УДК 534.87

Разработка сверточной нейронной сети для классификации амплитудно-частотных характеристик аудиосигналов*

В.Н. Попов, П.С. Ладыгин, В.В. Карев, Я.И. Борцова

Алтайский государственный университет (Барнаул, Россия)

Development of Convolutional Neural Network for Classification of Amplitude-Frequency Characteristics of Audio Signals

V.N. Popov, P.S. Ladygin, V.V. Karev, Ya.I. Bortsova

Altai State University (Barnaul, Russia)

Применена технология глубокой сверточной нейронной сети к обработке аудиофайлов, в частности для классификации амплитудно-частотных характеристик аудиосигналов. Сопоставление аудиофрагментов между собой сведено к решению задачи верификации дикторов. В качестве набора данных для обучения глубокой сверточной нейронной сети собрана большая репрезентативная выборка аудиосигналов и дополнена удовлетворяющим требованиям набором данных Free Music Archive, который содержит свободно распространяемые аудиозаписи музыкальных произведений. Для предотвращения переобучения предсказательной модели применены четыре типа аугментации, в том числе гауссовый шум, реверберация, изменение частоты основного тона, изменение темпа аудиосигнала. В качестве предсказательной модели была взята архитектура CQT-Net. Для сравнения векторов признаков используется косинусное сходство. Качество верификации было протестировано на двух выборках, состоящих из 1500 аудиозаписей, которые не использовались во время обучения. Построены Det-кривые для наборов данных, в том числе тестовых с измененным темпом и с измененным питчем. В роли метрики качества модели использован равный уровень ошибок (коэффициент Equal Error Rate). Оценена вероятность выявления наиболее применяемых искажений аудиосигналов в амплитудно-частотной области (не менее 92 %), что говорит о надежности полученной системы.

Ключевые слова: сверточная нейронная сеть, Det-кривые, классификация, косинусное сходство, предсказательная модель.

This paper studies the application of deep convolutional neural networks for the processing of audio files, particularly for classifying amplitude-frequency characteristics of audio signals. The mapping of audio fragments to each other is reduced to verifying objects by their representation. A large representative sample of audio signals was collected and supplemented with a satisfying Free Music Archive dataset to produce a dataset for training a deep convolutional neural network. The CQT-Net architecture is taken as a predictive model with cosine similarity being used to compare feature vectors. Four types of augmentation, including Gaussian noise, reverberation, change in pitch frequency, and change in tempo of the audio signal, are used to prevent retraining of the predictive model. The verification quality of the predictive model is tested on two separate datasets consisting of 1500 audio recordings excluded from the training dataset. Detection error tradeoff curves are plotted for all datasets, including testing ones with a changed pace and with a changed "pitch." Equal Error Rate is used as a model quality metric. The probability of identification of commonly used distortions of audio signals in the amplitude-frequency domain is evaluated to be higher than 92%. It signifies the reliability of the developed model.

Keywords: convolutional neural network, detection error tradeoff curve, classification, cosine similarity, predictive model.

DOI: 10.14258/izvasu(2022)1-19

*Исследование выполнено в рамках реализации Программы поддержки научно-педагогических работников ФГБОУ ВО «Алтайский государственный университет».

Введение

Современные компьютерные алгоритмы, в том числе алгоритмы машинного обучения, позволяют развиваться области аудиоанализа [1]. Он включает в себя автоматическое распознавание речи, цифровую обработку, классификацию, поиск, тегирование, обнаружение и генерацию акустических сигналов. Анализ аудиальной информации все чаще становится задачей для различного рода программных средств, использующих математический аппарат принятия решений.

Тем не менее в настоящее время идентификация, классификация и сравнение аудиосигналов и их фрагментов друг с другом не являются повсеместным автоматизированным процессом. Например, за доказательную базу (в ходе экспертных исследований) специалисты используют методы, которые включают в себя непосредственное исследование аудиоданных с помощью органов слуха [2], а применение технических средств ограничивается наложением одной записи на другую. Подобные действия несут в себе существенную долю субъективизма, зависят от квалификации специалиста и не всегда позволяют справиться с техническими приемами воздействия на аудиоданные (например, изменение питча, ритма, внесение эффектов и искажений) [3], а также с естественным преобразованием звука под влиянием распространения акустических волн в разных средах.

