Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ, Информатика и системы управления

КАФЕДРА ИУ7, Программное обеспечение ЭВМ и информационные технологии

Научно-исследовательская работа

НА ТЕМУ:

Исследование звука: Удаление шумов.

Студент ИУ7-54Б _____ **А.А. Андреев**
(Группа) (Подпись, дата) (И.О.Фамилия)

2021 г.

Оглавление

Введение.

Обработка звука - это процесс исследования динамической/статической звуковой дорожки при помощи применения определенного набора линейных и нелинейных алгоритмов с целью получения необходимой информации.

Алгоритмы динамической обработки звука работают с потоковым аудио, когда статически обрабатывают уже готовую звуковую дорожку.

Данный процесс происходит с использованием компьютерных программ и зачастую сопровождается трудными техническими вычислениями, которые ложатся на вычислительные мощности компьютера или на отдельные его комплектующие части.

Процесс исследования и обработки звука так или иначе присутствует в разных сферах профессиональной деятельности, будь то голосовые помощники, встроенные в мобильные устройства или любые другие устройства, индустрия профессионального бизнес-сообщества для фиксирования необходимой информации или же специальные службы, использующие самые современные технологии для расследования преступлений.

Если мы говорим о задаче обработки звука, то чаще всего имеем в виду применение к звуковой дорожке определенного набора стандартных и собственных алгоритмов, которые позволяют получить определенный срез информации о дорожке или же получить новую трансформированную аудио дорожку.

Цель данной работы – исследовать алгоритмы удаления посторонних шумов из аудио дорожки.

Такое программное обеспечение будет полезно для автоматических субтитров во время онлайн-конференций, логирования бизнес-встреч, работы с глухонемыми и слабослышащими.

Чтобы достигнуть поставленной цели, требуется решить следующие задачи:

1) проанализировать шумоподавление в ежедневной жизни;
2) разобраться в классификации шумов;
3) исследовать различные методы шумоподавления;
4) сформулировать преимущества и недостатки каждого из методов;
5) сделать вывод о проделанной научно-исследовательской работе.

1. Шумоподавление в ежедневной жизни

В данной части производится анализ функционального применения технологии шумоподавления в современной жизни.

1.1 Применение технологии на сформированной аудио-дорожке

Методы шумоподавления используются при очистке аудио от лишних звуковых событий для последующего повторного воспроизведения. При монтаже фильмов, музыки, подкастов и прочих медиа зачастую требуется избавляться от лишних звуковых событий.

При таких задачах может также потребоваться общее улучшение качества записи. Это включает в себя не только удаление шума, но и модифицирование сигнала, которое может улучшить восприятие записанной речи. Подобные инструменты обычно доступны в редакторах аудио и программах-микшерах для создания треков.

Например, в одном из самых известных аудиоредакторов Audacity используется подход, который называется “шумовые ворота” (noise gate), вернее, их конкретная спектральная разновидность, используемая после быстрого преобразования Фурье (FFT). Помимо этого, в Audacity есть оконные механизмы по сглаживанию сигнала и удалению его небольших артефактов. Инструменты в Audacity по шумоподавлению особенно хорошо подходят для восстановления микрокассетных записей.

1.2 Применение технологии на потоковой аудио-дорожке

Популярной и сложной задачей является шумоподавление на лету – шумоподавление и воспроизведение одновременно с записью речи. Преследуемая цель – это маскировка звуков, которые не имеют отношения к произносимой человеком информации и мешают ее восприятию. Чаще всего такое шумоподавление используется для аудиоконференций в Skype, Zoom, Discord и пр. Шумоподавление на лету как правило использует те же принципы “шумовых ворот”, но помимо этого применяются методы машинного обучения для очистки сигнала на лету. Например, компания Microsoft по результатам соревнования DNS-Challenge адаптировала наилучшие решения под свои разработки Skype и Teams. Эти решения основаны на рекуррентных нейронных сетях с LSTM блоками и на сверточных нейронных сетях. В результате новейшие версии Skype и Teams способны в режиме реального времени транслировать чистый голос при наличии агрессивных шумов: дрели, вентилятора или ветра.

1.3 Применение технологии на для использования распознавания речи

Третья интересная область использования методов шумоподавления – предобработка и чистка звукового сигнала перед применением методов автоматического распознавания речи, чтобы результат генерировался правильно. В этой области много подводных камней, так как сигнал не должен содержать искусственных артефактов речи, иначе такая «чистка» может ухудшить результат. Например, в этой работе уточняется, что системы шумоподавления на основе масок не способны улучшить результат распознавания речи и только ухудшают метрики из-за неестественности спектральных характеристик итогового сигнала. С другой стороны, алгоритмы по улучшению сигнала на основе глубоких нейронных сетей показали неплохой результат при препроцессинге в пайплайне распознавания речи.

Вывод

В данном разделе были описаны основные направления применения технологии шумоподавления в ежедневной жизни: шумоподавление аудио-дорожки, потоковое шумоподавление и шумоподавление для использования алгоритмов распознавания речи.

2. Классификация шумов

В данной части производится анализ классификации шумов для избавления от них в записи человеческой речи.

2.1 Множество классификаций шумов

Существует множество различных классификаций шумов [1], среди них выделяют, по характеру спектра и частоте волн.

