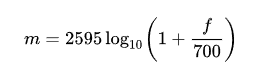
**Мел-спектрограмма**

Мел-спектрограмма — это спектрограмма, где частота выражена в мелах. В стандартной спектрограмме частотная ось линейна и измеряется в герцах (Гц). Однако слуховая система человека более чувствительна к изменениям на низких частотах, чем на высоких, и эта чувствительность уменьшается логарифмически с увеличением частоты. Шкала Мэл - это перцептивная шкала, которая аппроксимирует нелинейную частотную характеристику человеческого уха.

Первое упоминание о Мел-шкале датируется 1 января 1937 года в публикации [A Scale for the Measurement of the Psychological Magnitude Pitch](https://github.com/embatbr/graduation-project/blob/master/docs/references/A%20Scale%20for%20the%20Measurement%20of%20the%20Psychological%20Magnitude%20Pitch.pdf).

В этой статье авторы описывают, как они создали субъективную шкалу измерения высоты тона на основе полузначений частот при различных частотах. Она отличается от музыкальной и частотной шкалы, не являющихся субъективными. Пять наблюдателей анализировали звуки 10 различных частот при уровне громкости 60 дБ. На основе их результатов была построена числовая шкала, пропорциональная воспринимаемой субъективной высоте тона. В этой шкале тон с частотой 1000 гц был назначен высотой 1000 единиц (мелов). Близкое соответствие шкалы высоты тона интегралу дифференциальных порогов показывает, что, в отличие от порогов громкости, все пороги высоты тона имеют одинаковую субъективную величину. Это также подразумевает, что высота тона и дифференциальная чувствительность к нему являются линейными функциями распределения на базилярной мембране. Соответствие шкалы высоты тона и экспериментально определенных областей резонанса базилярной мембраны подтверждает, что, уменьшая высоту тона вдвое, наблюдатель настраивает его до стимуляции позиции, находящейся на полпути от исходного местоположения к концу мембраны. Измерение субъективного размера музыкальных интервалов (например, октав) в терминах шкалы высоты тона показывает, что интервалы увеличиваются с ростом частоты средней точки интервала (за исключением двух самых высоких октав). Этот результат подтверждает ранее сделанные суждения о относительном размере октав в различных частотных диапазонах.

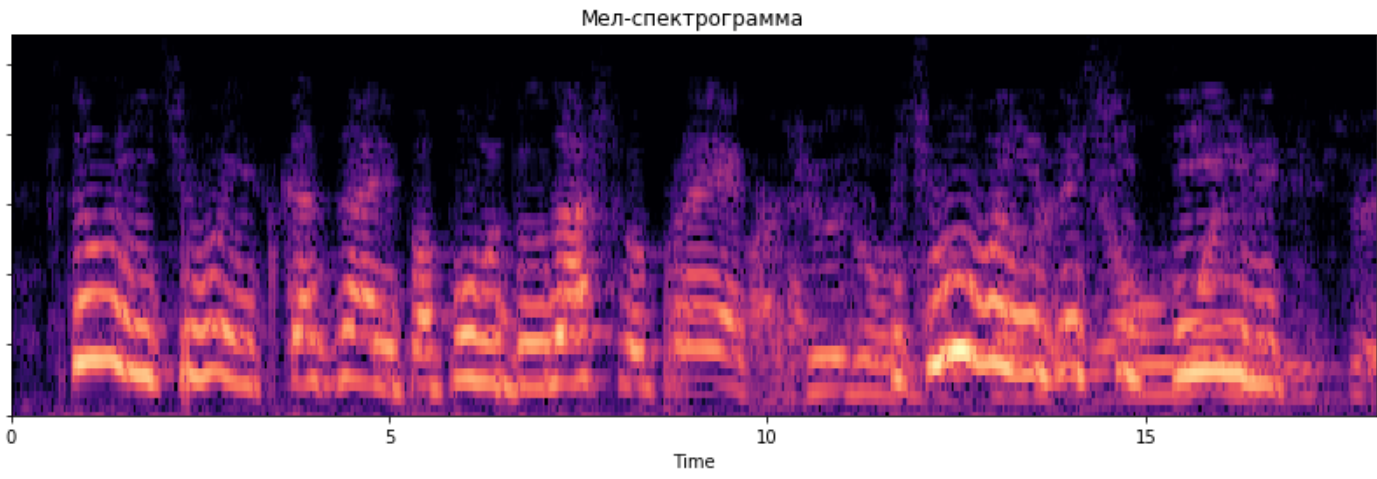
Для конвертации из *f* герца в *m* мел О'Шонесси в 1987 году предложил свою формулу:



Мел спектрограмму вычисляют по обычной спектрограмме, построенной с помощью оконного преобразования Фурье:



Суть этой операции в последовательном применении преобразования Фурье к коротким кусочкам речевого сигнала, домноженным на некоторую оконную функцию. Результат применения оконного преобразования — это матрица, где каждый столбец является спектром короткого участка исходного сигнала. Каждый столбец исходной спектрограммы скалярно умножается на каждый мел-фильтр (размещенный на частотной шкале), после чего получается вектор чисел, по размеру равный количеству фильтров. В результате таких преобразований значения из низких частот спектрограммы остаются практически неизменными на мел-спектре, а в высоких частотах происходит усреднение значений из более широкого диапазона. На мел-спектрограмме сохраняется больше информации, которая хорошо воспринимается и различается человеком, чем на обычной спектрограмме.