В связи с этим актуально развитие методов машинного обучения в области аудиоанализа. Глубокие нейронные сети зарекомендовали себя в задаче распознавания изображений и могут быть применены к широкому спектру других задач, в частности, к обработке аудиофайлов [4, 5], что позволит проводить кластеризацию и классификацию исследуемых акустических сигналов, может быть использовано для анализа зарегистрированных аудиосигналов в задачах акустического контроля различных объектов.

Сопоставление аудиофрагментов

Задача сопоставления аудиофрагментов между собой может быть сведена к задаче верификации дикторов, в основе которой используется система верификации объектов (рис. 1.).

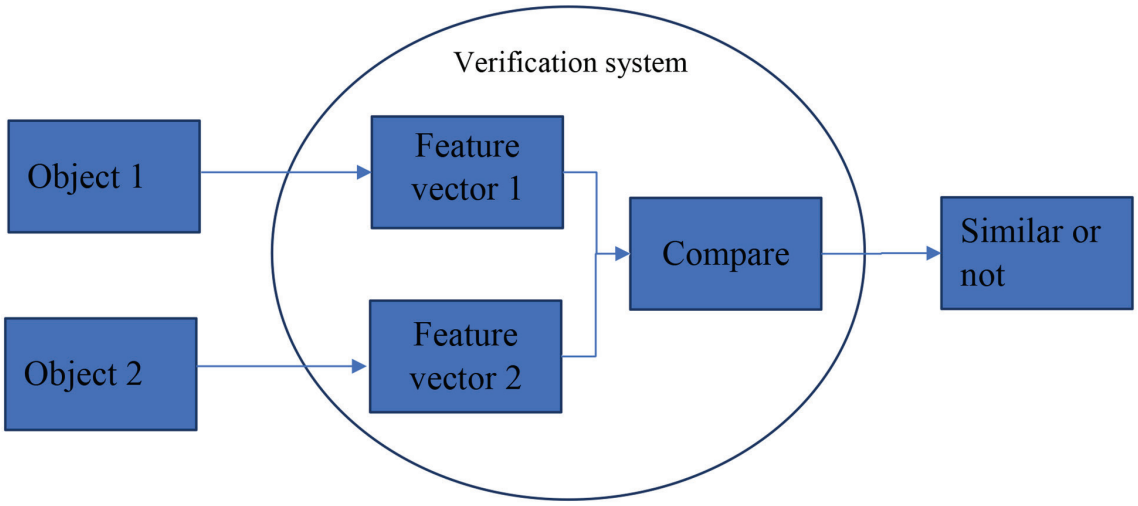

Рис. 1. Система верификации объектов

Системе верификации на вход подаются два объекта (Object 1, Object 2), представленные аудиофрагментами. При получении объектов система извлекает из каждого некоторое его представление. В качестве представления взят вектор признаков (Feature vector 1, Feature vector 2) с выхода предсказательной модели, в роли которой выступает глубокая сверточная нейронная сеть (CNN) [6]. Далее на выходе система верификации дает положительный или отрицательный ответ на вопрос, можно считать два объекта в достаточной степени схожими или нет.

При решении задачи верификации дикторов каждая пара объектов, подаваемая в систему верификации, имеет истинную метку, указывающую на то, принадлежит речь в двух аудиофрагментах одному человеку или двум разным. В данной работе истинная метка пары аудиофрагментов будет указывать на то, является один из аудиофрагментов измененной версией другого или это аудиофрагменты из двух разных записей.

Предсказательная модель изначально обучается на задаче классификации объектов. После этого последние несколько слоев сети удаляются, после чего на выходе оставшегося последнего слоя извлекается вектор признаков.

В качестве набора данных для обучения нейросети требуется большая репрезентативная выборка аудиозаписей, содержащих различные амплитудно-частотные характеристики. Для выполнения представленной работы был собран необходимый набор данных [7] и дополнен набором данных FMA [8], который также удовлетворяет требованиям. В качестве разметки для классификации использовались идентификаторы аудиозаписей, т.е. каждая аудиозапись являлась представителем своего уникального класса.

