Содержание

1 Аннотация

В современном мире нейронные сети достаточно развиты, чтобы сгенерировать музыку практически любого жанра и стиля. Но можно ли при помощи нейронных сетей сгенерировать композиции, вызывающие у людей определенные эмоции? В этой работе будут рассмотрены способы генерации музыки различной эмоциональной окраски и геометрические структуры, возникающие в человеческой и сгенерированной музыке.

Ключевые слова

Глубинное обучение, генерация музыки, вариационные автокодировщики, ЭЭГ, неоримановы преобразования

2 Введение

Многие исследования современных нейробиологов подтвердили, что прослушивание музыкальных композиций стимулирует различные зоны человеческого мозга, и, в том числе, провоцирует человека испытать широкий спектр эмоций. В это же время, активно развиваются генеративные нейронные сети, способные генерировать музыку по любому человеческому запросу: можно выбрать инструмент, эпоху, стилистику и множество других параметров. Возникает закономерный вопрос: можно ли «научить» нейронные сети создавать мелодии и композиции, которые будут вызывать у человека какой-либо эмоциональный отклик? Попытке ответить на этот вопрос и будет посвящено моё исследование: его **целью** будет являться создание кода, способного генерировать музыку, призванную вызвать определенные эмоции. В частности, меня будет интересовать использование различных геометрических структур (а именно нео-римановых преобразований) в генерации музыки и в том, как именно эти преобразования связаны с эмоциональными категориями.

Для достижения цели мной были поставлены следующие **задачи**:

- Изучить имеющиеся публикации (в частности, статьи и книги) по выбранной теме
- Определиться с форматом входных данных (спектрограммы, midi-файлы, текстовое представление композиций) и собрать датасет
- Провести несколько экспериментов с разными видами нейронных сетей (рекуррентные, сверточные и др.): при помощи каждой из моделей сгенерировать набор композиций

- Провести эксперимент с реальными слушателями: дать испытуемым послушать композиции, сгенерированные моделями из предыдущего шага, собрать данные ЭЭГ.
- По полученным на предыдущем шаге данным ЭЭГ определить, какую эмоцию испытывал (и испытывал ли) человек во время прослушивания каждой из сгенерированных композиций (используя регрессор, по данным ЭЭГ определить точные значения valence и arousal)
- Для предложенной модели и сгенерированных при её помощи треков треков понять, совпали ли испытанные людьми эмоции с ожидаемыми

Моё исследование **актуально** по нескольким причинам: во-первых, музыка оказывает глубокое влияние на человеческий мозг, и понимание механизмов, лежащих в основе этого воздействия, может дать ценную информацию о работе мозга в целом. Во-вторых, существуют исследования, доказывающие, что музыкальная терапия эффективна при коррекции различных состояний, включая депрессию, тревогу и болезнь Паркинсона: модель, генерирующая музыку с определенным эмоциональным откликом, может быть полезна для работы с этими состояниями. В-третьих, изучая взаимосвязь между мозгом и музыкой, исследователи могут глубже понять, как люди именно воспринимают музыку, что может помочь в разработке более совершенных музыкальных технологий.

Объектом моего исследования являются нейронные сети, способные генерировать музыку, а **предметом** исследования – генерация (при помощи нейронных сетей) композиций, вызывающих у людей определённые эмоции.

3 Related work

К генерации symbolic music (то есть музыки, записанной в определённой нотации, в частности, MIDI) существует множество подходов. Наличие нотации позволяет свести генерацию музыки к задаче генерации последовательностей из естественного языка, поэтому методы, используемые в NLP (natural language processing), могут быть применимы и к генерации музыки. Попытки использовать обычные RNN для генерации музыки не увенчались успехом из-за проблемы затухающих градиентов, и, как следствие, невозможности генерировать длинные последовательности. Для решения этой проблемы предпринимались попытки использовать LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit), но даже с этими модификациями возникала проблема – при увеличении длины последовательности модель «забывала» старые входные данные. Успешным примером применения моделей из NLP

к генерации музыки оказалось применение архитектуры трансформера и механизма внимания (attention), как, например, в статье «Symbolic music generation conditioned on continuous-valued emotions» [7]. Кроме того, в этой работе генерация была условной (conditional): valence и arousal были представлены в виде вещественного числа от -1 до 1.

Другим подходом, использующим архитектуру трансформера, был подход, описанный в статье о датасете EMOPIA [2]. В этом случае генерация тоже была условной, но valence и arousal были уже не вещественными числами, а бинарными категориями, сами треки при этом делились по low / high valence и low / high arousal.

Кроме этого, существуют различные подходы к безусловной генерации, которые можно было применить к генерации условной. К примеру, модель вариационного автокодировщика (VAE), подходящая для безусловной генерации, описана в статье «An Introduction to Variational Autoencoders», D. P. Kingma, M. Welling, 2019, [3]. VAE состоит из энкодера и декодера. Основная идея VAE состоит в том, чтобы сначала закодировать при помощи энкодера музыкальные произведения в представление более низкой размерности, называемое скрытым (или латентным) представлением. Затем латентное представление используется в качестве входных данных для декодера, который генерирует новое музыкальное произведение. Одним из преимуществ использования VAE для создания музыки является то, что их можно обучить генерировать новые произведения, похожие на произведения в наборе обучающих данных, но при этом являющиеся уникальными и новыми. Это позволяет модели распознавать основные паттерны и структуры переданных на вход композиций, но при этом иметь возможность генерировать новые фрагменты, которые не являются просто копиями обучающих данных. Модификация VAE для генерации музыки была описана в статье «Music generation with variational recurrent autoencoder supported by history» [8]. В этом случае был использован вариационный автокодировщик с модифицированным декодером: в декодер помимо объекта из латентного пространства подаётся информация об уже сгенерированных нотах, и предыдущие ноты влияют на последующую генерацию. Этот автокодировщик (названный авторами VRASH – Variational Recurrent Autoencoder Supported by History) обеспечивает хороший баланс между глобальной и локальной структурой трека. Стандартная структура VAE позволяет обнаружить паттерны в макроструктура трека, а использование VRASH позволяет добиться более разнообразных и интересных паттернов на локальном уровне. Именно поэтому эту архитектуру есть смысл применить к условной генерации.

Существует также подход, связанный с генерацией исполнений (performance generation) – этот подход описан в статье «This Time with Feeling: Learning Expressive Musical Performance»,

S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, 2018, [5]. Генерация музыки традиционно была сосредоточена либо на сочинении новых произведений (то есть генерации нот или сочетаний нот), либо на исполнении существующих (то есть генерации характеристик звука). В этой статье авторы исследуют концепцию прямого создания исполнения (performance generation), когда ноты, их характеристики и стиль исполнения генерируются одновременно. Кроме того, изучаются характеристики набора данных, необходимого для этой задачи, и выделяются преимущества работы с такими данными. Авторы представляют модель рекуррентной сети на основе LSTM, которая хорошо работает в задаче performance generation.

4 Основная часть

4.1 Датасет и преобразование midi-файлов

MIDI-нотация – это специальная форма записи звуковых файлов, в которой данные представлены не в форме звуковых волн, а в форме некоторой информации о конкретных “событиях“ в каждый конкретный момент, к примеру “мягко сыграть ноту Фа, используя в качестве инструмента фортепиано, а в качестве длительности – четверть (в текущем размере и темпе)“. Всего существует 413 различных событий: [5]

- 128 NOTE-ON событий: по одному для каждого из 128 MIDI-pitches (высота ноты). Каждое событие означает начало ноты определенной высоты.
- 128 NOTE-OFF событий: по одному для каждого из 128 MIDI-pitches (высота ноты). Каждое событие означает конец ноты определенной высоты.
- 125 TIME-SHIFT событий: каждое описывает длительность ноты с шагом от 8 мс до 1 секунды.
- 32 VELOCITY событий: каждое меняет темп для всех последующих нот (до следующего velocity-события).

Наличие нотации, как указывалось ранее, позволяет работать с композициями, как с последовательностями естественного языка, в частности, использовать для генерации архитектуру трансформера.

Для создания датасета, содержащего пары MIDI-файлов и высокоуровневых меток, используется API Spotify for Developers и получаются аудио-характеристики для образцов

из набора данных Lakh MIDI (LMD). В частности, было использовано подмножество LMD-matched, поскольку его образцы соответствуют записям в наборе данных Million Song Dataset (MSD), поэтому метаданные из MSD можно использовать для поиска в базе данных Spotify. Используя идентификатор трека для каждого MIDI-файла, были получены название песни, имя исполнителя и идентификатор песни Echo Nest. Используя идентификаторы песен Echo Nest и другой набор данных с названием Million Song Dataset Echo Nest mapping archive, было проведено сопоставление MIDI-треков и идентификаторов из Spotify. Затем для каждого образца MIDI был проведён поиск с использованием API Spotify for Developers. Запросом для поиска был связанный идентификатор Spotify. Если идентификатор Spotify не был доступен, для получения идентификатора трека было использовано название трека и его исполнитель. API Spotify for Developers позволяет пользователям получить доступ к аудио-характеристикам для заданной песни из частной базы данных Spotify. Эти аудио-характеристики являются как низко-, так и высокоуровневыми, и включают в себя danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. Высокоуровневые характеристики, такие как valence и arousal, оцениваются с помощью алгоритмов машинного обучения, которые обучаются на данных, размеченных экспертами.

