Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Классификация методов определения ритмического рисунка и темпа цифровой музыкальной записи»

Студент ИУ7-76Б _____ А. А. Петрова
(Группа) (Подпись, дата) (И.О. Фамилия)

Руководитель _____ К. А. Кивва
(Подпись, дата) (И.О. Фамилия)

2022 г.

РЕФЕРАТ

Расчетно-пояснительная записка 32 с., 4 рис., 1 табл., 16 ист., 1 прил.

Рассмотрены понятия темпа и ритма музыки. Проанализирована проблема автоматического определения темпа и ритма. Проведена классификация основных существующих методов их автоматического определения. Определены критерии сравнения рассмотренных методов. Выделены достоинства и недостатки каждого метода, на основании чего проведено их сравнение.

КЛЮЧЕВЫЕ СЛОВА

темп музыки, ритм музыки, bpm, вейвлет, марковская модель, байесовская модель, нейросети.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Автоматическая транскрипция музыки (АТМ) — это процесс преобразования акустического музыкального сигнала в ту или иную форму нотной записи [1]. Данную задачу можно разделить на несколько подзадач, к которым в том числе относятся задачи выделения информации о ритме и темпе музыки. Несмотря на то, что задачу АТМ для монофонических сигналов можно считать решенной [1], проблема создания автоматизированной системы, способной транскрибировать полифоническую (многоголосую) музыку без ограничений по степени полифонии или типу инструмента, остается открытой.

Цель данной работы – изучить основные существующие методы определения ритмического рисунка и темпа цифровой музыкальной записи.

Чтобы достигнуть поставленной цели, требуется решить следующие задачи:

— провести анализ предметной области и сформулировать проблему;
— сформулировать критерии сравнения методов выделения информации о ритме и темпе музыки;
— классифицировать основные существующие методы.

1 Аналитическая часть

1.1 Темп, ритм и метр

Темп – мера времени в музыке, упрощенно – «скорость исполнения музыки» [2].

Существует несколько способов измерения темпа. В классической музыке чаще всего используется словесное описание (как правило, на итальянском). Этот метод является неточным и дает лишь примерное представление о «скорости» исполнения музыкального произведения. Примеры такого описания: адажио, ленто (медленные темпы); анданте, модерато (средние темпы); аллегро, виво (быстрые темпы).

Второй, более точный способ измерения темпа – это число ударов в минуту (beats per minute, сокращенно bpm). Данный метод напрямую связан с частотой колебания маятника в метрономе (устройстве, предназначенном для точного ориентира темпа при исполнении музыки). Стандартным темпом считается 120 bpm, т. е. 2 Гц.

В данной работе будет использоваться второй способ измерения темпа (в bpm).

Ритм – организация музыки во времени [3]. Ритмическую структуру музыки образует последовательность длительностей – звуков и пауз.

Ритм в музыке принадлежит к числу терминов, дискуссии о которых ведутся в науке последние два столетия. Единого мнения по вопросу его определения нет. Чаще всего ритм определяется как регулярная, периодическая последовательность акцентов. Такое понимание ритма фактически идентично метру.

Метр в музыке – это чередование сильных и слабых долей в определенном темпе [2]. Обычно метр фиксируется с помощью тактового размера и тактовой черты (рис. 1.1).

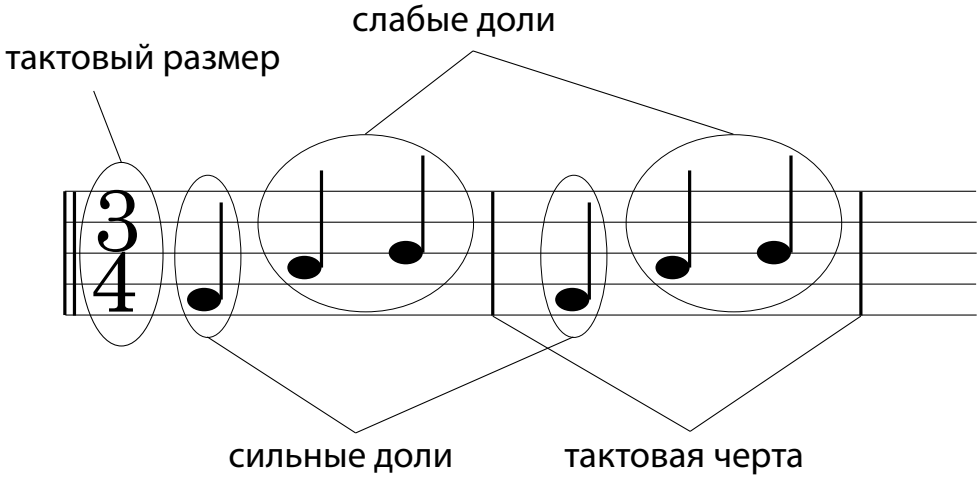


Рис. 1.1 – Обозначение метра

Размер задает относительную длительность каждой доли. Например, размер «3/4» говорит о том, что в такте 3 доли, каждая из которых представлена четвертной нотой. Можно сказать, что размер – числовое представление метра с указанием длительности каждой доли. Такт в свою очередь – единица метра, начинающаяся с наиболее сильной доли и заканчивающаяся перед следующей равной ей по силе (рис. 1.1).

