**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего образования**

**«КАЗАНСКИЙ ПРИВОЛЖСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**

ВЫСШАЯ ШКОЛА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Направление: 09.04.04 Программная инженерия

Профиль: Разработка программно-информационных систем

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

# Воспроизведение звуков при взаимодействиях 3D-объектов в виртуальных средах

Студент 2 курса

группы _____

«___»_____2019 г. Сидоров И.Н.

Научный руководитель
Старший преподаватель
кафедры программной инженерии

«___»_____2019 г. Кугуракова В.В.

Директор Высшей школы ИТИС

«___»_____2019 г. Хасьянов А.Ф.

Казань – 2019 г.

## Содержание

ВВЕДЕНИЕ ..... 5

**1 Сравнение с другими методами** ..... 7

2 Физическая, аудиовизуальная генерирующая модель ..... 8

**2.1 Обзор** ..... 8

**2.1.1 Физический движок** ..... 8

**2.1.1.1 Bullet Physics Library** ..... 9

**2.1.2 Графический движок** ..... 9

**2.1.2.1 Blender** ..... 10

**2.1.3 Звуковой движок** ..... 10

**2.1.3.1 Метод конечных элементов** ..... 10

3 Столкновение с вибрацией ..... 12

**3.1 Решение волнового уравнения** ..... 13

**3.2 Оффлайн-онлайн разложение** ..... 14

4 Набор данных Sound-20K ..... 15

**4.1 Формы** ..... 15

**4.2 Материалы** ..... 16

**4.3 Сценарии** ..... 16

**4.4 Набор данных** ..... 17

5 Анализ данных ..... 19

6 Восприятие объекта с аудио-визуальными данными ..... 21

**6.1 Данные** ..... 21

**6.1.1 Physics 101** ..... 21

**6.1.2 Greatest Hits** ..... 21

**6.2 Методы** ..... 22

**6.3 Распознавание материалов** ..... 24

**6.4 Распознавание атрибутов формы** ..... 25

7 Переход от синтетических к реальным данным ..... 27

**7.1 Распознавание материала по аудио данных** ..... 27

**7.2 Распознавание атрибутов формы** ..... 29

8 Заключение ..... 30

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ ..... 31

## Глоссарий

Психоакустика – научная дисциплина, изучающая психологические и физиологические особенности восприятия звука человеком. Мультимодальное обучение включает в себя связывание информации из нескольких источников.

ShapeNet – это крупномасштабный аннотированный набор данных трехмерных фигур.

Рендеринг – термин в компьютерной графике, обозначающий процесс получения изображения по модели с помощью компьютерной программы.

МКЭ (метод конечных элементов) – это численный метод решения дифференциальных уравнений с частными производными, а также интегральных уравнений, возникающих при решении задач прикладной физики.

Bullet – физический движок реального времени.

МГЭ (метод граничного элемента) – метод решения краевой задачи, в котором благодаря использованию формул Грина, она сводится к интегральному уравнению на границе расчетной области.

ResNet-18 – это сверточная нейронная сеть, которая обучается более чем миллиону изображений. Сеть состоит из 18 слоев и может классифицировать изображения по 1000 категориям объектов, таким как клавиатура, мышь, карандаш и многие животные.

SoundNet-8 – это сверточная нейронная сеть, состоящая из 8 слоев, которая обучена на не маркированных аудиофайлах.

Стохастический градиентный спуск – относится к оптимизационным алгоритмам и нередко используется для настройки параметров модели машинного обучения.

# ВВЕДЕНИЕ

В настоящее время активно развивается разработка визуального контента и тактильной составляющей виртуальной реальности. Но слуховая часть слабо разрабатывается либо игнорируется, а это 10% восприятия человеческого мозга, и мы теряем возможность полного погружения человека в виртуальную среду.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Изучить ранее применяемые методы.
2. Определить подходы к созданию генерирующей модели.
3. Определить дизайн генерирующей модели.
4. Изучить методы для реализации генерирующей модели.
5. Реализовать генерирующую модель.
6. Оценить результативность методов.

В данной дипломной работе предлагается альтернативный подход к преодолению трудностей в синтезировании аудиовизуальных данных 3D объектов [9]. Синтезированные звуки будут естественно рассеиваться, а также можно генерировать звук каждого объекта независимо. Структура синтеза данных состоит из трех основных генеративных моделей: физический движок, графический движок и звуковой движок. Физический движок принимает формы объектов, свойства материала и начальные условия в качестве входных данных, а затем имитирует их последующие движения и столкновения. Графический движок отображает видео на основе имитируемого движения объекта. Звуковой движок основан на работе линии звуковой симуляции [5]. Он объединяет предварительно вычисленные данные о форме объектов и их столкновения для точного синтеза звука. Основанная на физике модель генератора контрастирует с недавними методами синтеза нейронных звуков [7].

Основной задачей синтеза данных является достижение подлинного звучания. Чтобы синтетический звук был реалистичным, предполагается проводить сравнения синтезированного звука с реальными записями полученных в экспериментальных условиях. Для этого будет использоваться база данных Sound20K.

# 1 Сравнение с другими методами

Существует много работ в области слухового восприятия и психоакустики. По ним было разработано много проектов связанные с озвучкой 3D графики.

Например, такой проект как ShapeNet [2], содержащий большое количество 3D-моделей из множества семантических категорий и организует их в соответствии с таксономией WordNet. Это набор данных, предоставляющий множество семантических аннотаций для каждой 3D-модели, таких как последовательные жесткие выравнивания, детали и плоскости двусторонней симметрии, физические размеры, ключевые слова, а также другие запланированные аннотации. Аннотации находятся в общедоступным веб-интерфейсе, позволяющий визуализировать данные атрибутов объектов, управлять анализом геометрических данных и обеспечивать эталонное тестирование для исследований в области компьютерной графики. Так же были попытки вывода свойств материала объекта от звука попадания по нему барабанной палочкой [7].

По сравнению с этими работами, в разрабатываемом методе будет изучен синтез как визуальных, так и аудио данных. Так же канал синтеза звука будет основан на нескольких работах по имитации звука [6, 4, 5, 1, 8] в компьютерной графике. Многие мультимодальные алгоритмы обучения фокусировались на совместном изучении визуальных и текстовых данных.

## 2 Физическая, аудиовизуальная генерирующая модель

### 2.1 Обзор

Предлагается проект генеративной модели, созданный на основе реального мира: когда объекты, двигаются на сцене, применяются физические законы и физические события, такие как столкновения. Мы видим изображение объекта, слышим его вибрации, вызванные физическими событиями. Поэтому генеративная модель состоит из физического движка, связанный графическим и аудио движками как показано на рисунке 1.