4.2 Предложенная модель

Изначально для генерации была предложена модель VRASH – Variational Recurrent Autoencoder Supported by History [8]. Эта модель имеет следующую структуру модифицированного вариационного автоэнкодера:

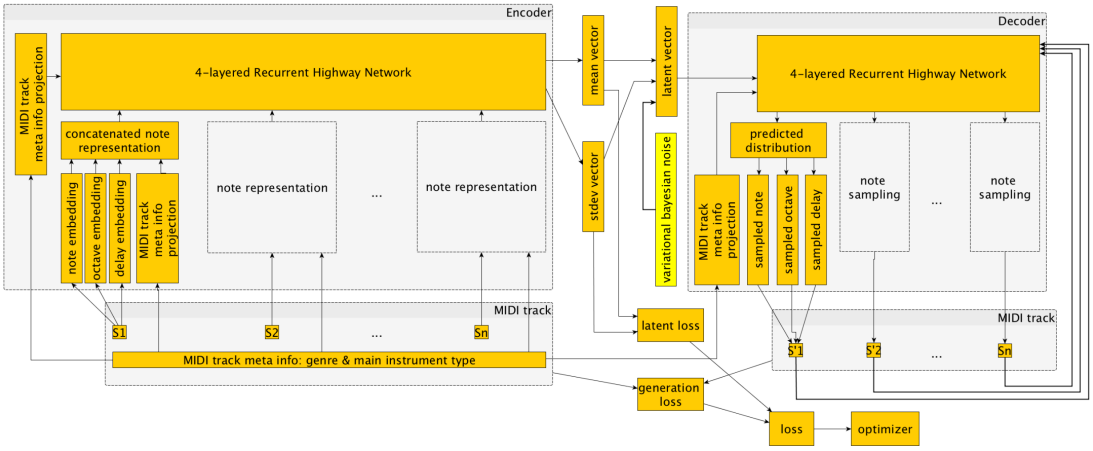

Рис. 4.1: Архитектура VRASH. Предыдущие результаты генерации учитываются при генерации последующих токенов.

Чтобы обеспечить генерацию на основе эмоций, модель была обучена на нескольких подмножествах датасета EMOPIA и датасета Lakh Pianoroll: треки из датасета Lakh Pianoroll на основе разметки по valence / arousal были разбиты на 4 категории в соответствии с категориями из EMOPIA (high / low valence, high / low arousal). Получившиеся треки были исключительно фортепианными и в целом более примитивными (в частности, потому что подход, использующие вариационные автокодировщики, является не устаревшим, но всё-таки не новейшим) чем треки, которые удалось получить далее при помощи другой модели. Для итоговой генерации была использована модель из статьи «Symbolic music generation conditioned on continuous-valued emotions» [7], а конкретно имплементация continious-concat-модели: в этом случае создаётся один вектор для двух нормализованных непрерывных значений valence и arousal, этот вектор повторяется в измерении последовательности и объединяется с эмбеддингами для каждого токена. Длины обуславливающих векторов и эмбеддингов составляют 192 и 576 соответственно, так что общая длина признаков входных данных преобразователя постоянна для всех моделей. Условная модель обучаются путем fine-tuning предварительно обученной безусловной модели – стандартного encoder-decoder трансформера с механизмом attention.

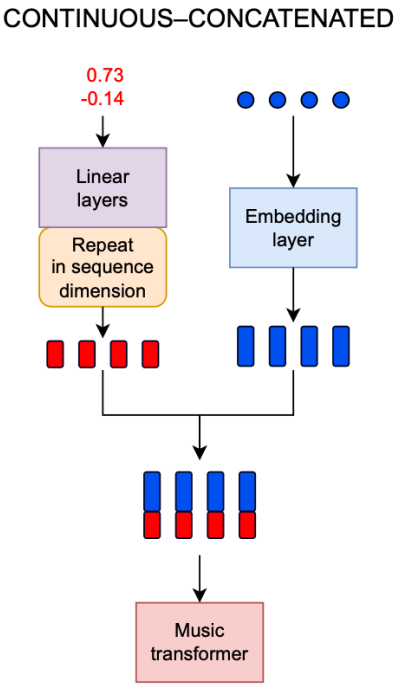

Рис. 4.2: Архитектура предложенной модели.

4.3 Человеческий мозг и музыка

4.3.1 Модель valence-arousal

Valence-arousal модель – это одна из самых распространённых моделей для описания эмоций, предложенная Джеймсом Расселом [6]. Модель является двумерной и состоит из измерений валентности (valence, описывает измерение неприятности и приятности – unpleasantness vs. pleasantness) и возбуждения (arousal, описывает измерение расслабления и возбуждения – relaxed vs aroused). Расселл также описывает разные эмоции в терминах валентности и возбуждения: к примеру, «приятное спокойствие» – пример высокой валентности и низких значений возбуждения, а «раздражение» – пример низкой валентности и высоких значений возбуждения. Например, при бинарном разделении музыкальных треков по низкой / высокой валентности и низкому / высокому возбуждению мы можем получить четыре категории: «грусть» – low valence, low arousal; «злость» – low valence, high arousal; «приятное волнение (excitement)» – high valence, high arousal; «приятное спокойствие» – high valence, low arousal.

4.3.2 Определение valence и arousal по ЭЭГ

Существует несколько исследований, посвящённых изучению связи музыки, эмоций и человеческого мозга [4]. Большинство этих исследований было посвящено только работе с измерением валентности: результаты ЭЭГ сравнивались на основе того, насколько приятную или неприятную музыку слушали участники экспериментов; либо же в этих исследованиях весь спектр эмоций (на плоскости valence-arousal) сводился к какому-то ограниченному набору категорий (грусть, счастье, злость и т.д.). Одним из последних исследований в этой области является исследование Filipe Galvão, Soraia M. Alarcão, and Manuel J. Fonseca, «Predicting Exact Valence and Arousal Values from EEG». В этом исследовании авторы обучают регрессоры на данных ЭЭГ, и эти регрессоры выдают точные значения valence и arousal с довольно маленькими значениями ошибок ($\text{MAE} < 0.06$, $\text{RMSE} < 0.16$), а так же распознают четыре эмоциональных класса (high / low valence, high / low arousal) с высокой долей правильно классифицированных объектов (accuracy ~ 0.844). В качестве регрессоров при этом используется регрессор на основе K ближайших соседей [1]. В оригинальной статье авторы обучали регрессор на датасетах DEAP, AMIGOS и DREAMER. Регрессор, предложенный в этой работе, был обучен на датасете DEAP.

ЭЭГ обеспечивает отличное временное разрешение (то есть позволяет оперировать короткими временными отрезками) и быстрое получение данных, будучи неинвазивным и недо-

рогим, что делает его хорошим кандидатом для измерения эмоционального состояния людей. Мозговые волны обычно подразделяются на пять разных частотных диапазонов: дельта (δ) 1–4 Гц; тета (θ) 4–7 Гц; альфа (α) 8–13 Гц; бета (β) 13–30 Гц; и гамма (γ) $>$ 30 Гц), каждый из которых более заметен в определенных состояниях сознания. Дельта – это самые медленные волны, которые наиболее выражены во время сна с небыстрым движением глаз (NREM). Тета-волны связаны с подсознательной деятельностью, такой как сновидения, и присутствуют в медитативных состояниях ума. Альфа-волны появляются преимущественно во время состояний бодрствования с закрытыми глазами и наиболее заметны над теменными и затылочными долями. Активность бета-волн, с другой стороны, связана с активным состоянием ума, более заметным в лобной коре во время интенсивной сосредоточенной умственной деятельности. Наконец, считается, что гамма-ритмы связаны с интенсивной мозговой активностью с целью запуска определенных когнитивных и двигательных функций. В большинстве исследований используется набор из тета-, альфа-, бета- и гамма-волн.

4.4 Нео-римановы преобразования: до и после генерации

4.4.1 Какие бывают преобразования?

Нео-риманова теория (названная в честь Хьюго Римана – теоретика музыки), предоставляет способ анализа не просто последовательностей из аккордов одной тональности, но аккордов, включающих в себя общие ноты (то есть более широкий и насыщенный спектр). Каждое нео-риманово преобразование представляет собой переход между одним мажорным и одним минорным аккордом (все аккорды обязаны быть терцовыми, в частности – трезвучиями). Базовыми и наиболее популярными нео-римановыми преобразованиями являются преобразования P, R и L. P – Parallel – переход между мажорным и минорным аккордом с общим основанием (к примеру, переход от До-мажора к До-минору). R – Relative – Переход между параллельными мажором и минором, то есть между тональностями, имеющими одинаковые ключевые знаки (к примеру, переход от До-мажора к Ля-минору). L – Leading tone exchange – переход, основанный на смене основания аккорда (к примеру, переход от До-мажора к Ми-минору – терция первого аккорда становится основанием второго).