В данной работе не будут учитываться тонкости различия ритма и метра. Соответственно, для измерения ритма будет использоваться числовое представление метра в виде тактового размера.

1.2 Проблема определения ритма и темпа

Основной проблемой автоматического определения ритма и темпа музыки является наличие некоторых особенностей в музыкальных записях с живыми инструментами, затрудняющих это определение. Одна из таких особенностей – это нечеткое попадание инструмента в ритмическую сетку. Такие небольшие отклонения на живых записях присутствуют всегда [4]. Они не заметны для уха человека, но могут осложнять автоматическое распознавание.

Также в некоторых случаях темп и ритм может изменяться в течение музыкального произведения. Пример переменного темпа приведен на рис. 1.2 (темп

обозначается числами сверху в bpm). На рис. 1.3 приведен пример переменного ритма (размера).

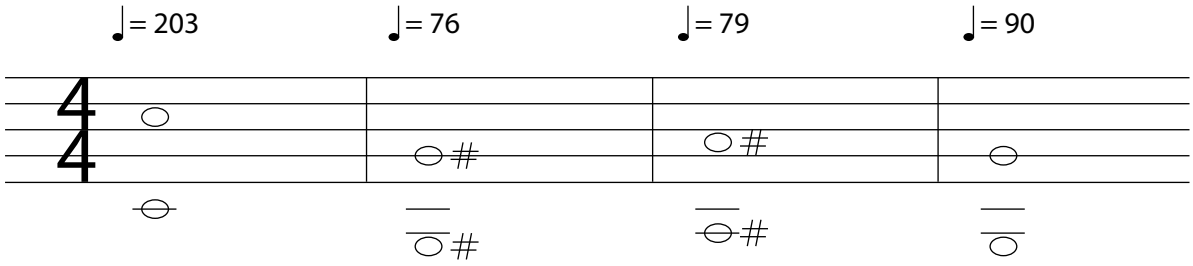


Рис. 1.2 – Пример переменного темпа (System of a down «Aerials»)

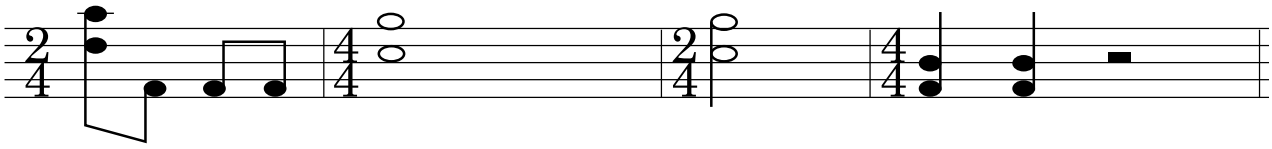


Рис. 1.3 – Пример переменного размера (Metallica «Master of puppets»)

В качестве критериев сравнения рассматриваемых далее методов выделены следующие:

— точность результатов применения метода;
— возможность определения переменного темпа и ритма;
— ограничения на формат входного аудиофайла;
— размер использовавшегося для обучения датасета (если обучение необходимо).

1.3 Дискретное вейвлет-преобразование

1.3.1 Общие сведения

Так как преобразование Фурье не позволяет получить частотно-временное представление сигнала, оно подходит только для стационарных сигналов (т. е. сигналов, частотное наполнение которых не меняется во времени). Большинство же реальных аудио-сигналов являются нестационарными. Основная же проблема оконного преобразования Фурье (ОПФ) заключается в невозможности получить произвольно точное частотно-временное представление сигнала,

то есть нельзя определить для какого-то момента времени, какие спектральные компоненты присутствуют в сигнале. Эта проблема называется проблемой разрешения.

В качестве альтернативы ОПФ было разработано вейвлет-преобразование.

Основная идея вейвлет-преобразования – это разделение сигнала на высокие и низкие частоты с помощью фильтров [5]. После применения фильтров полученные низкие частоты снова пропускаются через два фильтра и т. д. При этом высокие частоты остаются неизменными. Эта операция называется декомпозицией.

На высоких частотах лучше разрешение по времени, а на низких – по частоте.

Фильтры для высоких и низких частот определяются следующими уравнениями [6]:

$$y_{high}[k] = \sum_{n=-\infty}^{\infty} x[n]g[2k-n], \tag{1}$$

$$y_{low}[k] = \sum_{n=-\infty}^{\infty} x[n]h[2k-n], \tag{2}$$

где $x[n]$ – пропускаемый через фильтр сигнал (последовательность), $h[n]$ и $g[n]$ – импульсные характеристики (отклик на единичный импульс) низкочастотного и высокочастотного фильтров соответственно, k и n – целые числа, соответствующие отсчетам (теорема отсчетов [7]).

В целом, процедура пропускания сигнала через фильтр соответствует математической операции свертки сигнала $x[k]$ и импульсной характеристики фильтра $h[k]$, которая определяется как:

$$x[k] * h[k] = \sum_{n=-\infty}^{\infty} x[n]h[k-n]. \tag{3}$$

Выражение $2k-n$ в формулах 1 и 2 позволяет обрезать сигнал, тем самым увеличив его масштаб в два раза (т. к. половина частот удаляется в результате

фильтрации) [5].

