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

Синтез звуковых эффектов вокализации млекопитающих

Студент 2 курса магистратуры

группы 11-712

«___»_____2019 г. Шайхутдинов А. Л.

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 3

1. Обзор по теме диссертации 5

2. Средства разработки (инструменты) 11

3. Обзор синтезатора голоса млекопитающих 12

3.1. Тело животного 16

3.2. Легкие 17

3.3. Гортань 17

3.4. Голосовой тракт 18

4. Реализация 20

4.1. Структура программного обеспечения 20

4.2. Подпатч body (тело) 20

4.3. Подпатч lungs (лёгкие) 21

4.4. Подпатч larynx (гортань). 22

4.7. Подпатч animal (животное). 24

4.8. Подпатч sound-type (тип звука). 25

4.9. Подпатч emotion (эмоция). 25

4.10. Подпатч random. 25

4.11. Подпатч initialize (инициализировать). 26

4.12. Графический интерфейс пользователя (GUI) 27

5. Оценка и сравнение результативности методов 29

Заключение 30

Список литературы 31

Глоссарий 34

Приложение 35

Введение

Как правило, библиотеки звуковых эффектов состоят из множества засэмплированных звуков, хранящихся в виде несжатых wave файлов. Сэмплы в библиотеке между собой немного отличаются по звуку, чтобы у звукового дизайнера был широкий выбор, поэтому в библиотеке может находиться больше 100 сэмплов. На сегодняшний день существует огромное множество библиотек сэмплов, однако, у звуковых дизайнеров не всегда получается найти подходящий сэмпл для получения необходимого звука, и им приходится записывать собственный. Следует отметить, что процесс записи, например, реальных криков животных, соответствующих множеству ситуаций в видеоиграх, может оказаться очень непростым. Это ведет к потере большого количества ценного времени, а библиотеки сэмплов со временем становятся слишком большими и неорганизованными. Поэтому в видеоиграх крики животных обычно синтезируются и часто адаптируются для производства криков вымышленных существ, открывая широкие возможности в этой области.

В индустрии кино и видеоигр растет спрос на адаптируемые интерактивные звуковые эффекты, характеристики которых могут задаваться звуковыми дизайнерами, которые создают эффекты для фильма, и действиями пользователей, таких как игроки видеоигр. Это мотивировало множество исследований моделей синтеза на основе процедурного аудио.

Звуковые эффекты могут быть синтезированы из небольшого количества стартовых блоков, например, генератора случайных чисел и генератора синусоидальной волны. Таким образом можно создавать огромное количество контрастных звуков, изменяя различные параметры, предоставленные пользователю в физической модели. В статье [1], например, были предприняты попытки воссоздать и тиражировать большое количество сэмплов из библиотеки звуковых эффектов, используя небольшое количество физических моделей синтеза.

В последние годы процедурное аудио разрабатывалось в основном для использования в индустрии компьютерных игр, где память для хранения сэмплов ограничена, а повторение сэмпла, такого, как звук выстрела из оружия, вскоре становится утомительным и нереалистичным. Субъективные тесты, проведенные в статье [2], показали, что 64% пользователей предпочли синтезированные эффекты записанным звукам в простой игре, специально разработанной для теста. С помощью субъективных тестов в статье [3] показано, что физические модели могут быть эффективны при воспроизведении грубых, спектрально плотных звуков, таких как рёв льва, и могут воссоздать реалистичную артикуляцию. На сегодняшний день наиболее полная коллекция физических моделей содержится в книге [4]. Данная работа построена на одной из этих моделей.

Объектом исследования является создание модели синтеза звуковых эффектов вокализации млекопитающих. Предметом является синтез звуковых эффектов вокализации млекопитающих.

Цель данной диссертационной работы заключается в том, чтобы выяснить, является ли синтез достойной альтернативой библиотекам звуковых эффектов, основанным на сэмплах.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Изучение предметной области.
2. Изучение средств синтеза звука.
3. Разработка модели синтеза звука.
4. Оценка и сравнение результативности методов.

1. Обзор по теме диссертации

В исследовании голосового аппарата животных [5] предполагается, что методы человеческого речевого синтеза, такие как методология источника-фильтра, используемая в статье [6], могут быть применены и для синтеза криков животных.

Модели синтеза источника-фильтра для звуковых эффектов вокализации людей и животных подразделяются на две основные категории: физические модели, такие как в статье [1] и модели на основе сигналов, такие как в книге [4].

Физические модели тесно связаны с физикой звука. Их параметры носят физический характер и могут изменяться в соответствии с реальным миром. Такие модели могут быть точными и реагировать на изменения параметров, однако обладают значительной сложностью, что нежелательно, если возможности обработки ограничены.

По этим причинам основанные на сигналах подходы, направленные на воссоздание воспринимаемого качества сигнала через нефизические методы, обычно предпочтительны по причине их простоты. В то время как физические модели вокализации включают в себя сигнал, основанный на движении голосовых связок, модели, основанные на сигналах, не имеют отношения к физическому движению и создают сигнал с помощью других средств, таких как вэйвшейпинг. Аналогично, голосовой тракт, через который резонирует сигнал, часто моделируется в физической системе с использованием волноводного фильтра, сегменты задержки которого представляют собой распространение звуковых волн через голосовой тракт и коэффициенты фильтра которых непосредственно связаны с формой тракта [6]. Сигнальная система воспроизводит фильтрующий эффект голосового тракта без моделирования распространения волны, часто используя серию полосовых фильтров [4].

В статье [3] исследователи пытаются улучшить синтез вокализации млекопитающих путем моделирования нелинейности в гортани. Это исследование также направлено на то, чтобы показать, что точная конфигурация голосового тракта через линейное предсказание может воссоздать реалистичную артикуляцию, мотивируя дальнейшие исследования комплексной документацией системных параметров, позволяющих синтезировать огромное количество криков животных.

Большинство животных имеют общий метод производства звука [5], во многом похожий на человеческий голосовой аппарат, в котором происходит голосообразование с использованием мышечной диафрагмы, выводя воздух из легких через гортань. Существование эластичных мембран в гортани млекопитающих, называемых «голосовыми связками», дает возможность комплексного голосообразования.

Термин «голосовой тракт» используется для описания гортанной полости, глотки, оральной полости, носовой полости, через которые звук должен проходить. Длина и форма голосового тракта, которые млекопитающие могут изменять, определяют резонансные частоты, на которых воздух, проходящий через них, будет колебаться. Эти частоты называются формантами.

Благодаря анализу спектральной динамики вокализации млекопитающих было установлено, что первые две форманты F1 и F2 имеют значительный потенциал для изменения во время голосообразования, 23% и 61% соответственно [7]. Кроме того, считается, что формантная дисперсия (расстояние между F1 и F2) является наиболее ярким показателем размера тела.