Другими важными нео-римановыми преобразованиями являются преобразования, сохраняющие в аккордах одну общую ноту, и преобразующие две другие. Преобразование S – Slide – связывает мажорное и минорное трезвучие, имеющие общую терцию и разные основание и квинту, к примеру, До-мажор и До-диез-минор (Ми – общая нота, До сменяется на До-диез, Соль сменяется на Соль-диез). Преобразование N – Nebenverwandt – связывает

мажорное трезвучие с минорным на квинту ниже (К примеру, До-мажор и Фа-минор).

4.5 Изначальная идея

В начале работы над проектом присутствовала более сложная идея условной генерации: модель должна была генерировать музыку определенной эмоциональной окраски, обучившись на последовательностях нео-римановых преобразований. Для того, чтобы проверить, насколько эта идея разумна и реализуема, был проведён анализ небольшого датасета EMOPIA [2]. EMOPIA – это набор фортепианных треков, разделённых на сегменты относительно бинарных категорий valence и arousal, все треки записаны в формате MIDI. Треков в датасете всего 387, но после разделения их на сегменты получилось 1,087 midi-файлов. Для того, чтобы проанализировать содержание этих midi-файлов, была использована библиотека music21 и написан парсер для выделения сначала аккордов, а потом и нео-римановых преобразований. Оказалось, что тонических трезвучий или их обращений в предложенных треках было не более половины от всех аккордов в этих треках; более того, из пар последовательно идущих аккордов только в 20% случаев оказывалось, что аккорды пары представляют собой какое-либо из преобразований. Кроме того, терялась вся структура мелодии, а для фортепианных треков мелодия значительно важнее, чем аккордовый аккомпанемент. Из этого был сделан вывод, что генерация на основе исключительно нео-римановых преобразований не имеет смысла, и поэтому были несколько пересмотрены как план работы, так и набор моделей, которые можно было использовать для генерации: так, мы отказались от Riffusion – модификации Stable Diffusion, обученной по текстовому описанию генерировать спектрограммы (была идея дообучить модель на данных о преобразованиях).

Несмотря на то, что от генерации на основе нео-римановых преобразований мы отказались, преобразования в сгенерированных треках будут рассмотрены далее.

4.5.1 Преобразования в треках из датасета

Для визуализации преобразований была использована модель, предложенная в статье James S. Walker и Gary W. Don, A Geometric Analysis of the Harmonic Structure of In My Life [10]. В этой визуализации вершины графа – это различные аккорды (в случае, если аккорд мажорный, буква заглавная, и строчная иначе), а нео-римановы преобразования обозначены рёбрами между этими вершинами. Рёбра организованы таким образом, что между ними образуются шестиугольники. Каждому шестиугольнику сопоставляется заглавная буква, изображающая уже не аккорд, а ноту. Буквы расположены в центрах шестиугольников таким

образом, что вершина, находящаяся на пересечении трёх шестиугольников, соответствует аккорду, составленному из нот, соответствующих этим шестиугольникам. Именно этот способ визуализации показался наиболее подходящим (другой способ – образовать из ребёр не шестиугольники, а треугольники – по количеству преобразований, при этом в вершинах будут находиться ноты; используемый мной подход – chord-based, альтернативный – note-based).

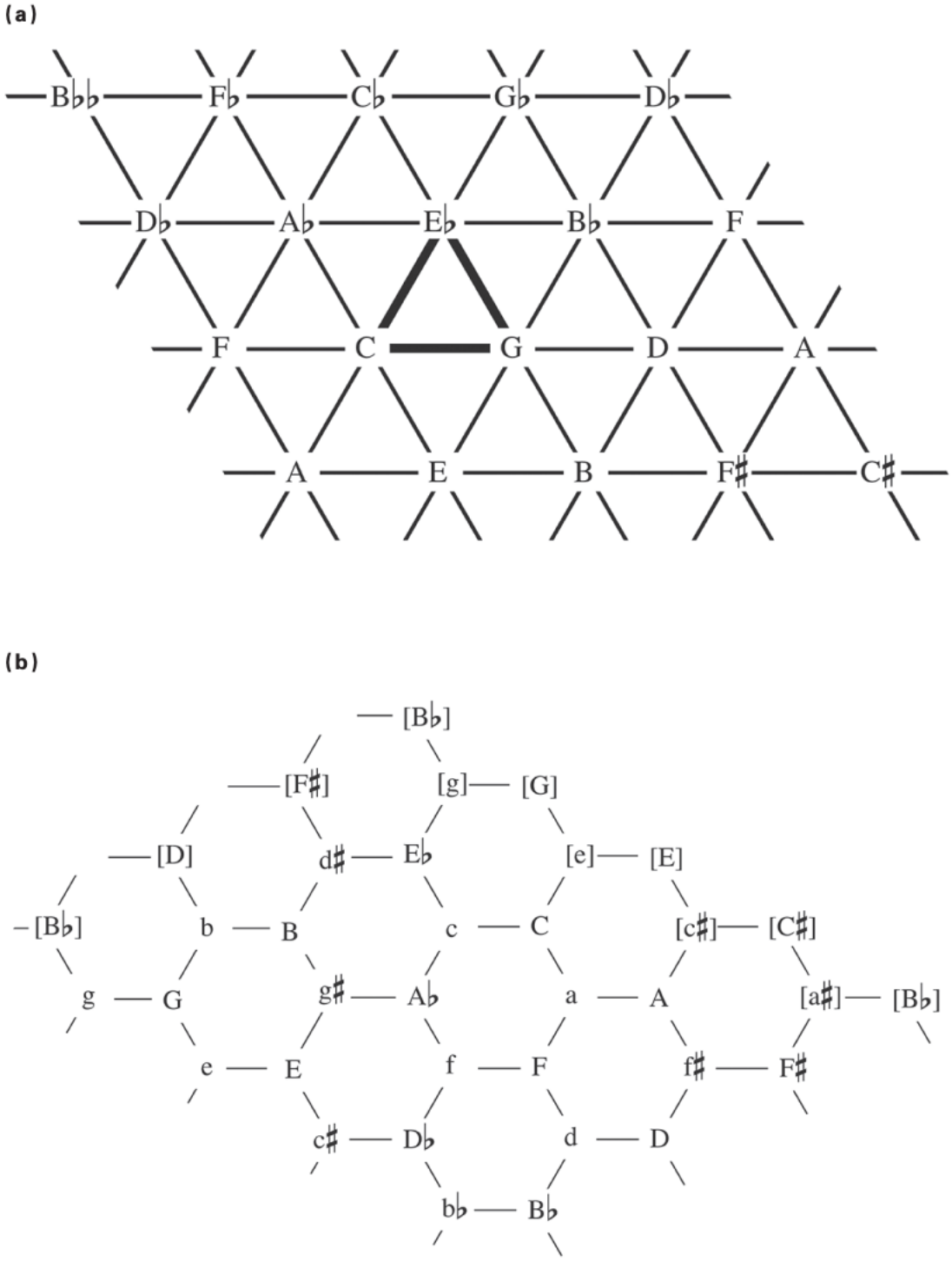

Рис. 4.3: Note-based и chord-based подходы. [9]

4.5.2 Преобразования в сгенерированных треках

Для того, чтобы разметить полученные сэмплы, было использовано приложение Sonic Visualizer и плагин Chordino, позволяющий получить лейблы аккордов прямо на спектрограмме.

