Серия «Математика»
2013. Т. 6, № 2. С. 2—17

Онлайн-доступ к журналу:
http://isu.ru/izvestia

ИЗВЕСТИЯ
Иркутского
государственного
университета

УДК 004.93

О задаче распознавания аккордов в цифровых звукозапиях

Н. Ю. Глазырин
Уральский федеральный университет

Аннотация. Описывается метод определения последовательности аккордов, содержащейся в музыкальной звукозаписи. По итогам ежегодного соревнования систем распознавания аккордов MIREX Audio Chord Estimation 2012 качество работы предлагаемого метода сопоставимо с результатами других участников. В отличие от большинства других систем, описанная в данной работе система не использует методы машинного обучения.

Ключевые слова: музыкальный информационный поиск; распознавание аккордов; спектрограмма; constant-Q преобразование.

1. Введение

В цифровом виде звук представляется как последовательность значений давления воздуха, измеренных через очень короткие промежутки времени. Задача распознавания аккордов в музыкальных звукозаписях состоит в получении из этой последовательности последовательности аккордов с указанием позиций начала и конца звучания каждого из них. Обычно это делается путем различных математических преобразований исходной последовательности. Представление звукозаписи в виде последовательности аккордов может являться промежуточным шагом в работе других алгоритмов, а также может представлять ценность само по себе. Так, информацию о последовательности аккордов в композиции можно использовать для определения структуры этой композиции, её разделения на более крупные сегменты. При помощи такого представления можно индексировать музыкальные произведения для поиска композиций по заданной последовательности аккордов, находить разные аранжировки одной и той же композиции. К данному виду информации может быть и чисто прикладной интерес, например, у людей, обучающихся игре на гитаре. Работающая в реальном време-

ни система распознавания аккордов в звуке позволит контролировать процесс обучения и немедленно фиксировать ошибки.

Начиная с 2008 г., в рамках ежегодной кампании по оценке алгоритмов музыкального информационного поиска (Music Information Retrieval Evaluation eXchange – MIREX) проводится соревнование алгоритмов, извлекающих последовательность аккордов в звукозаписи. В 2012 г. в данном соревновании принимали участие 6 команд, представивших в общей сложности 11 алгоритмов. Для оценки качества их работы использовались две музыкальных коллекции, в которых каждая композиция была предварительно размечена вручную. Итоговая оценка от 0 до 1 характеризует в среднем качество распознавания аккордов алгоритмом в рамках данной коллекции. Все показанные результаты находятся в промежутках $[0.7159, 0.8273]$ и $[0.6248, 0.7249]$ для каждой из коллекций соответственно, а алгоритм, представленный в данной работе, показал для этих коллекций результаты 0.7394 и 0.6248.

Статья имеет следующую структуру. В §2 формулируется задача распознавания аккордов в цифровых звукозаписях. §3 посвящен описанию предлагаемого автором метода, а в §4 приведены некоторые детали его реализации. В §5 описана методика оценки, используемая в соревновании MIREX Audio Chord Estimation. В заключительном §6 представлены результаты работы предложенного метода и проведено их сравнение с результатами других участников.

2. Постановка задачи

Воспринимаемый человеком звуковой сигнал можно представить как непрерывную неотрицательную функцию $x(t)$, которая показывает, как изменяется давление воздуха на барабанную перепонку человека в зависимости от времени t. Для любых реальных звуков эта функция отлична от 0 только на промежутке $[T_0, T_1]$. При анализе удобно полагать, что функция $x(t)$ является периодической с периодом $T_1 - T_0$, т. е. $x(t) = x(t + T_1 - T_0)$ для всех t. Любая такая функция может быть представлена в виде ряда Фурье:

$$x(t) = \sum_{k=-\infty}^{\infty} a(k) e^{i\omega k t}.$$

Здесь $\omega = \frac{2\pi}{T_1 - T_0}$, а $e^{i\omega k t} = \cos(\omega k t) + i \sin(\omega k t)$. Значения $|a(k)|$ составляют спектр звукового сигнала $x(t)$.

В цифровом виде функция $x(t)$ может быть представлена при помощи операций дискретизации и квантования. Для этого с некоторой частотой ν раз в секунду измеряется амплитуда функции $x(t)$ (дискретизация), после чего каждое полученное значение $x(t_i)$ заменяется на

$x_Q(t_i)$ – ближайшее из заданного множества X_Q возможных значений амплитуды (квантование). Как правило, это множество содержит 2^8, 2^{16} или 2^{24} элементов, чтобы каждое значение было представлено целым числом байт. Частота ν часто выбирается равной 44100 Гц. При этом ν называют частотой дискретизации, а значения $x_Q(t_i)$ называют отсчетами исходного сигнала $x(t)$. В соответствии с классической теоремой Котельникова, если спектр сигнала $x(t)$ ограничен частотой $\nu/2$ (т. е. $|a(k)| \neq 0$ только для $\omega|k| < \nu/2$), то исходный сигнал может быть восстановлен однозначно и без потерь по измеренным значениям $x(t_i)$. При квантовании эти значения заменяются на $x_Q(t_i)$, поэтому исходный сигнал может быть восстановлен из оцифрованного только с некоторой ошибкой, которая тем меньше, чем больше возможных значений амплитуды использовалось при квантовании. Для большинства звукозаписей эта ошибка незаметна на слух. Отметим также ещё раз, что спектр любых оцифрованных звуковых сигналов ограничен.