Само ДВП (дискретное вейвлет-преобразование) описывается формулой:

$$W(j,k) = \sum_j \sum_k x(k) 2^{-j/2} \psi(2^{-j} n - k), \tag{4}$$

где $\psi(t)$ – функция преобразования, называемая материнским вейвлетом, j и k связаны с параметрами сдвига τ (местоположение окна) и масштаба s (величина, обратная частоте). $s = s_0^j$, $\tau = k s_0^j \tau_0$. В данном случае $s_0 = 2$, $\tau_0 = 1$.

1.3.2 Определение ритма и темпа

Алгоритм определения ритма с помощью ДВП основан на обнаружении наиболее заметных периодов сигнала.

Сигнал сначала раскладывается на несколько частотных полос с помощью ДВП. Для этого сигнал «делится» пополам на высокие и низкие частоты, после чего низкие частоты снова разделяются пополам и т. д. Так продолжается до тех пор, пока не останутся два отсчета. Эта операция необходима, т. к. для высоких частот можно точнее указать их временную позицию, а для низких – их значение частоты [5]. После этого огибающая амплитуды во временной области каждой полосы извлекается отдельно. Это достигается за счет фильтрации нижних частот каждой полосы, применения полноволнового выпрямления и понижения частоты дискретизации [6]. Затем огибающие каждой полосы суммируются и вычисляется автокорреляционная функция. Пики автокорреляционной функции соответствуют различным периодам огибающей сигнала.

Фильтрация нижних частот:

$$y[n] = (1 - \alpha) x[n] - \alpha y[n], \tag{5}$$

где $\alpha = 0,99$.

Полноволновое выпрямление:

$$y[n] = abs(x[n]). \tag{6}$$

Понижение частоты дискретизации:

$$y[n] = x[kn]. \tag{7}$$

Нормализация в каждой полосе (удаление среднего значения) для исключения аномальных данных:

$$y[n] = x[n] - E[x[n]], \tag{8}$$

где $E[x[n]]$ – среднее значение последовательности $x[n]$.

Автокорреляция:

$$y[k] = \frac{1}{N} \sum_{n=0}^{N-1} x[n]x[n+k]. \tag{9}$$

Из результата берутся первые пять пиков автокорреляционной функции, после чего рассчитываются и добавляются в гистограмму соответствующие им периодичности в bpm. Этот процесс повторяется в процессе прохождения по сигналу. Периодичность, соответствующая наиболее заметному пику конечной гистограммы, является предполагаемым темпом аудиофайла в bpm.

Основными недостатками рассмотренного метода определения темпа являются неточные (в некоторых случаях даже ошибочные) результаты на музыке определенных жанров (например, на классической музыке), а также невозможность определить переменный темп.

1.4 Скрытые модели Маркова

1.4.1 Стохастическое моделирование

Как уже было упомянуто выше, практически во всех музыкальных записях имеет место небольшое отклонение нот от ритмической сетки. Рассматриваемый метод рассчитан именно на работу с такими случаями. Также в данном

методе подразумевается, что входные данные представлены в формате MIDI (Musical Instrument Digital Interface, стандарт обмена данными между цифровыми музыкальными инструментами). В MIDI файлах указывается информация о высоте ноты, ее длительности и силе нажатия [8].

Исследования показывают, что отклонения нот можно смоделировать с помощью распределения Гаусса относительно их идеальной длительности [9]. Тогда, если i – идеальная длительность ноты («намерение») в момент времени t, то ее исполненная длительность x_t моделируется функцией плотности вероятности $f_i(x_t)$.

Пусть $Q = \{q_1, q_2, ..., q_N\}$ – последовательность «намерений» в соответствующие моменты времени. Тогда наблюдаемая последовательность длительностей $X = \{x_1, x_2, ..., x_N\}$ определяется как:

$$P(X|Q) = \prod_{t=1}^{N} f_{q_t}(x_t). \tag{10}$$

В данном методе используются два типа моделей генерации ритмических рисунков для получения возможных ритмов:

— n-граммная модель (длина ноты предсказывается исходя из предыдущих n-1 нот в вероятностном смысле. Эта модель охватывает любые ритмические рисунки и может выдавать точную вероятность);

— «ритмический словарь» (состоит из всех известных ритмических рисунков за единицу времени. Хорошо представляет известные ритмические рисунки, в то время как неизвестные заменяются аналогичными существующими ритмами).

Обе модели можно представить в виде вероятностных сетей перехода состояний, где каждое состояние связано с предполагаемой длительностью ноты. Вероятность того, что номер состояния изменится в последовательности $Q = \{q_1, q_2, ..., q_N\}$ определяется как $P(Q) = p_{q_0} \prod_{t=1}^{N} a_{q_{t-1}q_t}$, где p_i – вероятность изначального нахождения в состоянии i, а a_{ij} – вероятность перехода из

состояния i в состояние j.

Колеблющиеся длительности и возможные последовательности нот могут быть объединены в рамках скрытой модели Маркова как вероятности перехода $A = \{a_{ij}\}$ и наблюдаемые вероятности $B = \{b_i(x_t)\}$ соответственно. В таком случае вероятность наблюдения последовательности длительностей X определяется как:

$$P(X) = P(X|Q)P(Q). \tag{11}$$

1.4.2 Определение ритма

Задача заключается в том, чтобы найти временную последовательность Q номеров состояний, которая дает максимальную апостериорную вероятность $P(Q|X)$ при заданной последовательности наблюдаемых длительностей X [9].

