УДК 004.8

# МЕТОДОЛОГИЯ ВЫДЕЛЕНИЯ ВОКАЛЬНОГО АУДИОПОТОКА ИЗ МЕДИАФАЙЛА

**К. В. Рубанов**

*Учреждение образования «Гомельский государственный технический университет имени П. О. Сухого», Республики Беларусь*

*Рассмотрены способы выделения вокального аудио потока из медиафайла. Результат исследования существующих методов показал недостаточное качество обработки исходного файла. В результате предложена методология выделения голоса с помощью сверточной нейронной сети.*

**Ключевые слова:** нейронные сети, выделение голоса, обработка звука, машинное обучение, сверточные сети.

# METHODOLOGY FOR EXTRACTING A VOCAL AUDIO STREAM FROM A MEDIA FILE

**K. V. Rubanau**

*Sukhoi State Technical University of Gomel, the Republic of Belarus*

*The methods for extracting a vocal audio stream from a media file are considered. The result of the study of existing methods showed the insufficient quality of the processing of the source file. As a result, a methodology for voice extraction using a convolutional neural network is proposed.*

**Keywords:** neural networks, voice extraction, sound processing, machine learning, convolutional networks.

Речь является повсеместно используемым и одним из самых доступных способов обмена информацией между людьми. Технологическое развитие позволило передавать, хранить и воспроизводить запись голоса без приложения человеческих усилий. Для улучшения средств манипуляции над существующими медиаданными требуется дальнейшее развитие технологий. Один из способов таких манипуляций – выделение вокального аудиоряда из медиафайлов, его дальнейшая очистка от посторонних шумов и использование во множестве сфер, например, обучение людей с нарушением зрения, озвучивание учебных материалов, расширение возможностей самовыражения людей с приобретенными или врожденными нарушениями речи. Частота голоса, выделенного из звукового потока, открывает возможность использования конкретного человеческого голоса для упомянутых выше нужд, не ограничиваясь специально созданными для этих целей экземплярами голосов дикторов.

Целью данной статьи является определение методологии выделения голоса из медиафайла.

Рассмотрим несколько уже существующих способов решения поставленной задачи.

Еще одним подходом к выделению вокального аудиоряда является подход, основаный на формализации знаний о природе звука и использовании физических формул для обработки сигналов. Алгоритм обработки аудиозаписи включает следующее [1]:

1) определить участки с вокалом. В исходном сигнале присутствует вокальная составляющая и звуки музыкальных инструментов, частоты которых могут коррелировать с частотой голоса. Необходимо сосредоточиться на участках, которые содержат вокальные частоты;

2) различить вокализированную и невокализированную речь. Невокализованная речь – речь, которая включает согласные, такие как ‘т’, ‘п’, ‘к’, ‘с’ (которые не производятся вибрацией голосовых связок). Все это проявляется в виде коротких всплесков в высокочастотной области. Поскольку эти виды речи отличаются, есть вероятность, что их придется обрабатывать по-разному;

3) оценить изменение фундаментальной частоты во времени;

4) на основании вывода 3 применить некую маску для захвата гармоник;

5) отдельно обработать элементы невокализированной речи.

Результатом является битовая маска, применение которой к амплитуде STFT (поэлементное умножение) дает приблизительную реконструкцию амплитуды STFT вокала. Затем необходимо объединить эту вокальную STFT с информацией о фазе исходного сигнала, вычислить обратный STFT и получить временной сигнал реконструированного вокала [1].

Также для решения поставленной задачи существует множество готовых программных продуктов, описание и оценка которых приведены далее.

Audacity – свободный аудиоредактор звуковых файлов, ориентированный на работу с несколькими дорожками. Стоит отметить, что полностью выделить только голос из аудиозаписи таким способом невозможно, поскольку приложение удаляет не только голос в диапазоне от 500 до 2000 Гц, но и все звуки, попадающие в этот диапазон.

iZotope RX7 Editer – система для подготовки и обработки музыкальных материалов, состоящая из набора плагинов.