Музыкальные звукозаписи обладают рядом специфических свойств. Укажем некоторые из них.

– *Наличие шума.* Наряду с шумом, вносимым на этапе квантования, это может быть шум пластинки или магнитной ленты, шум зала на концертных звукозаписях, другие шумы. При наличии шума некоторые спектральные компоненты $|a(k)|$ заменяются на искаженные $|\tilde{a}(k)|$.

– *Наличие инструментов с неопределенной высотой звучания.* К ним относятся многие ударные инструменты, в звучании которых невозможно выделить конкретную ноту. Спектр таких инструментов характеризуется большим количеством расположенных подряд существенно отличных от нуля значений, слабо отличающихся друг от друга, т. е. существуют такие положительные числа L и δ, что L существенно больше δ и $L < |a(k)| < L + \delta$ для всех k из некоторого промежутка $[k_0, k_1]$.

– *Наличие гармоник у инструментов с определенной высотой звучания.* В звучании таких инструментов можно выделить отдельную ноту. При этом наряду с частотой, соответствующей этой основной ноте, звучат другие частоты. Их звучание менее выражено, но они могут соответствовать частотам других нот. Математически это означает, что если k_0 таково, что $|a(k_0)|$ максимален, то существует по меньшей мере одно значение $k_1 > k_0$ такое, что $|a(k_1)|$ существенно отличен от 0. Соотношения между парами (k_0, k_1) задают тембр музыкального инструмента.

Следующие свойства являются менее общими. В большей степени они характерны для популярной музыки и могут не наблюдаться для других музыкальных стилей, например, для народной музыки.

– *Наличие ритма.* Очень часто ритм задается ударными инструментами без определенной высоты звучания. Моменты начала звучания отдельных нот обычно соответствуют моментам начала метрических долей в музыкальной композиции.

– *Наличие повторяющихся фрагментов*, например, припевов; неоднократное повторение мелодии.
– *Одновременное звучание нескольких инструментов.* Спектр звукового сигнала является суммой спектров отдельных инструментов. Различные инструменты могут воспроизводить звуки в разных полосах частот. При этом одна и та же нота может быть одновременно воспроизведена разными инструментами в разных октавах.

Человек воспринимает звуки, имеющие большую частоту, как более высокие, а звуки, имеющие меньшую частоту, как более низкие. Звуки с частотами f_0 и $2f_0$ воспринимаются как очень похожие и тесно связанные друг с другом. Будем называть тональным классом объединение всех звуков с частотами, кратными f_0, – частоте самого низкого из принадлежащих этому классу звуков.

Все частоты можно разделить на диапазоны, для каждого из которых начальная частота вдвое меньше частоты конца диапазона. Каждый такой диапазон называется октавой. При использовании наиболее употребимого в европейской музыке равномерно темперированного строя в каждой октаве выделяется 12 ступеней, называемых нотами. Ноты, принадлежащие одному тональному классу, имеют одинаковые названия. Таким образом, в равномерно темперированном строе можно выделить 12 тональных классов. Частоту каждой ноты можно вычислить по формуле $f_i = f_0 \cdot 2^{i/12}$, где f_0 – частота настройки. Обычно выбирают $f_0 = 440$ Гц, что соответствует ноте ля 1-й октавы. Расстояние между любыми двумя соседними ступенями равномерно темперированного строя составляет 1 полутон.

Одновременное звучание трёх и более различных нот создаёт аккорд. Наиболее заметная в звучании нота аккорда называется его основной нотой. Часто она является также самой низкой из нот аккорда. Тип аккорда задаётся интервалами между составляющими его нотами. Тип аккорда не меняется при добавлении к нему ноты из того же тонального класса, что и одна из составляющих его нот.

Для определения звучащего аккорда необходимо указать его основную ноту и тип или, что эквивалентно, все входящие в него ноты. При распознавании аккордов в цифровых звукозаписях, как правило, не требуется определять октавы для каждой из нот. Достаточно определить только их тональные классы.

Таким образом, для решения данной задачи необходимо построить последовательность преобразований, на вход которой подается последовательность квантованных значений амплитуды звука $x_Q(t_i)$. Результатом является последовательность, каждый элемент которой содержит основную ноту аккорда, тип аккорда, время начала звучания и время окончания звучания аккорда. При этом обычно полагают, что звукозапись сделана с использованием равномерно темперированного строя, об-

ладает перечисленными выше свойствами музыкальных звукозаписей и предназначена для прослушивания человеком.