По теореме Байеса:

$$P(Q|X) = \frac{P(X|Q)P(Q)}{P(X)}. \tag{12}$$

Значит, максимизация апостериорной вероятности эквивалентна нахождению $argmax P(X|Q)P(Q)$ среди всех возможных Q.

Оптимальная последовательность состояний находится с помощью алгоритма Витерби для поиска наилучшего пути в вероятностной сети переходов.

Основной недостаток представленного метода заключается в необходимости входных данных быть в формате MIDI. Также к недостаткам можно отнести периодические неточности в результатах. Например, музыкальные фрагменты с разным темпом (к примеру, 116 bpm и 127 bpm) могут быть определены как имеющие одинаковый темп (в данном случае 120 bpm [9]).

1.5 Байесовское иерархическое моделирование

1.5.1 Языковая модель

Байесовское иерархическое моделирование состоит из двух компонентов: языковой модели («language» model) и модели представления («performance»

model) [10].

Языковая модель построена на марковской модели нотных паттернов. В этой модели используется последовательность $B_k = z_{k,1}, ..., z_{k,l}$, где $k = 1, ..., K$ – индекс в множестве нотных паттернов длины K, $z_{k,l}$ – нота под номером l в нотном паттерне k, где $l = 1, ..., L$, а L – количество нот в паттерне. При этом вероятность последовательности паттернов $w_{1:I} = w_1, ..., w_I$, где $w_i \in \{B_k\}_{k=1}^{K}$ определяется как:

$$P(w_{1:I}) = P(w_1 = B_k) \prod_{i=2}^{I} P(w_i = B_{k'} | w_{i-1} = B_k). \tag{13}$$

Проблемой этой модели в чистом виде является обработка синкоп (смещения акцента с сильной доли такта на слабую [11]), поскольку синкопированная нота лежит за границей такта, которая обычно является границей нотных паттернов.

Модификация нотных паттернов

Пусть $z_{1:M}$ – последовательность нот, являющаяся результатом модели нотных паттернов, $y_{1:N}$ – модифицированная последовательность $z_{1:M}$ (со вставленными нотами), а $x_{1:N}$ – итоговая последовательность нот, получающаяся в результате языковой модели (содержащая в т. ч. синкопы).

Синкопы могут быть интегрированы в модель путем расширения пространства состояний базовой модели w_i до пары $(w_i; s_i)$, где s_i – степень синкопирования i-ой ноты (степень ее сдвига). Тогда выражение 13 изменяется как [10]:

$$P(w_{1:I}, s_{1:I}) = P(w_1 = B_k, s_1) \prod_{i=2}^{I} P(w_i = B_{k'}, s_i | w_{i-1} = B_k, s_{i-1}). \tag{14}$$

Процесс Дирихле

Поскольку используемые нотные паттерны и типы модификаций варьируются в зависимости от музыкальных произведений, для отдельных произ-

ведений учитываются разные значения параметров. В байесовской модели эти параметры считаются сгенерированными из предшествующих (априорных) моделей. Процесс Дирихле [12] может служить этой априорной моделью.

Пусть $\pi_{kk'} = P(w_i = B_{k'}|w_{i-1} = B_k)$. В случае конечных распределений процесс Дирихле для дискретного распределения π описывается базовым распределением ω и параметром концентрации α следующим образом:

$$\pi \sim Dir(\alpha\omega), \tag{15}$$

где $Dir()$ обозначает распределение Дирихле.

Когда параметр концентрации α мал, используется компактная грамматика, т. е. для каждого музыкального произведения будет использоваться небольшое количество паттернов нот.

1.5.2 Модель представления

Модель описывает два источника «колебаний» (неточностей) в музыкальном исполнении. Один из них – колебание ритма, а другой – колебание темпа [10]. Пусть $v_i = d_i/x_i$, где x_i – «формальная» длительность i-ой ноты, а d_i – фактический интервал между i-ой и (i+1)-ой нотами. Вариация v_i описывается марковским процессом. Предполагая, что колебания темпа и ритма являются гауссовскими, модель представления задается как:

$$\begin{cases} v_n|v_{n-1} \sim N(v_{n-1}, \sigma_v^2), \\ d_n|v_n, x_n \sim N(v_n x_n, \sigma_t^2), \end{cases} \tag{16}$$

где σ_v (σ_t) – стандартное отклонение для колебаний темпа (ритма).

Полная вероятность для модели представления задается как:

$$P(d_{1:N}, v_{1:N}|x_{1:N}) = \prod_{n=1}^{N} P(d_n|v_n, x_n)P(v_n|v_{n-1}), \tag{17}$$

где $P(v_1|v_0) \equiv P(v_1)$.

Таким образом, байесовское иерархическое моделирование позволяет немного увеличить точность определения ритма по сравнению с марковскими моделями (примерно на 2% [10]). Но остальные недостатки скрытых марковских моделей остаются прежними: работа только с MIDI форматами и определение только постоянного темпа.

