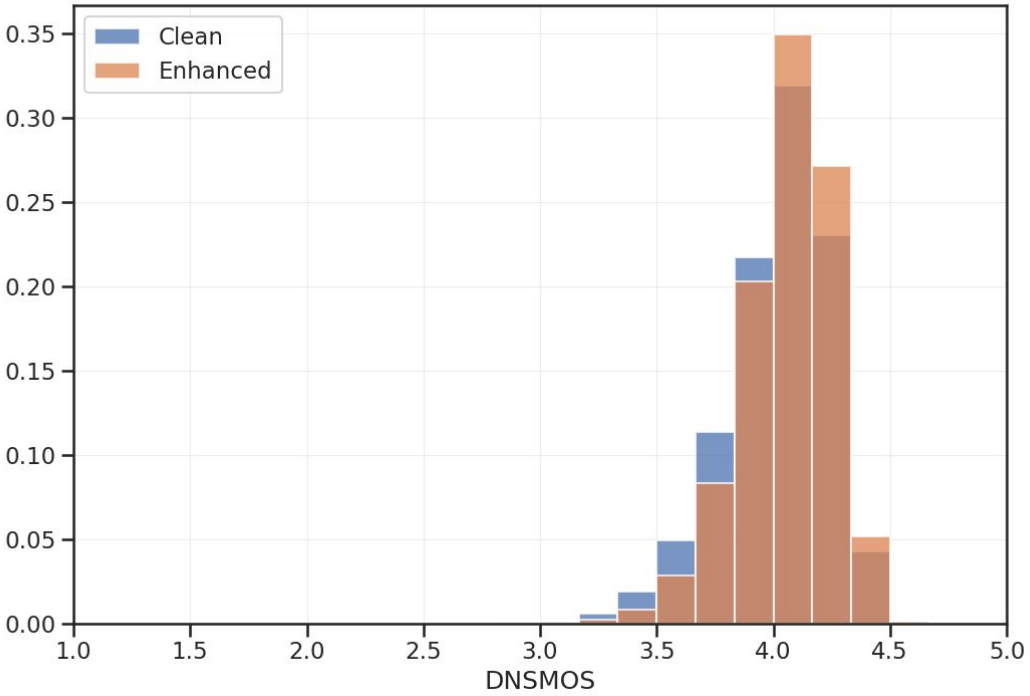
Акустические признаки	80-размерные мел-частотные спектральные коэффициенты
Скорость обучения	Warmup с пиком на 15% обучение и максимальным значением 10^{-3}
Функция потерь	CTC+Attention
Количество итераций	**100 ч** – $3.55 \cdot 10^5$ / **960 ч** – $1.7 \cdot 10^6$
Оптимизатор	Adam (стандартные параметры)
GPU	2 x Nvidia 1080 Ti

Обсуждение результатов 100 ч. Добавление Enhanced

Обучающая выборка	WER на тестовых выборках, %					
	Clean	Other	Clean Denoised	Clean Noisy	Other Denoised	Other Noisy
Clean	7.0	18.4	31.4	57.8	51.5	74.2
Clean + Enhanced	6.9	18.3	31.1	57.9	51.4	74.1
Clean_sp	6.7	17.5	30.9	57.4	50.8	73.2
Clean_sp + Enhanced	6.7	17.2	30.7	56.0	49.8	71.9

При добавлении 100-часовой выборки Enhanced не наблюдалось изменений в качестве распознавания как на чистых, так и на зашумленных и очищенных зашумленных тестах