3. Построение последовательности преобразований

3.1. Получение спектрограммы

Первым этапом во многих алгоритмах цифровой обработки звука является его представление в виде спектрограммы, которая показывает, как меняется распределение звуковой энергии по частотам со временем. При этом необходимо принять во внимание, что для распознавания аккордов наиболее интересны частоты, соответствующие частотам ступеней равномерно темперированного строя. Кроме того, на этом же шаге можно учесть наличие ритма в анализируемой звукозаписи.

Спектрограмма фактически является матрицей, состоящей из неотрицательных действительных чисел. Удобно представлять её в виде набора столбцов $C = (C_0, C_1, \ldots, C_M)$, каждый из которых представляет собой спектр короткого фрагмента звукозаписи. Позиции начала фрагментов $(t_0, t_1, \ldots, t_M)$ имеет смысл выбирать в точках начала метрических долей, для определения которых существуют специальные методы. Поэтому будем считать, что они заданы извне. Тогда каждый столбец C_i спектрограммы будет соответствовать фрагменту композиции между позициями $t_{i+1} - t_i$ (считаем, что последний фрагмент оканчивается в момент времени t_{M+1}).

Широко используемый при обработке звука алгоритм быстрого преобразования Фурье [6] не позволяет произвольным образом выбирать частоты своих компонент. Кроме того, в нём одно и то же количество отсчетов сигнала используется для получения всех компонент. Однако для вычисления высокочастотных компонент достаточно короткого фрагмента сигнала, в то время как для вычисления низкочастотных компонент нужен более длинный фрагмент. Constant-Q преобразование [2] является аналогом преобразования Фурье, в котором преодолен этот недостаток. Для него размер анализируемого фрагмента звукозаписи $N(k)$ зависит от частоты f_k соответствующей компоненты k. В свою очередь, f_k можно выбрать таким образом, что каждой ноте звукоряда будет соответствовать одинаковое число частотных компонент (одна или более). Пусть b – количество компонент в одной октаве, а f_{min} – частота наименьшей из компонент. Тогда частота k-й компоненты задается формулой $f_k = 2^{k/b} f_{min}$. Точно так же задаются частоты для нот равномерно темперированного строя, поэтому параметр f_{min} напрямую связан с частотой настройки музыкальных инструментов. Отношение $\frac{f_k}{f_{k+1} - fk} = \frac{1}{2^{1/b} - 1} = Q$ называется коэффициентом качества. При таком выборе частот Q не зависит от k. Отсюда происходит название

constant-Q преобразования. Оно задается формулой

$$X[k] = \frac{1}{N(k)} \sum_{n=0}^{N(k)-1} W[k,n] x_Q(t_n) e^{-\frac{i2\pi kn}{N(k)}}, \quad k = 0, 1, \ldots, K.$$

Здесь $N(k) = \nu/(f_{k+1} - f_k) = (\nu/f_k)Q$ – размер анализируемого фрагмента звукозаписи в отсчётах, $W[k,n]$ – так называемая оконная функция, отличная от нуля на некотором конечном промежутке, не превосходящем этого фрагмента. Выбор оконной функции влияет на форму искажений спектра, возникающих при переходе от анализа всей звукозаписи целиком к анализу короткого фрагмента.

Частота f_{min} наименьшей из компонент и количество K вычисляемых компонент constant-Q преобразования напрямую ограничивают диапазон частот звукозаписи, используемый для распознавания аккордов. Каждый столбец C_i спектрограммы является K-мерным вектором.

Важно учитывать (см. [10]), что музыкальные инструменты, участвующие в композиции, могут быть настроены на базовую частоту, отличную от 440 Гц. Поэтому предварительное определение частоты настройки позволяет с большей точностью выделить на спектрограмме информацию о звучании конкретной ноты для данной композиции.

Один из простых алгоритмов для определения частоты настройки был предложен в работе [15]. Звукозапись делится на короткие фрагменты между моментами времени t_i, $i = 0, 1, \ldots, M_1$, на каждом из которых выполняется constant-Q преобразование с f_{min}, совпадающей с частотой одной из нот равномерно темперированного строя, и достаточно высоким разрешением по частоте: $b = 12b_0$ компонент на октаву. Здесь количество фрагментов M_1 не обязано совпадать с M. На каждом фрагменте $C_i = C(t_i)$ определяется номер компоненты $C_i[j]$, $0 \le j < K$, которой соответствует максимальное значение спектра. Затем строится гистограмма значений функции $C_i[j]$, она состоит из K столбцов. Значения всех столбцов, номера которых сравнимы по модулю b_0, суммируются. В полученной гистограмме из b_0 столбцов номер столбца с наибольшим значением можно интерпретировать как наиболее вероятное отклонение от стандартной частоты настройки 440 Гц в диапазоне от -1/2 до +1/2 полутона с точностью до $1/b_0$ полутона. Если наибольшее значение приходится на 0-й столбец, то отклонения нет. Полученное значение частоты настройки используется для задания частоты наименьшей из компонент f_{min} основного constant-Q преобразования.