1.6 Использование сверточных нейросетей

1.6.1 Представление сигнала

Сигнал представляется в виде спектрограммы по шкале мела, чтобы снизить объем данных, который должен быть обработан нейросетью (мел, от слова «мелодия», - психофизическая (субъективная) единица высоты звука [13]). Шкала мела выбрана вместо линейной шкалы из-за ее связи с человеческим восприятием и диапазонами частот инструментов.

Чтобы создать спектрограмму, сигнал конвертируется в моно, его дискретизация понижается до 11025 Гц, после чего используются полуперекрывающиеся окна из 1024 отсчетов [14]. Это эквивалентно частоте кадров 21,5 Гц, что (согласно теореме отсчетов) достаточно для представления темпа до 646 bpm. Каждое окно преобразуется в 40-полосный спектр в шкале мел, охватывающий диапазон от 20 до 5000 Гц. В качестве длины спектрограммы выбрано 256 кадров, что примерно равняется 11,9 с.

1.6.2 Архитектура сети

Архитектура рассматриваемой сети представлена на рис. 1.4.

Сначала входные данные обрабатываются тремя сверточными слоями, каждый из которых состоит из 16 фильтров размера 1x5. С помощью этих фильтров сопоставляется ритмическая структура сигнала.

После этого идут четыре модуля с несколькими фильтрами. Каждый из модулей состоит из среднего слоя пулинга («avg pooling»), шести параллельных сверточных слоев с фильтрами разных размеров (от 1x32 до 1x256), слоя конкатенации и т. н. «узкого» («bottle-neck») слоя, предназначенного для уменьшения размерности. С помощью этих модулей достигаются две цели:

1) Пулинг по оси частот для суммирования диапазонов мел.
2) Сопоставление сигнала с различными фильтрами, способными обнаруживать длительные временные зависимости.

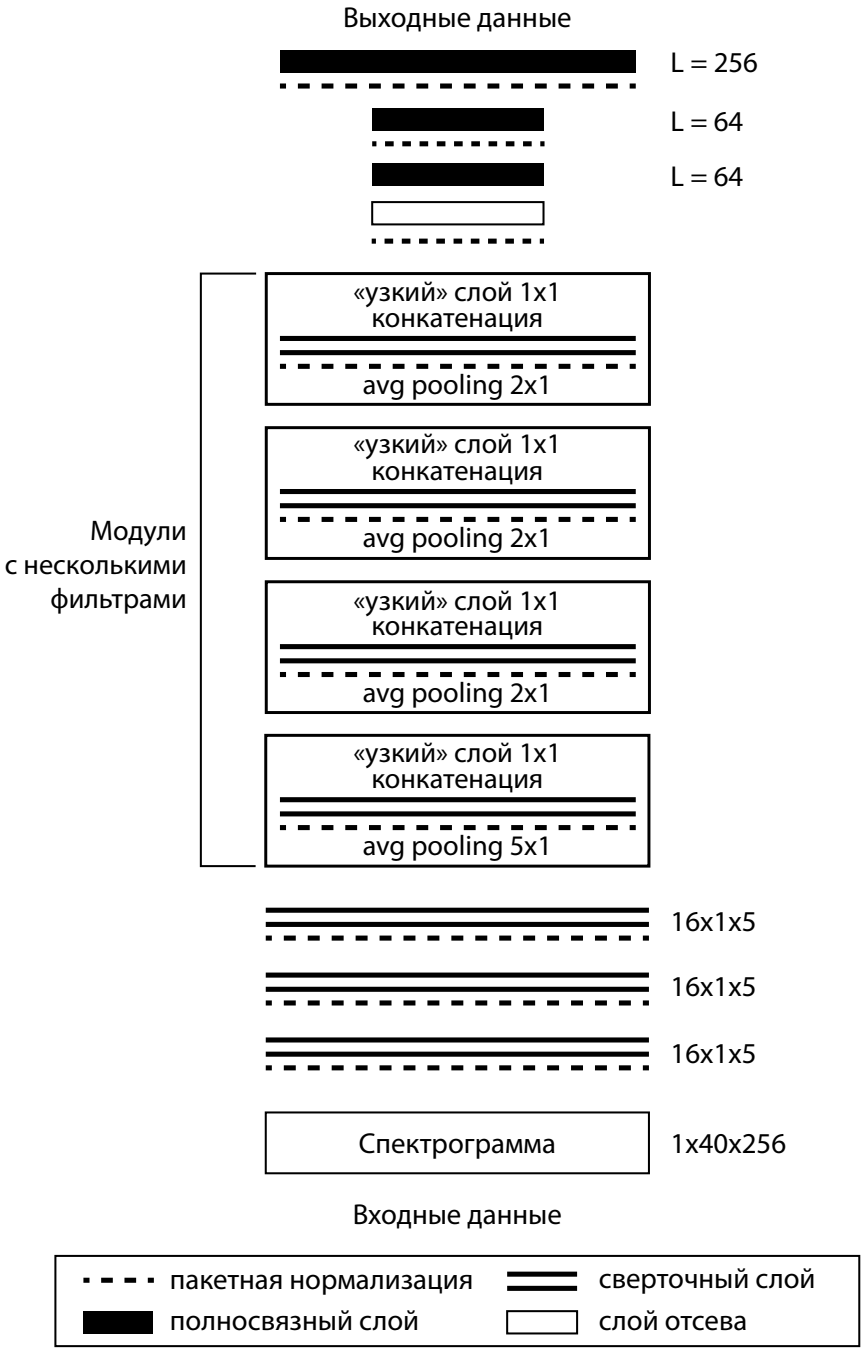


Рис. 1.4 – Схема архитектуры нейросети

Чтобы классифицировать свойства, полученные из сверточных слоев, добавляются два полносвязных слоя (по 64 единицы каждый), за которыми следует выходной слой с 256 единицами. Выходной слой использует softmax в качестве функции активации, а все остальные слои используют ELU [15]. Каждому сверточному или полносвязному слою предшествует пакетная нормализация [16]. Первому полносвязному слою также предшествует слой отсева с $p = 0,5$ («dropout») для противодействия переобучению.