DNSMOS. Clean vs Enhanced

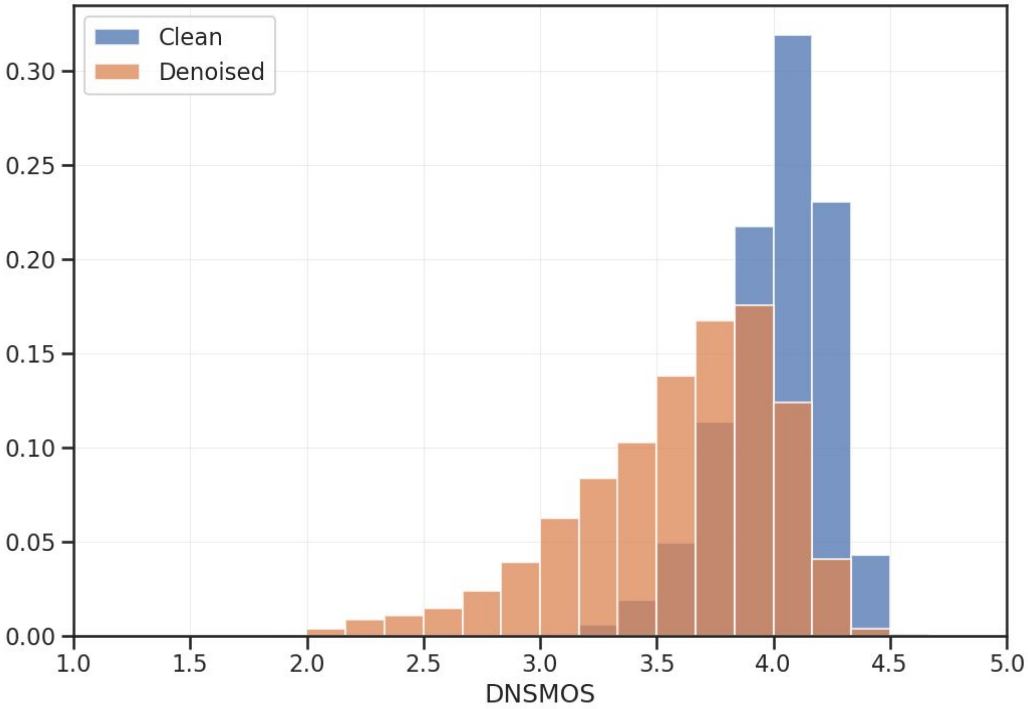


Обсуждение результатов 100 ч. Добавление Denoised

Обучающая выборка	WER на тестовых выборках, %					
	Clean	Other	Clean Denoised	Clean Noisy	Other Denoised	Other Noisy
Clean	7.0	18.4	31.4	57.8	51.5	74.2
Clean + Denoised	7.2	18.5	24.7	52.7	43.9	70.4
Clean + Denoised + Enhanced	6.8	17.9	24.0	51.4	43.2	68.5
Clean_sp + Denoised + Enhanced	6.6	17.5	23.1	49.8	42.3	68.0

Добавление Denoised привело к уменьшению ошибки на очищенных зашумленных тестах на 6–7%. Качество распознавания на чистых тестах ухудшилось. Добавление Enhanced и Denoised устранило отставание на чистых тестах, сохранив высокое качество на очищенных от шума