Голосовая мембрана простирается вверх от голосовой складки, которая покрывает мышцу щитовидной железы. В статье [8] сообщается, что голосовая мембрана может колебаться с очень различным поведением, что приводит к нелинейности.

Длина голосового тракта и площадь поперечного сечения челюсти млекопитающих изменяются во времени [9], и такие изменения в голосовом тракте приводят к динамической формантной структуре. Исследования функциональной морфологии тигровых и львиных голосовых связок заключают, что квадратная форма связок, которые способны растягиваться и смыкаться на низких уровнях энергии, позволяет регистрировать голосообразование (в том числе нелинейности) на очень низких частотах [10].

Были предприняты различные попытки моделирования потенциальных причин нелинейности, присутствующих в голосовом аппарате млекопитающих. Например, в статье [11] используется модель голосовых связок масса-пружина для реализации нелинейности с использованием осциллятора Ван дер Поля.

В книге [4] показана сигнальная модель для синтеза звуковых эффектов вокализации животных, которая состоит из параметрического вэйвшейпера, используемого для воссоздания широкополосных сигналов. Постоянная синусоидная пульсация модулируется другой косинусной функцией, представляющей «колебания голосовых связок», создавая боковые полосы вокруг гармонических составляющих сигнала. Этот сигнал пропускается через несколько параллельных полосовых фильтров для создания требуемой звуковой текстуры.

Существующие физически верные реализации системы человеческого голосового тракта, такие как в статье [6], основаны на волноводном фильтре Келли-Лохбаума [12], который моделирует левые и правые волны, распространяющиеся через участки голосового тракта, моделируемые как акустические трубки (цилиндры) разного размера, разделенные рассеивающими переходами, коэффициенты которых непосредственно соответствуют размеру сегмента голосового тракта по отношению к соседним сегментам.

В статье [13] указано, что результирующая форма волны голосового потока должна точно соответствовать физическому движению голосовых

связок. Для этого к параметрам должны быть применены ограничения для предотвращения возникновения голосового потока с нереалистичными значениями без введения слишком большой сложности модели.

В голосовом аппарате млекопитающих наблюдаются многочисленные нелинейные явления [8]. Причинами таких явлений являются голосовые связки, колеблющиеся в разных частотах или колеблющиеся асинхронно, колебание голосовой мембраны и растяжение и сужение поверхности тракта. В физических моделях, представленных статье [2], такие нюансы воссоздаются с использованием синтеза амплитудной и частотной модуляции (ЧМ). Был создан пользовательский параметр управления, содержащий 6 особенностей, реализованный следующим образом:

1. Ограничение циклов. Не вводится амплитудная модуляция или частотная модуляция, представляющая биологическую систему, в которой осцилляторы (голосовые связки) синхронизированы.

2. Шиммер. Амплитудная модуляция вводится с низкой частотой модулятора, чтобы воссоздать эффект естественного изменения колебаний голосовых связок.

3. Дрожание и субгармоники. Воссоздание эффекта синхронного колебания голосовых связок на частотах с целым отношением друг к другу.

4. Негармонические боковые полосы частот. Частотная модуляция воссоздает эффект рассинхронизированных голосовых связок, вибрирующих на частотах, которые не имеют целочисленного отношения друг к другу.

5. Множественные нелинейности. Частотная модуляция с серией из 2 модуляторов создают сложную волну со многими боковыми полосами частот, воспроизводящую взаимодействие голосовых связок с голосовой мембраной.

6. Детерминированный хаос. В серию частотной модуляции вводится третий модулятор, чтобы сигнал воссоздавал детерминированный хаос, возникающий благодаря растяжению и колебаниям голосовой мембраны, голосовых связок и стенок голосового тракта одновременно.

Физический голосовой тракт в статье [2] реализован с использованием модели волноводного фильтра, как в [6], с пятнадцатью сегментами, разделенными рассеивающими переходами, коэффициенты которых контролируются пользователем и относятся непосредственно к радиусам сегментов голосового тракта. Вход в волноводный фильтр представляет собой форму волны голосового потока. Контролируемая динамика включает амплитуду, изменение фундаментальной частоты, управление сложностью, управление шумом и длительность.

Предложенная модель голосового тракта полностью представляет собой систему без потерь, в результате чего форманты имеют очень узкую полосу пропускания. Были предприняты попытки моделировать потери, которые включают в себя такие факторы, как затухание из-за вибрационных стенок, текучесть и потери теплопроводности. В этой статье упрощенная версия такого эффекта достигается путем кроссфейдирования между нефильтрованными и отфильтрованными сигналами. Это позволяет пользователю управлять изменениями соотношения сигнал/шум и параметрами формант.

В сигнальной модели исходный сигнал подается в модель голосового тракта, состоящую из нескольких параллельных полосовых фильтров, представляющих сигнальные форманты, каждый из которых имеет управляемую пользователем центральную частоту, резонанс и коэффициент усиления. Кривые частоты и амплитуды подвергаются воздействию пользователя, что позволяет независимо изменять форманты.

Характеристики голосового тракта животных во время голосообразования все еще относительно неизвестны. Для синтеза речи человека конфигурации голосового тракта были получены с помощью анализа линейного предсказательного кодирования (LPC). В статье [6] цифровой фильтр разработан на основе желаемой формантной структуры, коэффициенты которой могут быть использованы для рекурсивного

вычисления радиуса голосового тракта с помощью алгоритма, такого как рекурсия Левинсона-Дурбина.

Вокализации млекопитающих менее периодичны и менее предсказуемы в их формантной структуре, чем человеческая речь, что затрудняет проектирование набора данных конфигураций голосового тракта. LPC был выполнен, чтобы продемонстрировать, что такие методы могут также работать для модели млекопитающих, а некоторые конфигурации хранятся в качестве примеров, которые будут использоваться и редактироваться.

Анализ LPC выполняется следующим образом: дискретизация входного аудиофайла понижается до фиксированного значения для достижения большего разрешения фильтра на низких частотах. Для учета поведения низкочастотного сигнала применяется фильтр верхних частот. Затем рекурсия Левинсона-Дурбина выполняется для расчета коэффициентов фильтра, которые впоследствии используются для вычисления радиусов сегмента голосового тракта. Следует отметить, что параметр задержки в модели синтеза должен быть установлен равным скорости декодирования диапазона для достижения формантной точности.

Чтобы представить изменяющуюся форму голосового тракта по всему голосообразованию, анализ LPC выполнялся на образце файла, разбитого на несколько разделов. Затем коэффициенты фильтра для этих разделов сохраняются и отправляются в модель голосового тракта во время голосообразования. Разделение между состояниями осуществляется путем линейной интерполяции коэффициентов.