3.2. Преобразования спектрограммы

На этом этапе делаются более специфические преобразования, направленные на подавление спектра инструментов с неопределенной высотой звучания. Как было отмечено выше, их спектр характеризуется

большим количеством расположенных подряд значений, существенно отличных от нуля. Также здесь делается попытка учесть наличие повторов в музыкальной композиции. Один и тот же аккорд с большой вероятностью может встретиться несколько раз в разных метрических долях. А значит, ему могут соответствовать разные столбцы спектрограммы.

3.2.1. *Подавление инструментов с неопределенной частотой звучания*

Описанный ниже процесс был предложен в [12] и успешно применен в алгоритме распознавания аккордов [4].

Чтобы соотношения между компонентами спектрограммы лучше соответствовали человеческому восприятию интенсивности звука, каждая компонента $C_i[j]$ заменяется на $\log_{10}(1000C_i[j]+1)$. Затем к каждому столбцу $C_i = (C_i[0], \ldots, C_i[K-1])$ применяется дискретное косинусное преобразование:

$$DC_i[k] = \sum_{j=0}^{K-1} C_i[j] \cos\left[\frac{\pi}{K}\left(j+\frac{1}{2}\right)k\right], \quad k = 0, \ldots, K-1.$$

В полученном векторе обнуляются компоненты $DC_i[0], \ldots, DC_i[\xi-1]$ (где ξ – параметр), после чего к нему применяется обратное дискретное косинусное преобразование. Таким образом, если рассматривать столбец C_i спектрограммы как сигнал (количество энергии, приходящейся на данную частоту, в зависимости от частоты), данный процесс удаляет его низкочастотные компоненты. Это позволяет убрать спектр инструментов с неопределенной частотой звучания, который характеризуется большим количеством расположенных подряд отличных от нуля значений, слабо отличающихся друг от друга.

3.2.2. *Применение самоподобия*

На данном шаге делается попытка скомбинировать информацию о разных случаях появления одного и того же аккорда в звукозаписи. Для столбцов $\{C_i\}_{i=0}^{M}$ полученной спектрограммы строится матрица самоподобия s_{ij}, элементами которой являются евклидовы расстояния между всеми парами (C_i, C_j). В работах [11] и [4] аналогичная матрица строится для векторов хроматических признаков, но не для столбцов спектрограммы. Кроме того, в них рассматриваются только диагонали этой матрицы, параллельные главной диагонали.

Матрица s_{ij} имеет нули на главной диагонали. Она нормализуется таким образом, чтобы $0 \le s_{ij} \le 1$ для всех i, j. Затем в каждой строке сохраняются PM наименьших значений ($0 \le P \le 1$), а все остальные заменяются на 1.

При помощи этой матрицы корректируются все столбцы C_i:

$$\widehat{C}_i = \frac{\sum\limits_{j=0}^{M}(1 - s_{ij}^3)C_j}{\sum\limits_{j=0}^{M}(1 - s_{ij}^3)}.$$

Значения s_{ij} возводятся в куб, чтобы ослабить влияние менее похожих на данный столбцов спектра.

3.2.3. *Применение тональных классов*

По условиям задачи не требуется указывать октаву для каждой из составляющих звучащий аккорд нот. Кроме того, как было отмечено, одна и та же нота может одновременно звучать в разных октавах. Поэтому имеет смысл скомбинировать значения спектра, принадлежащие одному тональному классу, но находящиеся в разных октавах. Сделать это можно простым их суммированием по каждому из столбцов. Ко всем значениям $\widehat{C}_i[j]$, $0 \le j < b$, прибавляются значения $\widehat{C}_i[j+b]$, $\widehat{C}_i[j+2b]$, $\widehat{C}_i[j+3b]$, ... для каждого $i = 0, \ldots, M$, что дает в результате последовательность b-мерных векторов $\{B_i\}_{i=0}^{M}$.

В равномерно темперированном строе можно выделить 12 тональных классов. Поэтому удобно выбирать b равным или кратным 12. В случае $b = 12b_0$ каждый вектор B_i преобразуется в 12-мерный вектор D_i:

$$D_i[j] = \sum_{h=-\lfloor b_0/2 \rfloor}^{\lfloor b_0/2 \rfloor} B_i[b_0 j - h] d^{|h|}, \quad i = 0, \ldots, M, \quad j = 0, \ldots, 11.$$

Параметр d регулирует вклад спектральных компонент, частоты которых не соответствуют в точности частотам нот. Для вычисления $D_i[0]$ вместо $B_i[-h]$ используются $B_i[60-h]$. В полученных векторах $\{D_i\}_{i=0}^{M}$ каждая компонента соответствует одному тональному классу. Векторы с таким свойством называют векторами хроматических признаков или, для краткости, *хроматическими векторами*.