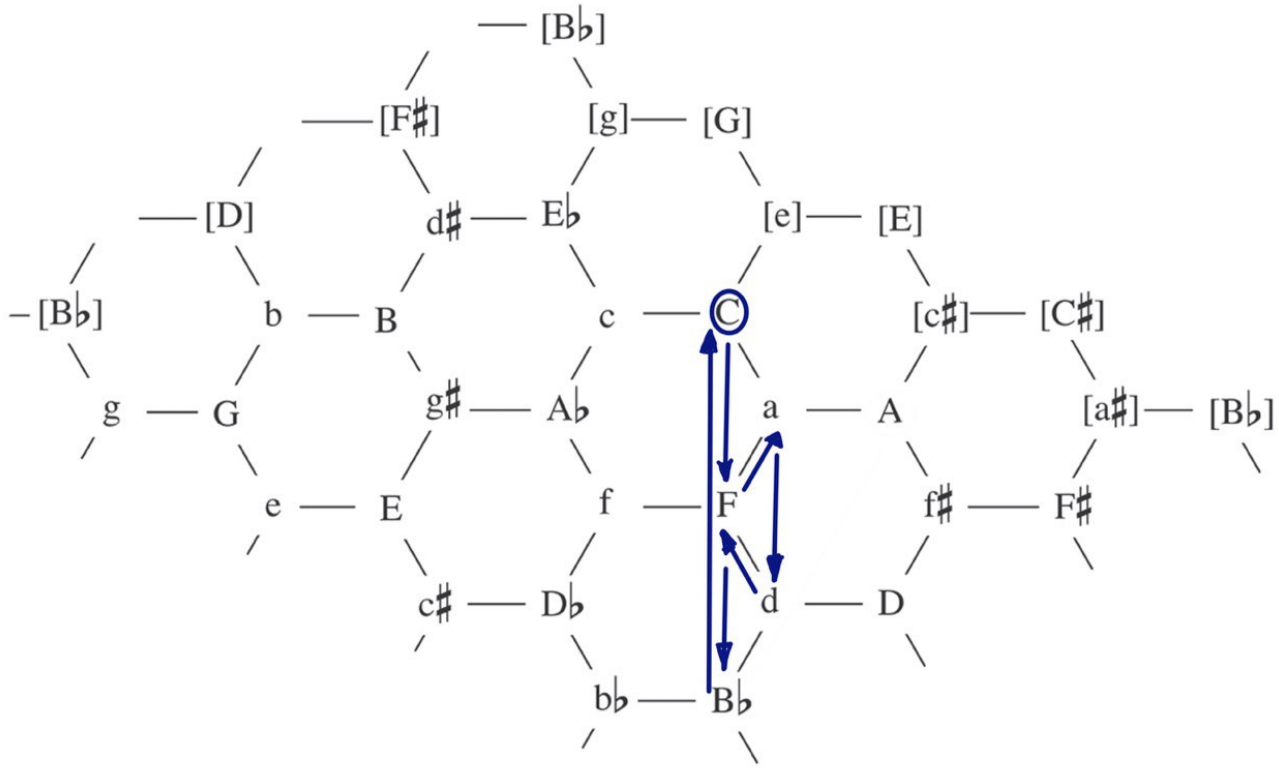

Рис. 4.4: Valence = -0.8, arousal = -0.8. Спокойный трек с печальной мелодией. Много LL-переходов – прыжков от текущей тоники к текущей доминанте (d-a, Bb-F, F-C). Также присутствуют L и LLLL-переходы.

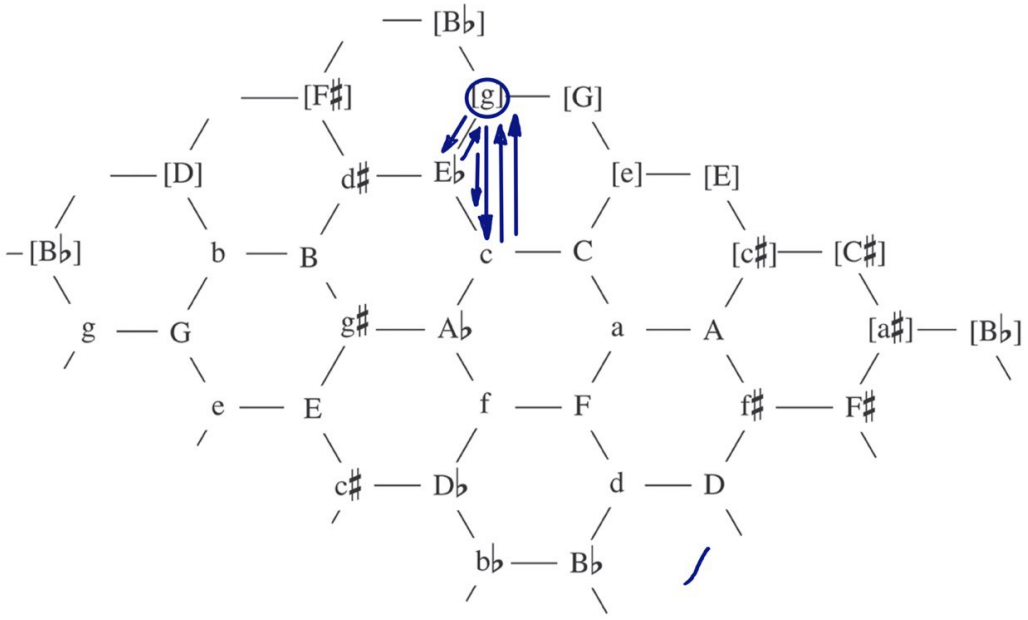

Рис. 4.5: Valence = -0.8, arousal = 0.8. Энергичный трек с печальной мелодией. Много LL-переходов (есть считать, что до – это тоника, с будет тоническим трезвучием, Eb – побочное трезвучие для терции (3 ступени), g – доминантовое трезвучие.)

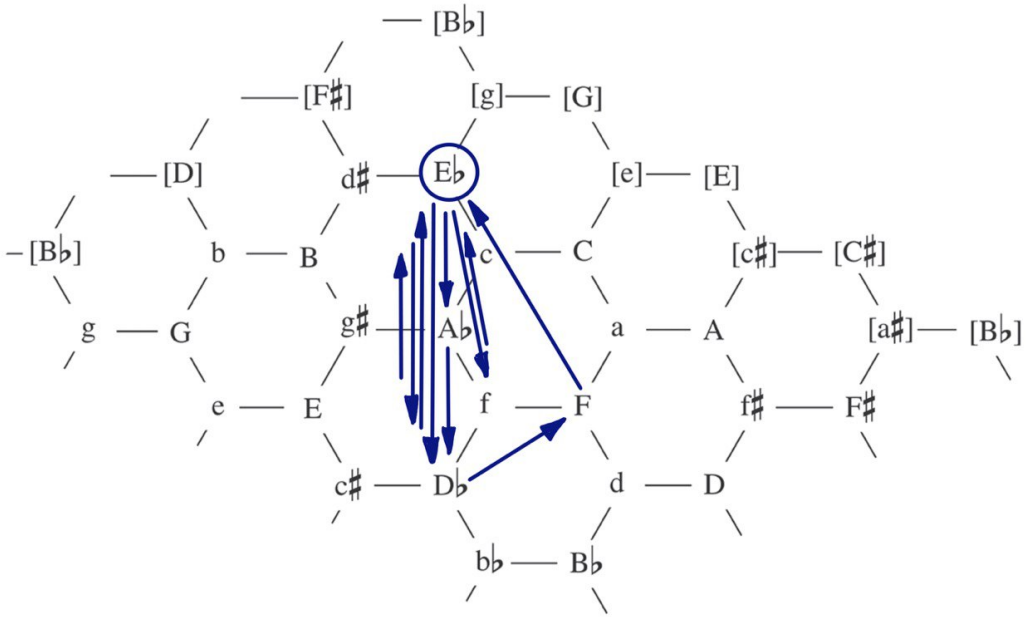

Рис. 4.6: Valence = 0.8, arousal = -0.8. Спокойный трек с жизнерадостной мелодией. Снова много LL- и LLLL-переходов, также присутствует составное преобразование LP (от Db к f, от f к F).

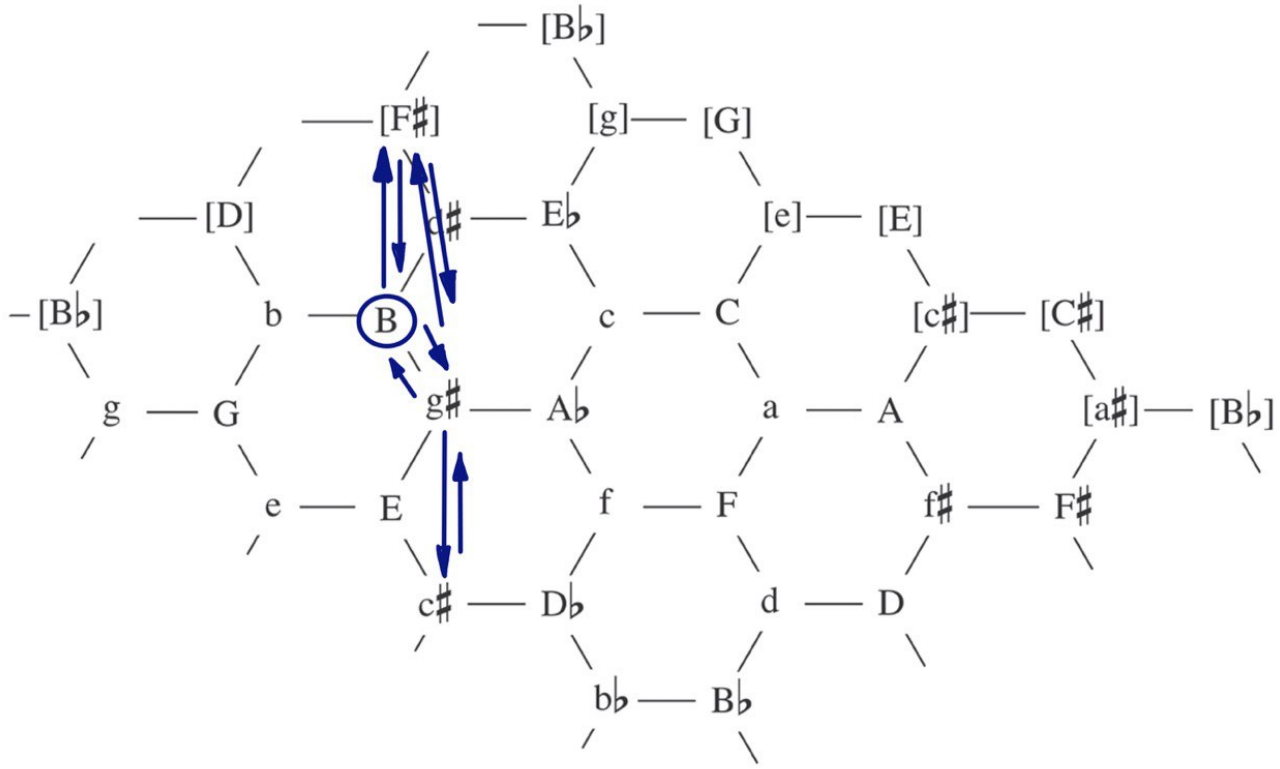

Рис. 4.7: Valence = 0.8, arousal = 0.8. Энергичный трек с жизнерадостной мелодией. Снова присутствуют L- и LL-переходы (g#-B, B-F#, c#-g#).