Каждая категория шума отличается своими особенностями, поэтому для лучшего шумоподавления стоит учитывать категоризацию шумов по временным характеристикам. Именно временные характеристики шума тесно связаны со способом образования шума: стационарный или колеблющийся шум, как правило, образованы какими-то постоянными процессами (естественными или искусственными), тогда как прерывистый и импульсный – резкими одноразовыми процессами. Прерывистый шум для простоты можно воспринимать как повторяющийся с некоторой периодичностью импульсный шум.

2.2 Категоризация шумов

В научном сообществе принято разделять шум на три большие категории [2]: Стационарный, Импульсный и Нестационарный, где в группу первого относитя белый шум, ко второму чих, хлопок и скрип, а третью группу делят на две категории: Прерывистый (Гедки телефона, Сигнализация, Стук молотка) и Колеблющийся (Шум вентилятора, Шум ветра, Двигатель автомобиля).

Также для определения сложности проводимых работ по удалению шума, принято разделять его на три категории [3]: Стационарный, Колеблющийся, Прерывистый и Импульсный. Первую категорию можно решить вычислительными методами, когда остальные три только при помощи методов машинного обучения. Если вычислительные методы решают задачи избавления сигнала от определенного шума, то нейросетевые методы обучаются решать задачу выделения только релевантной речевой информации из всего аудиопотока.

Вывод

В данном разделе были описаны основные классификации задачи шумоподавления, основанные на сложности производимых вычислений, которые, в свою очередь, базируются на области происхождения.

3. Методы шумоподавления

В данной части производится исследование различных методов технологии шумоподавления.

2.1 Традиционные методы

Самые простые традиционные методы шумоподавления используются в условиях, когда мы программно не знаем, какой характер шума и речи. Такое отсутствие информации также наблюдается, когда мы хотим избавляться от шума на лету. При таком шумоподавлении используются обычные или спектральные пороги – заглушаются любые отзвуки, если они не превышают определенного порога по громкости.

В основе других традиционных методов лежит моделирование распределения чистой речи или шума. Делается это с помощью нахождения спектральной плотности мощности (громкости) сигнала. Плотность мощности сигнала – вариант описания распределения значений сигнала в разные моменты времени. Спектральная плотность мощности сигнала, в свою очередь, – функция, которая описывает распределение мощности сигнала в зависимости от частоты, а именно – возможную мощность в различные единицы частоты. В таком случае, имея спектральную плотность мощности шума, можно использовать метод спектрального вычитания (spectral subtraction).

Спектральные мощности различных источников звука отличаются и порой значительно. (см. Рисунок 2.1)

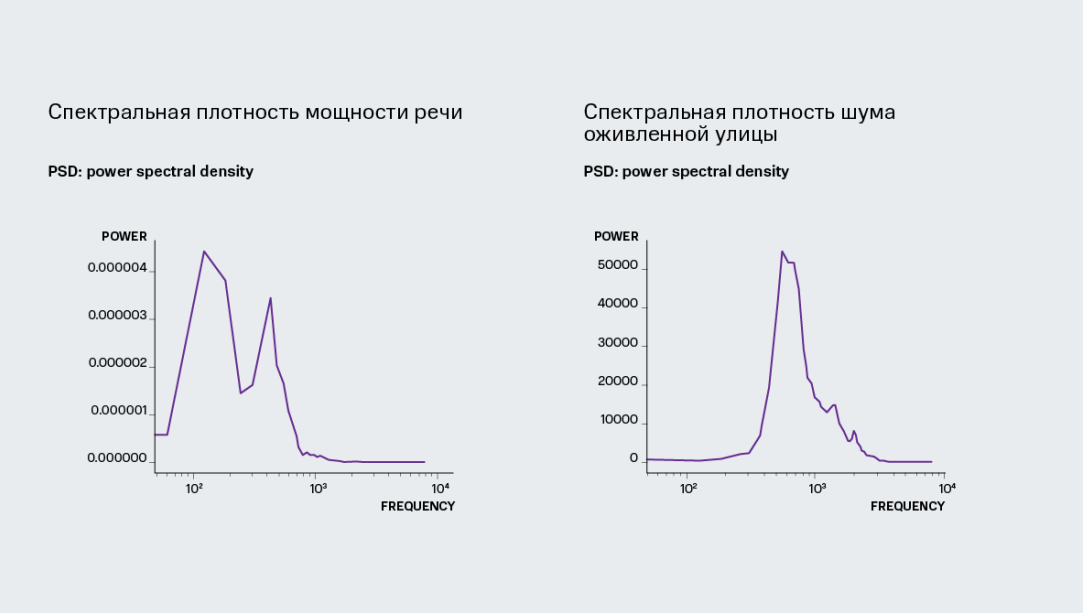

Рисунок 2.1: Пример сравнения спектральных мощностей.

Винеровское оценивание (Wiener filter) используется в качестве одного из традиционных обучаемых способов шумоподавления, отчасти похожий на метод спектрального вычитания. Этот подход основан на оптимальном подборе такого фильтра, который бы минимизировал разницу между чистым сигналом и улучшенным сигналом. Подобно некоторым алгоритмам машинного обучения, при вычислении винеровского фильтра минимизируется метрика Mean Square Error (MSE).

$$H(w) = \frac{P_{SS}(w)}{P_{yy}(w)} = \frac{P_{yy}(w) - P_{dd}(w)}{P_{yy}(w)} \quad (1)$$

где $P_{SS}(w)$ - спектр чистого сигнала, $P_{yy}(w)$ - спектр зашумленного сигнала, $P_{dd}(w)$ - спектр шумного сигнала.