Анализ литературы позволил определить несколько основных проблем изоляции вокала из музыкальной композиции:

– возможность существования на записи нескольких голосов. Фокусировка на удалении заранее заданных частот из композиции может привести к ситуации, когда на аудиозаписи присутствует несколько голосов с разной высотой. Удаление частот для одного голоса в этом случае может уменьшить качество выделения другого. Эта проблема может быть решена применением различных фильтров для каждого голоса, обрабатывая аудиозапись несколько раз, что может быть трудозатратно;

– способность многих исполнителей петь с принципиально разной высотой в разные моменты композиции. Эта проблема так же как и первая решается с помощью динамических фильтров;

– наличие инструментального аудиоряда, который может совпадать по высоте с человеческим голосом, что приведет к наличию шумов в результирующей аудиозаписи;

– представленные выше методы изменения голоса в исходной аудиозаписи будут обрабатывать не только вокальный аудиоряд, но и инструментальный, что нежелательно для решения задачи замены голоса.

Решением вышеперечисленных проблем может стать использование нейронных сетей, поскольку при правильном обучении сети она сможет подбирать правильные фильтры для каждого сэмпла аудиозаписи, что не под силу человеку в виду трудоемкости задачи.

*Проектирование пространства признаков*. Известно, что звуковые сигналы, такие как музыка и человеческая речь, основаны на временных зависимостях [2]. Иными словами, ничто не происходит изолированно в данный момент времени. Для определения присутствия голоса на конкретном фрагменте звукозаписи необходимо смотреть на соседние регионы. Такой временной контекст дает хорошую информацию о том, что происходит в интересующей области. В то же время желатель-

но выполнять классификацию с очень малыми временными приращениями, чтобы распознавать человеческий голос с максимально возможным разрешением по времени.

Для вычисления пространства признаков необходимо искользовать следующие параметры: частота дискретизации (*fs*): 22050 Гц (можно понизить с 44100 до 22050), дизайн STFT: размер окна = 1024, hop size = 256, интерполяция мел-шкалы для взвешивающего фильтра с учетом восприятия. Поскольку входные данные настоящие, можно работать с половиной STFT, сохраняя компонент DC, что дает 513 сэмплов [3], целевое разрешение классификации: один кадр STFT (~11,6 мс = 256 / 22050) [4], целевой временной контекст: ~300 миллисекунд = 25 кадров STFT, целевое количество обучающих примеров: 350 тыс, предполагая, что используется скользящее окно с шагом в 1 таймфрейм STFT для генерации учебных данных, нужно около 1,6 ч размеченного звука для генерации 350 тыс. образцов данных [1]. Таким образом, входной тензор признаков должен иметь размеры [350000, 513, 25]. Для достижения объема в 35000 фрагментов необходимо около 4-х трехминутных записей, поскольку трехминутная аудиозапись имеет около 8000 фрагментов. Визуальное представление обучающей выборки представлено на рис. 1.

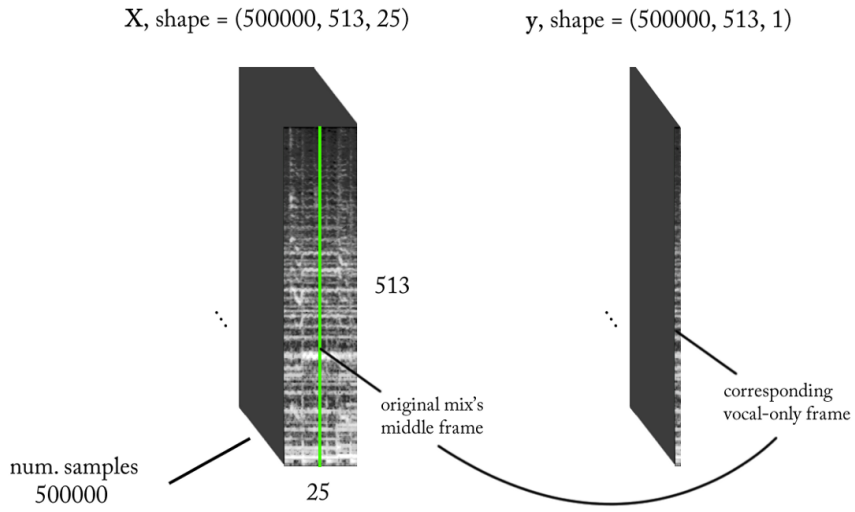


*Рис. 1.* Визуализация обучающей выборки для решения задачи многомерной регрессии [1]