DNSMOS. Clean vs Denoised

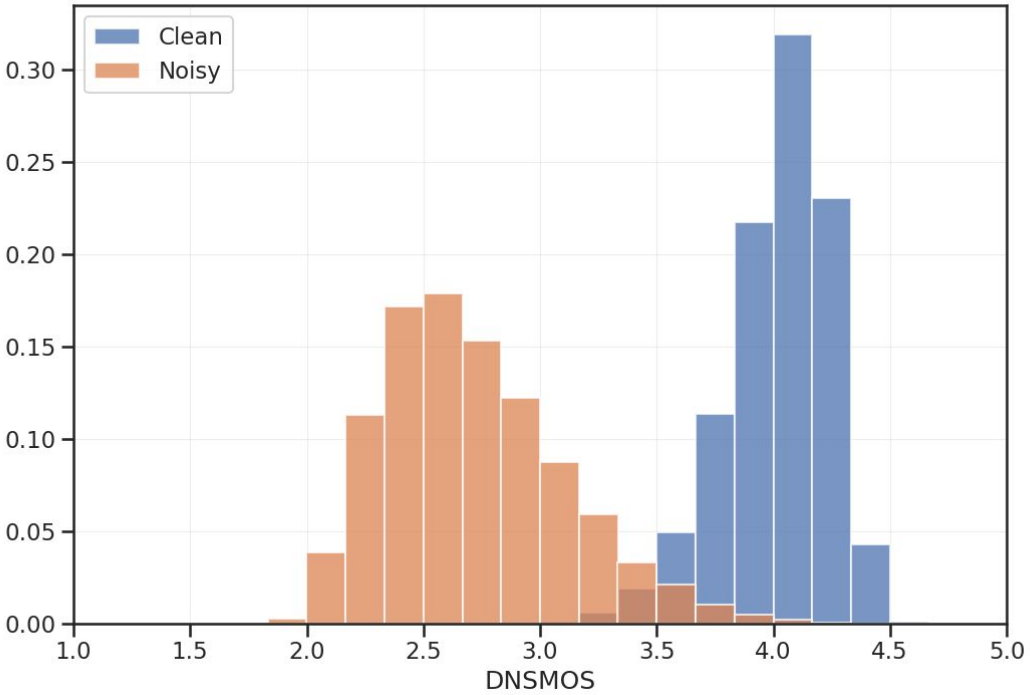


Обсуждение результатов 100 ч. Добавление Noisy

Обучающая выборка	WER на тестовых выборках, %					
	Clean	Other	Clean Denoised	Clean Noisy	Other Denoised	Other Noisy
Clean_sp + Denoised + Enhanced	6.6	17.5	23.1	49.8	42.3	68.0
Clean_sp + Noisy	7.5	17.9	26.5	26.6	45.8	46.7
Clean_sp + Noisy + Denoised + Enhanced	6.9	17.3	22.5	27.3	41.6	47.7

Добавлением Noisy привело к деградации качества на чистых тестах. на очищенных от шума тестах качество распознавания оказалось несколько ниже, чем в предыдущей серии экспериментов. При объединении Clean_sp, Noisy, Denoised и Enhanced качество на чистых тестах стало достаточно высоким. Наблюдалось значительное улучшение (на 4–4.2%) на очищенных тестах

DNSMOS. Clean vs Noisy



Сводка результатов на 100-часовой выборке

Обучающая выборка	WER на тестовых выборках, %					
	Clean	Other	Clean Denoised	Clean Noisy	Other Denoised	Other Noisy
Clean	7.0	18.4	31.4	57.8	51.5	74.2
Clean + Enhanced	6.9	18.3	31.1	57.9	51.4	74.1
Clean_sp	6.7	17.5	30.9	57.4	50.8	73.2
Clean_sp + Enhanced	6.7	17.2	30.7	56.0	49.8	71.9
Clean + Denoised	7.2	18.5	24.7	52.7	43.9	70.4
Clean + Denoised + Enhanced	6.8	17.9	24.0	51.4	43.2	68.5
Clean_sp + Denoised + Enhanced	6.6	17.5	23.1	49.8	42.3	68.0
Clean_sp + Noisy	7.5	17.9	26.5	26.6	45.8	46.7
Clean_sp + Noisy + Denoised + Enhanced	6.9	17.3	22.5	27.3	41.6	47.7

Результаты экспериментов на 960-часовой выборке

Обучающая выборка	**WER на тестовых выборках, %**					
	Clean	**Other**	**Clean Denoised**	**Clean Noisy**	**Other Denoised**	**Other Noisy**
Clean_sp	2.8	6.8	21.9	40.6	37	56.4
Clean_sp + Noisy + Denoised + Enhanced	2.9	6.9	12.8	14.8	25.9	28.9

Выводы

- В данной работе предложен новый подход к аугментации речевых образцов для обучения нейросетевой модели распознавания речи. Его суть заключается в добавлении к обучающей выборке сигналов, обработанных моделью улучшения качества речи.
- Показано, что добавление очищенной от искусственного модельного шума выборки позволило уменьшить ошибку WER в сценарии совместного использования улучшения речи и ASR
- Работоспособность предложенного подхода продемонстрирована как для сценария использования выборки ограниченного размера в случае low-resource языков, так и на большом объеме обучающих данных

Предложенный подход к аугментации может найти применение при обучении систем распознавания речи, предназначенных для работы в сложных условиях шумного окружения совместно с современными методами улучшения качества речи