Таким образом, оптимальный винеровский фильтр можно найти в случаях, когда нам известна «чистая версия» зашумленного сигнала, либо если нам известен конкретный шум, который встречается в аудиозаписях и который мы хотим убрать.

Зачастую после операций по фильтрации шума применяется сглаживание, чтобы избавиться от артефактов сигнала – «музыкального» шума – после чистки. Для сглаживания применяются различные фильтры, например, Гауссовый фильтр (или размытие по гауссу [5]).

2.2 Нейросетевые методы

Любая речь аудиодорожка - это совокупность двух независимых процесса, возникающих одновременно во времени, как отдельные инструменты в музыкальной композиции: шум и чистая речь.

В зависимости от способа решения задачи шумоподавления, разграничения спикеров или улучшения сигнала алгоритмы машинного обучения разделяют на две категории: На основе масок и Генеративные.

Нейросетевые методы, основанные на масках предсказывают маски для каждого спикера / инструмента или чистого сигнала. Эти маски накладываются на оригинальный текст.

Генеративные методы предсказывают новый сигнал для каждого спикера / инструмента или чистый сигнал.

Однако подходы, которые основаны на маскировании спектрограмм, имеют некоторые недостатки. Например, фаза волны в чистом сигнале может отличаться от фазы волны в зашумленном сигнале. Поэтому даже при вычислении идеальной маски для спектрограммы, восстановленная из

грязного сигнала фаза может вносить какие-то элементы шума и портить итоговое качество шумоподавления.

Еще одним недостатком такой системы является сложность вычисления частотных характеристик сигнала с помощью быстрого преобразования Фурье. Окно для такого преобразования должно быть достаточно большим для лучшего качества декомпозиции на частоты, что увеличивает количество вычислений. Большое количество вычислений приводит к низкой скорости работы алгоритма и его становится сложно применять в реальном времени.

2.2.1 Conv-TasNet

Данная технология базируется на методе масок, его основа - сверточные нейронные сети Conv-TasNet [8].

Многие современные подходы шумоподавления часто сравниваются с его архитектурой, как с одной из наиболее робастных реализаций. Он основан на наложении 1D свёрток на чистый сигнал без разложения на частоты.

Предшественник этой архитектуры – TasNet. Архитектура TasNet [6] состоит из сверточных энкодера и декодера с некоторыми особенностями:

- выход энкодера ограничен значениями от нуля до бесконечности $[0, \infty)$;
- линейный декодер конвертирует выход энкодера в акустическую волну;
- подобно многим методам-предшественникам на основе спектрограмм, на последнем этапе система аппроксимирует взвешивающую функцию (в данном случае LSTM) для каждого момента времени.

Conv-TasNet [7] – модификация алгоритма TasNet, которая использует в качестве взвешивающей функции сверточные слои с расширением (dilation). Это модификация была сделана после того, как свертки с расширением показали себя эффективным алгоритмом при одновременном анализе и генерации данных переменной длины, в частности, для синтеза в таких решениях, как WaveNet.

Подход для разделения аудио/шумоподавления Conv-TasNet состоит из 3-х компонентов (см. Рисунок 2.2.1.1): энкодер, разделение, декодер.

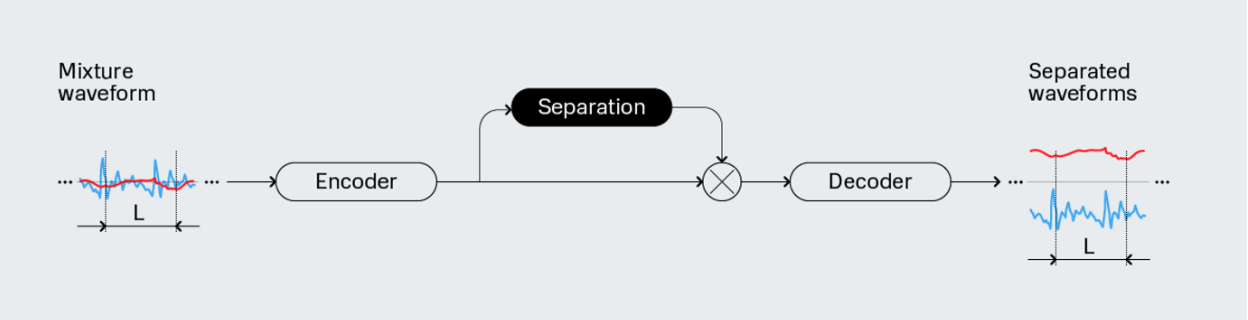

Рисунок 2.2.1.1: Подход к разделению аудио.

Основной компонент в схеме на Рисунке 2.2.1.1 – этап разделения. Этот этап решает проблему приближенного исчисления источников, смесь которых мы рассматриваем в качестве «грязных» примеров. Формально предположение о «смешанности» нашего сигнала можно выразить следующим образом:

$$x(t) = \sum_{i=t}^{C} S_i(t) \tag{2}$$

Где $x(t)$ - смесь в определенный момент времени, C - количество источников, несущих вклад в смесь, $S_1(t) \ldots S_c(t)$ - источники в определенный момент времени.

Задача алгоритма машинного обучения – определить источники s1(t), … , sc(t), зная заранее количество источников C и смесь x(t).

Разделение в алгоритме происходит не сразу, а только после извлечения признаков из сигнала с помощью «1D блоков» (1-D Conv см. Рисунок 2.2.1.2), он имеет особую структуру (см. Рисунок 2.2.1.3).