Чтобы реализовать задержку обратной связи с точностью до сэмпла для модели физического голосового тракта в Pure Data, потребовалась техника повторной блокировки, в которой размер блока для модели голосового тракта был уменьшен в зависимости от значения задержки, выбранного пользователем.

2. Средства разработки (инструменты)

Для моих целей больше всего подошла среда разработки Pure Data (http://tv-22.ru/vvedenie-v-puredata.html/).

Итак, Pure Data, сокращённо Pd — это среда и язык визуального программирования, изначально созданный для программирования звука. Пакет кроссплатформенный и бесплатный. Разрабатывается он довольно немалым сообществом, а автором ядра языка является Miller S. Puckette (он же автор известного Max/MSP, с которым у Pure Data очень много общего). Pd довольно популярен, его использует множество людей, начиная от игровых саунд дизайнеров, и заканчивая музыкантами.

При запуске Pure Data открывается консоль, которая выдает сообщения для пользователя. При создании нового файла открывается окно «паттерн», в котором мы будем создавать патч. Патч в pd состоит из блоков и их соединений, описывая таким образом поток данных. Блоки можно добавлять через меню Put. Есть несколько видов блоков:

- Object (объект)
- Message (сообщение)
- Number (число)
- Comment (комментарий, на функциональность не влияет)

В Pure Data есть два режима работы: режим редактирования, и режим исполнения. Переключаются они через пункт меню Edit -> Edit Mode или с клавиатуры по Ctrl+E. Если переключиться в режим исполнения, курсор изменит вид на обычную стрелку, не будет возможности редактировать патч, но будет возможность им управлять.

3. Обзор синтезатора голоса млекопитающих

В последнее время наблюдается растущий интерес к использованию роботов, подобных животным, в таких применениях, как общение и лечение. Например, PARO [13] - это интерактивный роботизированный тюлень, который особенно популярен для терапевтического использования в больницах и учреждениях по уходу, где настоящее животное будет проблематичным. Как и PARO, большинство таких «зооморфных» устройств спроектированы для поддержки конкретных вариантов использования, и зачастую только внешний вид робота имитирует данное животное.

В статье [14] представлены проектирование и реализация параметрического универсального синтезатора голоса млекопитающих в реальном времени. Используемый подход состоял в том, чтобы разложить всю систему звукообразования на соответствующие анатомические компоненты (такие как легкие, голосовые складки, язык и рот) и реализовать моделирование в реальном времени в «Pure Data», языке программирования потока данных с открытым исходным кодом. Программное обеспечение было успешно использовано для создания подходящего голоса млекопитающего для биомиметического робота MiRo, но существует множество других областей, где оно может найти применение. Робот MiRo (разработанный и созданный Consequential Robotics Ltd. в сотрудничестве с Университетом Шеффилда) является первым коммерческим роботом-биомиметиком, который основан на аппаратной и программной архитектуре, смоделированной по образцу биологического мозга. В статье [15] описывается, как была разработана система вокализации MiRo, использующая не предварительно записанные звуки животных, а параметрический универсальный синтезатор голоса млекопитающих в реальном времени, адаптированный к конкретным физическим характеристикам робота. Представлены принципы биомиметики,

лежащие в основе системы вокализации MiRo, и описано, как они были интегрированы в общую архитектуру робота.

Идея о том, что речь может быть создана искусственным устройством, имеет долгую и успешную историю. От механической говорящей машины фон Кемпелена до голоса Siri от Apple, исследования и разработки в области «синтеза речи» привели к большому разнообразию практических решений, таких, как симуляция голосового аппарата человека, связывание сегментов реальной речи и построение статистических моделей. Тем не менее, хотя основополагающие принципы хорошо установлены, было приложено относительно мало усилий для создания синтезаторов голоса животных.

Интерес к синтезу человеческой речи был вызван спросом на практические приложения, такие как голосовые помощники или читающие машины для слепых. Тем не менее, существует потенциальная польза от синтеза голоса животных: например, в сфере образования, робототехники или экологических полевых работ. Последнее представляет особый интерес, поскольку общепринятой практикой является использование методов «проигрывания» (на основе записанных сэмплов) [16], но такой подход не облегчает параметрический контроль над ключевыми экспериментальными переменными.

Как хорошо известно, разные виды живых существ производят звук по-разному; многие насекомые натирают части тела (процесс, известный как «стридуляция»), птицы создают свои песни, используя голосовой орган, известный как сиринкс, а млекопитающие обычно генерируют звук с помощью гортани [17]. Описанная в статье [14] работа была посвящена моделированию механизма звукообразования у наземных млекопитающих. Цель состояла в том, чтобы создать универсальный вокальный синтезатор млекопитающих, который можно настроить так, чтобы он звучал как любое конкретное животное путем выбора соответствующих настроек параметров. Кроме того, поскольку целевой робот MiRo должен был быть

«биомиметическим» по образовательным соображениям, важно, чтобы синтез голоса был основан на аппарате физического производства звука (а не на записях животных).

Голосовой тракт для типичного наземного млекопитающего состоит из гортани, глотки, полости рта и полости носа (см. Рис. 1). Эти анатомические особенности развивались главным образом для дыхания и приема пищи. Однако со временем они были набраны для создания и формирования звука. Основным источником звука является гортань, которая содержит набор «голосовых складок» (иногда их называют «голосовые связки» или «голосовые щели») - пару эластичных мембран, которые разъединены во время дыхания, но соединяются во время глотания (чтобы не допустить попадания еды в легкие). Следствием такого расположения является то, что воздух из легких, который проходит через закрытые голосовые складки, заставляет их вибрировать. Это создает акустическую энергию в форме насыщенного гармониками гудящего звука с отчетливой основной частотой (воспринимаемой как «высота» голоса). Мышцы в гортани контролируют длину и напряжение голосовых складок, которые, в свою очередь, определяют высоту и тембр генерируемого звука.

Рисунок 1, Схема голосового тракта млекопитающих

Остальную часть голосового тракта можно рассматривать как совокупность взаимосвязанных акустических трубок. Глотка лежит непосредственно над гортанью, и в ней находится надгортанник: эластичный хрящ, который контролирует вход в трахею (для дыхания) или в пищевод (для глотания). Над глоткой акустическая трубка разделяется на две части: полость рта (содержащую язык и оканчивающуюся у рта и губ) и полость носа (оканчивающуюся у носа). Поток воздуха в полость носа контролируется мягким нёбом, клапаноподобной структурой в задней части рта. Все эти трубки резонируют на разных частотах в зависимости от их размера и формы, поэтому каждая из них фильтрует спектр богатых гармониками возбуждений, генерируемых вибрацией голосовых складок; энергия увеличивается на одних частотах (резонансах) и подавляется на других. Резонансы голосового тракта известны как форманты. Форманты, возникающие из-за резонансов в ротовой полости, имеют особое значение, поскольку их можно изменить, перемещая язык и открывая/закрывая рот.