Всего сеть имеет 2921042 обучаемых параметра.

В результате выбирается один из 256 вариантов темпа от 30 до 285 bpm.

Таким образом, сверточные нейросети позволяют определять темп с достаточно высокой точностью (процент правильных оценок с допустимой погрешностью в 4%) (до 92% на основе комбинированной выборки, состоящей из аудиофайлов различных жанров с темпом от 44 до 216 bpm [14]). Также данный метод можно использовать и при определении глобального темпа, не только для фрагментов. Но он по-прежнему не позволяет определить переменный темп, а также не предназначен для определения ритма. Помимо этого нейросетевые методы имеют такие недостатки, как необходимость обучающих датасетов больших объемов, зависимость от исходных данных и долгое время обучения.

1.7 Сравнение методов

По результатам рассмотрения перечисленных выше методов была составлена таблица 1.1.

Как видно из таблицы, ни один метод в своем изначальном варианте не предполагает определение переменного темпа и ритма. Однако метод скрытых марковских моделей при небольшой модификации может позволить определить переменный темп и ритм [9].

Также стоит заметить, что все методы, кроме ДВП, содержат обучаемые параметры. В скрытых марковских моделях – это множество $\{a_{ij}\}$, а в байесовском иерархическом моделировании – множество $\{\pi_{kk'}\}$. В обоих методах обучение происходит с помощью статистической оценки. Размеры датасетов в

таблице были указаны исходя из данных, использовавшихся для обучения соответствующих моделей в исследованиях.

Таблица 1.1 – Сравнение рассмотренных методов

Метод	Точность результатов	Переменный темп и ритм	Формат входного аудиофайла	Размер использовавшегося датасета (кол-во аудиофайлов)
ДВП	$\sim$ 65% (13 верных из 20) [6]	Не определяются	Нет ограничений	Обучение не требуется
Скрытые марковские модели	$\sim$ 80% (при допустимой погрешности 4%)	Могут определяться при модификации метода	MIDI	88 [9]
Байес	Выше марковских примерно на 2%	Не определяются	MIDI	100 [10]
Сверточная нейросеть	до 92%	Не определяются	Нет ограничений	8596 [14]

Выводы

В этом разделе была проанализирована предметная область и сформулирована проблема. А также была проведена классификация и сравнение основных существующих методов решения поставленной задачи.

ЗАКЛЮЧЕНИЕ

В результате работы были рассмотрены понятия предметной области, такие как темп и ритм музыки, и проанализирована проблема автоматического определения темпа и ритма.

Были определены критерии сравнения методов.

Были рассмотрены основные методы автоматического определения темпа и ритма музыки: дискретное вейвлет-преобразование, скрытые марковские модели, байесовское иерархическое моделирование и сверточные нейронные сети. После чего было произведено сравнение изученных методов по выделенным ранее критериям.

По результатам сравнения можно сделать вывод, что ни один из рассмотренных методов без каких-либо модификаций не позволяет определять переменный темп и ритм. При этом сверточные нейросети позволяют добиться достаточно высокой точности результатов в сравнении с другими методами, не имея особых ограничений на формат входного аудиофайла, но для этого требуется выборка достаточно больших размеров и значительное время на обучение.

Таким образом, цель работы – изучить основные существующие методы определения ритмического рисунка и темпа цифровой музыкальной записи – была достигнута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Benetos E. / Automatic music transcription: challenges and future directions / Benetos E., Dixon S., Giannoulis D., Kirchhoff H., Klapuri A. // Journal of Intelligent Information Systems. – 2013. – С. 407-434.

2. Музыкальный словарь Гроува. // Москва. – 2007. – С. 858.

3. Чехович Д. О. Ритм музыкальный // Большая российская энциклопедия. Москва. – 2015. – Том 28. – С. 541.

4. Cemgil A.T., Desain P., Kappen B. Rhythm quantization for transcription // Computer Music Journal. – 2000. – С. 60-76.

5. Polikar R. The wavelet tutorial // 2-е изд. – 2001. – 67 с.

6. Tzanetakis G., Essl G., Cook P. Audio analysis using the Discrete Wavelet Transform. – 2001.

7. Биккенин Р. Р., Чесноков М. Н. Теория электрической связи. – 2010. – 329 с.

8. Pleshkova S., Panchev K., Bekyarski A. Development of a MIDI synthesizer for test signals to a wireless acoustic sensor network. – 2020.