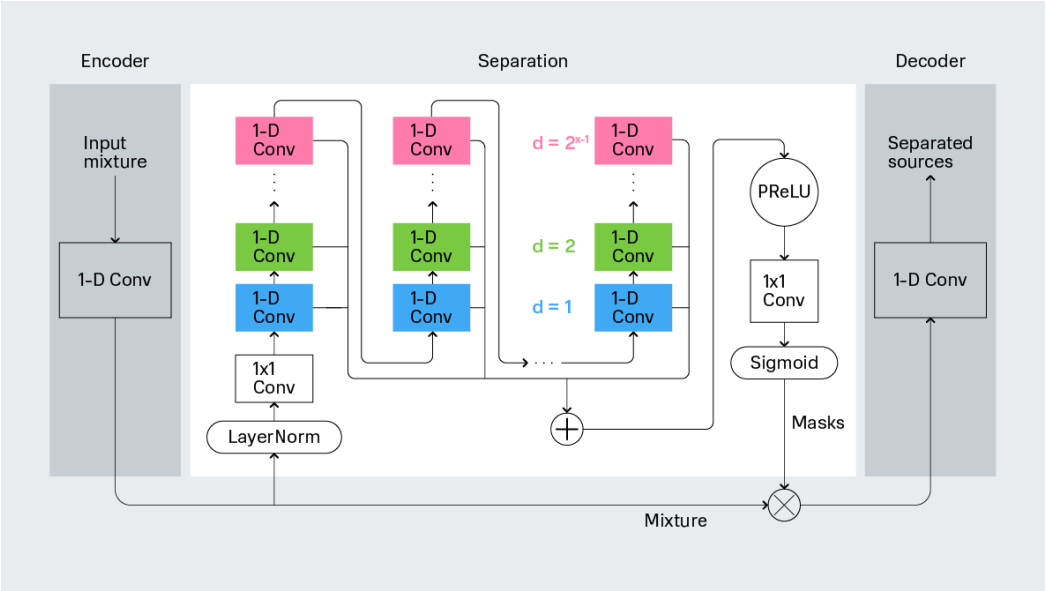

Рисунок 2.2.1.2: Преобразование сигнала смеси в набор отдельных источников.

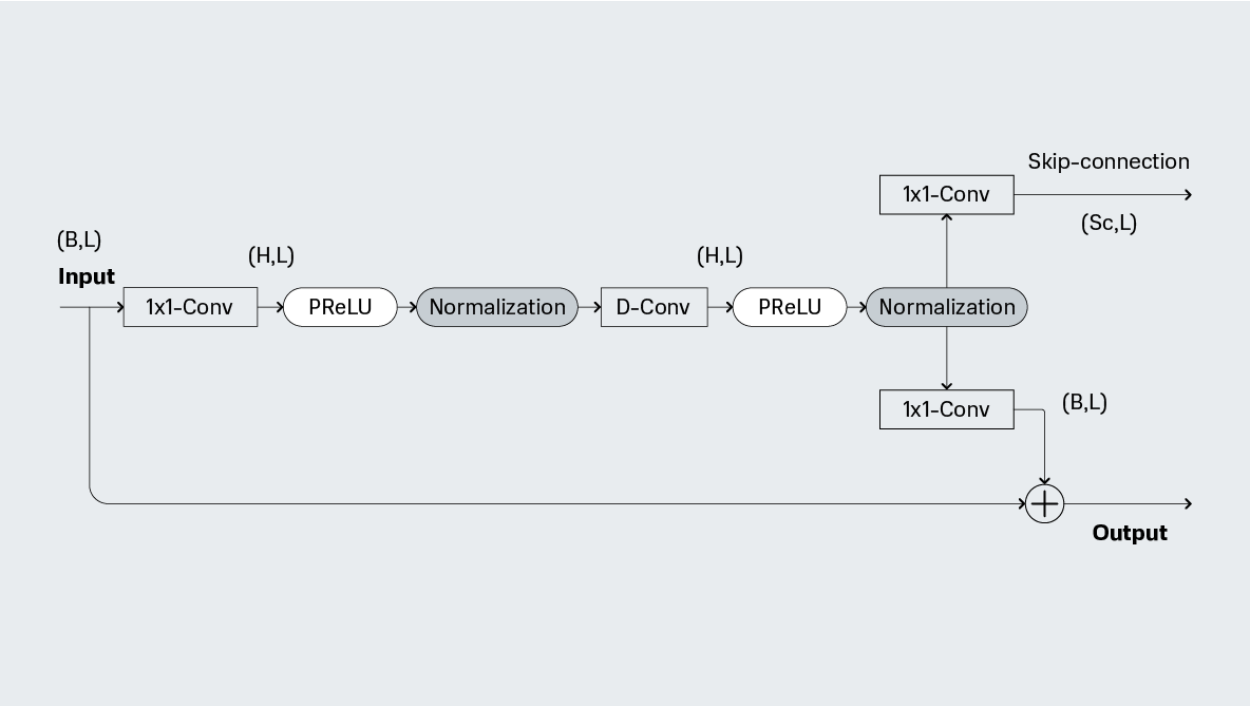

Рисунок 2.2.1.3: Структура 1-D Conv.

2.2.2 DEMUCS

Данный метод основывается на генерации. Алгоритм DEMUCS [9] или глубокое извлечение музыкальных источников (Deep Extractor for Music Sources) также используется для задач разделения источников в сигнале и шумоподавления. В отличие от предшественника Conv-TasNet, этот алгоритм напрямую генерирует источники из исходного сигнала, минуя промежуточное предсказание масок.

Создатели этого алгоритма вдохновились существующей архитектурой для сегментации изображений U-Net. U-Net архитектура представляет собой кодировщик и декодировщик, между которыми находится бутылочное горлышко. В отличии от обычного автокодировщика, слои между собой связаны «соединениями быстрого доступа», в результате итоговый сигнал не ухудшается после сжатия (см. Рисунок 2.2.2.1).

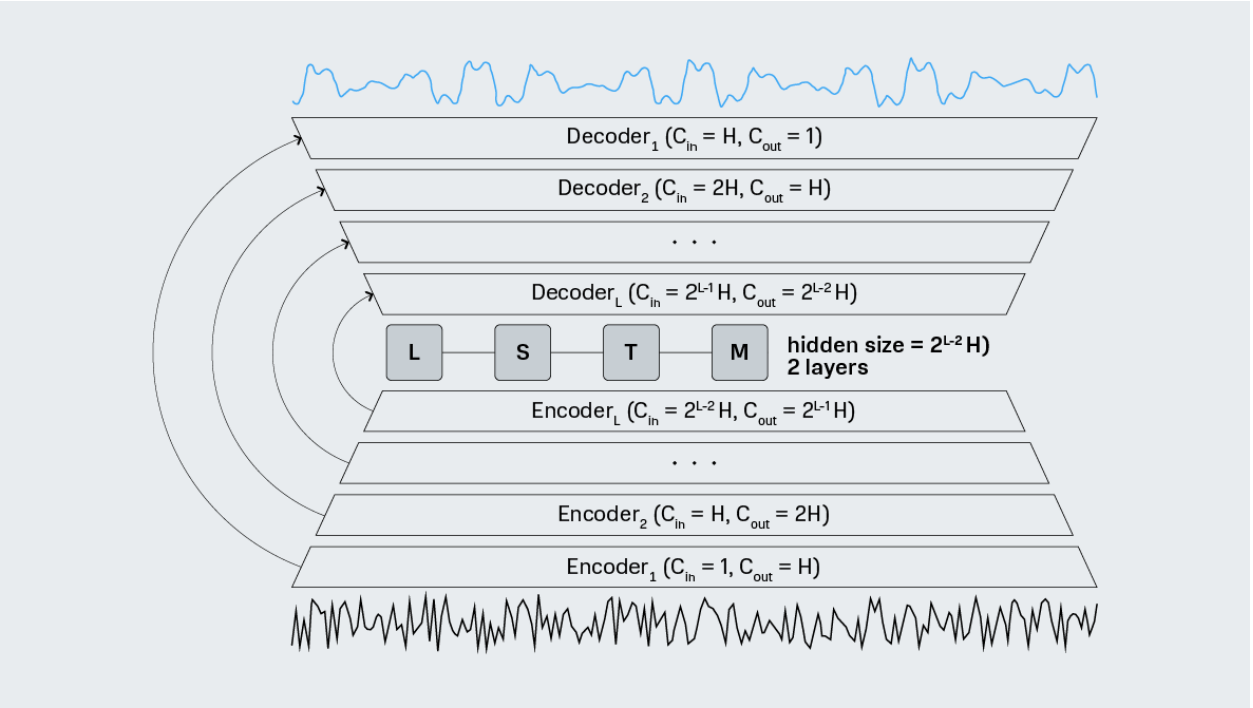

Рисунок 2.2.2.1: Структура U-Net для шумоподавления.

В качестве бутылочного горлышка в DEMUCS – однонаправленный LSTM слой. Это позволяет эффективно использовать алгоритм для анализа потоковых данных. Кодировщик и декодировщик (см. Рисунок 2.2.2.2) сформированы из блоков, которые составлены из сверточных слоев (1D, 1x1 и 1D Transpose) и функций активации (Gated Linear Unit и Rectified Linear Unit).

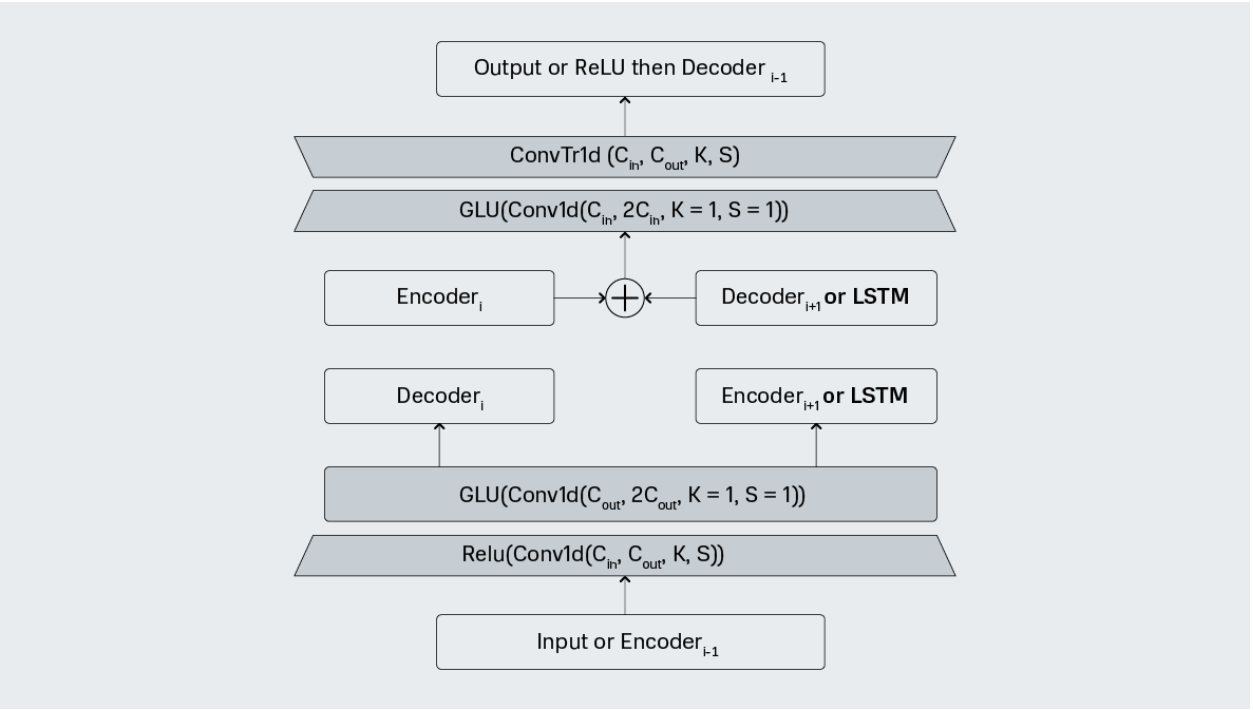

Рисунок 2.2.2.2: Компенсированность 1D, 1x1, 1D-T.

В качестве функции потерь при шумоподавлении достаточно использовать L1 Loss между предсказанной записью и эталонной, но для улучшения сходимости авторы статьи используют также STFT Loss разного масштаба (STFT с разными параметрами при подсчете функций потерь), который является суммой двух функций потерь – сходимости (spectral convergence) и амплитуд (magnitude):

$$L_{STFT}(y, \hat{y}) = L_{sc}(y, \hat{y}) + L_{mag}(y, \hat{y}) \tag{1}$$

$$L_{sc}(y, \hat{y}) = \frac{\| |STFT(y)| - |STFT(\hat{y})| \|_F}{\|STFT(y)\|_F} \tag{2}$$

$$L_{mag}(y, \hat{y}) = \frac{1}{T} \| log|STFT(y)| - log|STFT(\hat{y})| \|_1 \tag{3}$$

где y и $\hat{y}$ – эталонный сигнал и предсказанный сигнал соответственно, T – длина сигнала, $\|\cdot\|_F$ – норма Фробениуса, а $\|\cdot\|_1$ – L1 «норма» (абсолютная ошибка).

2.2.3 HiFi-GAN

Данный метод [10] основывается на генерации. В отличие от предшественников, генеративно-состязательная сеть высокой точности (High Fidelity Generative Adversarial Network) хорошо справляется с генерацией аудио подобно студийной записи без артефактов искусственной генерации (см. Рисунок 2.2.3.1).

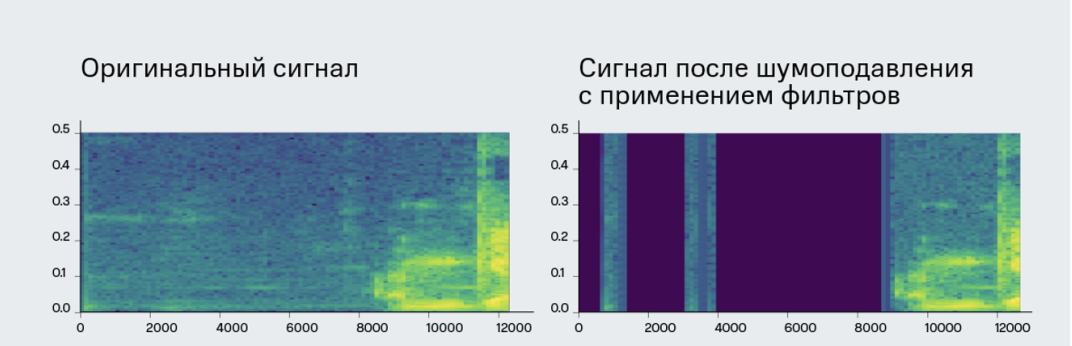

Рисунок 2.2.3.1: Результат шумоподавления HiFi-GAN.

Алгоритма шумоподавления HiFi-GAN состоит из трех основных частей (см. Рисунок 2.2.3.2): Wavenet, Postnet и GAN. За генерацию чистого сигнала на основе зашумленного отвечает блок WaveNet, этот алгоритм изначально успешно использовался для синтеза речи (текст → аудио). При модификации задачи для анализа аудио эта архитектура также показала себя эффективной. Особенность WaveNet-а для шумоподавления в том, что генерация нового сигнала происходит для всей записи целиком, а не для каждого момента времени t_n, как это делается в исходном алгоритме WaveNet. Это позволяет улучшать скорость генерации за счёт параллелизации процессов, которые могут выполняться одновременно.