Для того чтобы предотвратить переобучение предсказательной модели, применялись следующие аугментации: гауссов шум со случайной амплитудой в диапазоне [0.001; 0.015]; реверберация со случайными параметрами реверберации, коэффициента демпинга и размера комнаты в диапазоне [0; 20]; в качестве общей меры предотвращения переобучения и повышения устойчивости, а также ключевой идеей было добавление изменения частоты основного тона (f, Гц) в диапазоне $\left[\frac{f}{4*^{12}\sqrt{2}};\ 4f^{12}\sqrt{2}\right]$ и изменения темпа (beats per minute) в k раз, где k выбирается в диапазоне [0.8; 1.2]. Благодаря этим двум типам аугментации предсказательная модель обучалась так, чтобы быть устойчивой к такого рода видоизменениям и, следовательно, определять такие аудиозаписи как одинаковые.

Предсказательная модель

В качестве предсказательной модели была взята архитектура CQT-Net [9]. На вход модели подается CQT отображение из частотной энергии в заранее заданные частотные области (в музыке — ноты). CQT спектр извлекался из аудиофрагмента с помощью функции из программного пакета для обработки аудиофайлов librosa [10] с окном Ханна длинной 512 отсчетов. Количество признаков (бинов) на одну октаву было установлено 12. Учитывая, что всего октав 7, вектор признаков, представляющий одно окно, будет иметь длину 84 чисел. На вход модели подавался тензор размером [bs, 1, 84, T], где bs — размер группы примеров (batch size), 84 — количество признаков в одном окне Ханна, T — количество окон в одном аудиофрагменте, которое зависит от длины аудиофрагмента. Например, для аудиофрагмента длиной 20 секунд T равно 862.

Нейронная сеть состоит из сверточных слоев, а также Max Pooling [11] слоев. Высота ядер сверток на первых слоях сети равняется 12 или 13, как числу признаков (бинов) на октаву в CQT-спектре. Это позволяет нейронам на слое Conv2 иметь активационное поле высотой 36, что соответствует 36 полутонам. Также стоит обратить внимание на то, что сверточные слои обладают расширением для того, чтобы модель могла обращать внимание на длительные «мелодии» (последовательности частот) в аудиофрагменте. Что еще более важно, модель не включает в себя никаких операций объединения субдискретизации в частотном измерении. Другими словами, вертикальное смещение во всех слоях равно 1. В связи с этим модель работает с картами признаков большего разрешения, что положительно влияет на качество.

Полная архитектура сети представлена на рисунке 2.

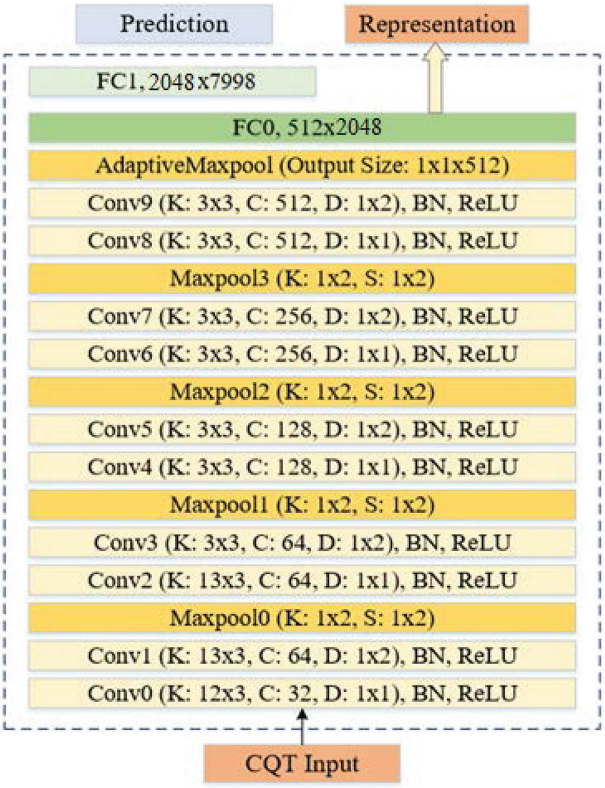

Рис. 2. Архитектура полученной сверточной нейронной сети

После обучения сети на задаче классификации векторы признаков извлекались со слоя FC0. Стоит заметить, что предсказательная модель во время обучения принимала на вход аудиофрагменты фиксированной длины. Это необходимо для упаковки аудиофрагментов в группы (батчи) данных.

Для того чтобы на вход полносвязного слоя FC0 всегда поступал вектор длиной 512 чисел, перед слоем FC0 используется слой AdaptiveMaxPool, принцип работы которого схож со слоем MaxPool, за исключением того, что ядро и смещение выбирается динамически, чтобы итоговая карта признаков была указанного заранее размера. На выходе слоя FC0 всегда будет вектор признаков длиной 2048 чисел. Единый размер вектора признаков необходим для того, чтобы можно было сравнивать фрагменты разных длин с помощью одного модуля сравнения, структуру которого рассмотрим ниже.

Сравнение представлений

После обучения сети вектор признаков, извлекаемый с выхода слоя FC0, будет являться представлением аудиофрагмента. Для измерения сходства двух представлений используется косинусное сходство векторов признаков. Для принятия решения выбирается порог (число) в диапазоне [-1; 1]. Если косинусное сходство двух векторов больше порога, то объекты считаются одинаковыми, иначе объекты считаются разными.

Детали эксперимента

Сеть обучалась на видеокарте NVIDIA Tesla V100 в облачном сервисе DataSphere. Одна эпоха обучения занимала в среднем 72 минуты, всего модель обучалась 30 эпох. Размер группы данных (батча) — 64 аудиофрагмента. В качестве функции ошибки использовалась кросс-энтропия. В качестве оптимизатора использовался стохастический градиентный спуск с импульсом. Начальный шаг обучения был равен 0.001, а коэффициент сохранения импульса 0.9. В качестве стратегии уменьшения шага обучения применялся метод ReduceLROnPlateau, который уменьшает шаг обучения в 0.3 раза, если среднее значение функции потерь не менялось значительно 3 эпохи подряд.

Тестирование модели

Для тестирования качества верификации был составлен тестовый набор данных, состоящий из 1500 аудиозаписей, которые не использовались во время обучения. Из этих аудиозаписей было составлено 3000 пар аудиофрагментов. 1500 «положительных» и 1500 «отрицательных». «Положительной» парой является пара фрагментов одной аудиозаписи, где к одному из фрагментов были применены аугментации. «Отрицательной» парой является пара, в которой аудиофрагменты были взяты из двух разных аудиозаписей.

Таких тестовых выборок было составлено две. В одной из них применялось только изменение темпа в 0.85–1.35 раза, во второй применялось только изменение тональности в диапазоне от -4 до +4 полутонов.

Det-кривые для каждого из наборов данных представлены на рисунках 3 и 4.

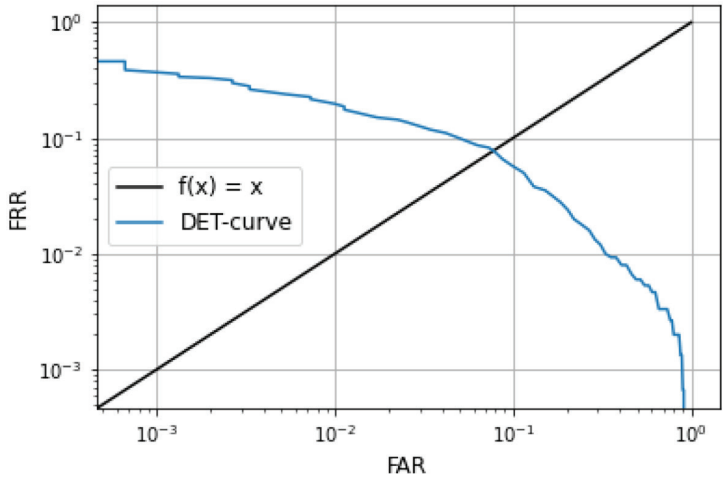

Рис. 3. Det-кривая для тестового набора данных с измененным темпом

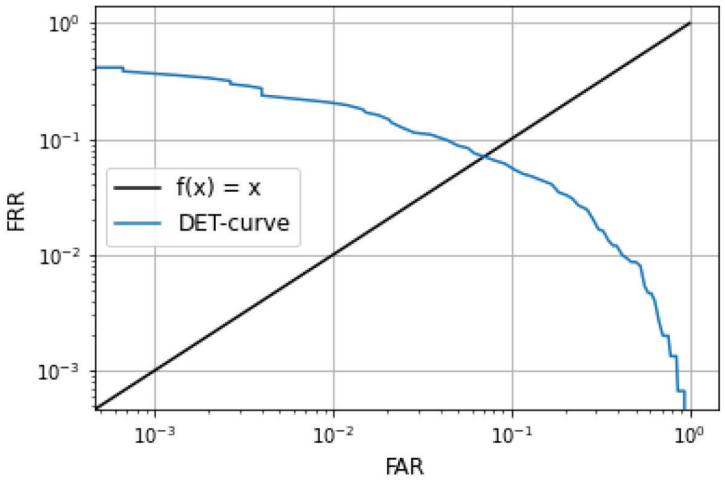

Рис. 4. Det-кривая для тестового набора данных с измененным питчем

В роли метрики качества модели используется коэффициент EER (Equal Error Rate) — процент ошибок при таком пороге, при котором FAR и FRR равны. EER в тестовых наборах данных равны соответственно:

- EER_tempo=7.77 %;
- EER_tone=7.07 %.

Интерпретировать значение коэффициента EER можно следующим образом: в среднем модель будет ошибаться не более чем в 7.77 % случаев, что говорит о надежности полученной системы.

Заключение

На данном этапе исследования показана эффективность применения современных технологий автоматического сопоставления аудиофрагментов в задачах выявления внесенных изменений в оригинальный аудиосигнал. В ходе работы была получена обученная сверточная нейронная сеть, которая с высокой долей вероятности (не менее 92 %) способна выявлять наиболее часто применяемые искажения аудиозаписей в амплитудно-частотной области.

Библиографический список

1. Бринк Х., Ричардс Дж., Феверолф М. Машинное обучение. СПб., 2017.

2. Furui, S., Rosenberg, A.E. Speaker Verification. Digital Signal Processing Handbook. CRC Press LLC, 1999.

3. Bimbot F. et al. A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Advances in Signal Processing. 2004. № 4.

4. Kim J.W. Salamon J. Li P. Crepe: A Convolutional Representation for Pitch Estimation. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2018. URL: https://arxiv.org/pdf/1802.06182.pdf (дата обращения: 13.12.2021).

5. Böck S. Krebs F., Widmer G. Accurate Tempo Estimation Based on Recurrent Neural Networks and Resonating Comb Filters ISMIR. 2015. URL: http://www.cp.jku.at/research/papers/Boeck_etal_ISMIR. (дата обращения: 13.12.2021).

6. Li Z., Yang W., Peng Sh., Liu F. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. Hohai University, Nanjing, China. 2020. URL: https://arxiv.org/ftp/arxiv/papers/2004/2004.02806.pdf (дата обращения: 13.12.2021).

7. Попов В.Н., Ладыгин П.С., Борцова Я.И., Карев В.В. Подготовка набора данных для обучения нейронной сети, используемой в задачах сравнения аудиофайлов // Проблемы правовой и технической защиты информации. Барнаул, 2021. Вып. IX.

8. Defferrard M., Benzi K., Vandergheynst P., Bresson X. FMA: A Dataset For Music Analysis. 18th International Society for Music Information Retrieval Conference, Suzhou, China. 2017. URL: https://arxiv.org/pdf/1612.01840.pdf (дата обращения: 13.12.2021).

9. Yu Zh., Xu X., Chen X., Yang D. Learning a Representation for Cover Song Identification Using Convolutional Neural Network. 2019. URL: http:// https://arxiv.org/abs/1911.00334 (дата обращения: 13.12.2021).

10. McFee B., Raffel C., Liang D., PW Ellis D., McVicar M., Battenberg E., and Nieto O. Librosa: Audio and music signal analysis in python. Proc. of the 14th python in science conf. 2015. URL: http://conference.scipy.org/proceedings/scipy2015/pdf. (дата обращения: 13.12.2021).

11. Goodfellow I., Bengio Y., Courville A. Deep learning: The MIT Press, 2016.