**Различные публикации связанные с мел-спектрограммами**

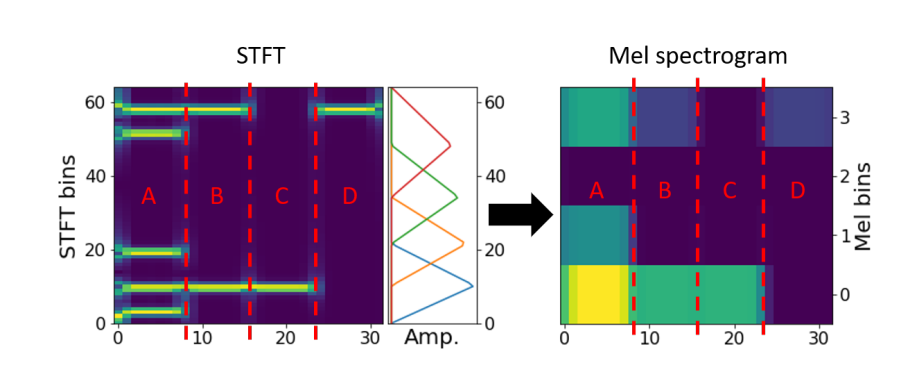
1. [The Impact of Audio Input Representations on Neural Network based Music Transcription](https://arxiv.org/pdf/2001.09989.pdf)

В данной статье подробно рассматривается влияние различных представлений входных данных на транскрипцию полифонической музыки с использованием нескольких инструментов.

Сравнивалось влияние четырех различных представлений входного сигнала:

* линейно-частотной спектрограммы (LinSpec)
* лог-частотная спектрограмма (LogSpec)
* CQT (The Constant Q Transform)
* Mel-спектрограмма (MelSpec).

Для MelSpec авторы использовали версию, реализованную в инструментарии HTK Speech Recognition toolkit. Затем для преобразования частотных бинов k из кратковременного преобразования Фурье (STFT) в мел-бины в мел-спектрограмме используются мел-фильтры.



*Мел-фильтры могут быть использованы для преобразования вывода STFT в мел-спектрограмму. Каждый мел-фильтр покрывает несколько бинов STFT и сокращает их до одного мел-бина.*

Авторы провели три исследования:

* Исследование 1: Разрешение спектрограмм.

В этом эксперименте мы изучаем влияние изменения количества бинов спектрограммы от примерно 256 до 2 048 для каждого из представлений входных данных. Пространства между бинами были тщательно выбраны так, чтобы первый бин начинался с 50 Гц, а последний бин заканчивался на 6 000 Гц для всех представлений входных данных. Для MelSpec используется фиксированный размер окна кратковременного преобразования Фурье (STFT) равный 4 096, а количество мел-фильтров варьируется от 128 до 2 048.

Модели, основанные на MelSpec, показывают лучшие результаты, когда количество мел-фильтров равно 1024. Если используется больше мел-фильтров, некоторые из них будут пустыми. Поэтому MelSpec с слишком большим количеством мел-бинов показывает немного худшие результаты по сравнению с логарифмическими и линейно-частотными спектрограммами. Этот результат также показывает, что мел-фильтры можно рассматривать как эффективный алгоритм сжатия для спектрограмм.

В целом, наилучшим представлением входных данных в этом исследовании оказался LogSpec с размером окна STFT равным 4096 и 2048 бинами.

* Исследование 2: Размер окна STFT

Во втором эксперименте мы анализируем влияние размера окна STFT вместе с изменением размера бина на точность транскрипции для наших моделей. Поскольку размер окна STFT остается постоянным в исследовании 1 (4096), временное разрешение в предыдущем эксперименте фиксировано. Однако в этом эксперименте временное разрешение меняется, чтобы мы могли изучить влияние высокого временного разрешения на точность транскрипции. Для MelSpec исследуется различное количество мел-фильтров для каждого изменяющегося значения размера окна STFT. Результаты дополнительно показывают, что мел-фильтры действуют как сжатие спектрограммы, уменьшая количество частотных бинов из результатов STFT с минимальной потерей информации. Улучшение точности транскрипции становится минимальным, когда размер окна STFT достигает 8 192; поэтому наш эксперимент был остановлен на 4 096.

Общая лучшая производительность все еще достигается с использованием LogSpec с размером окна STFT 4 096 и 2 048 бинами, плотно сопоставляемой с LinSpec.

* Исследование 3: Длина входного размера кадра

В предыдущих двух экспериментах длина входного аудиофайла фиксирована на уровне 16 384 сэмплов. Однако в этом эксперименте мы изучаем влияние изменения длины входного аудио от 0.14 до 0.45 секунды. Все остальные параметры спектрограммы, такие как размер окна STFT (4096), количество мел-фильтров (512) и количество бинов на октаву (72), остаются постоянными. Длина входного аудиофайла имеет лишь незначительное влияние на точность транскрипции музыки. Длина входного аудио влияет на ширину (временные шаги) спектрограммы и точность меток высоты звука. Поскольку метка высоты звука соответствует ноте, появляющейся в середине кадра, более длинные кадры могут запутать нейронную сеть. В целом лучший результат достигается при использовании моделей на основе LogSpec и LinSpec с n\_fft 4096 и 2048 частотными бинами.