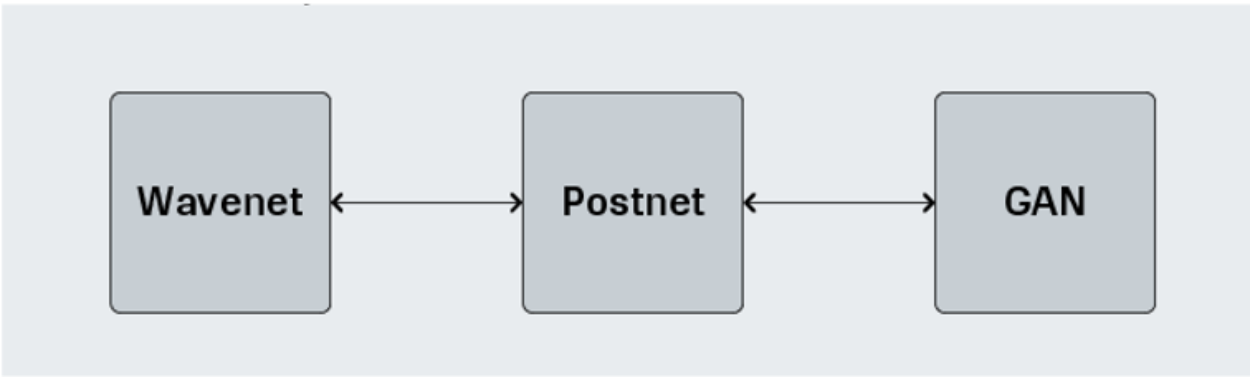

Рисунок 2.2.3.2: Составляющие алгоритма.

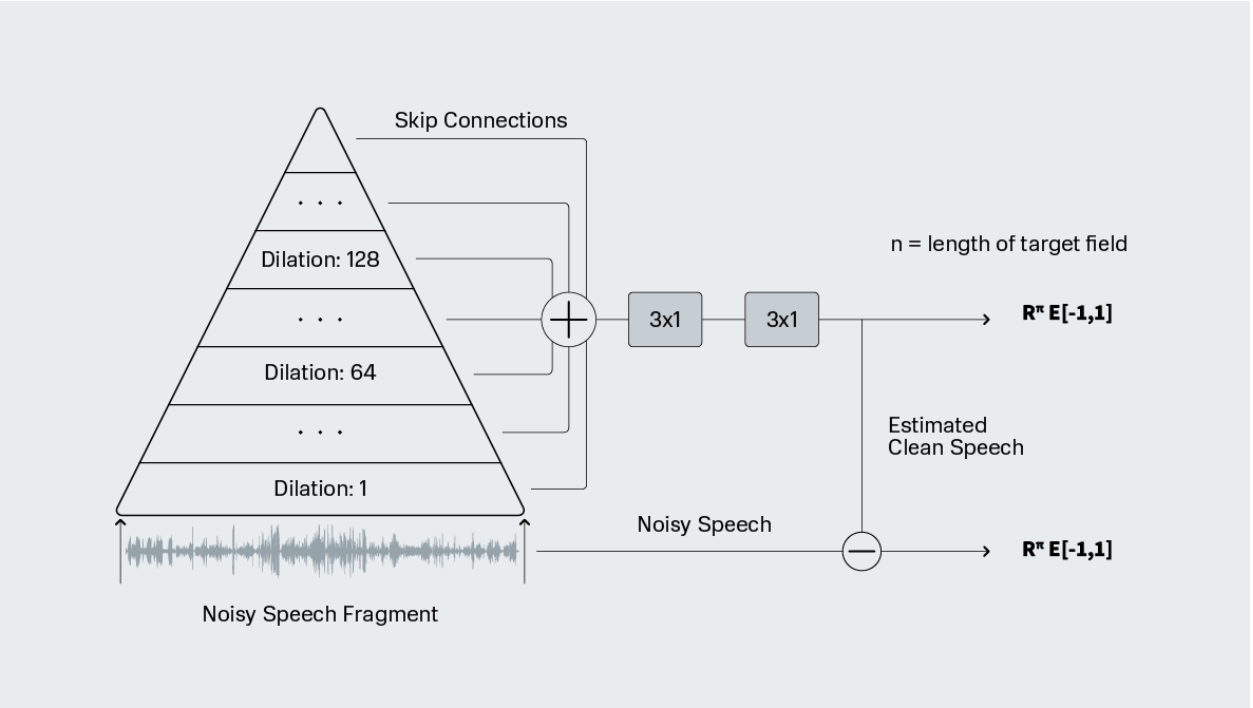

Рисунок 2.2.3.3: Результат генерации.

После генерации (см. Рисунок 2.2.3.3) WaveNet-ом сигнал проходит через несколько сверточных слоев, этот этап называется Postnet. Postnet нужен, чтобы исправлять и уточнять грубое и приближенное предсказание WaveNet-а. Кроме Postnet-а регулирующее действие дополнительно оказывают четыре разных дискриминатора, которые обучены отделять чистые оригинальные записи от сгенерированных. Каждый дискриминатор принимает выход Postnet-а в разном формате (см. Рисунок 2.2.3.4):

- Сигнал в исходном виде с разной частотой дискретизации: 16 кГц, 8 кГц, 4 кГц

- Mel-спектрограмму сигнала.