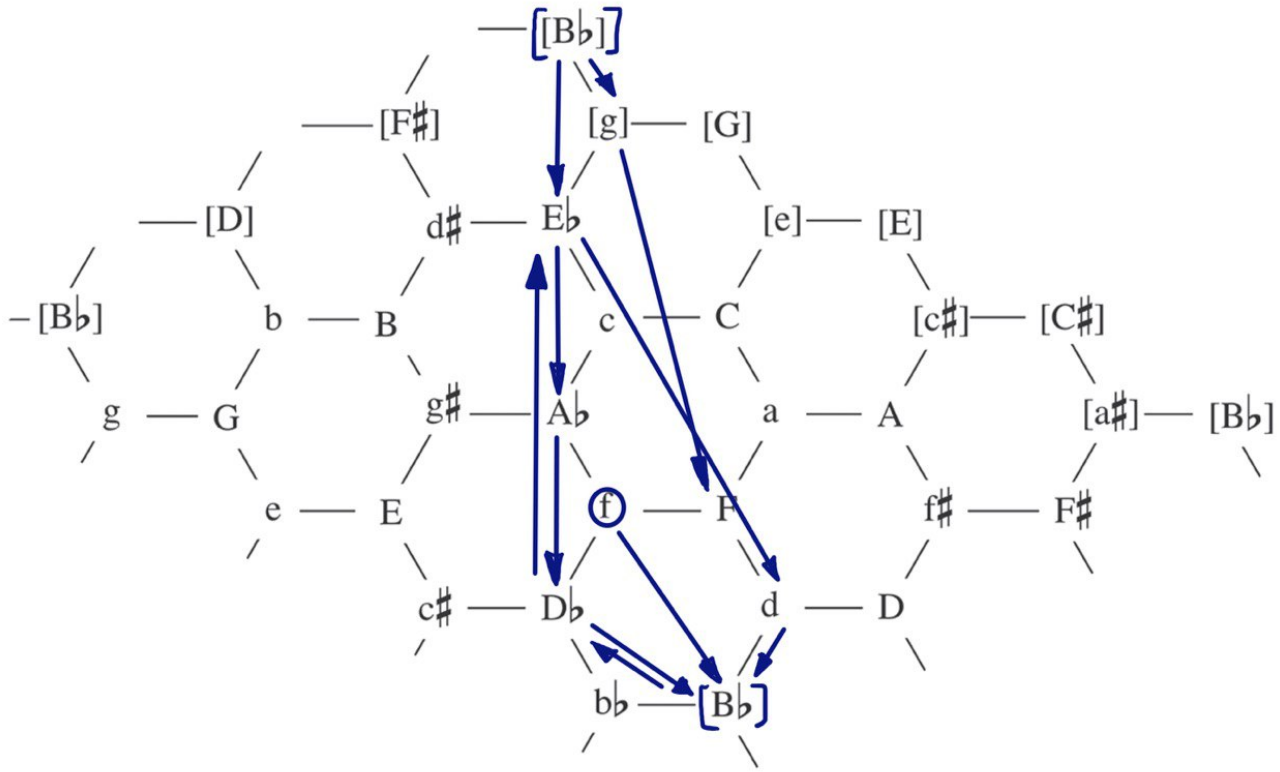

Рис. 4.8: Valence = 0.0, arousal = 0.0. Нейтральный по энергичности и позитивности трек. Большое количество обратных LL- и L- переходов (на квинту и на терцию вниз, а не вверх), также присутствуют LP и PL (Bb-Db, Db-Bb).

Выяснилось, что большая часть преобразований, обнаруженных в сгенерированных треках – это L- или LL-переходы. Это объяснимо – в музыке, написанной людьми, LL-переход также встречается очень часто, поскольку он является прыжком через квинту – то есть прыжком в соседнюю по кварто-квинтовому кругу тональность. L-переход тоже присутствует нередко: это прыжок в тональность, параллельную соседу по кварто-квинтовому кругу, – популярный переход в «человеческой» музыке.

4.6 Эксперимент

В целом структура эксперимента была следующей: в результате работы модели были получены несколько треков, сгенерированных по разным значениям valence и arousal (конкретно: valence 0.8, arousal 0.8; valence 0.8, arousal -0.8; valence -0.8, arousal 0.8; valence -0.8, arousal -0.8; valence 0, arousal 0). Эти треки были прослушаны человеком, сидящим в шлеме для ЭЭГ, а полученные электроэнцефалограммы переданы в модель-регрессор. Далее было произведено сравнение исходных значений valence и arousal (то есть те, что были переданы в модель-генератор) и тех, что были получены в результате работы регрессора на ЭЭГ-данных.

Для того, чтобы правильно подобрать количество электродов и их монтаж, были проанализированы уже существующие подходы к сбору ЭЭГ-данных: DREAMER, DEAP, AMIGOS (все в закрытом доступе) и DENS (в открытом доступе). Результаты получились следующими:

Название датасета	Количество электродов	Sampling rate, Hz	Дополнительная информация
DEAP	32	512	Электроды были расположены в соответствии с системой 10-10,
DREAMER	16 (14, если не считать A1 и A2)	128	Электроды были расположены в соответствии с системой 10-20: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4, A1 и A2
AMIGOS	14	128	Emotiv EPOC Neuroheadset, электроды были расположены в соответствии с системой 10-10: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4
DENS	128	250	Geodesic EEG System 400, данных о конкретном расположении электродов нет

Таблица 4.1: Датасеты с размеченными ЭЭГ-данными

Чтобы расположение электродов было детерминированным, а эксперименты были воспроизводимыми, учёными были созданы системы 10%-20% и 10%-10% (модификация системы 10%-20%). Система строится следующим образом: для определения расположения элек-

тродов используется медианная линия, которая проходит от переносицы до затылка и делится на 10 равных отрезков. Первый и последний электроды располагаются на расстоянии 10% от общей длины линии от иниона или назиона (см. рис. 4.2). Каждый следующий электрод располагается на расстоянии 20% от общей длины линии от предыдущего электрода. По медианной линии накладывают 5 электродов. Кроме того, на центральной линии, проходящей через наружные слуховые проходы, накладывают по два электрода на каждое полушарие и макушечный электрод. Линии, параллельные медианной и центральной линиям, называются парасагиттальными и височными. На парасагиттальной линии накладывают 5 электродов, а на височных — по 3 электрода. Всего на поверхность головы накладывается 21 электрод.

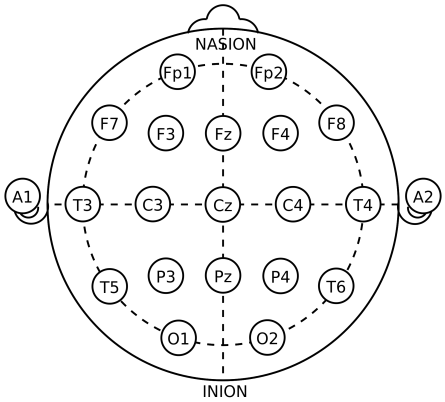

Рис. 4.9: Международная система 10-20 – стандартная система размещения электродов на поверхности головы.

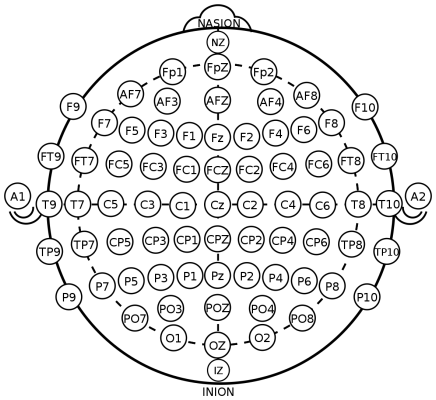

Рис. 4.10: Международная система 10-10 – модификация системы 10-20. Каждый следующий электрод располагается на расстоянии 10% от общей длины линии от предыдущего электрода. Всего на поверхность головы накладывается 73 электрода.

В итоге в эксперименте была использована конфигурация электродов из DEAP – стандартная модель из 32 электродов, а частота дискретизации – 1000 Гц. Испытуемым были прослушаны 32 стимула, каждый длиной 30 секунд; из них 16 сэмплов были сгенерированы моделью, а 16 взяты из датасета DEAP. Между стимулами была 10-секундная пауза для того,

чтобы «впечатления» от стимулов не накладывались друг на друга. После прослушивания сэмпла испытуемому предлагалось самостоятельно оценить valence и arousal стимула, эти оценки сравнены с реальными значениями valence и arousal и предсказаниями регрессора.