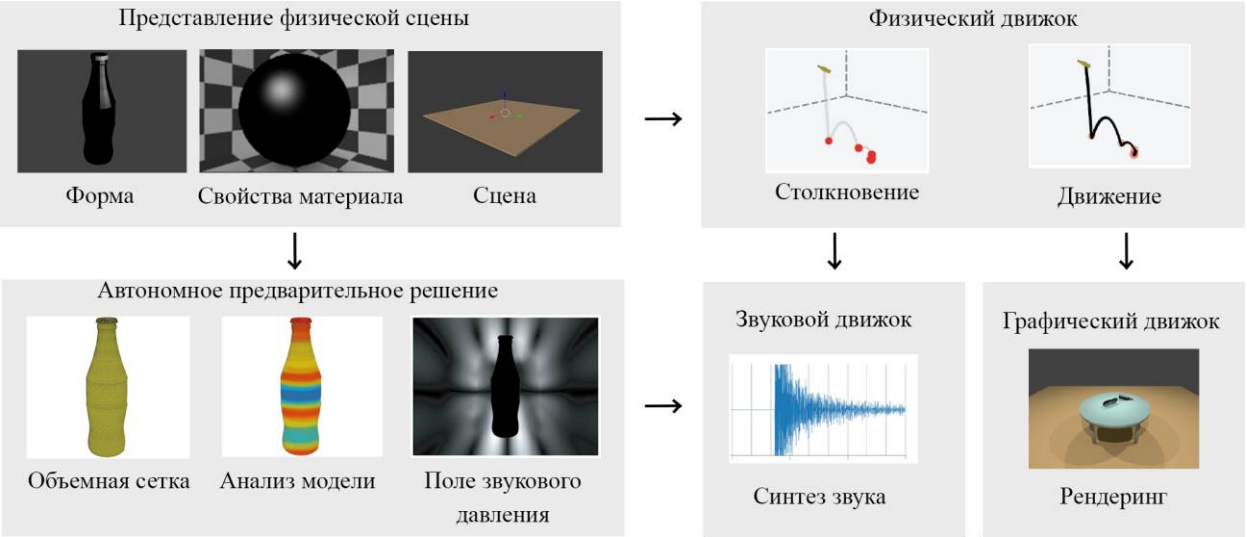



Рисунок 1 – Генеративная модель для аудиовизуальных сцен.

#### 2.1.1 Физический движок

Физический движок служит основным компонентом для генерации физических событий. Для аудио и визуальных данных, важными событиями являются движение и столкновение. Движение определяет положение объектов в любой момент времени, а столкновение определяет, как реагируют вибрации в объектах. С этой целью будет использоваться механизм физического моделирования с открытым исходным кодом Bullet [3], который обеспечивает работу графического движка и звукового движка. Для точности,

время шага для физического моделирования будет установлено на 1/300 с. Выходные данные физического движка - это информация о движении, которая состоит из центральной позиции объекта, и информация о столкновении, которая состоит из положения, направления и амплитуды столкновения.

#### 2.1.1.1 Bullet Physics Library

Это физический движок реального времени, который распространяется под свободной лицензией zlib. Скомпилированная библиотека и исходный код движка размещены в свободном доступе на хостинге Google Code. Bullet активно используется в компьютерных играх, фильмах, программах трёхмерного моделирования, как компонент других игровых движков и во множестве других более специфичных случаях. «Bullet» используется компанией AMD в рамках программы Open Physics Initiative, которая предполагает создание нового физического движка на базе «Bullet» и Digital Molecular Matter, который будет активно использовать аппаратное ускорение расчётов физических вычислений посредством OpenCL и DirectCompute.

### 2.1.2 Графический движок

В качестве графического движка будет использоваться Blender и его Cycles ray tracer [22] (этот метод позволяет упростить создание материалов, поддерживает глобальное освещение, и, наконец, более реалистичен в результате) как инструмент рендеринга. Конвейер рендеринга принимает информацию о движении, создаваемая физическим движком, как вход, т. е. центральные позиции объекта и его вращение при заданном времени, затем каждый объект будет преобразован соответственно. Для генерации видео графический движок настроен на рендеринг 30 кадров в секунду.

#### 2.1.2.1 Blender

Профессиональное свободное и открытое программное обеспечение для создания трёхмерной компьютерной графики, включающее в себя средства моделирования, анимации, рендеринга, постобработки и монтажа видео со звуком, компоновки с помощью «узлов» (Node Compositing), а также для создания интерактивных игр. В настоящее время пользуется наибольшей популярностью среди бесплатных 3D-редакторов в связи с его быстрым и стабильным развитием, которому способствует профессиональная команда разработчиков.

### 2.1.3 Звуковой движок

Звуковой движок, параллельный графическому движку, принимает информацию о столкновении объектов в качестве входных данных и звук слышен в положении камеры. Далее преобразует информацию о столкновении объекта с его вибрацией, используя МКЭ (метод конечных элементов), обнуляет вибрацию в качестве граничного условия для решения уравнения распространения волны, которое дает давление воздуха, то есть звук, в положении камеры.

#### 2.1.3.1 Метод конечных элементов

Суть метода заключена в его названии. Область, в которой ищется решение дифференциальных уравнений, разбивается на конечное количество подобластей (элементов). В каждом из элементов произвольно выбирается вид аппроксимирующей функции. В простейшем случае это полином первой степени. Вне своего элемента аппроксимирующая функция равна нулю. Значения функций на границах элементов (в узлах) являются решением задачи и заранее неизвестны. Коэффициенты аппроксимирующих функций обычно

ищутся из условия равенства значения соседних функций на границах между элементами (в узлах). Затем эти коэффициенты выражаются через значения функций в узлах элементов. Составляется система линейных алгебраических уравнений. Количество уравнений равно количеству неизвестных значений в узлах, на которых ищется решение исходной системы, прямо пропорционально количеству элементов и ограничивается только возможностями ЭВМ. Так как каждый из элементов связан с ограниченным количеством соседних, система линейных алгебраических уравнений имеет разрежённый вид, что существенно упрощает её решение.

## 3 Столкновение с вибрацией

Для преобразования столкновения в вибрацию объекта необходимо конвертировать поверхностную сетку объекта в объемную тетраэдрическую сетку с использованием поверхностей, проходящих через точки с одинаковым значением какой-либо величины и характеризующие распределение этой величины в пространстве (изоповерхность) [10], которые представляют собой исходную форму с N тетраэдрами. Затем применяется МКЭ для решения модальных форм объекта в диапазоне k-звуковые частоты. В частности, уравнение вибрации может записываться как:

$$M\frac{\partial^2 U}{\partial t^2} + Ku = t \quad (1)$$

где u ∈ $R^{3N}$ обозначает смещение элементарных узлов в 3D; M, K ∈ $R^{3N\times 3N}$ обозначает массу и жесткость матрицы системы соответственно, а f ∈ $R^{3N}$ означает применяемую внешнюю силу. Затем находим модальную матрицу Φ ∈ $R^{3N\times k}$ путем решения обобщенной задачи на собственные значения:

$$\Phi^T M\Phi = I,\ \Phi^T K\Phi = D \quad (2)$$

где D ∈ $R^{k\times k}$ является диагональной матрицей. Модальные формы затем определяются как столбцы матрицы Φ. Согласно процедурам модального анализа, в МКЭ [15] u может быть разложено на множество модальных коэффициентов c ∈ $R^k$ под модальным базисом Φ, т.е.

$$\Phi c = u \quad (3)$$

Тогда вызванная вибрация представляет собой линейную комбинацию модальной формы с модальными коэффициентами $c$. Чтобы убедиться в этом, необходимо заменить уравнение 1 в уравнение 3, которое дает:

$$\frac{\partial^2 c}{\partial t^2} + K\phi c = f \Rightarrow \frac{\partial^2 c}{\partial t^2} + Dc = \phi^T f \quad (4)$$

Поскольку D диагональна, система может быть преобразована в независимое синусоидное решение, а фактическая вибрация задается формулой Φc.