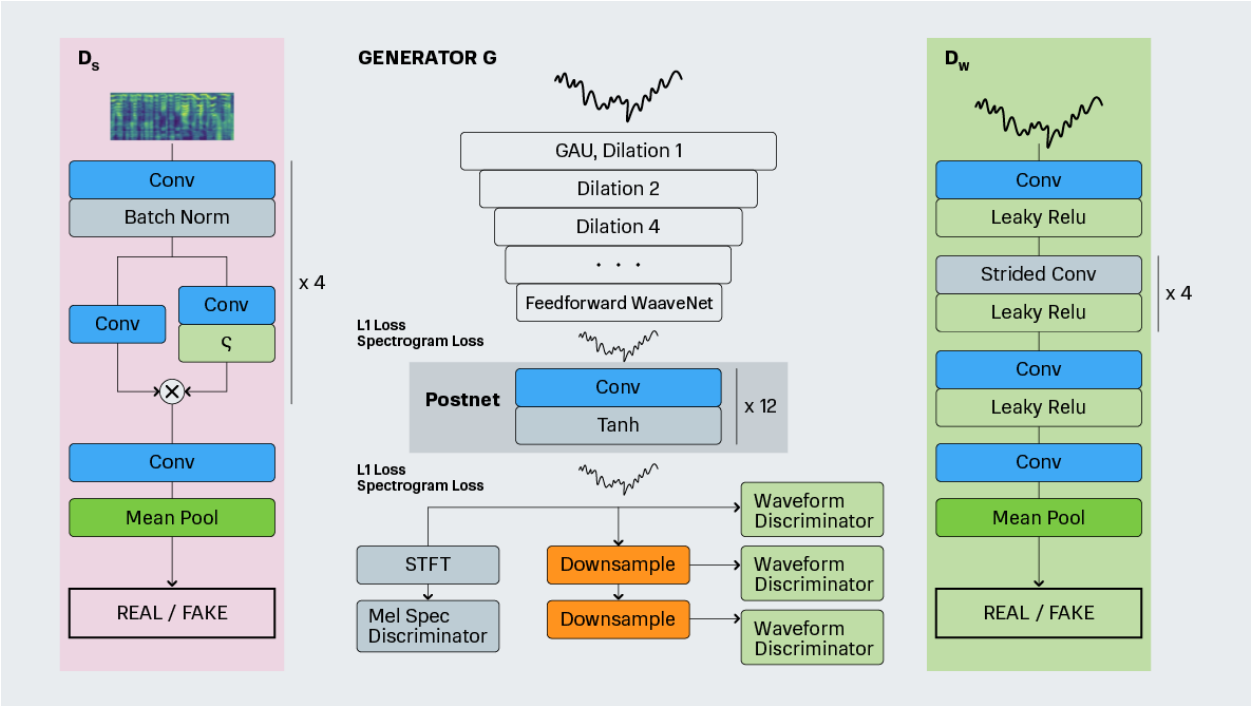

Рисунок 2.2.3.4: Общая архитектура после генерации

В итоге для обучения используются следующие функции потерь (ФП):

1. L1 (абсолютная ошибка на сигнале);
2. ФП на лог-спектрограммах предсказаний и чистого сигнала после преобразования Фурье со следующими параметрами:
   - размер окна 2048 и шаг 512,
   - размер окна 512 и шаг 128
3. Состязательная ФП (adversarial loss) для обучения Postnet-а;
4. ФП глубинных признаков (deep feature loss) для обучения дискриминаторов.

.

Заключение.

В данной научно-исследовательской работе были исследованы алгоритмы удаления посторонних шумов из аудио дорожки:

1) проанализировано шумоподавление в ежедневной жизни;
2) разобраны классификации шумов;
3) исследованы различные методы шумоподавления;
4) сформулированы преимущества и недостатки каждого из методов.

Алгоритмы шумоподавления активно развиваются из-за их необходимости в ежедневном использовании человеком. И поэтому скорее всего, в скором времени мы увидим совершенно новые технологии, основанные на иных процессах и структурах, которые позволять совершать данную работу, затрачивая меньший объем ресурсов.

Список использованной литературы

1. Noise reduction [Электронный ресурс] - Режим доступа: https://wiki.audacityteam.org/wiki/How_Audacity_Noise_Reduction_Works#algorithm
2. Chandan K. A. Reddy , Vishak Gopal , Ross Cutler , Ebrahim Beyrami1, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, Johannes Gehrke, The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results [Электронный ресурс] - Режим доступа: https://arxiv.org/pdf/2005.13981.pdf
3. FullSubNet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement [Электронный ресурс] - Режим доступа: https://arxiv.org/abs/2010.15508
4. Yanxin Hu, Yun Liu2, Shubo Lv1, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu1, Bihong Zhang, Lei Xie, DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement, сер. 10, 2009, вып. 3 [Электронный ресурс] - Режим доступа: https://arxiv.org/pdf/2008.00264v4.pdf
5. С. А. Соловьев, Решение разреженных систем линейных уравнений методом Гаусса с использованием техники аппроксимации матрицами малого ранга. – *Выч. мет. программирование*, **15**:3 (2014), 441–460
6. Yi Luo, Nima Mesgarani Department of Electrical Engineering, Columbia University, New York, NY, Dual-Signal Transformation TASNET: TIME-DOMAIN AUDIO SEPARATION NETWORK FOR REAL-TIME, SINGLE-CHANNEL SPEECH SEPARATION [Электронный ресурс] - Режим доступа: https://arxiv.org/pdf/1711.00541.pdf
7. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

[Электронный ресурс] - Режим доступа: https://arxiv.org/pdf/1609.03499.pdf

8. Yi Luo, Nima Mesgarani, Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation [Электронный ресурс] - Режим доступа: https://arxiv.org/pdf/1809.07454.pdf
9. Alexandre Defossez, Gabriel Synnaeve, Yossi Adi, Facebook AI Research INRIA PSL Research University, Real Time Speech Enhancement in the Waveform Domain [Электронный ресурс] - Режим доступа: https://arxiv.org/pdf/2006.12847.pdf
10. Jiaqi Su, Zeyu Jin, Adam Finkelstein, Princeton University Adobe Research, HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks [Электронный ресурс] - Режим доступа: https://arxiv.org/pdf/2006.05694.pdf