3.3. Последовательность аккордов

Простейшим методом для перехода от последовательности 12-мерных векторов $\{D_i\}_{i=0}^{M}$ к последовательности аккордов является сопоставление с векторами-шаблонами. Каждая компонента вектора D_i соответствует одному тональному классу. Каждый аккорд в рамках задачи можно рассматривать как набор тональных классов. Естественным образом можно определить так называемые бинарные шаблоны, в которых на позициях, соответствующих входящим в аккорд тональным классам, стоят 1, а на остальных позициях стоят 0. Встречаются

шаблоны с другими соотношениями компонент. Удобно масштабировать шаблоны таким образом, чтобы сумма всех компонент или сумма квадратов всех компонент равнялась 1.

Для каждого аккорда, который может быть распознан в звукозаписи, необходимо задать шаблонный вектор. Результатом сопоставления является аккорд, шаблон которого является наиболее похожим на данный вектор D_i. В качестве меры схожести здесь может выступать косинусное расстояние, евклидово расстояние, расстояние Кульбака – Лейблера и др. Сравнение различных мер схожести, а также различных шаблонов аккордов можно найти в [14].

Однако простое пошаговое сравнение хроматических векторов с шаблонами не дает возможности учитывать тональность музыкальной композиции, а также характерные для данного стиля музыки последовательности аккордов. В [1] используется сложная модель тональной гармонии, построенная на основе джазовых композиций. Такая модель позволяет вовлечь теорию музыки в процесс распознавания аккордов. Но при этом она наилучшим образом будет работать при анализе композиций тех жанров, на основе которых она была построена. Так, у авторов упомянутой работы не получилось добиться существенного улучшения качества работы алгоритма с применением этой модели к анализу композиций *The Beatles*.

Наиболее популярным способом учесть часто используемые последовательности аккордов является применение скрытых марковских моделей (см. [5]). При этом обычно названия аккордов соответствуют скрытым состояниям, а хроматические векторы – наблюдениям. Матрица переходов между состояниями и матрица вероятностей появления наблюдаемых символов обычно строятся путём обучения на различных композициях. Поэтому выбор обучающих композиций влияет на качество работы и дальнейшую применимость таких моделей. Наиболее вероятная последовательность аккордов, соответствующая данной последовательности хроматических векторов, восстанавливается при помощи алгоритма Витерби.

4. Реализация описанного метода

Описанная в предыдущем параграфе последовательность преобразований позволяет определить последовательность аккордов по заданной последовательности значений звукового давления. Эти преобразования опираются на указанные выше свойства музыкальных звукозаписей и равномерно темперированного строя. Конкретные наилучшие значения параметров описанных преобразований могут быть получены только путем экспериментов на реальных музыкальных звукозаписях.

4.1. Спектрограмма

Для определения частоты настройки музыкальных инструментов звукозапись делится на короткие фрагменты между позициями t_i, $i = 1, \ldots, M_1$, $M_1 \neq M$, длиной 0.5 с каждый. На каждом из них выполняется constant-Q преобразование с разрешением по частоте b_0, равным 10 компонент на ноту или 120 компонент на октаву. При этом преобразование охватывает 4 октавы, начиная с частоты 440 Гц. При таких параметрах преобразование выполняется очень быстро, в пределах 1 с на одну композицию.

Моменты начала метрических долей $(t_0, t_1, \ldots, t_M)$ определяются при помощи программы *BeatRoot*[1]. Затем последовательность этих моментов делается в T раз более частой путем вставки равномерно $T - 1$ промежуточного значения, где T – параметр. Это значительно увеличивает объем вычислений, но дает возможность получить спектрограмму с большим разрешением по времени.

Для каждого из полученных моментов времени выполняется constant-Q преобразование над фрагментом файла с центром в данной точке. Преобразование охватывает 4 октавы, имеет разрешение 60 компонент на октаву. Частота первой компоненты соответствует частоте ноты *до* большой октавы (65.41 Гц при частоте настройки 440 Гц). Таким образом, на охватываемый частотный диапазон от 65.41 Гц до 987.77 Гц приходится 240 компонент преобразования.

Затем к каждой строке спектрограммы применяется скользящий медианный фильтр с размером окна w значений. После этого удаляются столбцы, соответствующие промежуточным значениям, добавленным ранее. Каждый из оставшихся столбцов соответствует одной метрической доле. Важность сглаживания при помощи скользящих фильтров отмечается во многих работах, например, в [9]. Оно делает результат менее зависимым от отдельных значений. При этом обычно сглаживание производится на спектрограммах с относительно низким временным разрешением, что приводит к размыванию границ аккордов. При сглаживании предложенным способом предварительно повышается разрешение по времени, поэтому после удаления дополнительных столбцов спектрограммы она получается существенно менее размытой.