4.6.1 Выделение признаков из данных ЭЭГ

Данные ЭЭГ после эксперимента были получены в формате .edf – European Data Format. Чтобы работать с этими файлами, была использована библиотека MNE, позволяющая выделять из файла информацию о сигналах, полученных с отдельных электродов, а также позволяющая выделить только сигналы отдельных частот. Кроме того, есть возможность визуализировать как сами сигналы, так и, к примеру, дополнительную информацию об этих сигналах:

- Изображение топологических карт для отображения распределения активности в заданном временном окне или частотном диапазоне
- Изображение спектрограмм для анализа изменений мощности в зависимости от частоты и времени
- Изображение силы связи между различными мозговыми регионами на основе коэффициента корреляции или когерентности
- 3D-визуализация мозговой активности на поверхности коры мозга с помощью метода Курлова-Нетцеля

Сами данные ЭЭГ, то есть сигналы, полученные с разных электродов, перед анализом нужно отфильтровать – избавиться от артефактов и отобрать только интересуемый диапазон частот. После этого данные можно преобразовывать некоторым образом и получать новые признаки, полезные для обучения регрессоров, полносвязных нейронных сетей и других моделей, предназначенных для предсказания valence и arousal по данным ЭЭГ. К преобразованиям, которые можно применить для ЭЭГ-данных, относятся:

- Преобразование Фурье (Fourier Transform) – это метод, который позволяет преобразовать сигнал в частотный домен, представив его в виде суммы гармонических функций различных частот. Это позволяет анализировать сигнал в частотной области, что может быть полезно для обнаружения особенностей сигнала (к примеру, частот, доминирующих в определенных состояниях).

- Вейвлет-преобразование (Wavelet Transform) – это метод, который позволяет анализировать сигналы, учитывая как частотные, так и временные характеристики. Он преобразует сигнал в пространство коэффициентов, которые соответствуют коэффициентам разложения вейвлет-функций (график вейвлет-функций выглядит как волнообразные колебания с амплитудой, уменьшающейся до нуля вдали от начала координат). Это позволяет анализировать сигнал во временной и частотной областях одновременно.

- Амплитудно-частотная модуляция (Amplitude Modulation Frequency Modulation, AM-FM) – это метод, который позволяет анализировать сигналы во временном и частотном доменах. Он основан на модуляции частоты и амплитуды сигнала, что позволяет извлечь информацию о фазовых и амплитудных модуляциях в сигнале.

- Преобразование Хилберта (Hilbert Transform) – это метод, который позволяет извлекать амплитуду и фазу из сигнала. Он преобразует сигнал в комплексную функцию, которая может быть использована для анализа амплитуды и фазы.

- Декомпозиция на компоненты низкого и высокого порядков (Independent Component Analysis, ICA) – это метод, который позволяет извлекать из сигнала независимые компоненты, каждая из которых может представлять собой отдельное физиологическое событие в сигнале.

Также важнейшим методом для анализа является Power Spectral Density (PSD). Это функция, которая показывает распределение мощности сигнала по частотам. PSD является одним из основных инструментов анализа частотных характеристик сигнала в обработке ЭЭГ.

Power Spectral Density может быть получена путем применения преобразования Фурье (FFT) к временному сигналу ЭЭГ. Результатом FFT является комплексный спектр, который может быть преобразован в PSD. PSD отображает мощность сигнала в зависимости от его частоты. Обычно ее измеряют в микроваттах в квадрате на герц или в децибелах. PSD также может использоваться для оценки изменений в частотных характеристиках ЭЭГ в ответ на различные стимулы или условия. Например, PSD может использоваться для изучения изменений в ЭЭГ во время сна и бодрствования, в ответ на лекарства или нарушения мозгового кровообращения.

Другим полезным преобразованием является оконное преобразование Фурье – модификация обычного преобразования. ЭЭГ-данные не-стационарные: это значит, что статистическая характеристика сигнала может меняться со временем. Если эти сигналы преобразовать в частотную область (frequency domain) с использованием обычного преобразования

Фурье, то это дает информацию о частоте, которая усредняется по всему сигналу ЭЭГ – теряются важные данные о том, когда именно проявляется данная частота. Если разделить сигнал на небольшие сегменты таким образом, что его можно рассматривать как стационарный, и сосредоточиться на свойствах сигнала в определенном сегменте, который называется окном, и применить на него преобразование Фурье, получится оконное преобразование Фурье (Short-Time Fourier Transform, STFT). Окно перемещается по всему сигналу и применяет преобразование Фурье для нахождения спектрального содержания отдельного сегмента. Это дает представление о характере изменяющихся во времени спектральных характеристик сигнала. Посмотрим на зависимость STFT и DFT – дискретного преобразования Фурье. Рассмотрим сигнал (дискретный относительно времени), где L – длина сигнала:

$$\mathrm{x} : [0 : \mathrm{L} - 1] = \{0, 1, \ldots, L - 1\} \to \mathrm{R}$$

Уравнение для дискретного преобразования Фурье можно записать следующим образом.

$$\widehat{x}(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi nk/N}$$

Где k $\in [0,\ \mathrm{K}]$ и K – это индекс частоты относительно частоты Найквиста – величина, равная половине частоты дискретизации. N – это длительность фрейма. Уравнение возвращает комплексные коэффициенты Фурье для k-го частотного бина. Эти коэффициенты предоставляют два параметра: фазу и амплитуду. Для STFT рассматривается дополнительный параметр – размер шага окна (H), который определяет размер сдвига окна. Функция выборки окна ω определена на $\omega : [0,\ \mathrm{N} - 1] \to \mathrm{R}$. Тогда STFT может быть определено как:

$$S(m, k) := \sum_{n=0}^{N-1} x(n + mH)\omega(n)e^{(-i2\pi kn/N)}$$

Где m $\in [0, \mathrm{M}]$ и M это максимальный индекс для фрейма $\mathbf{M} = \left\lfloor \frac{L-N}{H} \right\rfloor$. STFT – не только функция от k, но и от m, то есть от момента времени. Следовательно, функция возвращает коэффициенты Фурье для k-й частоты (k-го бина частоты) в m-й момент времени (m-й бин времени). Спектрограммы – это просто мощность сигнала из STFT, взятая в квадрате.

$$\chi(m, k) := |S(m, k)|^2$$

Спектрограмму можно представить в виде 2D-изображения, где горизонтальная ось пред-

ставляет время (m), а вертикальная ось – частотные бины (k). Количество частотных бинов – это $\frac{N}{2} + 1$. Количество временных фреймов – это $\frac{L-N}{H} + 1$. Само значение функции $\chi(m, k)$ представляет собой интенсивность или цвет в точке (m, k). Спектрограммы можно использовать для обучения различных моделей, в частности, свёрточных нейронных сетей.

4.6.2 Предложенный регрессор и результаты эксперимента

Подходы для определения valence и arousal по ЭЭГ-данным могут быть разными. Во-первых, могут быть использованы классические методы машинного обучения – градиентный бустинг (XGBoost, CatBoost, и др.) или случайный лес (можно обучить один регрессор предсказывать valence, а второй – предсказывать arousal). Кроме того, можно использовать подходы из глубинного обучения – различные нейронные сети. Для анализа полученных ЭЭГ-данных была использована модель, обученная на признаках, выделенных из ЭЭГ-данных по принципам, описанным в предыдущей подглаве: было применено оконное преобразование Фурье и PSD – Power Spectral Density. Архитектура модели – сверточная рекуррентная нейронная сеть (CRNN). CRNN сочетает в себе сильные стороны CNN и RNN за счет включения как сверточных, так и рекуррентных слоев. Она может фиксировать локальные закономерности в спектрограммах, используя сверточные слои, а также моделировать временные зависимости, используя рекуррентные слои. Обучалась модель на датасете DEAP, описанном ранее – данные в этом датасете описывают 40 экспериментов (trials), содержат информацию с 32 каналов (электродов) и каждый канал содержит информацию о 8064 сигналах. Сама модель состоит из двух слоёв (Conv2D, MaxPooling, ReLU) – свёрточная часть – и юнита GRU (Gated Recurrent Unit) – рекуррентная часть. Обучение модели длилось 30 эпох, была использована кросс-валидация (всего экспериментов 40, на 35 модель учится и на 5 происходит валидация). В качестве функции потерь была использована MSELoss, в качестве optimizer – Adam с learning rate = 0.001, в качестве scheduler – CosineAnnealingLR с $\mathrm{T}_{\max} = 30$.