## 3.1 Решение волнового уравнения

Затем, учитывая вибрацию объекта, необходимо решить для фактического профиля давления воздуха в положении камеры. Базовое уравнение для описания этого процесса является волновым уравнением,

$$\left(\nabla^2 - \frac{1}{V^2}\frac{\partial^2}{\partial t^2}\right)p(x,t) = 0 \quad (5)$$

где p (x, t) обозначает давление воздуха в момент времени t и положение x. v обозначает скорость звука в воздухе. Предположим, что $P_i$ (x, t) решение граничного условия Неймана вибрация объекта:

$$\frac{\partial}{\partial n}P_i(x,0) = \phi_i, x \in s \quad (6)$$

где S - поверхность объекта. Решение уравнения 6 $\Sigma_i$ $c_i p_i$(x, $t$).. Поскольку каждая модальная форма имеет свою естественную частоту $\omega_i$, $p_i$(**x**, $t$) можно записать в виде $q_i(x)e^{-j\omega \boldsymbol{i} t}$, где j =√-1. Тогда волновое уравнение можно превратить в уравнение Гельмгольца для каждой модальной формы,

$$(\nabla^2 + k^2)q_i(x) = 0 \ \frac{\partial}{\partial n}q_i(x) = \phi, x \in S \quad (7)$$

Звуковой движок решает вышеупомянутое уравнение Гельмгольца с использованием метода граничных элементов (МГЭ) [11]. Построены два решения для уравнения 7: прямой интеграции с использованием библиотеки с открытым исходным кодом NiHu [12], и итеративное решение с быстрым многополюсным методом [13] ускоренным использованием FMM3D библиотеки, позволяющей оценивать потенциальные поля, управляемые уравнениями Лапласа, Гельмгольца в свободном пространстве.

Библиотека NiHu способна вычислять коэффициенты дискретных граничных интегральных операторов общим способом с произвольно определенными ядрами и функциональными пространствами. Шаблонное ядро NiHu определяет рабочий процесс общего алгоритма БЭМ, независимого от конкретного приложения. Ядро обеспечивает выразительный синтаксис, основанный на нотации оператора БЭМ, отражающей математику за граничными элементами в исходном коде C ++. Настраиваемая библиотека

компонентов содержит элементы, характерные для конкретных приложений, такие как различные методы численного интегрирования и методы регуляризации. Библиотека может быть использована для создания автономного приложения C ++ с использованием внешних библиотек с открытым исходным кодом или для компиляции набора инструментов Matlab через интерфейс MEX. Благодаря массовому использованию метапрограммирования шаблонов C ++, NiHu генерирует оптимизированные коды для конкретных приложений, в том числе разнородных задач.

## 3.2 Оффлайн-онлайн разложение

Простая реализация из вышеуказанной структуры будет неэффективной. [5] Предполагается ускорить её путем разложения аудио движка в онлайновую и автономную фазу. В автономной фазе вычисляются модальные формы объекта и давления воздуха на его поверхность, используя ускоренный МГЭ FMM3D библиотекой. Затем поверхностное давление воздуха для каждого режима аппроксимируется (научный метод, состоящий в замене одних объектов другими) давлением, создаваемым набор вибрирующих точек внутри объекта, его местоположение и амплитуда колебаний предварительно вычисляются и сохраняются.

# 4 Набор данных Sound-20K

“Sound-20K” – это набор данных состоящий из 20 378 аудио-видео пар и других связанных данных.

## 4.1 Формы

Отобраны 39 3D фигур из ShapeNet [2]. Все они монолитны, имеют многоячеистые сетки с последовательными внешними нормалями. Объекты подразделяются на 6 суперкатегорий и 21 категорию. Все фигуры в наборе данных установлены с длиной диагонали в единицах и были перемещены в их геометрический центр.

В Sound-20K аннотированы трехмерные атрибуты объектов. Но три из двенадцати атрибутов принадлежат для фигур, состоящих из несколько частей, поэтому они не были использованы. Оставшиеся атрибуты имеют: плоскость, отсутствие плоскости, цилиндр, шероховатость, пустоты, отверстия, тонкие стенки, зеркальную симметрию и кубическое соотношение сторон. [16]

Также для обеспечения реалистичности в базе данных учтены размеры, материал, форма объектов. Наиболее распространенный материал имеет маркировку на соответствующей текстуре. Образцы форм показаны на рисунке 2.

Рисунок 2 – Примеры объектов из набора данных Sound-20K

## 4.2 Материалы

Представлены семь часто используемых материалов и параметры их физических свойств для генерации звука и физического моделирования. Семь материалов включают: керамику, полистирол, сталь, древесноволокнистые плиты средней плотности (МДФ), дерево, поликарбонат и алюминий. Каждый материал имеет свои механические и акустические свойства: модуль Юнга (E), коэффициент Пуассона (ν), плотность (ρ), коэффициент трения (μ), коэффициент восстановления (e) и коэффициенты демпфирования Рэлея (α, β) [1]. ρ, μ, e каждого объекта подаются в физический движок для моделирования твердого тела; E, ν, ρ, α, β используются для предварительного расчета и синтеза звука в Интернете.

## 4.3 Сценарии

Выбраны 22 сценария с различными уровнями сложности. Каждый сценарий не зависит от обозначения формы или материала. Эти сценарии основаны на 9 базовых платформах трех уровней сложности. Как показано на

рисунке 2b, предварительно содержат несколько простых геометрий, в том числе лестницы, пандус, плоскую поверхность, угол между стенами и полусферическими оболочками. У более продвинутых сцен есть стол на земле и неровный пол. Для каждого каркаса были назначены несколько параметров инициализации с различным числом объектов, линейными и угловыми скоростями, которые контролируют потенциальные взаимодействия между объектами в одной и той же сцене во время моделирования.

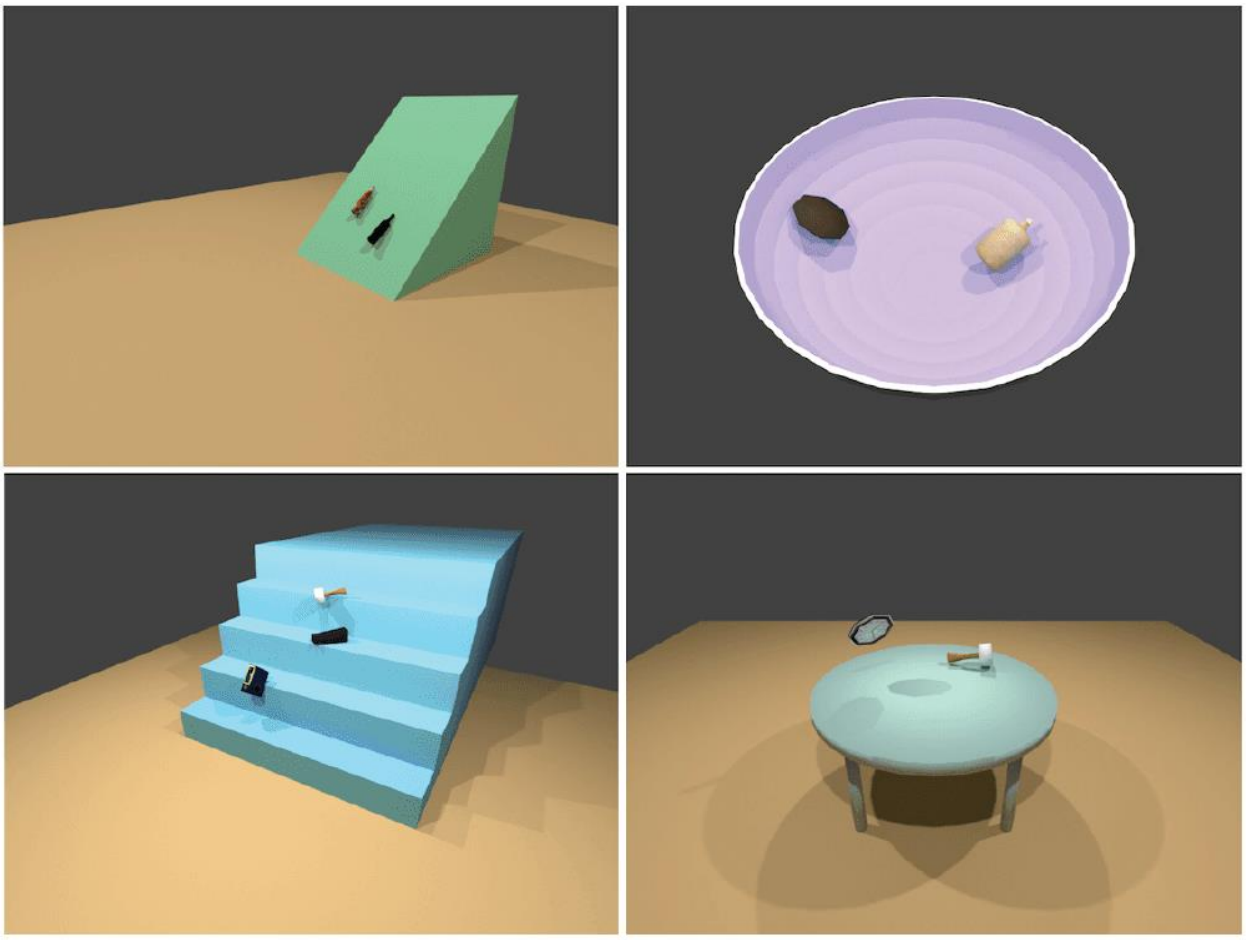


Рисунок 3 – Образцы сценариев из набора данных Sound-20K

### 4.4 Набор данных

Синтезировано 20 378 видео с аудио посредством выборки из множества возможных комбинаций пар форм и материала в соответствии с индивидуальными требованиями. Наложено два ограничения следующим

образом. Только формы с одинаковыми фактическими размерами могут появляться совместно в определенном сценарии. Кроме того, формы выбираются только в том случае, если их относительные размеры относительно установки сцены реалистичны. Например, стул и стол, которые считаются большими по размеру, разумно упасть на землю вместе, но падение на обеденный стол было бы нереально. Учитывая эти ограничения, можно автоматически генерировать записи сцен в пакетах в качестве входных данных для генеративной модели. Примеры видео показаны на рисунке 4.

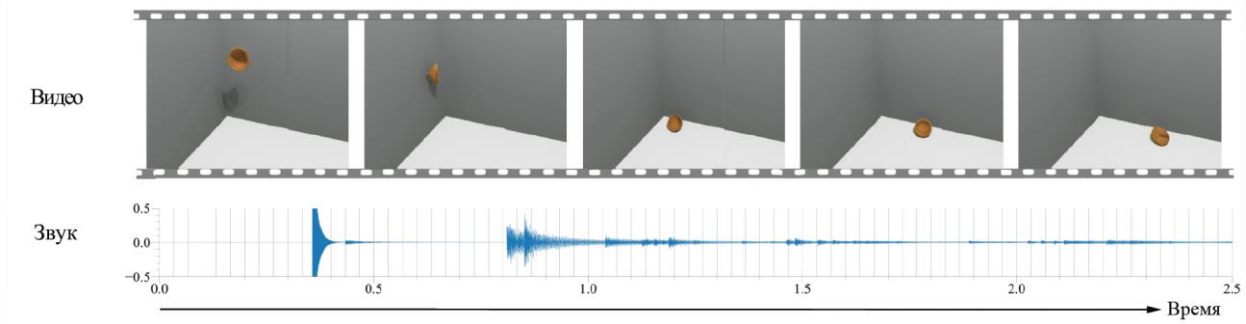



Рисунок 4 – Образцы аудиовизуальных данных из набора данных Sound-20K

## 5 Анализ данных

На рисунке 5а показано, что, хотя исходный набор данных ShapeNet имеет сильный перекос в сторону стульев, бутылок и некоторых других конкретных категорий, наш выбор пытается выровнять это распределение по большинству подкатегорий. После того, как были объединены некоторые подкатегории с похожими функциями в основную категорию рисунок 5б. Как показано на рисунке 5в, распределение объектов по всем атрибутам является относительно равномерным, что указывает на широкий охват характеристик формы. Была построена статистика для числа объектов, назначенных каждому материалу на рисунке 5г.

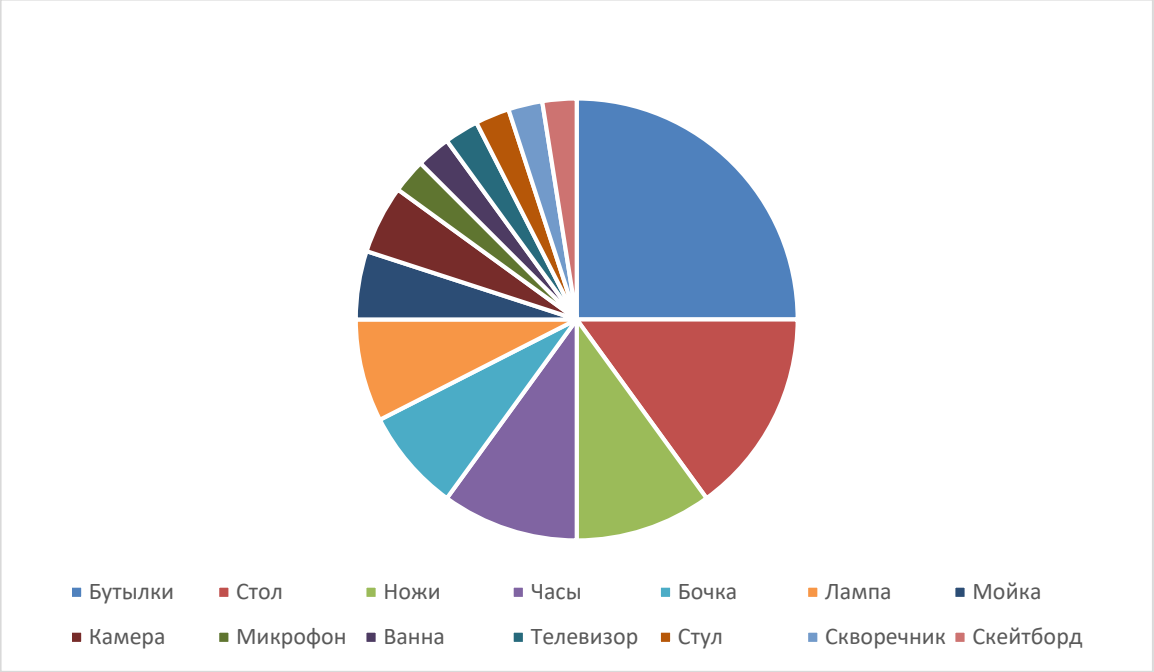



Рисунок 5а – Частота появлений подкатегорий.