Для работы с описанным выше объемом данных нейронная сеть должна иметь слои: сверточный слой размерностью [25, 513], несколько промежуточных слоев, формат и количество которых должно быть определено на этапе реализации, финальный полносвязный слой. Исходя из этого, результатом работы нейронной сети будет одномерный тензор, а его результатом – чистый фрагмент записи голоса.

Таким образом, определена методология выделения голоса из медиафайла и разработана примерная архитектура нейронной сети для решения поставленной задачи.

Литература

1. Koretzky, A. Audio AI: выделяем вокал из музыки с помощью сверточных нейросетей / A. Koretzky. – 2019. – Режим доступа: https://habr.com/ru/post/441090/. – Дата доступа: 20.02.2021.
2. Волковец, В. И. Создание и обработка звука при разработке интерактивных приложений / В. И. Волковец // Докл. БГУИР. – 2018. – С. 4.

3. Адаптивное разделение источников звука в режиме реального времени : заявка 15/434, 419 Соединенные Штаты Америки / Пилл Р. ; пат. поверенный Коретский [и др.] ; заявл. 12.01.17 ; опубл. 17.08.17 ; приоритет 16.02.16, № US201662295497P (США).
4. Как работает сверточная нейронная сеть: архитектура, примеры, особенности. – 2018. – Режим доступа: https://neurohive.io/ru/osnovy-data-science/glubokaya-svertochnaja-nejronnaja-set/. – Дата доступа: 27.02.2021.

УДК 620.178.16:620.178.3

## К ВОПРОСУ ОБ АВТОМАТИЗИРОВАННОМ ПРОЕКТИРОВАНИИ УЗЛОВ ТРЕНИЯ

**С. А. Тюрин, Н. Н. Малык**

*Учреждение образования «Гомельский государственный технический университет имени П. О. Сухого», Республика Беларусь*

*Рассмотрена задача и предложен автоматизированный метод расчета и проектирования узлов трения, в основу которого положена механика контакта, механика деформирования и разрушения и механика усталостного разрушения.*

**Ключевые слова:** усталость, трение, элемент конструкции, узел трения, проектирование.

## TO THE QUESTION OF AUTOMATED DESIGN OF FRICTION UNITS

**S. A. Tyurin, N. N. Malyk**

*Sukhoi State Technical University of Gomel, the Republic of Belarus*

*The problem is considered and a automated method for calculating and designing friction units is proposed, which is based on the mechanics of contact, the mechanics of deformation and fracture, and the mechanics of fatigue fracture.*

**Keywords:** fatigue, friction, structural element, friction unit, design.

Изучая некоторые закономерности объемного разрушения при механической усталости, мы имеем дело с отдельным элементом конструкции, который называли (деформируемым твердым) телом, деталью, образцом либо просто объектом.

Поверхностное повреждение при трении в трибологии возникает при относительном движении, по меньшей мере, двух взаимодействующих тел, например, при скольжении либо качении; в обоих случаях они образуют пару трения. Говорят, что пару (или узел) трения составляют образец и контробразец, либо иначе: тело и контртело. Их силовое взаимодействие обусловлено специфической – контактной нагрузкой.

**Автоматизированный метод проектирования узлов трения.** При механической усталости критерием предельного состояния служит усталостное (объемное) разрушение детали (элемента конструкции (рис. 1)), например, разделение ее на части. Условие прочности:

$$\sigma = \frac{M}{W} \le [\sigma] = \frac{\sigma_{-1}}{\tilde{n}_\sigma}, \qquad (1)$$

где $\sigma$ – циклическое напряжение; $\sigma_{-1}$ – предел выносливости при механической усталости; $[\sigma]$ – допустимое напряжение; $M$ – изгибающий момент; $W$ – момент сопротивления; $n_\sigma$ – коэффициент запаса.

По условию (1) решаются три задачи: