УДК 004.23

И. А. Веялкин, Д. В. Шиман
Белорусский государственный технологический университет

ОБЗОР СТРУКТУР АЛГОРИТМОВ ПОИСКА НА ОСНОВЕ ОТПЕЧАТКОВ АУДИОДАННЫХ

Выполнен обзор алгоритмов аудиопоиска и этапов реализации таких алгоритмов. Выявлены ключевые особенности существующих методик аудиопоиска и указаны подходы к решению задачи распознавания образа в контексте аудиоданных. Проведен анализ параметров алгоритмов аудиопоиска, характеризующих область эффективного применения (контекст задачи). В данной работе отмечены критерии устойчивости конкретного алгоритма или включенных в него этапов к искажениям. Представлена информация об эффективности поиска и ее зависимость от объема данных, а также экстремумы. Оценена скорость расчета отдельных параметров, в частности метрик сравнения, необходимых для оценки подобия искомого фрагмента и конкретного вхождения во множество поиска. Предложены методики определения степени достоверности результатов поиска, основанные на коэффициенте ложных срабатываний и коэффициенте ложных отказов.

Для проведения анализа за основу брались применяемые на практике в коммерческих продуктах (таких как Shazam, MusicBrainz) алгоритмы. В статье описаны современные тенденции в развитии существующих подходов, основанные на методах анализа данных, в том числе с применением нейронных сетей, а также алгоритмов распознавания изображения. Последние, в свою очередь, требуют преобразования способа представления данных. Двоичное представление оцифрованного входного аудиосигнала можно интерпретировать как набор аудиопризнаков, либо изображений с набором визуальных признаков.

Ключевые слова: отпечаток, аудиопоиск, частота, спектр, нечеткий поиск, признак, распознавание образа.

I. A. Veyalkin, D. V. Shiman
Belarusian State Technological University

REVIEW OF AN ALGORITHM STRUCTURE OF AUDIO RETRIEVAL WITH FINGERPRINTING

The paper contains a review of numerous audio retrieval algorithms and particular retrieval stages. The most valuable features of existing audio retrieval approaches are shown. The article describes a solution implementation for a pattern recognition problem in terms of audio data. In this paper we prepared the analysis of different audio retrieval algorithms key parameters and showed the most effective applications for a particular implementation. The paper contains the analysis of numerous characteristics like distortion invariance, robustness, scalability, critical points, computation performance (i.e. distance metrics), reliability, false positive rate, false negative rate, etc. of reviewed audio retrieval approaches.

The analysis is based on the commercial implementations of reviewed algorithms (such as Shazam, MusicBrainz) and takes into account modern development trends and latest researches in audio retrieval and pattern matching problem. The paper refers to the conversions required for particular approaches (such as raw binary data, image data, etc.). We point out how data analysis approaches, image pattern recognition algorithms, artificial neural networks are applicable to build a robust audio retrieval with a fingerprinting system for a particular application.

Key words: fingerprint, frequency, spectrum, feature, pattern recognition.

Введение. Отпечаток аудиоданных – это компактная сигнатура на основе данных аудиосигнала, которая обобщает аудиозапись. В процессе получения отпечатков аудиоданных (в системах идентификации аудио на основе содержимого аудиосигнала) извлекают перцепционный дайджест (перцепционный дайджест – это свертка на основе признаков) части аудиосигнала, т. е. отпечаток, и сохраняют его в базе данных.

Идеальная система поиска на основе отпечатков должна быть:

– способна точно идентифицировать элемент независимо от уровня сжатия и искажений или помех в канале передач;

– в состоянии идентифицировать целые аудиозаписи по отрывкам в несколько секунд (в зависимости от применения) – свойство, известное как детализация, или устойчивость к обрезке. Это требует методов для работы со

смещениями, т. е. недостаточной синхронизации между полученным отпечатком и теми, которые хранятся в базе данных;

– способна иметь дело с искажениями различной природы, такими как изменение высоты звука (изменение скорости воспроизведения аудио быстрее или медленнее), частотная коррекция, фоновый шум, цифроаналоговое и аналогово-цифровое преобразование, кодирование речи и аудио (такие как GSM или MP3), и т. д.;

– вычислительно эффективной. Это связано с размером отпечатков, сложностью алгоритма поиска и сложностью извлечения отпечатка.

Когда поступает фрагмент неизвестного аудиосигнала, его отпечаток рассчитывается и сравнивается с теми, что сохранены в базе данных. Используя отпечатки и алгоритмы поиска соответствия, которые характеризуются набором параметров и методами проверки и оценки соответствия, даже существенно искаженные версии записи могут быть идентифицированы и сопоставлены с эталонным аудиофайлом из предварительно составленной базы данных аудиозаписей.

Системы на основе алгоритмов отпечатков аудиоданных относятся к классу задач распознавания образа, включают в себя, в частности, задачу музыкального поиска и применяются для решения различного рода проблем, среди прочих: мониторинг радиовещания в автоматическом режиме с целью контроля прав на ротируемые музыкальные произведения; поиск аудиофайлов по записанному с микрофона фрагменту; поиск дубликатов в музыкальной коллекции.

Проблема автоматической идентификации аудиосигнала происходит из-за большой размерности и значительного варьирования звуковых данных для схожего на слух содержимого [1].

Основная часть. Алгоритмы включают два основных процесса: извлечение отпечатка (см. рисунок); алгоритм сопоставления.

При извлечении отпечатка получается набор соответствующих перцепционных характеристик записи в сжатой и устойчивой форме. Требования к отпечаткам: отличительная мощность среди огромного числа других отпечатков; инвариантность к искажениям; компактность; простота вычисления.

Предлагаемые решения, выполняющие вышеуказанные требования, предполагают компромисс между сокращением размерности и потерей информации. Извлечение отпечатка состоит из внешнего интерфейса и блока формирования отпечатка. Внешний интерфейс формирует набор из замеров параметров сигнала. Блок формирования отпечатка определяет окончательное представление отпечатка.

Обобщенные этапы алгоритмов: первичная обработка (предварительная обработка, разбиение на кадры и выбор перекрытия, преобразования, извлечения признаков, постобработка); формирование отпечатков; поиск; проверка.

Этапы построения аудиоотпечатка

Останавливаясь подробнее, стоит отметить, что в процессе предварительной обработки формат представления записанного сигнала приводят к общему виду (например, PCM 16 бит, один канал, 44,1 кГц), а также проводят дополнительную обработку, выполняя нормализацию амплитуды (привязка динамического диапазона от –1 до 1), предварительную коррекцию или предыскажение (выделение или усиление определенных частот), GSM-кодирование или декодирование (в случае передачи сигнала по мобильной сети).

При разбиении на кадры сигнал делится на кадры с размером, сопоставимым с изменением скорости лежащих в основе акустических событий. Ключевым предположением при измерении характеристик является то, что сигнал можно рассматривать как стационарный в течение интервала в несколько миллисекунд. Вводится параметр «частота кадров в секунду», влияющий на эффективность дальнейшего поиска. Важной особенностью является применение перекрытия на данном этапе. Это необходимо для повышения устойчивости к так называемому «сдвигу», т. е. когда входные данные не выравнены идеально. К каждому кадру применяется коническая оконная функция, чтобы

минимизировать разрывы в начале и в конце. Существует компромисс при выборе вышеуказанных значений между частотой изменений в спектре и сложностью системы.

Идея, лежащая в основе линейных преобразований, состоит в преобразовании набора замеров в новый набор признаков. Если выбраны подходящие преобразования, то это позволяет существенно уменьшить избыточность. Существуют оптимальные преобразования в смысле упаковки информации и свойств декорреляции, такие как Карунена-Лоэва или сингулярного разложения [1]. Однако эти преобразования зависят от проблематики и вычислительно сложны. По этой причине общеупотребимы преобразования низкой сложности, использующие фиксированные базисные векторы. Поэтому большинство методов поиска и идентификации аудио по данным его сигнала применяют стандартные преобразования из временного в частотный интервал для облегчения эффективного сжатия, удаления шума и последующей обработки.

Наиболее распространенным преобразованием является преобразование Фурье:

$$f(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ixw}\, dx. \qquad (1)$$

Обычно это быстрое преобразование Фурье:

$$F(t,\omega) = \int_{-\infty}^{\infty} f(\tau - t) W(\tau - t) e^{-iwt}\, d\tau, \qquad (2)$$

где $W(\tau - t)$ – оконная функция.

В случае дискретного преобразования Фурье:

$$F(m,\omega) = \sum_{n=-\infty}^{\infty} f[n] w(n-m) e^{-j\omega n}. \qquad (3)$$

Также было предложено дискретное косинусное преобразование:

$$F : R^{2N} \to R^{N}, \qquad (4)$$

где $2N$ – вещественные числа, преобразующиеся в соответствии с формулой

$$X_k = \sum_{n=0}^{2N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right]. \qquad (5)$$

Встречаются варианты алгоритмов с преобразованием Хаара или преобразованием Уолша-Адамара [2]. На практике было выявлено, что дискретное преобразование Фурье в общем менее чувствительно к смещению, чем преобразование Уолша-Адамара [3]. Существует МКПП преобразование, демонстрирующее практическую инвариантность к сдвигу [4].

На этапе извлечения признаков в общем случае используют знания об этапах трансдукции слуховой системы человека, чтобы извлечь более значимые для восприятия параметры.

Часто применяемые в алгоритмах характеристики, относящиеся к восприятию: средний коэффициент перехода через нуль (например, классификация шума и музыки, широко распространена в распознавании речи); рассчитанный темп, ритм; среднее значение спектра; равномерность спектра; главные тона; ширина диапазона частот в сигнале.

Так, равномерность спектра определяется как:

$$F_s = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} = \frac{\exp\left(\frac{1}{N}\sum_{n=0}^{N-1} \ln x(n)\right)}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)}. \qquad (6)$$

Многие системы извлекают несколько признаков, выполняя анализ критической полосы частот спектра, формируя векторы признаков. В некоторых работах используются коэффициенты кепстра в мел масштабе [5], когда как кепстр выражается следующим образом:

$$C_s(q) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \ln\left[S(w)\right]^2 e^{i\omega q}\, d\omega. \qquad (7)$$

Ряд исследователей утверждают, что когда происходит искажение звукового канала, спектральные оценки и только относящиеся сюда признаки являются недостаточными. Они предлагают анализ частотной модуляции для характеристики изменяющегося во времени поведения звуковых сигналов. В этом случае признаки соответствуют среднему геометрическому оценки частоты модуляции энергии 19 размещенных по баркам полосовых фильтров.

Стоит отметить, что обычно используемые признаки являются эвристическими и, таким образом, могут не быть оптимальными. По этой причине применяется модифицированное преобразование Карунена-Лоэва, ориентированный метод главных компонент, чтобы найти оптимальные признаки «неконтролируемым» способом.

На заключительном этапе первичного преобразования к модели сигнала добавляют производные по времени более высокого порядка для того, чтобы лучше охарактеризовать временные вариации в сигнале. Некоторые системы применяют только производную от признаков, а не абсолютные признаки. Использование производной от измерений сигнала имеет тенденцию усиливать шум, но в то же время фильтровать искажения. Довольно распространено применять

к признакам квантование очень низкого разрешения (троичное или двоичное). Целью квантования является получение устойчивости к искажениям, нормализация, легкость аппаратных реализаций, снижение требований к памяти и удобство в последующих частях системы.

После получения набора признаков следует этап формирования отпечатка. Блок формирования отпечатков, как правило, принимает последовательность из векторов признаков, рассчитанных на основе отдельных кадров. Использование избыточностей во временных окрестностях кадра, внутри записи и по всей базе данных полезно для дальнейшего уменьшения размера отпечатков. Тип выбранной модели определяет пригодную метрику расстояния, а также конструкцию алгоритмов индексации для быстрого поиска.

Некоторые системы включают в себя значимые атрибуты высокого уровня, такие как ритм или главную высоту.

Так, например, алгоритм получения отпечатка MusicBrainz TRM включает в вектор: среднюю частоту прохождения через нуль, рассчитанный ритм (количество ударов в минуту), среднее значение спектра и некоторые другие признаки для представления части аудио (соответствующего 26 с). Данный подход был спроектирован для таких приложений, как установление связи между mp3-файлами и их метаданными (название, исполнитель и др.), и лучше подстроен для малой сложности, чем для устойчивости.

Некоторые из существующих алгоритмов базируются на так называемых «глобальных избыточностях» в пределах песни. Если мы предположим, что признаки данного аудиообъекта сходны среди них, то компактное представление может быть получено путем кластеризации векторов признаков. Ряд проанализированных реализаций алгоритмов поиска используют модель отпечатков, которая еще глубже применяет глобальную избыточность. Обоснование в большей степени вдохновлено распознаванием речи. В речи алфавит звуковых классов, т. е. звуки речи, может быть использован, чтобы сегментировать коллекцию необработанных речевых данных в текст, достигая большего сокращения избыточности без «большой» потери информации. Аналогичным образом мы можем рассматривать музыкальную совокупность как предложения, построенные объединением звуковых классов конечного алфавита. «Одинаково воспринимаемые» барабанные звуки, например, встречаются в большом количестве поп-песен. Это приближение дает отпечаток, который состоит из последовательности индексов к набору звуковых классов образца коллекции аудиоэлементов. Звуковые классы оцениваются с помощью неконтролируемой кластеризации и моделируются путем скрытых марковских моделей. Статистическое моделирование зависимости сигнала от времени позволяет сократить местную избыточность. Представление отпечатков в виде последовательности индексов к звуковым классам сохраняет информацию о процессе эволюции аудио с течением времени.

Методика поиска и сравнения подразумевает определение метрики расстояния между двумя отпечатками и непосредственно алгоритма поиска и сравнения.

Метрики расстояния очень сильно связаны с типом выбранной модели. При сравнении векторных последовательностей корреляция является распространенным явлением. Используют следующие метрики: эвклидово расстояние или слегка модифицированные версии, которые работают с последовательностями различной длины; классификация по методу ближайшего соседа на основе оценки перекрестной энтропии; манхэттенское расстояние, либо расстояние Хэмминга, когда используется двоичное квантование, применяется в системах, где последовательности векторов признаков квантуются.