Рисунок 5б – Распределение основных категорий.

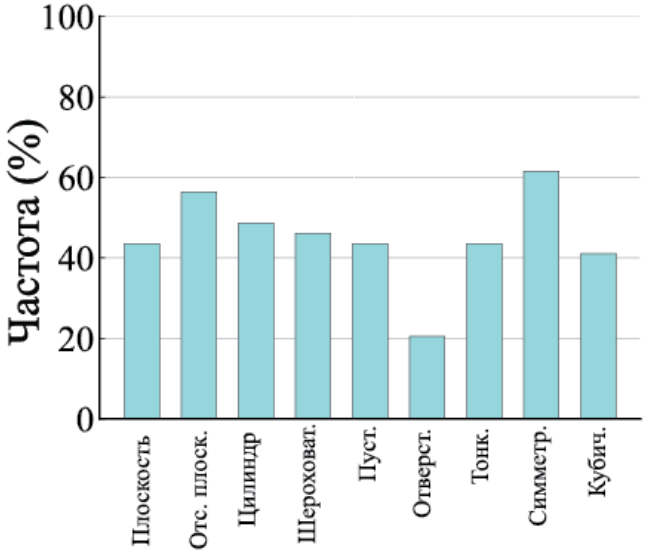



Рисунок 5в – Распределение атрибутов формы.

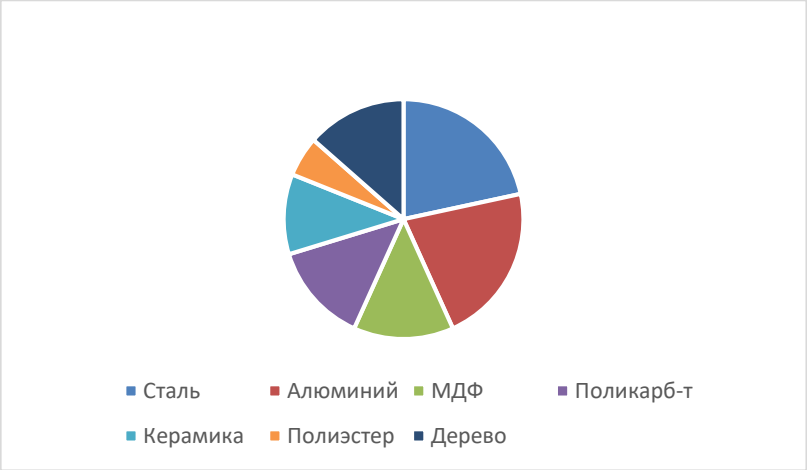



Рисунок 5г – Распределение материалов.

# 6 Восприятие объекта с аудио-визуальными данными

Генеративная модель генерирует данные в двух модальностях: аудио и видео. В данной главе проверяется, что как аудио, так и визуальные данные способствуют восприятию физического объекта, но дополняют друг друга. Рассмотрены две задачи восприятия объекта: распознавание материала и распознавание атрибутов трехмерной формы. Наблюдается, что аудио данные содержат больше информации об материале объекта, в то время как визуальные данные помогают лучше распознавать атрибуты 3D-формы.

## 6.1 Данные

В дополнение к синтетическому Sound-20K, рассматривается два реальных набора видеоданных со звуком.

### 6.1.1 Physics 101

Physics 101 [17] - это набор данных, содержащий 15 190 видео 101 объектов, которые ведут себя в пяти сценариях. Сценарии включают в себя объекты, падающие на поверхность определенного материала, объекты, разбрызгиваемые в воду и т. д. Каждый объект снабжен аннотацией к своему основному вещественному материалу, массе и объему.

### 6.1.2 Greatest Hits

Greatest Hits [21] - это набор данных, содержащий 977 видеофильмов о человеке, исследующем окружение барабанной палочкой. Каждое видео, в среднем, содержит 48 действий (попаданий и царапин), а всего 46,577 действий. Набор данных также содержит метку материала целевого объекта в каждом видео. Рисунок 6.