Значение вклада спектральных компонент, частоты которых не соответствуют в точности частотам нот, было выбрано произвольным образом: $d = 0.6$. Оно не оказывало видимого влияния на результат.

[1] См. `http://www.eecs.qmul.ac.uk/~simond/beatroot/`.

4.2. Последовательность аккордов

Используются 12-мерные шаблонные векторы для всех употребляемых в музыке сочетаний из трёх нот: мажорных, минорных, увеличенных и уменьшенных трезвучий. Мажорные и минорные трезвучия применяются значительно чаще, чем увеличенные и уменьшенные. Шаблоны для аккордов, состоящих из 4 и более нот, не используются в рамках данного метода, а значит, результатом распознавания такого аккорда будет одно из трезвучий. Например, шаблон для аккорда до мажор имеет вид (1.3, 0, 0, 0, 1, 0, 0, 1.3, 0, 0, 0, 0). Аналогичным образом строятся шаблоны для остальных мажорных и минорных трезвучий. Для уменьшенных и увеличенных трезвучий используются бинарные шаблоны. В качестве результата выбирается аккорд, расстояние Кульбака–Лейблера от шаблона которого до данного хроматического вектора является наименьшим. Перед вычислением расстояния и шаблон, и вектор D_i масштабируются таким образом, чтобы сумма квадратов их компонент была равна 1. Только фрагменты композиции до первой определенной метрической доли и после последней определенной метрической доли считаются не содержащими аккордов.

В полученной последовательности аккордов выделяются участки, состоящие из аккордов с одной и той же основной нотой, но разных типов (например ре мажор и ре минор подряд). Такие последовательности практически не встречаются в музыке. Для каждого участка соответствующие хроматические векторы суммируются в один вектор, а ближайший к нему аккорд присваивается всему участку.

5. Методика оценки

В рамках соревнования MIREX Audio Chord Estimation в 2012 г. алгоритмы оценивались на двух музыкальных коллекциях. Первая коллекция (MIREX09) состоит из 180 композиций группы *The Beatles* и 38 композиций групп *Queen* и *Zweieck*. Вторая коллекция (McGill) состоит из лучших композиций американских чартов за период с 1958 по 1991 годы [3]. Для каждой композиции вручную была определена соответствующая последовательность аккордов с моментами начала и окончания звучания каждого из них. Подготовка такой разметки даже для одной композиции весьма трудоемка, поэтому общее число композиций относительно невелико. Для большинства из участвовавших алгоритмов требовалось обучение, поэтому каждая коллекция была поделена на обучающую и тестовую выборки. Итоговые результаты подсчитывались только на основе тестовой выборки.

Для каждой композиции вычислялось так называемое относительное перекрытие: отношение совокупной длины участков композиции,

для которых был правильно указан звучащий аккорд, к совокупной длине участков композиции, на которых звучит трезвучие (назовем её эффективной длительностью композиции). Участки, на которых звучали более сложные аккорды, исключались при подсчете. Затем для всей коллекции вычислялись средний коэффициент перекрытия (AOR – Average Overlap Ratio) и взвешенный средний коэффициент перекрытия (WAOR – Weighted Average Overlap Ratio) по формулам:

$$AOR = \frac{1}{C}\sum_{m=1}^{C} r_m, \quad WAOR = \frac{\sum\limits_{m=1}^{C} \ell_m r_m}{\sum\limits_{m=1}^{C} \ell_m}$$

Здесь C – количество композиций в коллекции, r_m – относительное перекрытие для композиции m, ℓ_m – её эффективная длительность.

6. Результаты

Результаты соревнования MIREX Audio Chord Estimation 2012 показаны в таблице 1. Описанный в данной работе алгоритм – NG1. Наилучшие результаты выделены полужирным шрифтом. Во всех остальных алгоритмах, за исключением показавшего похожий результат DMW1 [1], использовались различные варианты скрытых марковских моделей.

Таблица 1

Результаты MIREX ACE 2012

	MRX	MRX	McG	McG
Алгоритм	AOR	WAOR	AOR	WAOR
CCSS1	0.7940	0.7791	0.6736	0.6619
DMW1	0.7368	0.7199	0.6433	0.6249
KO1	0.8285	0.8163	0.7128	0.6980
NG1	0.7603	0.7394	0.6418	0.6248
NMSD1	**0.8351**	**0.8273**	0.7239	0.7140
NMSD2	0.8104	0.8007	0.7302	0.7206
NMSD3	0.8210	0.8121	0.7329	0.7233
NMSD4	0.8272	0.8198	**0.7347**	**0.7249**
PMP1	0.7470	0.7342	0.6695	0.6556
PMP2	0.7367	0.7241	0.6609	0.6478
PMP3	0.7290	0.7159	0.6532	0.6423

Заметная разница в результатах алгоритмов на двух коллекциях объясняется тем, что коллекция MIREX09 используется для оценки

алгоритмов уже несколько лет и не содержит неточностей. Подготовка новых тестовых данных весьма трудоемка, и до недавнего времени это был единственный доступный набор данных, использовавшийся при разработке всех систем распознавания аккордов. Поэтому алгоритмы, использующие машинное обучение, могут быть в той или иной степени переобучены на этом наборе. Коллекция McGill впервые была использована в 2012 году. Она содержит музыкальные записи разных стилей. Кроме того, в ней ещё встречаются неточности в разметке, которые в одинаковой степени влияют на результаты всех систем.

Вне рамок соревнования автором были вычислены те же самые метрики на коллекциях из 180 песен *The Beatles* и на коллекции *RWC Popular Music* [13] из 100 песен. Они приведены в таблице 2. По сравнению с предыдущей версией метода, представленной в [7], удалось добиться существенного улучшения в качестве распознавания.

Таблица 2

Собственные оценки

Collection	AOR	WAOR
2 коллекции вместе	0.7355	0.7287
RWC Popular Music	0.6469	0.6513
The Beatles	0.7847	0.7765
The Beatles (пред. версия метода)	0.7046	0.7125

Предложенный метод зависит от нескольких параметров. В данном параграфе исследуется влияние каждого из них на качество распознавания аккордов. Поскольку пространство возможных значений параметров многомерно, будем рассматривать различные значения каждого из параметров в отдельности при фиксированных остальных.

Таблица 3. Суммарное время обработки двух коллекций

T	Время работы
1	около 4700 с
2	около 5500 с
4	около 7200 с
8	около 10500 с

Таблица 4. Влияние параметра w

T	w	AOR	WAOR
1	1	0.6544	0.6450
1	3	0.7078	0.7027
2	3	0.7171	0.7087
2	5	0.7276	0.7203
4	7	0.7226	0.7172
4	9	0.7339	0.7267
4	11	0.7346	0.7281
4	13	0.7277	0.7209
8	17	0.7342	0.7265
8	19	0.7339	0.7269
8	**21**	**0.7355**	**0.7287**
8	23	0.7333	0.7261

– T, *количество столбцов спектра*, приходящихся на одну метрическую долю. Этот параметр в наибольшей степени влияет на скорость работы алгоритма. В табл. 3 показаны времена обработки обеих коллекций (280 композиций) при разных значениях T. При весьма существенной разнице в скорости работы качество распознавания аккордов различается незначительно, как следует из табл. 4. Отметим, что в указанную длительность работы входит также время, затраченное на определение моментов начала метрических долей библиотекой *Beatroot*.
– w, *размер окна* скользящего медианного фильтра. В табл. 4 показана зависимость качества распознавания аккордов от значения параметра w при фиксированных значениях T. В обоих случаях наилучший результат получается при размере окна между 2 и 3 метрическими долями.
– ξ, *количество первых обнуляемых значений* после дискретного косинусного преобразования при подавлении информации о тембре. Как видно из табл. 5, наилучший результат получается при относительно небольших значениях ξ (10, 15), тогда как в работе [4] использовалось значение $\xi = 25$.
– P, *доля учитываемых значений* в каждой строке матрицы самоподобия. Как показывает табл. 6, этот параметр слабо влияет на результат. Первая строка этой таблицы соответствует отсутствию коррекции спектрограммы с помощью матрицы самоподобия. Таким образом, данная процедура улучшает результат, но разница для значений P в интервале от 0.1 до 0.2 практически незаметна.

Таблица 5. Влияние параметра ξ

ξ	AOR	WAOR
0	0.6653	0.6581
5	0.7310	0.7241
10	0.7329	0.7266
15	**0.7355**	**0.7287**
20	0.7333	0.7261
25	0.7075	0.7006

Таблица 6. Влияние параметра P

P	AOR	WAOR
0.00	0.7126	0.7038
0.05	0.7331	0.7251
0.10	0.7360	0.7276
0.15	0.7355	0.7279
0.20	**0.7355**	**0.7287**
0.25	0.7323	0.7259
0.30	0.7283	0.7162

Предложенный в данной работе метод не использует информацию о тональности. В итоговой последовательности аккордов исправляются только очевидно негармоничные, и потому практически невозможные подпоследовательности. Однако, показанные по итогам MIREX 2012 результаты практически не отличаются от результатов метода [1], в котором используется модель тональной гармонии. Несомненно, такого рода модель должна улучшить качество распознавания аккордов в некоторых случаях. Однако любая дополнительная информация о сочетаемости аккордов должна использоваться аккуратно. Композиторы

могут отступать от общепринятых правил, модель может быть неполна или может не подходить к данному стилю музыки.

По этой же причине алгоритмы, основывающиеся на скрытых марковских моделях и других методах машинного обучения, могут не годиться для анализа некоторых композиций. На текущий момент существует не так много композиций, в которых размечены позиции начала и конца звучания аккордов. Обычно исследователи работают с несколькими сотнями композиций. Очевидно, такого объема данных недостаточно для охвата музыки разных эпох и направлений. А значит, скрытые марковские модели будут в большей степени подходить для анализа музыки, похожей на уже известную. Возможно, данную проблему удастся преодолеть с помощью методов глубокого обучения (например, [8]), которые могут показывать хорошие результаты на абсолютно незнакомых данных.

Список литературы

1. Bas De Haas W. Improving audio chord transcription by exploiting harmonic and metric knowledge / W. Bas De Haas, J. P. Magalhães, F. Wiering // Proceedings of the 13th International Society for Music Information Retrieval Conference. Porto, Portugal, October 8–12, 2012. – Porto, 2012. – P. 295–300.
2. Brown J. C. Calculation of a constant q spectral transform. / J. C. Brown // Journal of the Acoustical Society of America. – 1991. – Vol. 89, N 1. – P. 425–434.
3. Burgoyne J. A. An expert ground truth set for audio chord recognition and music analysis / J. A. Burgoyne, J. Wild, I. Fujinaga // Proceedings of the 12th International Society for Music Information Retrieval Conference. Miami, USA, October 24–28, 2011. – Miami, 2011. – P. 633–638.
4. Cho T. A feature smoothing method for chord recognition using recurrence plots / T. Cho, J. P. Bello // Proc. of the 12th International Society for Music Information Retrieval Conference, Miami, USA, October 24–28, 2011. – Miami, 2011. – P. 651–656.
5. Cho T. Exploring common variations in state of the art chord recognition systems / T. Cho, R. J. Weiss, J. P. Bello // Proceedings of the 7th Sound and Music Computing Conference (SMC 2010). – Barcelona, Spain, 2010.
6. Cooley J. An algorithm for the machine calculation of complex Fourier series / J. Cooley, J. Tukey // Mathematics of Computation. – 1965. – Vol. 19. – P. 297–301.
7. Glazyrin N. Chord recognition using Prewitt filter and self-similarity / N. Glazyrin, A. Klepinin // Proceedings of the 9th Sound and Music Computing Conference. Copenhagen, Denmark, July 11-14, 2012. – Copenhagen, 2012. – P. 480–485.
8. Humphrey E. J. Learning a robust tonnetz-space transform for automatic chord recognition / E. J. Humphrey, T. Cho, and J. P. Bello // Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE. Kyoto, Japan, 2012. – Kyoto, 2012. – P. 453–456.
9. Analyzing chroma feature types for automated chord recognition / N. Jiang, P. Grosche, V. Konz, M. Müller // Proc. of the AES 42nd International Conference: Semantic Audio. Ilmenau, Germany, 2011, July 22–24. – Ilmenau, 2011.

10. Lerch A. On the requirement of automatic tuning frequency estimation / A. Lerch // Proceedings of ISMIR 2006 7th International Conference on Music Information Retrieval. Victoria, Canada, 8-12 October 2006. – Victoria, 2006. – P. 212–215.
11. Mauch M. Using musical structure to enhance automatic chord transcription / M. Mauch, K. C. Noland, S. Dixon // Proc. of the 10th International Conference on Music Information Retrieval. Kobe, Japan, 26–30 October 2009. – Kobe,2009. – P. 231–236.
12. Müller M. Making chroma features more robust to timbre changes / M. Müller, S. Ewert, S. Kreuzer // Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '09, Washington, DC, USA. – IEEE Computer Society, 2009. – P. 1877–1880.
13. RWC Music Database: Popular, Classical, and Jazz Music Databases / T. Nishimura, M. Goto, H. Hashiguchi, R. Oka // Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002). Paris, France, October 13-17, 2002. – Paris, 2002. – P. 287–288.
14. Oudre L. Chord recognition using measures of fit, chord templates and filtering methods / L. Oudre, Y. Grenier, C. Févotte // Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New York, USA, 2009. – N. Y., 2009. – P. 9–12.
15. Zhu Y. Music key detection for musical audio / Y. Zhu, M. S. Kankanhalli, S. Gao // Proceedings of 11th International Multimedia Modelling Conference (MMM'05). Melbourne, Australia, January 12-14, 2005. – Melbourne, 2005. – P. 30–37.

N. Y. Glazyrin
Towards the task of audio chord esitmation

Abstract. This paper describes a method of audio chord recognition. In the annual MIREX Audio Chord Esitmation 2012 contest this method achieved results comparable to other participants' results. It does not employ any machine learning algorithms, unlike majority of other systems.

Keywords: music information retrieval; chord recognition; chord estimation.

Глазырин Николай Юрьевич, аспирант, Институт математики и компьютерных наук, Уральский федеральный университет, 620000, Екатеринбург, ул. Тургенева, 4; тел.: (343)3507579 (nglazyrin@gmail.com)

Glazyrin Nikolay, graduate student, Ural Federal University, 4, Turgeneva St., Ekaterinburg, 620000; Phone: (343)3507579 (nglazyrin@gmail.com)