Базовый подход в алгоритмах поиска заключается в использовании вычислительно простого расстояния, чтобы быстро сузить пространство поиска и далее задействовать методы на основе индексов как альтернативу полному перебору для досконального сопоставления с применением более вычислительно дорогого расстояния.

Заключительным этапом является проверка, которая направлена на то, чтобы решить, представлены ли аудиообъекты по запросу в хранилище объектов для идентификации или нет. Оценочные результаты получаются в процессе сравнения извлеченного отпечатка с базой данных отпечатков результатов. Для того чтобы принять решение на счет успеха проведенного поиска, оценка должна быть выше определенного порога. Выбор порогового значения зависит от используемой модели отпечатков, характерной информации запроса, схожести отпечатков в базе данных, размера базы данных.

Чем больше база данных, тем выше вероятность ошибочной идентификации, т. е. ложно положительный результат. Частоту ложно положительных срабатываний также называют коэффициентом ложной идентификации, или частотой ложных тревог. Ложно отрицательная частота также фигурирует под названием коэффициент ложного отказа.

В общем случае система оценивается по характеристикам оценок эффективности информационного поиска: точность, полнота.

Заключение. В статье рассматривается общий подход и структура алгоритмов поиска на

основе отпечатков аудиоданных. Описываются проблемы поиска, специфичные контексту рассматриваемой задачи, и возможные комбинации методов для их преодоления.

Изложенные алгоритмы имеют широкое применение в решении различных практических задач: распознавание аудиофайла по фрагменту, записанному на микрофон; распознавание музыкального произведения по напеванию; поиск нечетких дубликатов аудиофайлов; поиск каверверсий и ремиксов музыкальных произведений; выделение мелодии из полифонического сигнала (подзадача предыдущего пункта); классификация музыки; поиск похожих аудиофайлов; поиск музыкальных рекомендаций в контексте музыкальной библиотеки.

Литература

1. Theodoris S., Koutroumbas K. Pattern Recognition. New York: Academic Press, 1999. 189 p.

2. Transformbased indexing of audio data for multimedia databases / S. Subramanya [et al.] // Conf. on Computational Intelligence and Multimedia Applications. New Delhi, 1999. P. 134–138.

3. Short-term sound stream characterisation for reliable, real-time occurrence monitoring of given sound-prints / G. Richly [et al.] // Proc. 10th Mediterranean Electrotechnical Conference. New York, 2000. P. 526–528.

4. Mihak M., Venkatesan R. A perceptual audio hashing algorithm: a tool for robust audio identification and information hiding // 4th Workshop on Information Hiding. Pittsburg, 2001. 414 p.

5. Content-based identification of audio material using mpeg-7 low level description / E. Allamanche [et al.] // International Symposium on Music Information Retrieval. Indiana, 2001. P. 197–204.

References

1. Theodoris S., Koutroumbas K. Pattern Recognition. New York, Academic Press, 1999. 189 p.

2. Subramanya S., Simha R., Narahari B., Youssef A. Transformbased indexing of audio data for multimedia databases. *Conf. on Computational Intelligence and Multimedia Applications*. New Delhi, 1999, pp. 134–138.

3. Richly G., Varga L., Kovacs F., Hosszu G. Short-term sound stream characterisation for reliable, real-time occurrence monitoring of given sound-prints. *Proc. 10th Mediterranean Electrotechnical Conference.* New York, 2000, pp. 526–528.

4. Mihak M., Venkatesan R. A perceptual audio hashing algorithm: a tool for robust audio identification and information hiding. *4th Workshop on Information Hiding*. Pittsburg, 2001. 414 p.

5. Allamanche E., Herre J., Helmuth O., Froba B., Kasten T., Cremer M. Content-based identification of audio material using mpeg-7 low level description. *International Symposium on Music Information Retrieval*. Indiana, 2001, pp. 197–204.

Информация об авторах

Веялкин Игорь Александрович – ассистент кафедры информационных систем и технологий. Белорусский государственный технологический университет (220006, г. Минск, ул. Свердлова, 13а, Республика Беларусь). E-mail: igor.veyalkin@gmail.com

Шиман Дмитрий Васильевич − кандидат технических наук, доцент кафедры информационных систем и технологий. Белорусский государственный технологический университет (220006, г. Минск, ул. Свердлова, 13а, Республика Беларусь). E-mail: d.shiman@belstu.by

Information about the authors

Veyalkin Igor Aleksandrovich − assistant, the Department of Information Systems and Technologies. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: igor.veyalkin@gmail.com

Shiman Dmitriy Vasil'yevich − Ph. D. (Engineering), Assistant Professor, the Department of Information Systems and Technologies. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: d.shiman@belstu.by

Поступила 06.04.2015