9. Takeda H., Saito N., Otsuki T. Hidden Markov model for automatic transcription of MIDI signals. – 2002. – С. 428-431.

10. Nakamura E., Itoyama K., Yoshii K. Rhythm transcription of MIDI performances based on Hierarchical Bayesian Modelling of repetition and modification of musical note patterns. – 2016.

11. Hoffman M. Syncopation. – 2009.

12. Kotz S., Balakrishnan N., Johnson N.L. Continuous Multivariate Distributions. Volume 1: Models and Applications (Chapter 49: Dirichlet and Inverted Dirichlet Distributions). – New York: Wiley. – 2000. – Том 1.

13. Stevens S.S., Volkmann J., Newman E.B. A scale for the measurement of the psychological magnitude pitch // Acoustical Society of America. – 1937. – С. 188.

14. Schreiber H., Muller M. A single-step approach to musical tempo estimation using a convolutional neural network // 2018. – С. 98-105.

15. Clevert D.A., Unterthiner T., Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). – 2015.

16. Ioffe S., Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. – 2015.

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н.Э. БАУМАНА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Классификация методов определения ритмического рисунка и темпа цифровой музыкальной записи

СТУДЕНТ: ПЕТРОВА АННА АЛЕКСЕЕВНА
ГРУППА: ИУ7-76Б
РУКОВОДИТЕЛЬ: КИВВА КИРИЛЛ АНДРЕЕВИЧ

ПРИЛОЖЕНИЕ А

Цель и задачи

Цель – изучить основные существующие методы определения ритмического рисунка и темпа цифровой музыкальной записи.

Задачи:

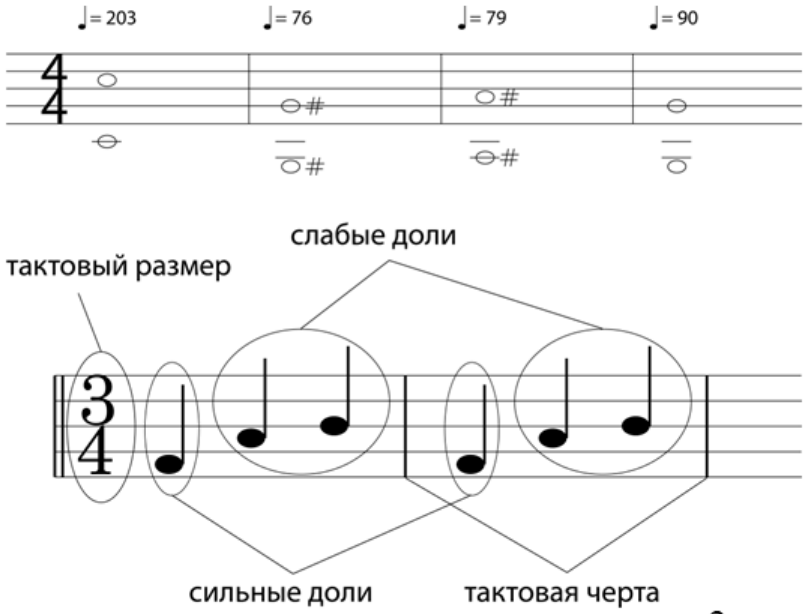
- провести анализ предметной области и сформулировать проблему;
- сформулировать критерии сравнения методов выделения информации о ритме и темпе музыки;
- классифицировать основные существующие методы.

Основные понятия

Темп – мера времени в музыке, упрощенно – «скорость исполнения музыки». Измеряется в bpm (число ударов в минуту).

Ритм – регулярная, периодическая последовательность акцентов. Такое определение ритма фактически идентично метру.

Метр – чередование сильных и слабых долей в определенном темпе. Численно фиксируется с помощью тактового размера.



Проблема определения ритма и темпа

- нечеткое попадание в ритм и темп на живых записях;
- переменный ритм и темп.

Критерии сравнения методов:

- точность результатов;
- определение переменного темпа и ритма;
- ограничения на формат входного аудиофайла;
- размеры датасетов (если обучение необходимо).

Дискретное вейвлет-преобразование

Преобразование Фурье (ПФ) => оконное ПФ => вейвлет-преобразование

Основная идея – разделение сигнала на высокие и низкие частоты с помощью фильтров.

$$y_{high}[k] = \sum_{n=-\infty}^{\infty} x[n[g[2k-n], \quad (1)$$

$$y_{low}[k] = \sum_{n=-\infty}^{\infty} x[n[h[2k-n]\,. \quad (2)$$

Скрытые модели Маркова

- n-граммная модель (длина ноты предсказывается исходя из предыдущих n-1 нот в вероятностном смысле);
- «ритмический словарь» (состоит из всех известных ритмических рисунков за единицу времени).