Вывод: Когда разрешение спектрограммы достаточно высокое (например, большее количество частотных бинов), мы обнаружили, что представление логарифмической частоты обеспечивает лучший результат транскрипции, за счет вычислительной сложности. Если вычислительная сложность является проблемой, MelSpec может сжать спектрограмму в более компактное представление, сохраняя при этом относительно высокую точность транскрипции. Нейронные сети могут лучше учиться на более низком уровне входного представления, таком как логарифмическая или линейно-частотная спектрограмма. В будущем исследования также будут затрагивать влияние дополнительных входных представлений, таких как обученное представление, полученное путем end-to-end обучения транскрипции.

1. [MEL SPECTROGRAM INVERSION WITH STABLE PITCH](https://arxiv.org/pdf/2208.12782.pdf)

В этой статье авторы предлагают новую модель вокодера, специально разработанную для музыки. Ключевым моментом для улучшения стабильности тона является выбор пространства целей, инвариантного к сдвигам, которое состоит из спектра амплитуд и градиента фазы.

Логарифмическая мел-спектрограмма (просто мел-спектрограмма) - это временно-частотное представление в низком разрешении, получаемое из спектрограммы мощности M2. Сначала происходит искажение оси частот с использованием мел-шкалы, а затем значения масштабируются до логарифмического амплитудного значения. Для оценки исходного аудиосигнала x из мел-спектрограммы требуется восстановление потерянной информации, включая фазовую информацию и линейно-распределенное и более высокое разрешение по частоте спектра амплитуды.

Большинство современных подходов пытаются обучать эту обратную трансформацию end-to-end, что является особенно сложным для полифонического музыкального сигнала. Вместо точного воспроизведения удерживаемой ноты модель end-to-end должна научиться:

1. различным шаблонам для каждой комбинации сдвига фазы и периода периодической волны, и

2. как активировать их в правильной последовательности.

Вместо восстановления сигнала во временной области авторы предлагают использовать в качестве выходного пространства промежуточное временно-частотное представление, состоящее из трех каналов: спектра амплитуды и двух компонент фазового градиента. Обученная модель выходного представления не нуждается в обучении:

1. вариаций сдвига периодических волн, так как они явно моделируются обратным STFT, и

2. в последовательности фазы, которая обрабатывается с помощью алгоритма интеграции фазы. В отличие от фазового спектра, производная фазы по времени инвариантна к сдвигам, поэтому она более подходит для цели инвариантной к сдвигам архитектуры, такой как сверточная нейронная сеть.

В данной работе предложена новая модель инверсии мел-спектрограммы, специально разработанная для музыки, которая достигает улучшенной реконструкции удерживаемых нот и аккордов по сравнению с передовыми моделями по синтезу речи.

1. [Mel-FullSubNet: Mel-Spectrogram Enhancement for Improving Both Speech Quality and ASR](https://arxiv.org/abs/2402.13511)

В данной работе предлагается Mel-FullSubNet - сеть для удаления шума и устранения эха в мел-спектрограммах одноканального аудиосигнала для улучшения качества речи и производительности автоматического распознавания речи (ASR). Mel-FullSubNet принимает на вход зашумленную мел-спектрограмму и предсказывает соответствующую чистую мел-спектрограмму. Улучшенную мел-спектрограмму можно преобразовать в речевую волну с помощью нейронного вокодера или использовать напрямую для ASR.

По сравнению с линейной частотной спектрограммой или временной формой сигнала, мел-частоты представляют речь более компактным образом (но все еще воспринимаемо эффективно) и имеют более низкую размерность признаков (количество частот) с точки зрения машинного обучения, что приводит к меньшей ошибке предсказания улучшения мел-спектрограммы. Это выгодно как для качества речи, так и для ASR:

(i) Нейронные вокодеры были широко изучены в области текст в речь и способны эффективно преобразовывать мел-спектрограмму обратно в временную форму сигнала. Следовательно, свойство низкой ошибки улучшения мел-спектрограммы может быть сохранено нейронными вокодерами, и может быть достигнуто более высокое качество речи.

(ii) Улучшение полнодиапазонной речи требует восстановления всех деталей полнодиапазонной речи, а затем сжатия до мел-частот для ASR, что, очевидно, приведет к получению менее точной мел-спектрограммы, чем прямое улучшение мел-спектрограммы. Более точная мел-спектрограмма, безусловно, будет полезна для ASR.

Спектрограмма переорганизуется двумя способами, чтобы сформировать последовательности вдоль частоты и вдоль времени, которые будут обрабатываться соответственно следующими полнодиапазонными и поддиапазонными сетями LSTM. Последовательность вдоль частоты формируется для каждого временного кадра, и всего T последовательностей длиной Fmel. Входной вектор каждого частотного шага конкатенирует компонент сигнала одного временного кадра и его соседних Ntime−1 временных кадров, который затем преобразуется в D-мерный входной вектор с помощью линейного слоя. Последовательность вдоль времени формируется для каждой мел-частоты, и всего Fmel T-длинных последовательностей. Входной вектор каждого временного шага конкатенирует компонент сигнала одной мел-частоты и ее соседних Nfreq −1 мел-частот, который также преобразуется в D-мерный входной вектор с помощью линейного слоя. Mel-FullSubNet состоит из нескольких чередующихся полнодиапазонных и поддиапазонных LSTM сетей. Между одной полнодиапазонной сетью и ее последующей поддиапазонной сетью используется механизм ворот. Размер вектора для представления каждого T-F бина всегда остается D на протяжении всей сети. После окончательного поддиапазонного слоя один линейный слой преобразует D-мерный вектор в 1-мерный, что приводит к улучшенной логарифмической мел-спектрограмме Xˆ(t, fmel). Сеть обучается с использованием среднеквадратичной ошибки (MSE) улучшенной логарифмической мел-спектрограммы.