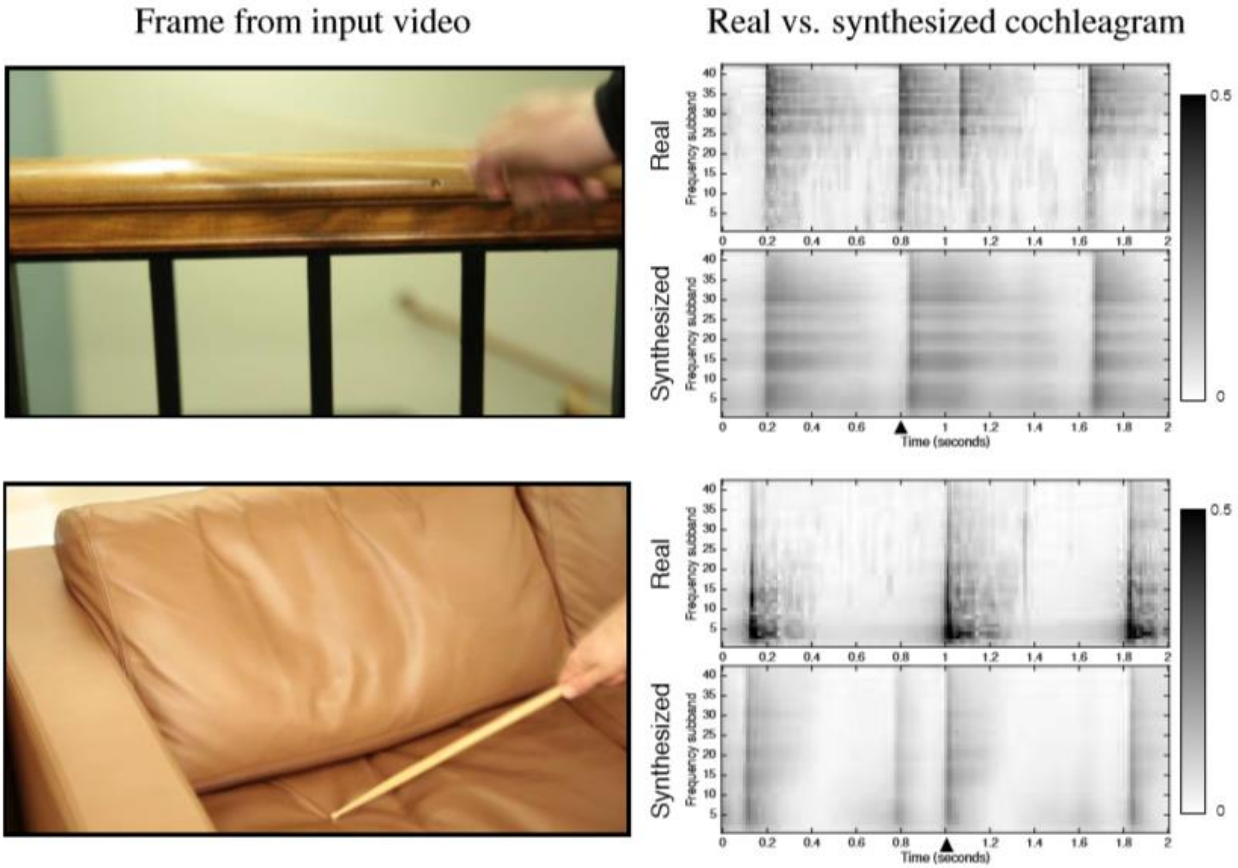



Рисунок 6 – Пример базы данных Greatest Hits

## 6.2 Методы

Для данного исследования используются стандартные сверточные нейронные сети. В сети есть аудио и визуальный поток, поскольку обучение проходит на аудио и видео данных. Структура сети показана на рисунке 7.

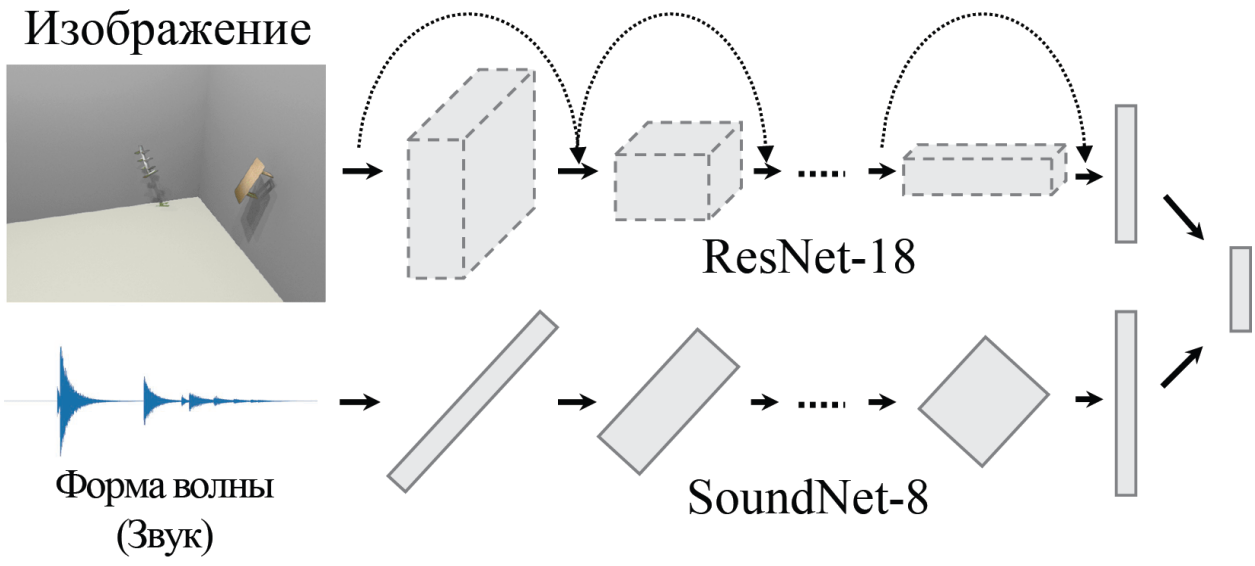



Рисунок 7 – Двухпотоковая сверточная нейронная сеть, где структура визуального потока представляет собой 18-слойную ResNet [19], а структура аудио потока - 8-слойная SoundNet [20].

В визуальном потоке используется ResNet-18 [19], который принимает изображение в качестве входного сигнала и создает 512-мерный вектор. Представленный аудио поток использует ту же структуру, что и SoundNet-8 [20], принимая необработанную звуковую последовательность в качестве входных данных и создавая 1024-мерный вектор. Для мультимодального обучения два вектора объединяются, что приводит к скрытному вектору измерения 1,536. Окончательный результат работы сетей и функция потерь зависят от задачи. Для распознавания материала это будут n k-мерный векторов с перекрестной потерей энтропии, где n - количество объектов в сцене, а k - количество типов материалов в наборе данных. Для распознавания атрибутов формы это будут n 9-мерных векторов с бинарной кросс-энтропийной потерей, как рассматривается экспериментах 9 атрибутов формы. Так же использовался стохастический градиентный спуск для обучения, с темпом обучения 0,001, импульсом 0,9 и размером партии 16.

## 6.3 Распознавание материалов

Эксперимент по распознаванию материалов будет проводится на основе Sound-20K и Physics 101, так как все они имеют метки материала с визуальными и / или аудио данными в качестве входных данных. Для Sound-20K использовались видео без текстур, чтобы исключить корреляции между внешним видом объекта и его материалом. Для обоих наборов данных используется 95% видео для обучения и 5% для тестирования. На рисунке 8 показаны матрицы путаницы классификации материалов в наборе данных Physics 101, где наблюдается, что аудио данные также содержат обширную информацию для вывода материала объекта по сравнению с визуальными данными. Результаты на Sound-20K, содержит мало информации о материале, потому что внешний вид объекта не имеет текстуры. Результаты по Physics 101 подтвердили, что эти эффекты существуют в реальных видео. В частности, на рисунках 8а и 8б показано, что объекты, которые трудно отличить от визуальных данных (маленькая металлическая монета и деревянная ручка), идентифицируются моделью по разным звукам, которые они издают.

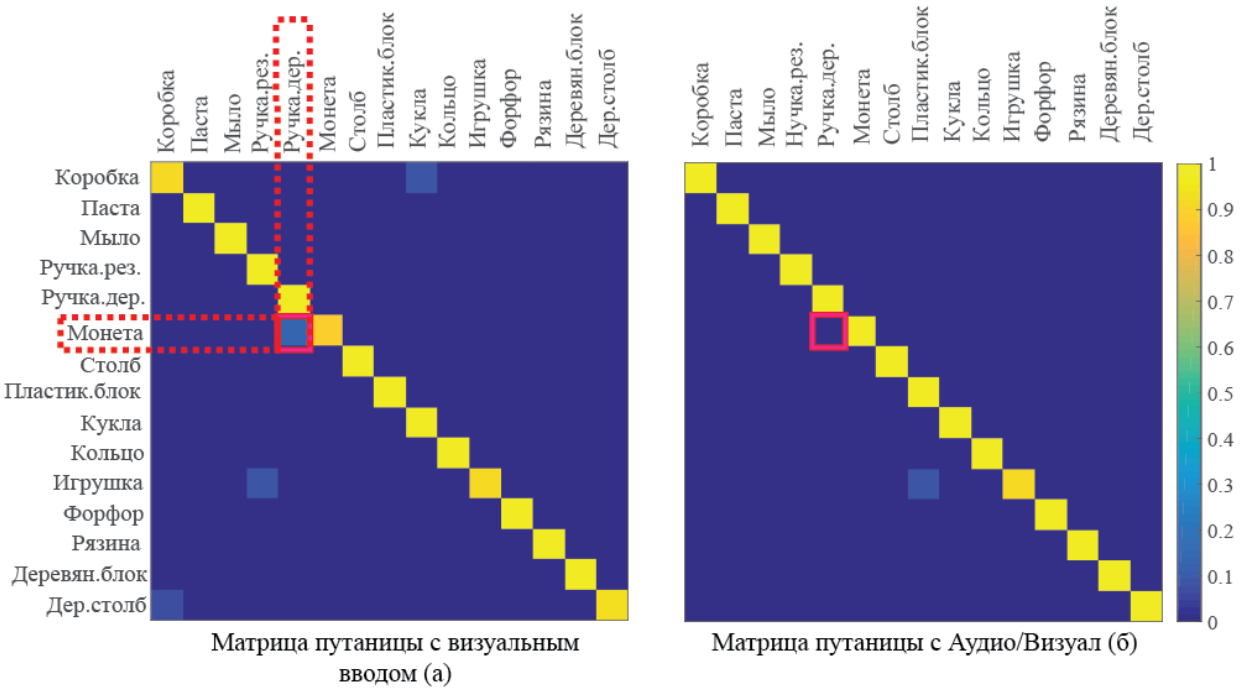



Рисунок 8 – Аудио и визуальные данные дополняют распознавание

материала. На данном рисунке показана точность классификации, где совместное обучение по двум методикам превосходит обучение по любому из них. В рисунке с визуальным вводом и аудиовизуальным вводом показаны матрицы путаницы на моделях Physics 101, использующих визуальные или аудиовизуальные данные. Хотя визуальные данные очень информативны, они допускают ошибки для плохо видимых объектов (монетка и деревянную ручку). Использование аудио данных устраняет неоднозначность.

## 6.4 Распознавание атрибутов формы

Вывод полной трехмерной формы из аудио данных практически невозможно. Однако выяснить какие данные о трехмерной форме можно восстановить из звука вполне реально. Как было описано ранее в главе 5, все объекты в Sound-20K и Physics 101 помечены девятью атрибутами. Поэтому было изучено, какие атрибуты можно извлечь из аудиовизуальных данных. Экспериментальная установка такая же, как в главе 6.3. В таблице 1 приведены результаты распознавания атрибутов трехмерной фигуры. Как и ожидалось, визуальные данные содержат богатую информацию о большинстве атрибутов формы. В то же время аудио данные также так же обладают информацией по многим атрибутам формы.

Таблица 1 – Как на Sound-20K, так и на Physics 101 [17] совместное обучение на основе аудиовизуальных данных позволяет достичь максимальной производительности. Для этой конкретной задачи визуальные данные играют основную роль, однако можно наблюдать, что аудио также содержит богатую информацию об атрибутах формы объекта.

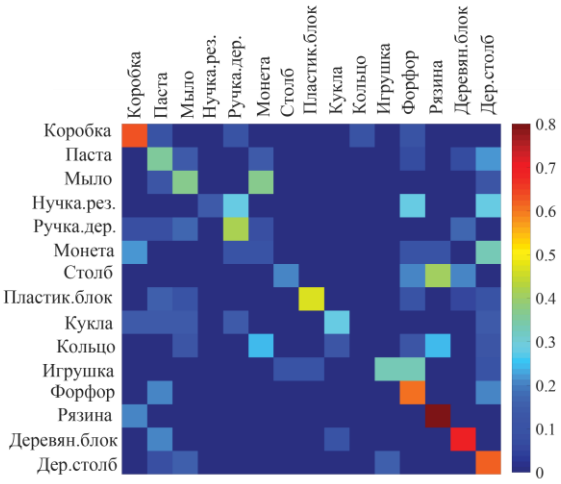
| Данные | Вход | Искривление | | | | Заполняемость | | | | | Среднее |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Плоск | Отс. Плоск. | Цил. | Шерох. | Пуст. | Отв. | Тонк. | Сим-я. | Куб. | |
| Sound-20K | Аудио | 78.1 | 77.9 | 69.9 | 72.2 | 78.8 | 86.9 | 77.8 | 69.7 | 73.2 | 76.1 |
| | Визуальный | 86.5 | 86.3 | 78.1 | 71.4 | 78.7 | 87.1 | 75.3 | 72.2 | 78.9 | 79.4 |
| | Аудио/Визуал | 86.9 | 86.9 | 81.1 | 80.6 | 84.0 | 90.8 | 84.7 | 79.0 | 81.4 | 83.9 |
| Physics 101 | Аудио | 85.2 | 85.9 | 83.1 | 88.0 | 76.1 | 97.2 | 92.3 | 88.7 | 82.4 | 86.5 |
| | Визуальный | 97.2 | 96.5 | 95.1 | 95.8 | 94.4 | 98.6 | 97.2 | 97.2 | 97.9 | 96.6 |
| | Аудио/Визуал | 97.2 | 97.2 | 97.2 | 97.9 | 97.2 | 97.9 | 96.5 | 96.5 | 95.8 | 97.0 |

# 7 Переход от синтетических к реальным данным

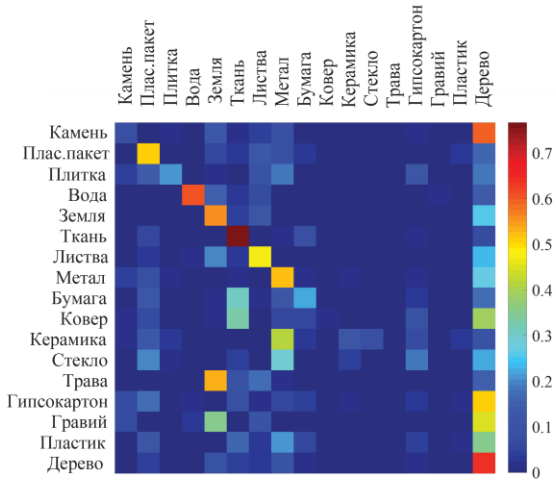
Чтобы осуществить переход от синтетических к реальным данным, будет использоваться раннее полученная информация. Как обсуждалось ранее, ацдио и визуальные данные содержат дополнительную информацию о физических объектах. Поэтому будет производиться вывод материалов объекта из аудио данных и вывод атрибутов формы из аудиовизуальных данных.

## 7.1 Распознавание материала по аудио данных

И Physics 101, и The Greatest Hits имеют метки материала объекта, однако они отличаются от набора меток, которые используются в Sound-20K. Поэтому для оценки функций, изученных в Sound-20K, перенесены в реальные наборы данных. В частности, для применения линейных SVM (метод опорных векторов) к этим функциям были извлечены элементы из средних уровней сети для классификации материалов.




| Методы | Точность (%) |
|---|---|
| Исходные данные | 58.5 |
| pool5 | 29.6 |
| conv6 | 35.9 |
| conv7 | 41.5 |
| MFCC | 40.7 |




| Методы | Точность(%) |
|---|---|
| Исходные данные | 44.2 |
| pool5 | 33.7 |
| conv6 | 37.9 |
| conv7 | 39.2 |
| MFCC | 38.1 |

Рисунок 9 – Классификация материалов по The Greatest Hits [21] с различными функциями, изученными на синтетическом Sound-20K и линейном SVM. Слева: матрица путаницы с функциями conv7. Справа: точность классификации

На рисунке 9 показаны результаты классификации материалов, где сравниваются функции из различных уровней аудио потока данной сети, обученной на синтетическом Sound-20K, с другими функциями, которые были разработаны вручную. Также показаны матрицы путаницы функций conv7. Видно, что разработанные функции, обученные на синтетических данных,

достигают сопоставимых результатов с классическими функциями MFCC (мел-кепстральные коэффициенты) [18].

## 7.2 Распознавание атрибутов формы

Для этой распознавания атрибутов формы использовалось 785 видео в Sound-20K с одним объектом на сцене для обучения; по одному сценарию из Physics 101 - падение объекта - для оценки. Так как эти сцены имеют схожий визуальный макет. Поскольку видео в этих двух наборах данных имеют очень разные виды, используется силуэты объектов вместо изображений RGB в качестве визуального ввода. Чтобы получить силуэты объекта, используется вычитание фона. Оценка изученной модели на реальных данных проводится без какой-либо подстройки. В таблице 2, показано что полученные синтетические данные, могут напрямую передаваться на ввод данных реального мира, а аудио и визуальные данные хорошо работают с различными наборами атрибутов.

Таблица 2 – Распознавание атрибутов формы в сценарии падения объекта в Physics 101 с использованием моделей, обученных на подмножестве аудио и силуэтов объектов из Sound-20K

| Вход | Плоск. | Отс. Плоск. | Цил. | Шерох. | Пуст. | Отвер. | Тонк. | Сим. | Куб. | Среднее |
|---|---|---|---|---|---|---|---|---|---|---|
| Ауидо | 48.1 | 50.0 | 53.9 | 86.5 | 78.9 | 83.7 | 37.1 | 41.8 | 79.8 | 61.6 |
| Визуальный | 70.2 | 69.2 | 58.7 | 43.3 | 38.5 | 61.5 | 42.3 | 71.2 | 63.5 | 57.6 |

# 8 Заключение

В результате выполнения дипломной работы был предложен метод использования синтетических аудиовизуальных данных для точной передачи восприятия физических свойств 3D объектов при их столкновении. Использование синтетических данных дает уникальное преимущество: их относительно легко собирать и расширять, и они полностью аннотированы. Так же был выбран основанный на физике дизайн конвейера, который синтезирует реалистичные аудиовизуальные данные в больших масштабах.

В процессе выполнения дипломной работы были решены следующие задачи:

1. Изучены ранее применяемые методы.
2. Определены подходы к созданию генерирующей модели.
3. Определен дизайн генерирующей модели.
4. Изучены методы для реализации генерирующей модели.
5. Реализована генерирующая модель.
6. Оценена результативность методов.

На данный момент реализована генеративная модель, которая способна привязывать аудио файлы к 3D объектам, основываясь на уже сформированных базах данных, созданных в экспериментальных условиях. Поэтому геймдизайнеры и разработчики виртуальной среды ограничены определенным набором данных. Добавление возможности разработчикам вносить свои уникальные аудио файлы для реализации творческих идей в создании более правдоподобного мира может стать отличной идеей для дальнейшего улучшения.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Bonneel, N., Drettakis, G., Tsingos, N. Viaud-Delmon, I. and James, D. Быстрые модальные звуки с масштабируемым синтезом в частотной области. ACM TOG, 27(3):24. – 2008.
2. Chang, A. X., Funkhouser T., Guibas L., Hanrahan P., Huang Q., Li Z., Savarese S., Savva M., Song S., Su H., et al. Shapenet: информационный репозиторий 3d-моделей. arXiv preprint arXiv:1512.03012. – 2015.
3. Coumans, E. Bullet physics engine. Open Source Software: http://bulletphysics. org, 1. – 2010. – 3 [11]
4. Davis, S. and Mermelstein, P. Сравнение параметрических представлений для распознавания односложных слов в непрерывно произносимых предложениях. IEEE TASSP, 28(4):357–366. – 1980.
5. Director-O'Brien, J. F. Синтезирование звуков от физического движения. In ACM SIGGRAPH 2001 видео обзор по программе анимационного театра, page 59. ACM. – 2001.
6. James D. L., Barbic, J. and Pai, D. K. Предварительно рассчитанная акустическая передача: чувствительная к выходу, точная генерация звука для геометрически сложных источников вибрации. ACM TOG, 25(3):987– 995. – 2006.
7. O'Brien, J. F., Shen, C., and Gatchalian, C. M.. Синтезирующие звуки из моделирования твердого тела. In SCA. – 2002.
8. Owens, A., Isola P., McDermott J., Torralba A., Adelson E. H., and Freeman, W. T. Визуально обозначенные звуки. In CVPR. – 2016
9. Van den Doel, K. and Pai, D. K. The sounds of physical shapes. Presence: Teleoperators and Virtual Environments, 7(4):382–395. – 1998. – 2 [40]
10. Wu, J., Lim, J. J., Zhang, H., Tenenbaum, J. B. and Freeman W. T. Physics 101: Изучение физических свойств объекта из немаркированных видео. In BMVC. – 2016.

11. Su, H., C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Оценка точек обзора в изображениях с использованием cnns, обученных с использованием визуализированных 3D-моделей. In ICCV. – 2015.
12. Labelle, F. and J. R. Shewchuk. Isosurface stuffing: быстрые четырехгранные сетки с хорошими двугранными углами. ACM TOG, 26(3):57. – 2007.
13. Ciskowski, R. D. and Brebbia, C. A. Boundary element meth- ods in acoustics. Springer. – 1991.
14. Fiala, P. and Rucz, P. Nihu: An open source c++ bem library. Advances in Engineering Software, 75:101–112. – 2014.
15. Liu, Y. Метод быстрого многополюсного граничного элемента: теория и приложения в технике. Cambridge university press. – 2009.
16. Gimbutas, L. G. Z. Fmmlib3d. Fortran li- braries for fast multiple method in three dimensions, http://www.cims.nyu.edu/cmcl/fmm3dlib/fmm3dlib.html., 1. – 2011.
17. Hughes, T. J. Метод конечных элементов: линейный статический и динамический анализ конечных элементов. Courier Corporation. – 2012.
18. Fouhey, D. F., Gupta, A. and Zisserman A. 3d shape attributes. In CVPR. – 2016
19. Wu, J., Lim, J. J., Zhang, H., Tenenbaum, J. B. and Freeman, W. T. Physics 101: Learning physical object properties from unlabeled videos. In BMVC. – 2016.
20. Davisand, S., Mermelstein, P. Comparisonofparametricrepresentations for monosyllabic word recognition in continuously spoken sentences. IEEE TASSP, 28(4):357–366. – 1980.
21. He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition. In CVPR. – 2016.
22. Aytar, Y., Vondrick, C., and Torralba, A. Soundnet: Learning sound representations from unlabeled video. In NIPS. – 2016.

23. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H. and Freeman, W. T. Visually indicated sounds. In CVPR. – 2016.
24. Valenza, E. Blender 2.6 Cycles: Сборник рецептов о материалах и текстурах. – 2013