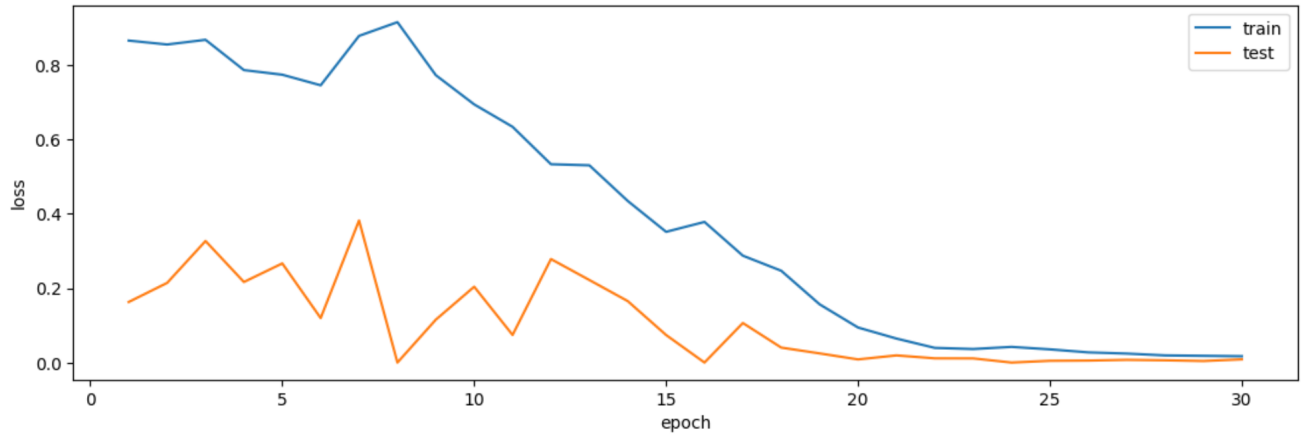

Рис. 4.11: Графики MSE для тренировочных и тестовых данных.

4.6.3 Результаты эксперимента

Разметка от испытуемого получилась следующей:

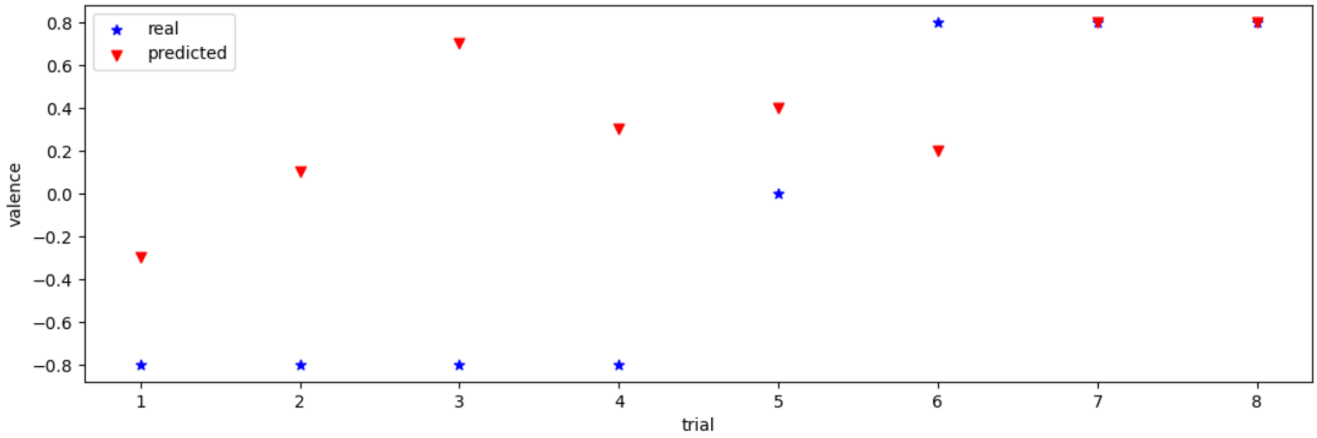

Рис. 4.12: График valence. Красным показаны предсказанные испытуемым значения, синим – параметры, переданные в модель.

Видно, что высокие значения valence были оценены испытуемым с небольшой погрешностью, оценка же для низких значений оказалась сильно завышенной. Это может быть связано с несколькими причинами: во-первых, модель обучалась на midi-файлах, полученных при обработке песен, то есть, данные о тексте, сильно влиявшие на valence, были потеряны. Написанная человеком музыка с низкой оценкой valence в отрыве от текста может быть воспринята как более позитивна. Во-вторых, midi-преобразование сопряжено с использованием soundfonts. Сами midi-файлы не содержат звуковых значений, но содержат лишь описание этого звука (инструкции о том, как этот звук проигрывать). Soundfont представляет собой набор предзаписанных «midi-синтезаторов», имитирующих инструменты. Поэтому, к примеру, гитара из рок-композиции с низкой valence и гитара из баллады с высокой valence могли

звучать примерно одинаково. В-третьих, представление о высокой валентности у людей может быть похожим: мажорные позитивные треки воспринимаются примерно одинаково, в то время как степень того, насколько звук кажется «грустным» или неприятным у разных людей может отличаться.

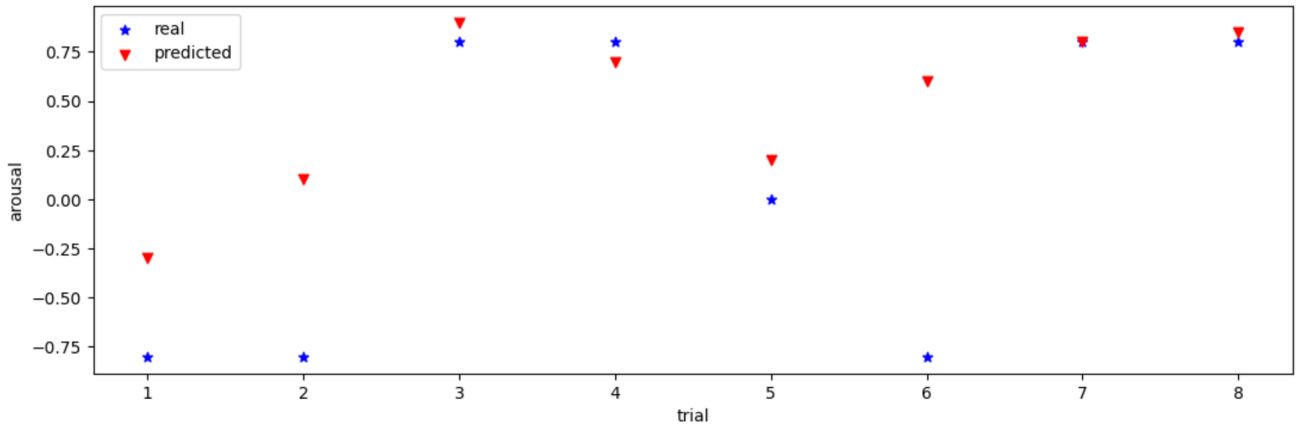

Рис. 4.13: График arousal. Красным показаны предсказанные испытуемым значения, синим – параметры, переданные в модель.

Видно, что, как и в случае с valence, высокие значения arousal были определены достаточно точно, для более низких же значений оценка испытуемого оказалась завышенной: модель склонна генерировать треки, которые ощущаются более энергичными. Возможно, это связано с тем, что arousal модель «учит» в том числе через плотность и количество нот в произведении и передаёт arousal не через медленный темп, а через более простую относительно количества инструментов и звуков композицию.

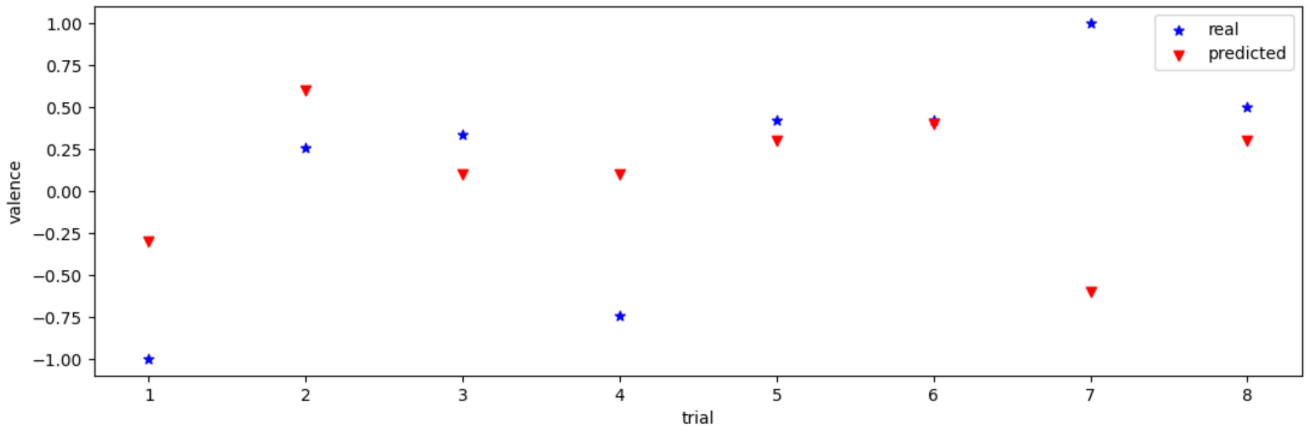

Рис. 4.14: График valence. Красным показаны предсказанные испытуемым значения, синим – значения из разметки датасета DEAP.

Видно, что valence определена точнее, чем в случае со сгенерированными треками, но погрешность всё ещё большая, а тенденция завышать значения тоже присутствует.