Объединяя мощную сеть полной и поддиапазонной фильтрации с стратегией улучшения мел-спектрограммы и нейронным вокодером, Mel-FullSubNet достигает значительного улучшения качества речи. Кроме того, улучшенную мел-спектрограмму можно непосредственно использовать для надежного автоматического распознавания речи (ASR). По сравнению с другими сетями улучшения мел-спектрограммы, предложенная сеть полной и поддиапазонной фильтрации может привести к меньшему количеству артефактов речи и, следовательно, более совместима с системами ASR. В целом, улучшение мел-спектрограммы с использованием сети полной и поддиапазонной фильтрации может открыть новые возможности для ASR-ориентированного улучшения речи с одним каналом.

1. [DIFFUSION-BASED MEL-SPECTROGRAM ENHANCEMENT FOR PERSONALIZED SPEECH SYNTHESIS WITH FOUND DATA](https://arxiv.org/abs/2305.10891)

В данной статье исследуется использование условных моделей диффузии для обобщенного улучшения речи, которое направлено на одновременное решение нескольких типов аудиоискажений. Улучшение выполняется в области логарифмической мел-спектрограммы для соответствия целям обучения TTS.

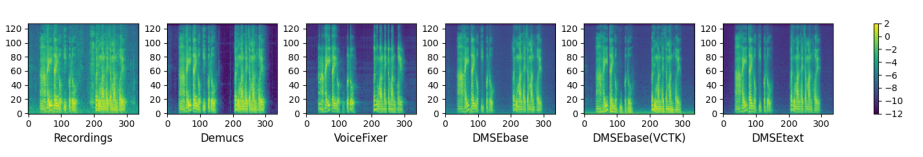
Авторы предлагают использовать условные модели диффузии для обобщенного улучшения речи и применять улучшение непосредственно к логарифмическим мел-спектрограммам для соответствия целям обучения TTS. Выбор модели диффузии обусловлен работой Palette, который использовал единственную общую модель диффузии для решения широкого круга задач трансляции изображений. Мел-спектрограммы являются временно-частотными представлениями, которые могут быть рассмотрены как изображения. Таким образом, авторы ожидаем, что модель диффузии будет эффективна в улучшении мел-спектрограмм. Для улучшения устойчивости модели к неизвестным формам аудиоискажений вводится текстовое содержание речи в качестве дополнительного условия, которое обычно доступно в контексте обучения TTS.

Улучшение мел-спектрограмм на основе диффузии:

Предположим, что у нас есть большая коллекция пар мел-спектрограмм, обозначенных как , где x(i) представляет собой мел-спектрограмму высококачественного образца речи, а y(i) представляет собой мел-спектрограмму соответствующего ухудшенного образца. y(i) может быть создан путем искусственного ухудшения аудио в x(i). Нас интересует изучение условного распределения P(x|y) на D. Если D сконструирован так, чтобы быть репрезентативным, улучшение мел-спектрограммы может быть достигнуто путем выборки из изученного условного распределения. Здесь используются модели диффузии для изучения параметрического приближения P(x|y).

Модель улучшения мел-спектрограммы опирается на синтетический набор данных парных образцов для изучения условного распределения P(x|y). Поэтому разрыв между обучающими образцами и реальными ухудшенными записями неизбежен. В результате обученная модель может переобучиться на внутренних данных и не сможет хорошо обобщаться на новые формы аудио-дефектов.

Предложенная модель улучшения называется DMSEtext, что сокращает Diffusion-based Mel-Spectrogram Enhancement с условием текста. Модель без условия текста называется DMSEbase. Они сравниваются с двумя базовыми моделями: Demucs и VoiceFixer. Demucs - это модель подавления шума. Используется она в качестве базовой модели, чтобы исследовать эффективность однозадачной модели подавления шума в предполагаемом применении. VoiceFixer - это модель регрессии, обученная для решения ряда аудио-дефектов, включая аддитивный шум, реверберацию, обрезание и низкоскоростную полосу пропускания. Мотивация выбора VoiceFixer в качестве базовой модели заключается в том, чтобы проверить, есть ли у модели улучшения речи на основе диффузии преимущества перед подходом на основе регрессии.

Сравнивая DMSEbase с DMSEbase(VCTK), мы можем заключить, что нет значительного прироста от использования набора данных на кантонском языке, за исключением доступности текстовой информации. На рисунке ниже приведен конкретный пример мел-спектрограмм, улучшенных различными системами. Мы видим, что аудио, улучшенное с помощью Demucs, сохраняет больше речевого содержания, а также помех, чем аудио, улучшенное другими системами.

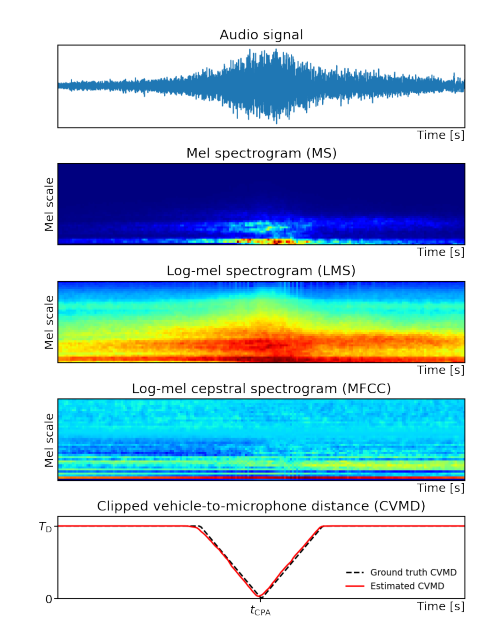
В этом проекте была представлена модель улучшения мел-спектрограмм на основе диффузии. Авторы эмпирически показали, что использование генеративного моделирования способствует выводу чистого речевого сигнала и что предпочтительна универсальная модель улучшения перед моделью однозадачного устранения шума для предварительного улучшения реальных записей для разработки модели TTS. Субъективные оценки от человеческих слушателей демонстрируют, что полученный синтетический голос производит речь высокого качества по сравнению с голосами, обученными на данных, улучшенных сильными базовыми моделями улучшения речи.

