Инженерный вестник Дона, №9 (2023)
ivdon.ru/ru/magazine/archive/n9y2023/8692

Препроцессинг речевых данных с целью обучения нейронной сети

Е.В. Щенникова, Д.Ю. Флёрина, Р.Е. Навошин

Национальный исследовательский Мордовский государственный университет

Аннотация: В данной статье рассматриваются проблемы препроцессинга аудиоданных для дальнейшего применения при обучении нейронной сети. В качестве решения ряда проблем выбран метод с мел-частотными кепстральными коэффициентами, что позволило уменьшить входные данные для обучения, увеличить производительность, улучшить четкость распознавания.

Ключевые слова: машинное обучение, препроцессинг данных, аудиоанализ, мел-кепстральные коэффициенты, извлечение признаков, спектр голосового сигнала, преобразование Фурье, окно Ханна, дискретное косинусное преобразование, короткое преобразование Фурье.

Сложность обработки голосовых данных

В последние годы голосовые данные стали объектом всё большего интереса в области машинного обучения. Эти данные представляют собой ценную информацию, которая помогает в решении различных задач, таких, как распознавание речи, эмоциональный анализ, автоматическое транскрибирование и другие [1].

Голосовые данные обладают уникальными особенностями, которые делают их сложными для обработки и анализа. Работа с голосовыми данными в машинном обучении является сложной по нескольким причинам:

– Высокая размерность данных: голосовые данные представлены как временные сигналы с высокой частотой дискретизации и длительностью. Это приводит к большому количеству точек данных в каждом голосовом сигнале. Обработка и анализ таких высокоразмерных данных требует больших вычислительных ресурсов и методов с высокой производительностью [2].

– Вариабельность и сложность данных: голосовые данные сильно варьируются в зависимости от различных факторов, таких, как возраст, пол, акцент, эмоциональное состояние и фоновый шум. Это осложняет извлечение информации из данных и построение моделей, которые обобщают эти различия [3].

– Проблемы с пропущенными значениями: голосовые данные часто содержат пропущенные значения или недостоверные фрагменты, вызванные фоновым шумом или другими факторами [4]. Обработка и корректное обращение с пропущенными значениями сложная задача, требующая специальных алгоритмов восстановления и очистки данных.

– Интерпретация эмоций и контекста: голос несет большое количество эмоциональной и контекстуальной информации, которая усложняет анализ и применение для подготовки моделей машинного обучения. Определение и интерпретация эмоций, интонаций и других аспектов голоса являются субъективными и сложными задачами [5].

– Ограниченная доступность размеченных данных: получение большого объема размеченных голосовых данных сложная задача. Необходимо обладать специализированными навыками или иметь доступ к размеченным наборам данных для подготовки эффективных моделей машинного обучения.

Все эти факторы делают работу с голосовыми данными сложной задачей в машинном обучении. Она требует использования специализированных алгоритмов и моделей, а также глубокого понимания особенностей голосовых данных. Однако, с применением правильных методов и архитектур моделей, можно достичь хороших результатов в распознавании речи, эмоциональном анализе и других задачах обработки голосовых данных.

Извлечение признаков из голосовых данных

На первом этапе обучения нейронной сети извлечении признаков – стояла задача преобразования голосовых данных из временного сигнала в числовые признаки, чтобы в дальнейшем использовать их для обучения нейронной сети.

Анализ мел-частотных кепстральных коэффициентов повсеместно используется в глубоком обучении для обучения моделей, таких, как сверточные нейронные сети и рекуррентные нейронные сети [6]. Анализ мел-частотных кепстральных коэффициентов является одним из наиболее распространенных и эффективных методов для анализа аудиосигналов, особенно в таких областях, как распознавание речи, музыкальная информатика, аудиоидентификация и классификация звуков.

Для выделения признаков с помощью мел-частотных кепстральных коэффициентов был загружен и обработан, посредством языка python, файл с голосовыми данными. На рис. 1 продемонстрировано временное представление аудио файла.

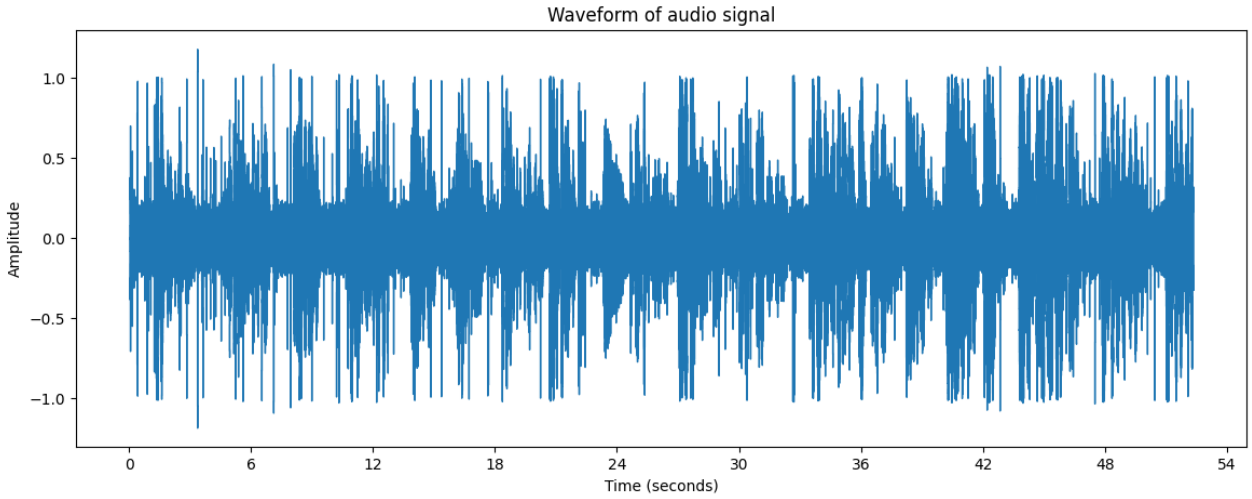

Рис. 1. Временное представление данных

Временное представление, такое, как временной график амплитудного сигнала, позволяет визуализировать изменение амплитуды звукового сигнала в зависимости от времени. Мы получили наглядное представление о временных характеристиках голосовых данных.

В контексте препроцессинга голосовых данных, временное представление позволило идентифицировать различные аспекты звука, такие, как длительность фонем, наличие тишины, а также обнаруживать артефакты и шумы. Это наблюдение временных характеристик является важной

составляющей в разработке методов предобработки для улучшения распознавания речи и обработки аудиосигналов.

На рис. 2 показан спектр исходного сигнала по всей временной оси, который мы получили с помощью кратковременного преобразования Фурье с использованием окна Ханна.

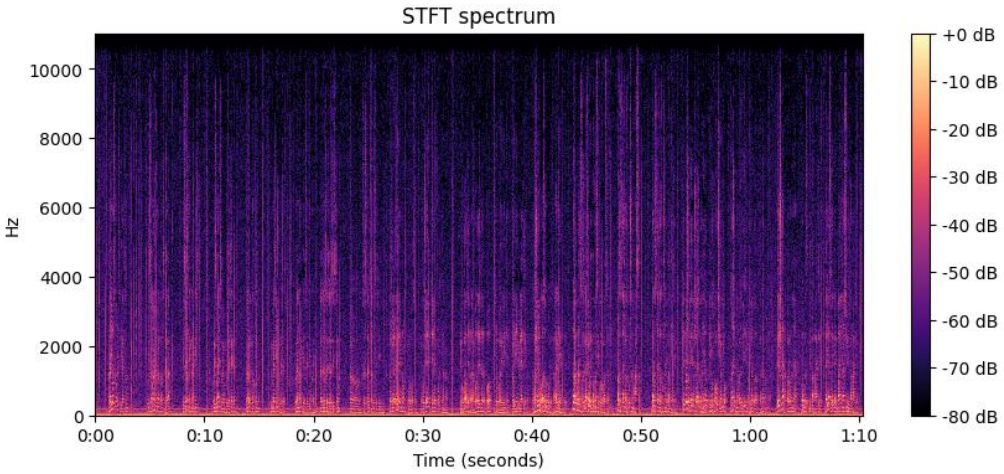

Рис. 2. – Спектр исходного сигнала

Быстрое преобразование Фурье через окно Ханна это способ применения оконной функции Ханна к сигналу перед выполнением короткого преобразования Фурье, что улучшает точность спектрального анализа.

Формула для вычисления коротких преобразований Фурье с использованием окна Ханна выглядит следующим образом:

$$X[m,n]=\sum_{k=0}^{N-1}x[n+k]\omega[k]e^{-j\frac{2\pi}{N}mk}, \quad (1)$$

где $X[m, n]$ – комплексное значение кратковременного преобразования Фурье в момент времени n и с частотой m; $x[n]$ – входной аудиосигнал; $\omega[k]$ – значением окна Ханна в момент времени k; N – размер окна; m – индекс частоты; n – индекс времени.

Оконная функция Ханна имеет следующую формулу:

$$\omega = 0{,}5 - \cos\left(\frac{2\pi k}{N-1}\right), \quad (2)$$

где N – размер окна, а k – текущий индекс времени.

Для последующего анализа мы перевели частоты на шкалу «мел», используя следующую формулу:

$$mel(f) = 1125\ln\left(1+\frac{f}{700}\right), \quad (3)$$

где f – частота в герцах, а $mel(f)$ – соответствующее значение на мел-шкале.

Чтобы расположить спектр на мел-шкале, использовали интерпретируемый язык python, а именно - функция librosa.feature.melspectrogram(), которая автоматически вычисляет мел-спектрограмму с помощью кратковременного преобразования Фурье (рис. 3).

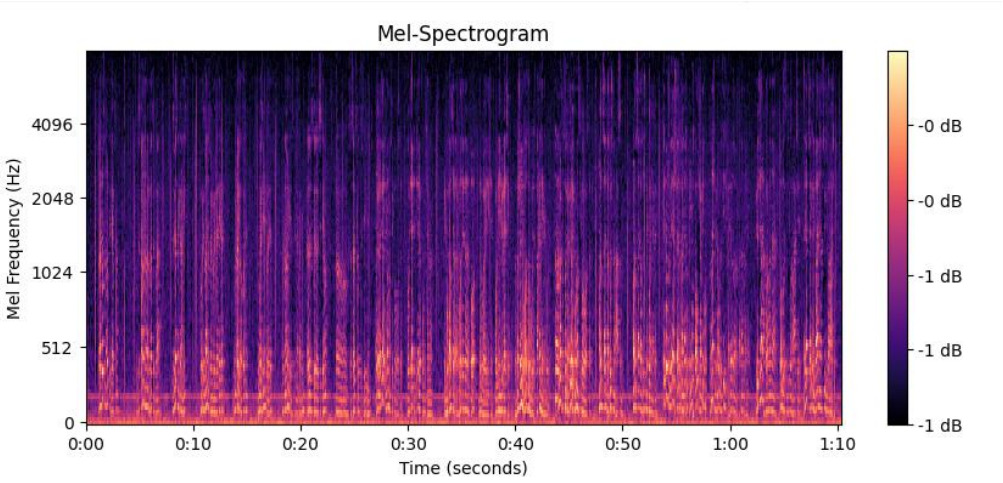

Рис. 3. – Мел-спектограмма

Мел-шкала была разработана для моделирования поведения человеческого слуха, поэтому она позволила более точно отобразить различия между низкими частотами, которые слышатся хорошо, и высокими частотами, которые менее различимы.

На рис. 4 отображен график переведенных мел-частот на обычную частотную шкалу по следующей формуле:

$$f(mel) = 700\left(\exp\left(\frac{mel}{1125}\right)-1\right), \quad (4)$$

где *mel* – значение на мел-шкале, а *f(mel)* – соответствующая частота в герцах.

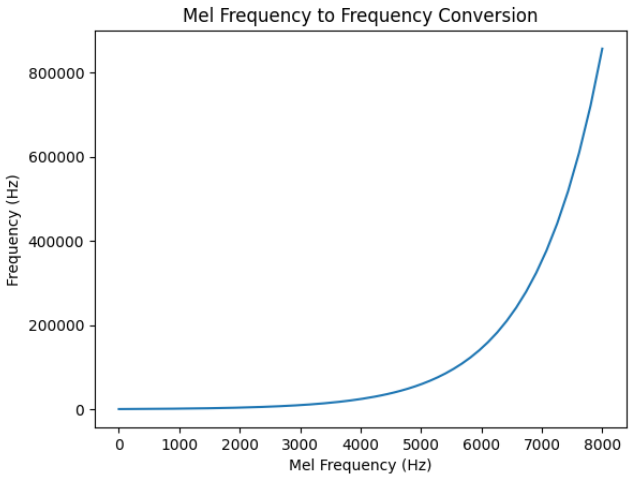

Рис. 4. – Преобразование mel-частоты в частотную шкалу

На представленном выше графике показано, что окна «собираются» в области низких частот, обеспечивая более высокое «разрешение» там, где оно необходимо для распознавания. Увеличение значения частоты со значениями мел-частоты говорит о том, что звук становится более высоким. Для восприятия звуков человеческий слух менее чувствителен к изменениям частот в более высоких диапазонах, чем в более низких диапазонах. Перевод мел-шкалы в обычную частотную шкалу позволил соотнести изменения на шкалах со способностью человеком воспринимать изменения частот.

Таким образом, так как значения частоты увеличиваются вместе со значениями мел-частоты, это означает, что человеческий слух воспринимает более высокий звук. Например, для мел-частоты на уровне 1000 Гц звук соответствует частоте около 1660 Гц, тогда как для мел-частоты 2000 Гц звук уже соответствует частоте около 3000 Гц, что значительно выше.

Простым перемножением векторов спектра сигнала и оконной функции мы нашли энергию сигнала, которая попадает в каждое из окон анализа. Мы получили некоторый набор коэффициентов, это так называемые мел-частотные спектральные коэффициенты. Возводим их в квадрат и

логарифмируем для приведения к кепстральному виду. Эти преобразования были выполнены специальными функциями: librosa.power_to_db для логарифмирования мел-спектра и librosa.feature.mfcc для вычисления мел-частотных кепстральных коэффициентов. Результаты представлены на рис. 5.

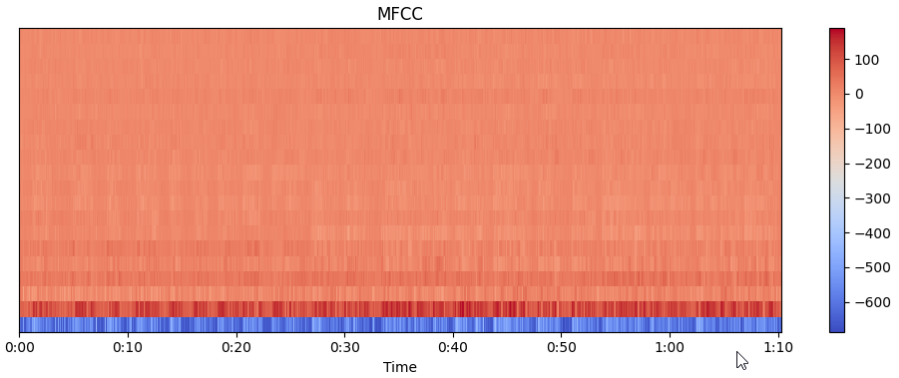

Рис. 5. – Мел-частотные кепстральные коэффициенты

Использование мел-шкалы помогает в задачах анализа звука и используется при обучении нейронных сетей при работе с речью. Использование мел-кепстральных коэффициентов улучшило качество распознавания за счет того, что позволило увидеть наиболее информативные коэффициенты. Эти коэффициенты уже были использованы как входные данные для нейронной сети.

Использование мел-частотных кепстральных коэффициентов улучшило производительность нейронной сети при анализе звука. Кроме того, мел-частотные кепстральные коэффициенты имеют меньшую размерность по сравнению с обычным представлением аудиоданных, что ускорило обучение и уменьшило количество параметров сети.

Заключение

Одним из перспективных направлений использования голосовых данных является анализ эмоционального состояния людей на основе голосовых данных [7-9]. Такой анализ может быть использован в психологии и психотерапии, например, для оценки эффективности терапии или оценки

симптомов психических расстройств [10]. Но также существует сложность в обработке голосовых данных, что требует применения новых методов обработки данных. В данной статье мы решили проблему использованием метода с мел-частотными кепстральными коэффициентами путем извлечения информации о спектре голосового сигнала, с использованием преобразования Фурье и шкалы частот, основанную на восприятии звука человеком.

Таким образом, использование мел-шкалы и мел-частотных кепстральных коэффициентов помогло облегчит облегчить обучение и улучшить производительность нейронной сети в задаче обработки и анализа звука.

Литература

1. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. Москва. Радио и связь. 1981. 496 с.

2. Oppenheim A.V., Schafer R. W., Buck J. R. Discrete-Time Signal Processing. Prentice Hall. 1998. 870 p.

3. Huang X., Acero A. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall. 2001. 980 p.

4. Бессмертный И.А. Искусственный интеллект. СПб. СПбГУ ИТМО. 2010. 132 с.

5. Сато Ю. Без паники! Цифровая обработка сигналов. Москва. ДМК Пресс. 2010. 176 с.

6. Пучков Е.В. Сравнительный анализ алгоритмов обучения искусственной нейронной сети // Инженерный вестник Дона. 2020. №4. URL: ivdon.ru/magazine/archive/n4y2013/2135.

7. Liao Y., Vakanski A., Xian M., Paul D., Baker R. Computers in Biology and Medicine // Elsevier. 2020. URL: pubmed.ncbi.nlm.nih.gov/32339122.

8. McCulloch W.S., Pitts W. A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics. 1943. 115 p.

9. Стебаков И.Н., Шутин Д.В., Марахин Н.А. Машинное обучение в реабилитационной медицине и пример классификатора движений пальцев для кистевого тренажера // Инженерный вестник Дона. 2020. №6. URL: ivdon.ru/ru/magazine/archive/N6y2020/6514.

10. Манкибаев Б. С. Основные направления внедрения искусственного интеллекта в медицине // Наука, образование и культура. 2019. №3. С. 69-71.

References

1. Rabiner L.R., Shafer R.V. Tsifrovaya obrabotka rechevykh signalov [Digital speech processing]. Moskva. Radio i svyaz. 1981. 496 p.

2. Oppenheim A.V., Schafer R. W., Buck J. R. Discrete-Time Signal Processing. Prentice Hall. 1998. 870 p.

3. Huang X., Acero A. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall. 2001. 980 p.

4. Bessmertnyy I.A. Iskusstvennyy intellekt [Artificial intelligence]. SPb. SPbGU ITMO. 2010. 132 p.

5. Sato Yu. Bez paniki! Tsifrovaya obrabotka signalovy [Don't panic! Digital signal processing]. Moskva. DMK Press. 2010. 176 p.

6. Puchkov E.V. Inzhenernyj vestnik Dona. 2020. №4. URL: ivdon.ru/magazine/archive/n4y2013/2135.

7. Liao Y., Vakanski A., Xian M., Paul D., Baker R. Elsevier. 2020. URL: pubmed.ncbi.nlm.nih.gov/32339122.

8. McCulloch W.S., Pitts W. A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics. 1943. 115 p.

9. Stebakov I.N., Shutin D.V., Marakhin N.A. Inzhenernyj vestnik Dona. 2020. №6. URL: ivdon.ru/ru/magazine/archive/N6y2020/6514.

10. Mankibaev B.S. Nauka, obrazovanie i kultura. 2019. №3. P. 69-71.