$Q = \{q_1, q_2, \dots, q_N\}$ – идеальные длительности нот, $X = \{x_1, x_2, \dots, x_N\}$ – наблюдаемые длительности нот.

$$P(Q) = p_{q_0} \prod_{t=1}^{N} a_{q_{t-1} q_t}, \quad (3)$$

$$P(Q|X) = \frac{P(X|Q)P(Q)}{P(X)}. \quad (4)$$

Байесовское иерархическое моделирование

- языковая модель;
  - модификация нотных паттернов (добавление синкоп в модель);
  - процесс Дирихле;

  $\pi_{kk'}$ - вероятность перехода от паттерна k к паттерну k'. $\pi \sim Dir(\alpha\omega)$.
- модель представления;
  - колебания (неточности) темпа;

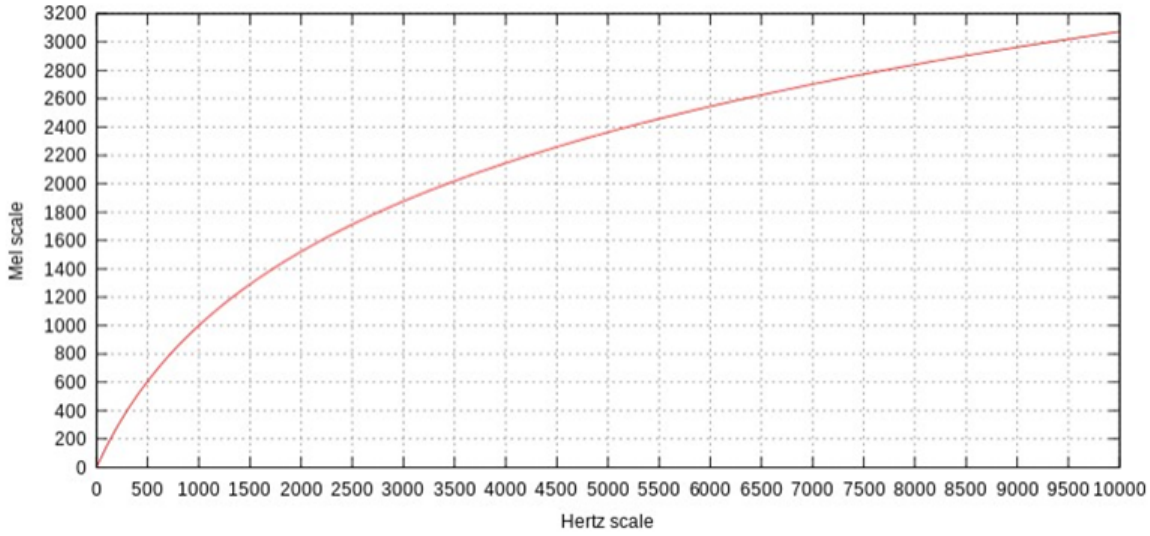
$$v_n|v_{n-1} \sim N(v_{n-1}, \sigma_v^2) \quad (5)$$

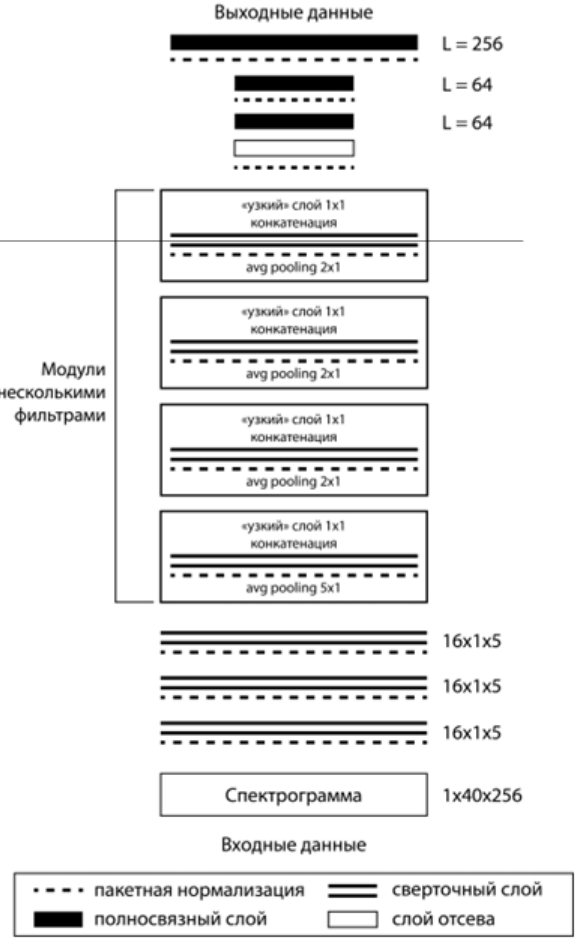
  - колебания ритма.

$$d_n|v_n, x_n \sim N(v_n x_n, \sigma_t^2) \quad (6)$$

7

Сверточные нейросети





Сравнение методов

Метод	Точность результатов	Переменный темп и ритм	Формат входного аудиофайла	Размер обучающего датасета*
ДВП	~ 65 % (13 верных из 20)	Не определяются	Нет ограничений	Обучение не нужно
Скрытые марковские модели	~ 80 %	Могут определяться при модификации метода	MIDI	88
Байес	Выше марковских примерно на 2%	Не определяются	MIDI	100
Сверточная нейросеть	До 92%	Не определяются	Нет ограничений	8596

**На основе данных из исследований*

Выводы

- рассмотрены понятия предметной области, такие как темп и ритм музыки, и проанализирована проблема автоматического определения темпа и ритма;
- определены критерии сравнения методов;
- рассмотрены основные методы автоматического определения темпа и ритма музыки: дискретное вейвлет-преобразование, скрытые марковские модели, байесовское иерархическое моделирование и сверточные нейронные сети;
- произведено сравнение изученных методов по выделенным ранее критериям.