1. [Mel-spectrogram features for acoustic vehicle detection and speed estimation](https://arxiv.org/pdf/2204.04013.pdf)

В статье рассматривается обнаружение и оценка скорости движущихся транспортных средств по данным с одного датчика. Авторы предсказывают момент проезда транспортного средства, минимизируя обрезанное расстояние от транспортного средства до микрофона, которое предсказывается из мел-спектрограммы входного аудиосигнала в рамках обучения с учителем. Кроме того, особенности, основанные на мел-спектрограмме, используются непосредственно для оценки скорости транспортного средства, без введения каких-либо промежуточных характеристик.

В данной статье рассматривается задача обнаружения транспортных средств и оценки их скорости на основе звуковых измерений с использованием единственного сенсора. Мы предлагаем метод обучения с учителем на основе кратковременного спектра мощности входного аудио. Обнаружение транспортных средств осуществляется путем минимизации обрезанного расстояния от транспортного средства до микрофона (CVMD), подход, представленный в [Robust Audio-Based Vehicle Counting in Low-to-Moderate Traffic Flow](https://arxiv.org/abs/2010.11716). Минимальная точка обрезанного расстояния представляет собой ближайшую точку подхода транспортного средства к микрофону (CPA). Затем момент CPA используется для локализации части кратковременного спектра мощности аудио, которая будет представлять входные характеристики для оценки скорости. Рассматриваются следующие представления кратковременного спектра мощности: (1) мел-спектрограмма (MS), (2) логарифмическая мел-спектрограмма (LMS) и (3) логарифмическая мел-кепстральная спектрограмма (MFCC). Предложенные методы обнаружения транспортных средств и оценки скорости обучаются и тестируются на наборе данных из 304 записей транспортных средств на улицах.

Мел-основанные аудиопредставления хорошо подходят в качестве признаков в приложениях классификации аудио [Acoustic Features for Environmental Sound Analysis](https://hal.science/hal-01575619/document). Многообещающие результаты по оценке скорости, полученные в [Acoustic vehicle speed estimation from single sensor measurements](https://trepo.tuni.fi/bitstream/handle/10024/137723/paper.pdf?sequence=1), стимулировали нас к дальнейшему изучению потенциала трех вариантов мел-основанных аудиопредставлений, а именно MS, LMS и MFCC. Мы анализируем влияние, когда они используются непосредственно в качестве входных признаков, без промежуточного признака ослабления скорости, предложенного в [Acoustic vehicle speed estimation from single sensor measurements](https://trepo.tuni.fi/bitstream/handle/10024/137723/paper.pdf?sequence=1). MS представляет собой спектр мощности в краткосрочной области, который отображает частоту исходного сигнала на мел-шкалу (логарифмическое преобразование частотных диапазонов). LMS получается как логарифм модуля MS. После применения дискретного косинусного преобразования к LMS мы получаем представление MFCC.



*Первый график: Оригинальный аудиосигнал.*

*Второй график: Мел-спектрограмма аудиосигнала.*

*Третий график: Логарифмическая мел-спектрограмма аудиосигнала.*

*Четвертый график: Логарифмическая мел-кепстральная спектрограмма аудиосигнала.*

*Пятый график: Фактическое значение и предсказанное значение CVMD. Аудиосигнал укорочен до 5 секунд и центрирован по tCPA.*

Изучение влияния аудиопредставлений, основанных на мел-шкале, показывает, что MS является наиболее надежным признаком оценки скорости, с лучшим средним значением RMSE 7,87 км/ч и лучшими средними значениями точности классификации 48,7% (∆ = 0) и

91,0% (|∆| ≤ 1). Однако LMS также рассматривается как важный признак классификации, поскольку он обеспечивает средние значения точности классификации, очень близкие к тем, которые получаются на основе MS. Представленные результаты квалифицируют мел-основанные признаки как важный компонент для будущих исследований обнаружения транспортных средств и оценки скорости.

Экспериментальные результаты показывают, что тщательно выбранные мел-признаки могут быть использованы непосредственно при оценке скорости, без промежуточных и созданных вручную признаков.

1. [Mel-Weighted Single Frequency Filtering Spectrogram for Dialect Identification](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9180347)

В данном исследовании предлагаются мел-взвешенные спектрограммы с фильтрацией по одной частоте (SFF) для идентификации диалектов. Спектр, полученный с использованием SFF, обладает высоким спектральным разрешением для гармоник и резонансов, одновременно сохраняя хорошее временное разрешение некоторых характеристик речевого возбуждения, таких как импульсоподобные события. Спектр SFF может лучше представлять характеристики речи, такие как моменты всплесков и закрытия гортани, чем спектр STFT. Гипотеза авторов заключается в том, что эти сложные представления в спектре SFF должны помочь в отличии диалектов. Для этого они создали систему идентификации диалектов, которая использует безнадзорное, сжатое представление признаков мел-взвешенной SFF спектрограммы (Mel-SFF спектрограмма) с помощью глубоких автоэнкодеров последовательности.

Эта статья исследует идентификацию диалектов независимо от текста без использования каких-либо предварительно обученных моделей или транскрипций. Предложенная система использует автокодировщик, который обучается с использованием спектрограммы с мел-взвешенным одноканальным фильтрованием по частоте (Mel-SFF) для получения бутстрапных нормализованных признаков, которые используются в классификации. Базовая система с аналогичной архитектурой обучается с использованием спектрограммы с мел-взвешенным преобразованием Фурье короткого времени (Mel-STFT).

Предложенная система состоит из трех этапов: извлечение спектрограммы Mel-SFF, получение безнадзорного представления из спектрограммы с использованием автоэнкодера и классификация.

В данном исследовании было исследовано использование спектрограммы, полученной с помощью взвешенного по Мелу одночастотного фильтра, для идентификации диалектов с использованием баз данных STYRIALECT и UT-Podcast. Диалекты были идентифицированы путем обучения автоэнкодера с использованием спектрограммы Mel-SFF и передачи характеристик узкого места автоэнкодера классификатору. Предложенная спектрограмма Mel-SFF показала лучшую производительность по сравнению с базовыми системами на основе i-векторов. Кроме того, объединение безнадзорных представлений (BNFs), вычисленных из спектрограмм Mel-SFF и Mel-STFT с использованием автоэнкодеров последовательности-последовательности, дало лучший показатель UAR (46.9%) для базы данных STYRIALECT. В UT-Podcast предложенная и системы фьюжена показали относительное улучшение UAR на 4.69% и 7.71% по сравнению с базовой системой, основанной на спектрограмме Mel-STFT, соответственно. Кроме того, предложенная система показала лучшую производительность, особенно на коротких фразах, по сравнению с базовой системой в экспериментах с данными UT-Podcast. Таким образом, мы приходим к выводу, что высокое спектральное и временное разрешение спектра SFF приводит к улучшению идентификации диалектов для изучаемых немецких и английских диалектов. Кроме того, мы заключаем, что предложенная система спектрограммы Mel-SFF лучше отличает диалекты от коротких фраз, чем ее базовая система на основе STFT.

1. [Cough Recognition Based on Mel-Spectrogram and Convolutional Neural Network](https://www.frontiersin.org/articles/10.3389/frobt.2021.580080/full)

В связи с пандемией COVID-19, в 2021 году возросла необходимость различать звук кашля для оценки подозрительных пациентов в населении. В этой статье предлагается метод распознавания кашля на основе мел-спектрограммы и сверточной нейронной сети, названной сетью распознавания кашля (CRN), которая может эффективно различать звуки кашля.

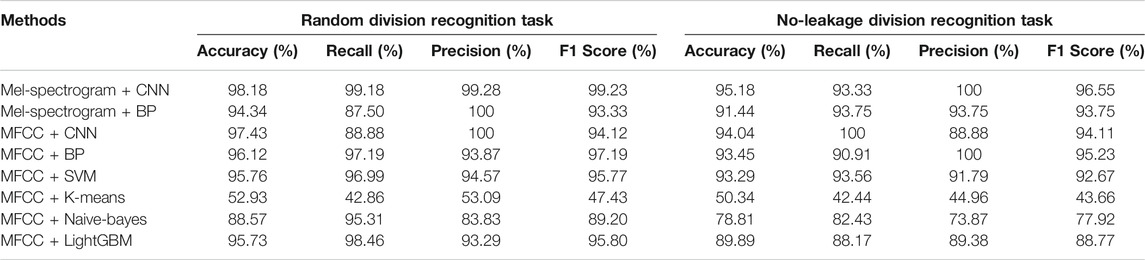
Мел-спектр содержит преобразование STFT для каждого кадра спектра (энергетический/амплитудный спектр), от линейной частотной шкалы к логарифмической мел-шкале, после чего проходит через фильтр-банк для получения собственного вектора. Эти собственные значения могут быть приблизительно выражены как распределение энергии сигнала на мел-шкале частот.

После того как аудиоданные обработаны и преобразованы в данные длиной 1 секунда, мы преобразуем все данные в мел-спектрограммы, чтобы обучить сверточные нейронные сети для распознавания.

Аудиоданные обычно имеют сложные характеристики, поэтому необходимо извлечь полезные признаки для распознавания аудио. Мел-спектрограмма является одним из эффективных методов для обработки звука, и для каждого аудиообразца используется дискретизация с частотой 8 кГц.

Мел-спектрограмма - это эффективный инструмент для извлечения скрытых характеристик из аудио и их визуализации в виде изображения. Модель CNN эффективно извлекает характеристики из изображений, а затем выполняет задачи, такие как классификация и распознавание. Поэтому мы используем модель CNN для эффективной классификации аудио и для достижения точного распознавания и обнаружения кашля.

Как показано в таблице , можно увидеть, что сочетание Мел-спектрограммы и CNN достигает наилучшего результата в распознавании кашля по сравнению с другими методами. Для случайно разделенных наборов данных правильная скорость распознавания составляет 98%. Можно заметить, что модель все равно способна достичь хорошего уровня распознавания, даже если в смесь входят различные звуки.



*Таблица*

Из результатов экспериментов на основе наборов данных с произвольным разделением и разделением без утечки мы видим, что предложенная модель CRN может достичь отличной производительности в распознавании кашля. По сравнению с другими методами точность CRN самая высокая, и большинство показателей являются лучшими. Чтобы оценить обобщающую способность модели, было собрано несколько звуков кашля, которые не использовались в обучении. Авторы обнаружили, что CRN также может эффективно их распознавать. Эксперименты показывают, что модель может эффективно распознавать кашель в сложных сценах и правильно распознавать кашель с различными другими звуками, что полезно для мониторинга кашля в повседневной жизни.

1. [AN EXPLORATION OF LOG-MEL SPECTROGRAM AND MFCC FEATURES FOR ALZHEIMER’S DEMENTIA RECOGNITION FROM SPONTANEOUS SPEECH](http://mile.ee.iisc.ac.in/publications/softCopy/SpeechProcessing/Camera_Ready_Version_SLT_2021.pdf)

В этой работе исследуется эффективность логарифмической мел-спектрограммы и коэффициентов мел-частотных кепстральных преобразований (MFCC) для распознавания болезни Альцгеймера (АД) на наборе данных ADReSS challenge. Используется три различные глубокие нейронные сети (DNN) для распознавания АД и прогнозирования баллов по мини-тесту состояния психического здоровья (MMSE): (i) сверточную нейронную сеть, за которой следует сеть долгой и краткосрочной памяти (CNN-LSTM), (ii) предварительно обученную сеть ResNet18, за которой следует LSTM (ResNet-LSTM), и (iii) пирамидальную двунаправленную LSTM, за которой следует CNN (pBLSTM-CNN). CNN-LSTM достигает точности 64.58% с использованием коэффициентов MFCC, а ResNet-LSTM достигает точности 62.5% с использованием логарифмических мел-спектрограмм.

Логарифмические мел-спектрограммы и коэффициенты мел-частотных кепстральных преобразований (MFCC) широко используются в глубоких нейронных сетях для различных задач, таких как распознавание эмоций, классификация аудио, выявление когнитивных нарушений/психических расстройств, таких как депрессия, и автоматическое распознавание речи (ASR). Однако они не слишком исследованы для оценки речевых нарушений, связанных с болезнью Альцгеймера. В этой работе мы оцениваем эффективность логарифмических мел-спектрограмм и коэффициентов MFCC для следующих задач на наборе данных ADReSS challenge с использованием глубоких нейронных сетей:

Бинарная классификация образцов спонтанной речи на классы с АД и без.

Прогнозирование баллов по мини-тесту состояния психического здоровья (MMSE), показательного измерения когнитивного нарушения.

Для генерации логарифмических мел-спектрограмм выбрали 224 мел-фильтра. Этот выбор позволяет использовать предварительно обученные модели CNN, такие как ResNet. Авторы использовали окно Ханнинга размером 2048 выборок (≈ 128 мс) и шаг (количество выборок между последовательными кадрами) 512 выборок (≈ 32 мс). Количество точек в вычислении преобразования Фурье составляет также 2048. Логарифмическая мел-спектрограмма, дельта и дельта-дельта изображения масштабируются до значений от 0 до 1 с использованием библиотеки scikit-learn. Архитектуры CNN в начале требуют изображений фиксированного размера на входе. Для достижения этого логарифмические мел-спектрограммы разделяются по времени на неперекрывающиеся сегменты по 224 кадра каждый, охватывающие длительность 7,264 секунды. Логарифмические мел-спектрограммы, дельта и дельта-дельта признаки поочередно подаются на каналы 1, 2 и 3 входного слоя CNN. Таким образом, вход в CNN имеет размерность 3x224x224 (каналы x высота x ширина). Этот выбор размерности признаков обусловлен тем, что предварительно обученные модели, такие как ResNet, должны иметь минимальную размерность 3x224x224 на входе.

Алгоритм pBLSTM-CNN использует признаки логарифмических мел-спектрограмм, сгенерированные с окном размером 400 выборок (25 мс) и шагом 160 выборок (10 мс). Используются 512-точечное FFT и 40 мел-шкал фильтра. Также используются дельта и дельта-дельта признаки, что приводит к размерности входа 120.

Результаты исследования показывают, что логарифмические мел-спектрограммы и коэффициенты мел-частотных кепстральных коэффициентов (MFCC) эффективны для задачи обнаружения болезни Альцгеймера (AD), когда они используются вместе с моделями глубокого обучения (DNN). Модель ResNet-LSTM, основанная на трансферном обучении, хорошо справляется как с задачей классификации AD, так и с предсказанием оценки мини-медицинского состояния (MMSE). Применение архитектуры ResNet18 перед LSTM позволяет захватить пространственные особенности в логарифмической мел-спектрограмме, важные для различения признаков AD.

1. [Mel Spectrogram-based advanced deep temporal clustering model with unsupervised data for fault diagnosis](https://www.sciencedirect.com/science/article/pii/S0957417423000520?via%3Dihub)

Диагностика неисправностей механического оборудования с помощью методов машинного обучения, основанных на данных, была разработана в последнее время как перспективный метод повышения надежности промышленных систем. Однако эти методы страдают от нехватки данных из-за сложности их сбора, что ограничивает возможности выделения аномалий. Для решения этой проблемы мы предлагаем модель глубокого временного кластерного анализа (ADTC), основанную на мел-спектрограмме, которая может извлекать и верифицировать признаки не размеченных данных через автоэнкодер на основе обучения без учителя и алгоритме K-средних.

Изображения, основанные на увеличенных мел-спектрограммах, были использованы в данном исследовании для преодоления проблемы разреженности данных и дисбаланса данных, характерных для промышленности и методов машинного обучения. Кроме того, разреженность данных можно преодолеть, извлекая особенности основной структуры данных заранее через безнадзорное обучение с использованием этих данных. Диагностическую производительность также можно улучшить путем коррекции особенностей наборов данных, которые содержат неточные характеристики, с использованием обучения с учителем.

Для преодоления проблемы несбалансированных временных рядов механического оборудования было проведено увеличение данных. Ограниченная способность извлечения признаков из данных механического оборудования с разреженностью была решена путем исследования извлечения признаков с использованием модели на основе безнадзорного обучения. Более того, диагностика неисправностей была выполнена путем включения извлеченных значений признаков необученных данных в модель обучения с учителем.

С использованием предложенного метода были оценены доменно-адаптивные характеристики механического оборудования и устойчивость к воздействию промышленного шума. Результаты показывают, что высокая предсказательная точность предложенной модели диагностики неисправностей делает ее устойчивой к шумным средам и проявляет адаптивные характеристики доменов в наборах данных с различной продолжительностью времени. Таким образом, было показано, что модель ADTC на основе мел-спектрограммы является эффективным подходом к диагностике неисправностей, а сбалансированный набор данных обладает лучшей производительностью по диагностике неисправностей. Однако метод K-средних имеет ограничение в том, что количество кластеров должно быть определено заранее.

1. [Analyzing Noise Robustness of Cochleogram and Mel Spectrogram Features in Deep Learning Based Speaker Recogntion](https://assets.researchsquare.com/files/rs-2218302/v1/313eab0f-6eed-4551-b835-0218ca71ee23.pdf?c=1667616266)

В данном исследовании был проведен анализ устойчивости к шуму признаков кохлеограммы и мел-спектрограммы в распознавании дикторов с использованием аддитивных шумов (например, болтовня, уличные и ресторанные шумы). Для эксперимента использовалась часть train-clean-100 набора данных LibriSpeech, состоящая из 251 диктора (126 мужчин и 125 женщин) и 28 539 фраз. Модель CNN использовалась для обучения и классификации дикторов по разным классам. Результаты оценки показывают, что кохлеограмма более устойчива, чем мел-спектрограмма, при низком уровне SNR. Оба признака кохлеограммы и мел-спектрограммы показывают примерно одинаковую точность при высоком SNR и отсутствии аддитивного шума.

**Encoder и Decoder**

Кодер-декодер - это тип архитектуры нейронной сети, которая используется для обучения по принципу "sequence-to-sequence". Она состоит из двух частей - кодера и декодера. Кодер обрабатывает входную последовательность для получения набора контекстных векторов, которые затем используются декодером для генерации выходной последовательности. Эта архитектура позволяет решать такие задачи, как машинный перевод, резюмирование текста, создание подписей к изображениям и другие. Идея заключается в том, чтобы иметь возможность получать данные в одной форме (например, текст) и преобразовывать их в другую (например, изображения). Таким образом, машины могут научиться понимать сложные взаимосвязи между различными типами данных и использовать их для более эффективной обработки.

Кодер - это первая часть архитектуры кодера-декодера. Он принимает входную последовательность и обрабатывает ее для создания набора контекстных векторов, которые затем используются декодером. То, как происходит процесс кодирования, зависит от типа используемого приложения. Например, для текстовых приложений, таких как машинный перевод или обобщение, слова в каждом предложении преобразуются в числовые значения, которые представляют их математически. Затем эти числа проходят через ряд слоев, которые уменьшают их размерность, сохраняя при этом необходимую информацию о том, как они соотносятся друг с другом в структуре предложения. Эта "закодированная" версия каждого предложения передается декодеру для дальнейшей обработки.

Декодер отвечает за то, чтобы взять это закодированное представление и восстановить его в исходную форму (или что-то похожее). Чтобы сделать это, должна существовать какая-то связь между тем, что было закодировано, и тем, что нужно восстановить; в противном случае это будет просто случайное угадывание. Чтобы установить эту связь, большинство современных архитектур используют механизмы внимания, которые позволяют определенным частям входной последовательности (например, отдельным словам) влиять на то, как последующие части обрабатываются или интерпретируются моделью - по сути, придавая больший вес или важность определенным элементам по сравнению с другими при генерации выходных последовательностей из входных данных кодирования. Таким образом, модели становятся гораздо более точными в создании выходных данных, которые точно отражают источники входных данных, и даже могут изучать различные модели на различных наборах данных без необходимости дополнительных циклов обучения или процедур настройки параметров.

Подробно архитектуру encoder-decoder объясняет Google Cloud Tech в своём видео:[Encoder-decoder architecture: Overview](https://www.youtube.com/watch?v=zbdong_h-x4)

Статья: [Understanding How Encoder-Decoder Architectures Attend](https://arxiv.org/pdf/2110.15253.pdf)

RNN Encoder-Decoder была представлена в 2014 году в публикации [Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation](https://arxiv.org/pdf/1406.1078.pdf).

Кодер и декодер предлагаемой модели совместно обучаются для максимизации условной вероятности целевой последовательности при исходной последовательности. Эмпирически установлено, что производительность системы статистического машинного перевода повышается при использовании условных вероятностей пар фраз, вычисляемых кодером-декодером RNN, в качестве дополнительной характеристики в существующей логлинейной модели. Качественно показано, что предложенная модель обучается семантически и синтаксически осмысленному представлению языковых фраз.