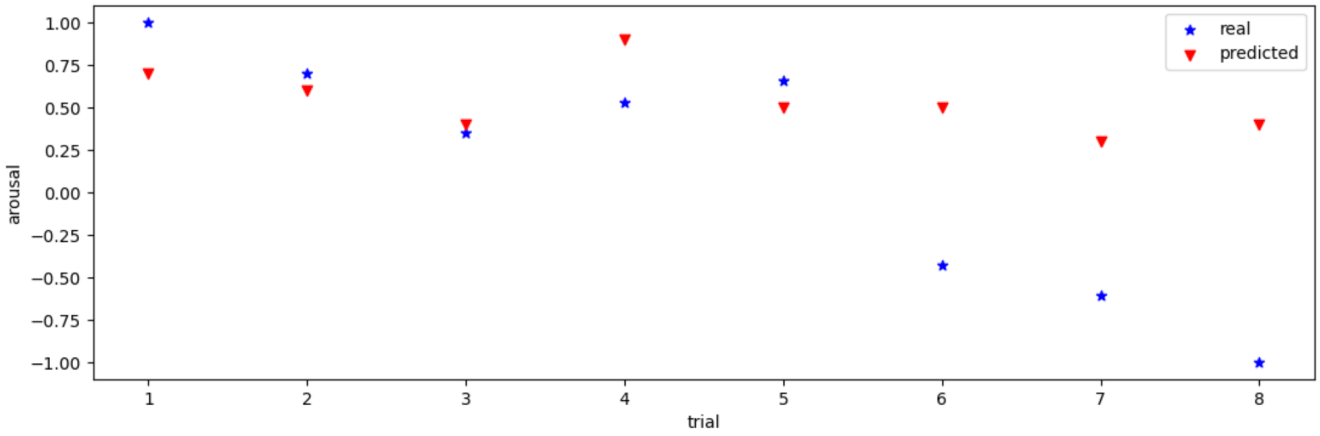

Рис. 4.15: График arousal. Красным показаны предсказанные испытуемым значения, синим – значения из разметки датасета DEAP.

Видно, что, как и в случае со сгенерированными треками, высокие значения arousal были определены с небольшой погрешностью, для более низких же значений оценка испытуемого оказалась завышенной.

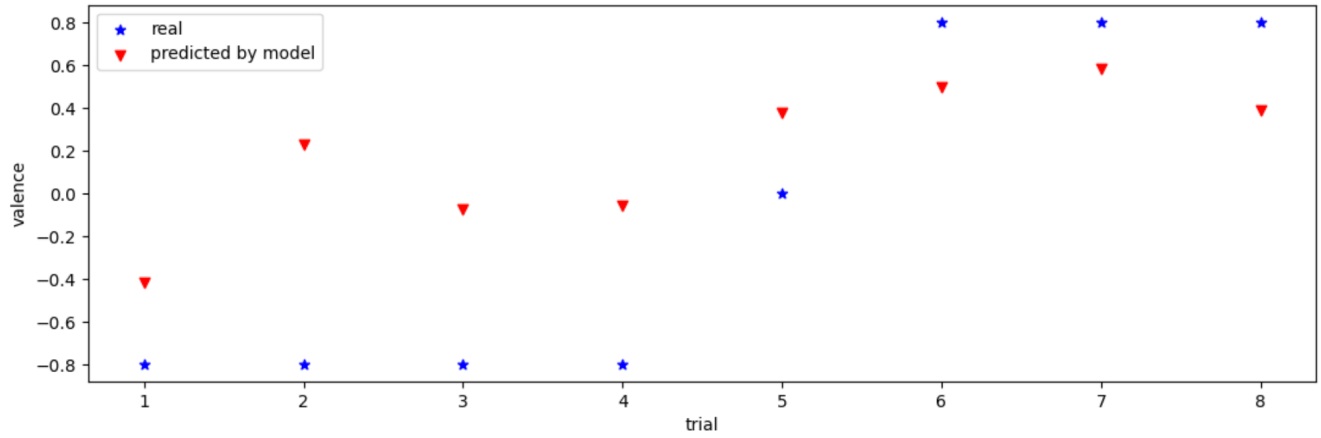

Рис. 4.16: График valence. Красным показаны предсказанные регрессором значения, синим – значения, переданные в генеративную модель.

Видно, что модель достаточно консервативная и предсказывает среднее значение valence для большинства треков. Кроме того, значения оценок снова завышенные.

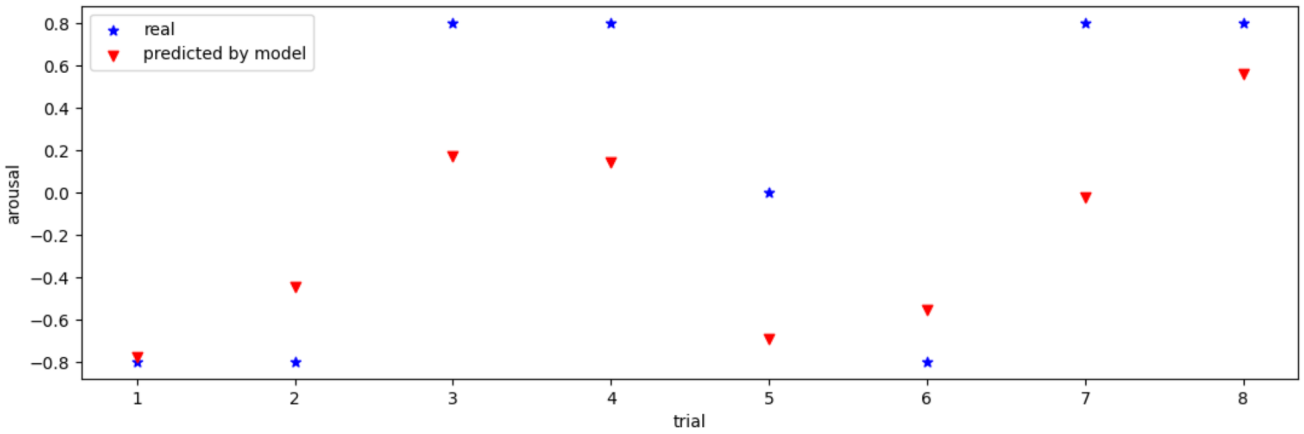

Рис. 4.17: График arousal. Красным показаны предсказанные регрессором значения, синим – значения, переданные в генеративную модель.

Видно, что модель достаточно консервативная и предсказывает среднее значение arousal для большинства треков. Оценки оказались скорее заниженными, а не завышенными.

4.7 Заключение и дальнейшая работа

В результате данной работы получена модель, позволяющая с некоторой погрешностью генерировать музыку определённой эмоциональной окраски, а также CRNN-регрессор, позволяющий по данным ЭЭГ предсказывать значения valence и arousal. Кроме того, были сгенерированы и проанализированы на предмет нео-римановых преобразований несколько сэмплов с разными исходными значениями valence и arousal. В качестве дальнейших исследований можно предложить несколько работ:

- Более глубокое или широкое исследование нео-римановых преобразований и других геометрических структур в музыке.
- Улучшение качества текущих моделей или предложение новых архитектур для генерации музыки.
- Предложение новых регрессоров для определения valence и arousal.
- Дальнейшая работа с полученными в результате эксперимента ЭЭГ-данными: более сложный или подробный анализ.
- Проведение новых экспериментов с моделями и ЭЭГ: изучение не только valence и arousal, но и других эмоциональных характеристик.

Список литературы

[1] Filipe Galvão, Soraia M Alarcão, and Manuel J Fonseca. Predicting exact valence and arousal values from EEG. *Sensors (Basel)*, 21(10):3414, May 2021.

[2] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. MOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In *Proc. Int. Society for Music Information Retrieval Conf.*, 2021.

[3] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. URL: https://doi.org/10.1561%2F2200000056, doi:10.1561/2200000056.

[4] S. Koelsch. *Brain and Music*, chapter 4.1, 8, 12.6, 12.7. John Wiley & Sons, Ltd, 2012.

[5] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This Time with Feeling: Learning Expressive Musical Performance. *arXiv preprint, arXiv:1808.03715*, 2018. URL: https://arxiv.org/abs/1808.03715, doi:10.48550/ARXIV.1808.03715.

[6] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.*, 17(3):715–734, 2005.

[7] Serkan Sulun, Matthew E. P. Davies, and Paula Viana. Symbolic music generation conditioned on continuous-valued emotions. *IEEE Access*, 10:44617–44626, 2022. doi:10.1109/ACCESS.2022.3169744.

[8] Alexey Tikhonov and Ivan P. Yamshchikov. Music generation with variational recurrent autoencoder supported by history. *CoRR*, abs/1705.05458, 2017. URL: http://arxiv.org/abs/1705.05458, arXiv:1705.05458.

[9] Dmitri Tymoczko. The generalized tonnetz. *Journal of Music Theory*, 56(1):1–52, April 2012. doi:10.1215/00222909-1546958.

[10] James S. Walker and Gary W. Don. A Geometric Analysis of the Harmonic Structure of «In My Life». *arXiv preprint arXiv:2008.11749*, 2020. URL: https://arxiv.org/abs/2008.11749, doi:10.48550/ARXIV.2008.11749.