Чтобы определить характеристики звуков, производимых голосовым трактом млекопитающего, необходимо смоделировать (а) «воздушный поток» (скорость, с которой воздух выходит из легких), (б) «возбуждение» (звук, генерируемый гортанью) и (в) «фильтрация» (возникающая в резонансных полостях).

Наиболее значимыми факторами, влияющими на выработку звука у большинства млекопитающих, являются физические характеристики, такие как размер тела, емкость легких, а также размер и форма голосового тракта. Из них размер тела является основным определяющим фактором, поскольку он влияет на акустические свойства соответствующих анатомических компонентов (легких, голосовых складок, языка и рта). Другие факторы связаны с динамикой организации и синхронизации поведения различных компонентов во времени. Таким образом, принятый подход состоял в том,

чтобы разложить всю систему производства звука на эти ключевые элементы, начиная с тела.

3.1. Тело животного

Очевидно, что млекопитающие сильно различаются как по физическим размерам, так и по форме. Основной тип тела имеет четыре конечности, приспособленные для использования на суше, но некоторые млекопитающие приспособлены для полета или плавания. Две последние категории млекопитающих имеют тенденцию быть крайними по размеру (синий кит имеет длину более 30 м, а свиноносая летучая мышь имеет длину всего 30 мм), и они также используют различные механизмы для создания звука. Цель здесь состояла в том, чтобы предназначаться для наземных млекопитающих размером от маленькой мыши до большого слона, при этом основной интерес был сосредоточен на животных размером с маленькую собаку. Масса тела для набора наземных млекопитающих приведена в таблице 1 [18].

Таблица 1, Типичная масса тела для набора наземных млекопитающих (данные взяты из [18])

Животное	**Масса тела** (кг)
Африканский слон	6654
Азиатский слон	2547
Лошадь	521
Корова	465
Свинья	192
Человек	62
Овца	55.5
Шимпанзе	52.16
Коза	27.66
Кошка	3.3
Кролик	2.5
Морская свинка	1.04
Мышь	0.023

3.2. Легкие

Основным источником энергии для вокализации млекопитающих являются легкие, а ключевым фактором является частота, с которой воздух может выходить из легких.

Поток воздуха связан с дыханием, и, согласно [19], частота дыхания определяется как:

$$B = 0{,}84 * M^{-0{,}26} \quad (1)$$

где B - частота дыхания (в герцах), а M - масса тела (в килограммах).

Тем не менее, дыхание (и вокализация) использует только часть воздуха в легких. Кроме того, во время вокализации расход воздуха ограничен действиями голосового тракта. Например, средний расход воздуха во время речи взрослого мужчины составляет 0,48 л/с [20], а средняя продолжительность - 4,88 с [21]. Таким образом, средний объем воздуха, используемый для речи человека, составляет 2,33 литра. Для взрослого мужчины весом 80 кг с объемом легких 5,56 л это соответствует 0,42 от общего количества. Учитывая, что обычное время выдоха (полученное из частоты дыхания 0,27 Гц) будет 1,86 с, это означает, что вокализация замедляет расход воздуха в 2,62 раза.

3.3. Гортань

Когда воздух пропускается через закрытые голосовые складки, он вырывается всплесками, поскольку складки на мгновение раздвигаются. После выхода каждого пузырька воздуха эффект Бернулли заставляет голосовые складки снова закрываться, и это действие генерирует импульс акустической энергии, который распространяется через остальную часть голосового тракта. Эта последовательность событий повторяется полурегулярными интервалами, вызывая периодический сигнал возбуждения с энергией на основной частоте вибрации и связанных с ней гармониках. В

первом приближении сигнал, генерируемый голосовой щелью, может быть смоделирован как «пилообразный» сигнал.

Основная функция гортани во время вокализации - контролировать длину и натяжение вибрирующих голосовых складок, определяя тем самым высоту и тембр звука, который возбуждает остальную часть голосового тракта. Согласно [22], связь между массой тела и частотой возбуждения вокализации у животных, варьирующихся по размеру от мышей до слонов, может быть адекватно смоделирована с помощью следующей формулы:

$$F = M^{-0,4} \tag{2}$$

где F - основная частота возбуждения (в кГц).

Тембр вокализации является функцией (а) регулярности колебаний голосовых складок, (б) отношения между основной частотой и ее гармониками и (в) степени турбулентности в потоке воздуха. Последнее означает, что в дополнение к полностью «голосовым» звукам гортань млекопитающих способна воспроизводить голос с придыханием или шипящие звуки (шепот в человеческой речи), удерживая голосовые складки близко друг к другу и обеспечивая небольшой непрерывный воздушный поток. Все эти аспекты могут быть смоделированы подходящим созданием формы волны возбуждения и введением соответствующего уровня случайного шума.

3.4. Голосовой тракт

Резонансы голосового тракта (форманты) действуют как «полосовые» фильтры, которые усиливают акустическую энергию на своих резонансных частотах и подавляют акустическую энергию на других частотах. Это оказывает формирующее влияние на богатый гармониками спектр сигнала возбуждения, исходящего из гортани. (Различное размещение формант приводит к возникновению различных гласных звуков в человеческой речи.)

Частоты разных формант связаны с размером и формой резонансных полостей. Это означает, что они зависят от общей длины голосового тракта - чем длиннее голосовой тракт, тем ниже частоты формант. Как и следовало ожидать, длина голосового тракта коррелирует с размером тела (например, у собак [23] и обезьян [24]), и статья [23] предлагает следующее соотношение:

$$\mathrm{L} = 3{,}15 + (11{,}53 * \mathrm{Log\ M}) \tag{3}$$

где L - длина голосового тракта (в см).

Это означает, что у крупных животных длинные голосовые пути и, следовательно, низкие формантные частоты (и наоборот). Это также означает, что распределение частот формант в вокализации предоставляет слушателю информацию о размере животного.

Частоты формант можно рассчитать, предполагая, что голосовой тракт представляет собой однородную акустическую трубку, которая закрыта в голосовой щели и открыта/закрыта во рту. Если рот закрывается, то форманты снижаются по частоте [25]. Следовательно, резонансная частота n-ой форманты R_n (в Гц) может быть аппроксимирована уравнением:

$$\mathrm{R}_n = \left(2 * \mathrm{n} - (\mathrm{m} + 1)\right) * \frac{\mathrm{c}}{4 * \mathrm{L}} \tag{4}$$

для n = 1; 2; 3; :, где m - степень открытия рта (0 = открыт, 1 = закрыт) и c - скорость звука (в см/с).

4. Реализация

4.1. Структура программного обеспечения

Общая структура программного обеспечения для синтеза основана на моделировании потока энергии через общий голосовой аппарат млекопитающих в соответствии с принципами, изложенными в предыдущем разделе. Ключевые объекты Pd соответствуют легким, гортани, голосовому тракту и постобработке (см. рис. 2).

Рисунок 2, Поток данных в синтезаторе голоса млекопитающих

4.2. Подпатч body (тело)

Команда издать звук запускает моделируемый поток воздуха из легких с продолжительностью, которая рассчитывается по массе тела. Длительность вокализации используется для определения периода всего произнесения (выраженного в Гц). Эти сигналы и сообщения передаются в гортань.

Исходя из массы тела, определяет частоту дыхания, его цикл, продолжительность выдоха, вокализации (параметр bdov, base duration of

vocalisation), основную частоту (параметр ff, fundamental frequency), а также длину голосового тракта (параметр vtl, vocal tract length).

Для разных пород и возрастов животных одного подвида был добавлен множитель массы тела (параметр bmm, body mass multiplier).

Для определения частоты дыхания используется формула (1), затем определяется цикл дыхания, продолжительность выдоха. Чтобы определить продолжительность вокализации, умножили продолжительность выдоха на 2,62.

Для определения основной частоты используется формула (2). Это значение так же выводится в окно программы.

Для определения длины голосового тракта используется формула (3). Для определенных животных был добавлен множитель длины голосового тракта (vtlm, vocal tract length multiplier). Значение длины голосового тракта выведено в окно программы.

Для определенных животных также был добавлен множитель fram (flow rate animal multiplier – множитель объемного расхода для разных животных), а для определенных типов звуков – frm (flow rate multiplier – множитель объемного расхода для разных типов звуков) и frmgui (flow rate multiplier gui – множитель объемного расхода для разных типов звуков для интерфейса).

Для возможности внести вариативность в звук был добавлен множитель объёмного расхода (параметр randfrm, random flow rate multiplier) для рандомизации продолжительности вокализации.

4.3. Подпатч lungs (лёгкие)

Исходя из параметра dov (duration of vocalisation - продолжительность вокализации) определяются значения параметра sr (syllabic rate – слоговая

скорость), отвечающего за частоту открытия рта и параметра mf (modulation frequency – частота модуляции), отвечающего за продолжительность фразы.

В данном подпатче во вход приходит сигнал от кнопки старта вокализации. Далее он идёт в генератор аудио, который запускается через 20 миллисекунд после нажатия на зелёную кнопку «голос», и останавливается через 20 миллисекунд после истечения времени продолжительности отрезка вокализации или 20 миллисекунд после нажатия на красную кнопку «стоп».

Для управления силой вокализации вручную в интерфейсе программы был добавлен слайдер «сила», управляющий параметром pow (power – сила). Параметр pow также управляет запуском и остановкой вокализации.

Для определенных животных был добавлен множитель частотной модуляции (mfm, modulation freq multiplier).

Энергия из этого подпатча идёт в следующий.

4.4. Подпатч larynx (гортань).

Модулирует поток энергии, используя смоделированное действие голосовых складок, вибрирующих на определенной основной частоте, которая модулируется периодом произнесения. При настройках по умолчанию это приводит к модели интонации подъема и спада.

Принимает энергию и параметр mf (modulation frequency – частота модуляции) из лёгких, а также параметры ff (fundamental frequency – основная частота) из подпатча body (тело), fmd (FM depth – глубина ЧМ) из подпатчей animal (животное) и emotion (эмоция), amd (AM depth – глубина AM) из подпатчей animal (животное) и initialize (инициализировать), asp (aspiration – придыхание) из подпатчей sound type (тип звука) и initialize (инициализировать).

В данном подпатче значение основной частоты голоса конвертируется в аудиосигнал, который изменяется по высоте. Частота модуляции определяет скорость, с которой изменяется высота голоса во время вокализации. Параметр глубины ЧМ определяет степень изменения высоты голоса во время вокализации. Глубина АМ определяет степень изменения громкости голоса во время вокализации.

Для определённых типов звука было добавлено придыхание с помощью примеси шума к основному сигналу.

Для возможности внести вариативность в звук был добавлен множитель глубины частотной модуляции (параметр randfmd, random FM depth).

В данном подпатче слайдер «сила» влияет на модуляцию основной частоты голоса.

4.5. Подпатч vocal tract (голосовой тракт).

В данном подпатче значение частоты язычка конвертируется в аудиосигнал, который модулирует основной сигнал. Голосовой тракт имитирует три акустических резонанса (форманты) с использованием полосовых фильтров, частоты которых определяются длиной голосового тракта и степенью открывания рта (с использованием уравнения 4). Параметр слоговой скорости управляет открытием и закрытием рта.

Принимает аудиосигнал из подпатча larynx (гортань), параметры uf (uvula frequency – частота язычка) из подпатча sound-type (тип звука), vtl (vocal tract length – длина голосового тракта) из подпатча body (тело), mco (mouth close/open – открытие рта), отвечающий за то, насколько открыт рот в начале и в конце вокализации, и dmc (degree of closure – предел открытия рта), отвечающий за то, насколько изменяется открытие рта во время вокализации, из подпатча sound-type (тип звука), sr (syllabic rate – слоговая скорость) из подпатча emotion (эмоция) и энергию из лёгких. Выводит аудиосигнал.

Для возможности внести вариативность в звук был добавлен множитель предела открытия рта (randdmc, random degree of closure).

В данном подпатче слайдер «сила» влияет на открытие рта.

4.6. Подпатч postproc (постобработка).

Наконец, результат работы голосового тракта отправляется объекту постобработки, который содержит дополнительный кольцевой модулятор и линию задержки (для того, чтобы ввести эффект эха).

Подпатч принимает аудиосигнал из подпатча vocal tract (голосовой тракт), параметры ringmod (ring modulation – кольцевая модуляция) и delay (задержка) из подпатча initialize (инициализировать).

4.7. Подпатч animal (животное).

Программное обеспечение также предоставляет ряд предустановленных настроек. Например, можно выбрать определенных животных (например, крысу, кошку, собаку, овцу или корову). Выбор одного из этих пресетов просто перемещает соответствующие ползунки в определенные заранее заданные положения. После выбора пресета все еще можно изменять любые параметры в соответствии с требованиями для достижения конкретной цели проектирования.

Для каждого пресета животного этот подпатч устанавливает значения параметров fram (flow rate animal multiplier – множитель объемного расхода для разных животных), mfm (modulation freq multiplier – множитель частотной модуляции), fmd (FM depth – глубина ЧМ), amd (AM depth – глубина АМ), vtlm (vocal tract length multiplier – множитель длины голосового тракта), mco (mouth close/open – открытие рта) и bm (body mass – масса тела).

4.8. Подпатч sound-type (тип звука).

Для каждого пресета типа звука (обычный, рык, дыхание, храп, смех/плач, чихание, кашель) устанавливает значения параметров asp (aspiration – придыхание), uf (uvula frequency – частота язычка), mco (mouth close/open – открытие рта), dmc (degree of closure – степень открытия рта), frm (flow rate multiplier – множитель объемного расхода), ringmod (ring modulation – кольцевая модуляция) и mfm (modulation freq multiplier – множитель частотной модуляции).

4.9. Подпатч emotion (эмоция).

Чтобы ввести некоторые эмоции в вокализацию, в синтезатор были включены параметры для возбуждения и тона. Параметр возбуждения влияет на продолжительность вокализации. Тон влияет на степень изменения основной частоты; высокая валентность приводит к выразительной вокализации, тогда как низкая валентность приводит к большей монотонности звуков.

Параметр arous (arousal – возбуждение) имеет пределы от 0,7 до 7, формирует значение параметра dov (duration of vocalisation - продолжительность вокализации).

Параметр val (valence – валентность, тон) имеет область значений от -1 до 1, формирует значение fmd (FM depth – глубина ЧМ) по формуле val * 0,25 + 0,3.

4.10. Подпатч random.

Периодически генерирует значения от 0 до 2 для параметров mf (modulation frequency – частота модуляции) и sr (syllabic rate – слоговая скорость).

А также генерирует в обозначенных пределах (в процентах) множители случайной величины для параметров глубины частотной модуляции, основной частоты, множителя объемного расхода для разных типов звуков и предела открытия рта.

4.11.Подпатч initialize (инициализировать).

При запуске программы и при нажатии на соответствующие кнопки в окне программы данный подпатч сбрасывает голосовые параметры на предустановленные значения.

Сбрасывает слайдер «сила» на значение 0, параметру anim (animal – животное) присваивается значение 1, stype (sound type – тип звука) – 0.

Сбрасывает настройки тела животного. Параметру bmm (body mass multiplier – множитель массы тела) присваивается значение 1.

Сбрасывает настройки лёгких. Параметрам frmgui (flow rate multiplier gui – множитель объемного расхода для разных типов звуков) и fram (flow rate animal multiplier – множитель объемного расхода для животных) присваивается значение 1.

Сбрасывает настройки гортани. Параметру amd (AM depth – глубина AM) присваивается значение 1, asp (aspiration – придыхание) – 0.

Сбрасывает настройки голосового тракта. Параметру dmc (degree of closure – степень открытия рта) присваивается значение 1, а параметрам uf (uvula frequency – частота язычка) и mco (mouth close/open – открытие рта) – 0.

Сбрасывает настройки эмоции. Параметру arous (arousal – возбуждение) присваивается значение 1, а параметру val (valence – валентность, тон) – 0.

Сбрасывает настройки постобработки. Параметрам ringmod (ring modulation – кольцевая модуляция) и delay (задержка) присваиваются значения 0.

А также сбрасывает настройки рандомизации.

4.12. Графический интерфейс пользователя (GUI)

Поскольку задачей было создание гибкого параметрического синтезатора, решено, что важно разрешить установку ключевых параметров управления через графический интерфейс с помощью соответствующих кнопок и слайдеров (см. рис. 3 и таблицу 2). Это облегчило настройку вокализации в реальном времени, и значительно улучшило процесс проектирования различных звуков. Однако, на практике существует ряд потенциальных зависимостей (как описано в предыдущем разделе). Таким образом, при установке размера тела в определенное значение также устанавливаются: слоговая скорость и частота модуляции (согласно уравнению 1), основная частота (согласно уравнению 2), длина голосового тракта (согласно уравнению 3).

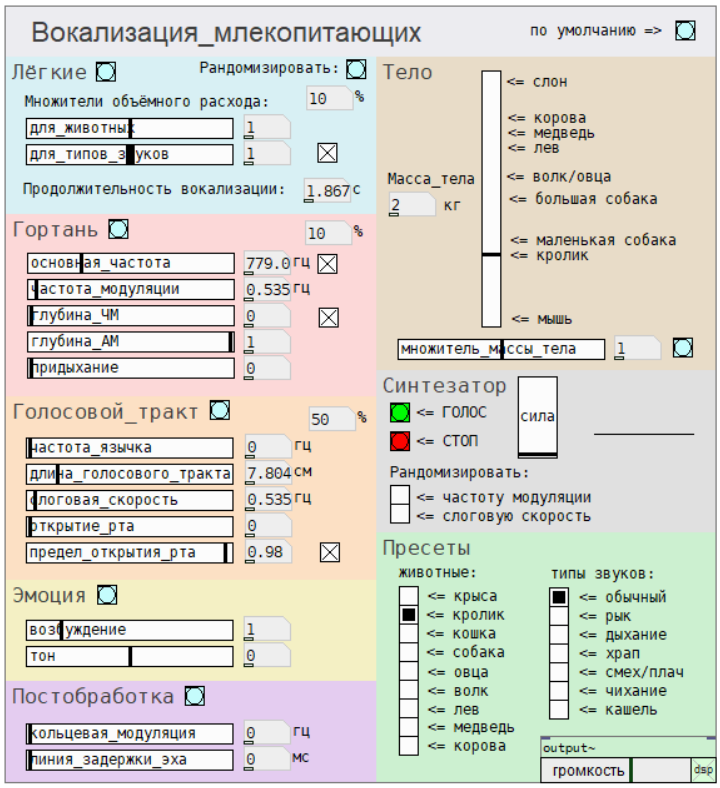


Рисунок 3, Скриншот интерфейса программы

Таблица 2, Параметры управления, предоставленные GUI программы

	Параметр	Значения
Тело	масса тела множитель массы тела	от 0.1 до 5000 кг от 0.5 до 2
Легкие	Множители объёмного расхода: для животных для типов звуков	 от 0 до 2 от 0 до 2
Гортань	основная частота частота модуляции глубина частотной модуляции глубина амплитудной модуляции придыхание	от 30 до 3000 гц от 0 до 20 гц от 0 до 1 от 0 до 1 от 0 до 1
Голосовой тракт	частота язычка длина голосового тракта слоговая скорость открытие рта предел открытия рта	от 0 до 100 гц от 1 до 50 см от 0 до 20 гц от 0 до 1 от 0 до 1
Эмоция	возбуждение тон	от 0.7 до 7 от -1 до 1
Постобработка	кольцевая модуляция линия задержки эха	от 0 до 100 гц от 0 до 100 мс

5. Оценка и сравнение результативности методов

На спектрограмме (рис. 4) можно увидеть, как звук синтезатора копирует особенности голосов животных, записанных на микрофон.

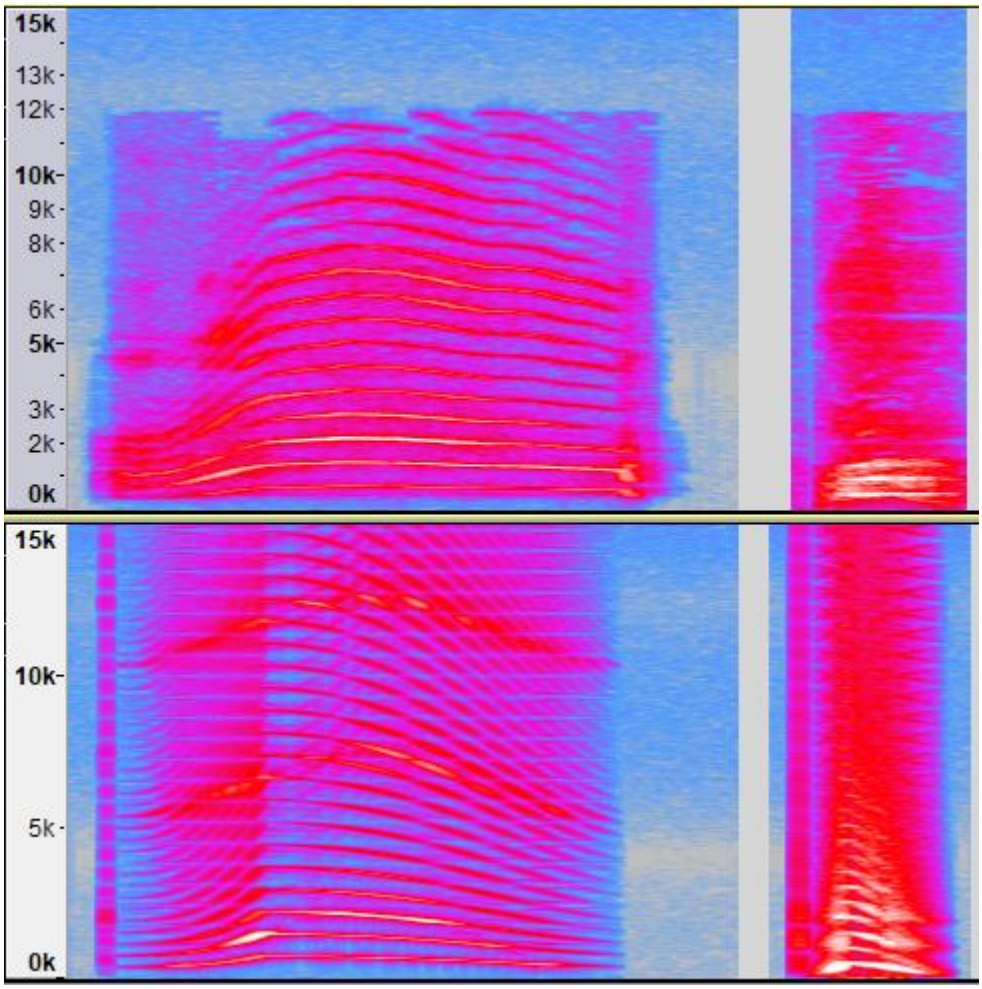


Рисунок 4, Слева кошачье мяуканье, справа лай собаки, сверху запись с микрофона, снизу синтезированный звук

Полученные результаты кажутся приемлемыми с импрессионистской точки зрения. Синтезатор имеет большой потенциал развития.

Заключение

В результате выполнения данной диссертационной работы было выяснено, что синтез может являться достойной альтернативой библиотекам звуковых эффектов, основанным на сэмплах.

Для достижения поставленной цели были выполнены следующие задачи:

1. Изучена предметная область.
2. Изучены средства синтеза звука.
3. Разработана модель синтеза звука.
4. Произведены оценка и сравнение результативности методов.

Список литературы

1. Hendry S. and Reiss J. D. Physical Modeling and Synthesis of Motor Noise for Replication of a Sound Effects Library // 129th AES Convention, San Francisco, 2010, 7 p.
2. Wilkinson W. J. and Reiss J. D. A Synthesis Model for Mammalian Vocalisation Sound Effects // 61st Audio Engineering Society Conference: Audio for Games, 2016, 8 p.
3. Bottcher N. and Serafin S. Design and evaluation of physically inspired models of sound effects in computer games // In Proc. of the Audio Engineering Society 35th International Conference, London, 2009, 6 p.
4. Farnell A. Designing sound // Cambridge: MIT Press, 2010, 689 p.
5. Fitch W. T. Production of vocalizations in mammals // Visual Communication 3, 2006, pp. 115-121.
6. Cook P. Identification of Control Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing // Center of computer research in music and acoustic, 1991, 266 p.
7. Taylor A. M. and Reby D. The contribution of source–filter theory to mammal vocal communication research // Journal of Zoology 280, 2010, pp. 221-236.
8. Fitch W. Tecumseh, Jürgen Neubauer and Hanspeter Herzel Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production // Animal Behaviour 63, 2002, pp. 407-418.
9. Ananthakrishnan, Gopal, et al. An acoustic analysis of lion roars. II: Vocal tract characteristics // Fonetik 2011. Royal Institute of Technology, Stockholm, Sweden, 2011, pp. 5-8.
10. Klemuk, Sarah A., et al. Adapted to roar: functional morphology of tiger and lion vocal folds // PloS one 6.11, 2011, 12 p.

11. Lucero, Jorge C. and Jean Schoentgen Modeling vocal fold asymmetries with coupled van der Pol oscillators // Proceedings of Meetings on Acoustics. Vol. 19. No. 1. Acoustical Society of America, 2013, 8 p.

12. Kelly, John L., and Carol C. Lochbaum Speech synthesis // Proc. Fourth Int. Congr. Acoustics, 1962, pp. 1-4

13. PARO Therapeutic Robot. - Режим доступа: http://www.parorobots.com, свободный.

14. Roger K. Moore A Real-Time Parametric General-Purpose Mammalian Vocal Synthesiser // INTERSPEECH. pp. 2636–2640. San Francisco, 2016.

15. Roger K. Moore and Ben Mitchinson A Biomimetic Vocalisation System for MiRo // Springer International Publishing, pp. 363-374, 2017.

16. P. K. McGregor, Ed., Playback and Studies of Animal Communication // Boston, MA: Springer US, 1992.

17. S. L. Hopp and C. S. Evans Acoustic Communication in Animals // Springer Verlag, 1998.

18. P. J. Rousseeuw and A. M. Leroy Robust Regression and Outlier Detection // New York: Wiley, 1987.

19. J. Worthington, I. S. Young, and J. D. Altringham The relationship between body mass and ventilation rate in mammals // Experimental Biology, vol. 161, pp. 533–536, 1991.

20. E. T. Stathopoulos Oral airflow during vowel production of children and adults // Cleft Palate Journal, vol. 21, no. 4, pp. 277–285, 1984.

21. Y.-T. Wang, J. R. Green, I. S. B. Nip, R. D. Kent, and J. F. Kent Breath group analysis for reading and spontaneous speech in healthy adults // Folia Phoniatrica et Logopaedica, vol. 62, no. 6, pp. 297–302, 2010.

22. N. H. Fletcher A simple frequency-scaling rule for animal communication // Journal of the Acoustical Society of America, vol. 115, no. 5, pp. 2334–2338, 2004.

23. T. Riede and T. Fitch Vocal tract length and acoustics of vocalization in the domestic dog (Canis familiaris) // Journal of Experimental Biology, vol. 202, no. 20, pp. 2859–2867, 1999.

24. W. T. Fitch Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques // Journal of the Acoustical Society of America, vol. 102, no. 2, pp. 1213–1222, 1997.

25. I. R. Titze Acoustic interpretation of resonant voice // Journal of Voice, vol. 15, no. 4, pp. 519–528, 2001.

Глоссарий

Амплитудная модуляция (**АМ**) – вид модуляции, при которой изменяемым параметром несущего сигнала является его амплитуда.

Вэйвшейпер (Waveshaper) – вид искажения, с помощью которого из простых звуков получаются сложные в результате изменения формы сигнала.

Голосовая складка – складка слизистой оболочки гортани, выступающая в ее полость, содержащая голосовую связку и голосовую мышцу.

Звуковой дизайн – вид творческой деятельности по моделированию звуковых пространств и созданию специальных звуковых эффектов для экранных и мультимедийных произведений.

Полосовой фильтр – фильтр, который пропускает частоты, находящиеся в некоторой полосе частот.

Процедурное аудио – нелинейный, зачастую синтетический звук, созданный в реальном времени в соответствии с набором программных правил и изменяющихся входных данных.

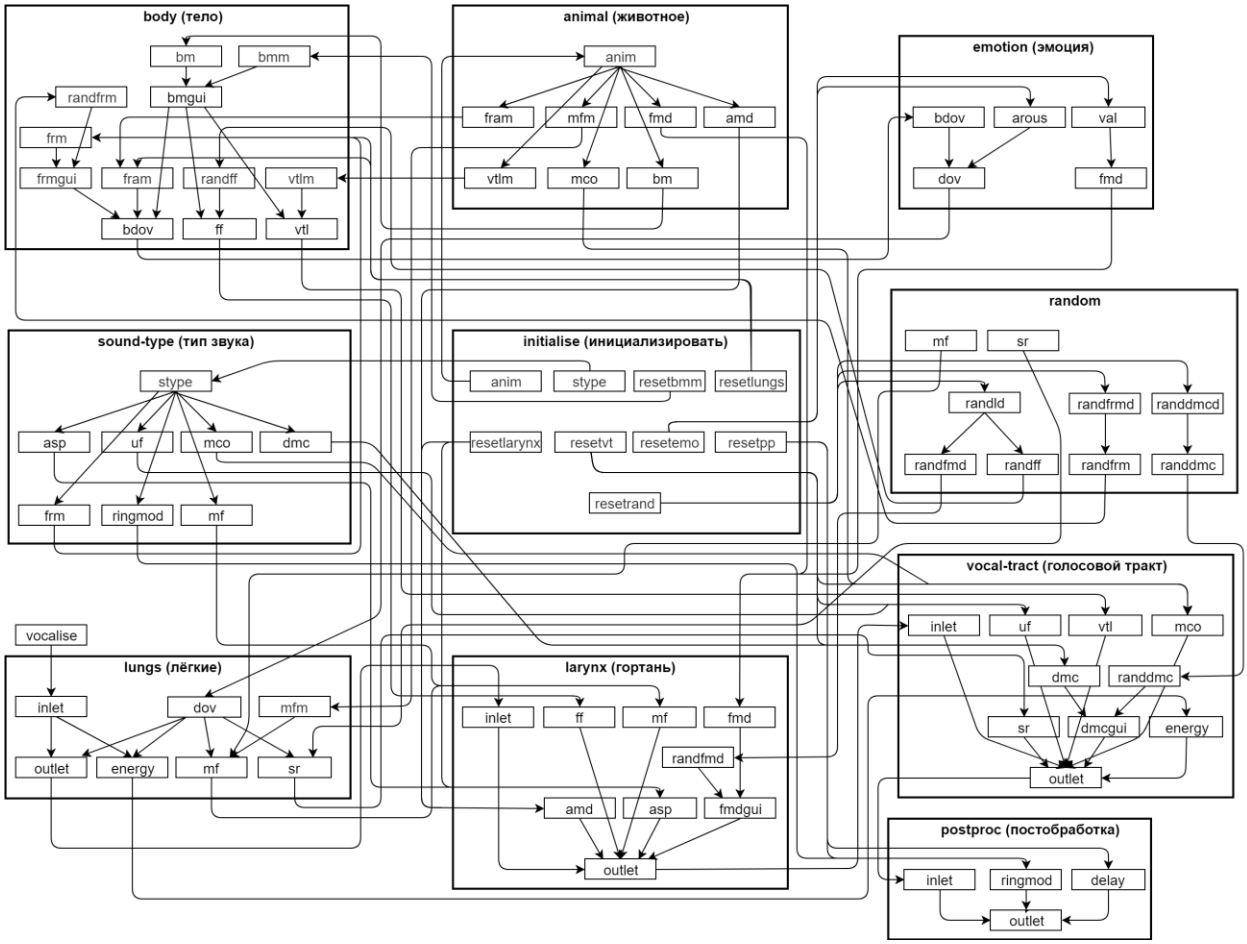
Субтрактивный синтез – это метод синтеза, основанный на вычитании элементов друг из друга. В синтезаторах определяющим элементом субтрактивного синтеза является наличие фильтров. Фильтр «вырезает» часть спектра из звука, формируя нужную тембральную окраску.

Форманта – характерный тон гласных звуков, определяющий их тембр, возникающий в резонаторе и изменяющийся благодаря изменению формы, объема и состояния стенок резонатора.

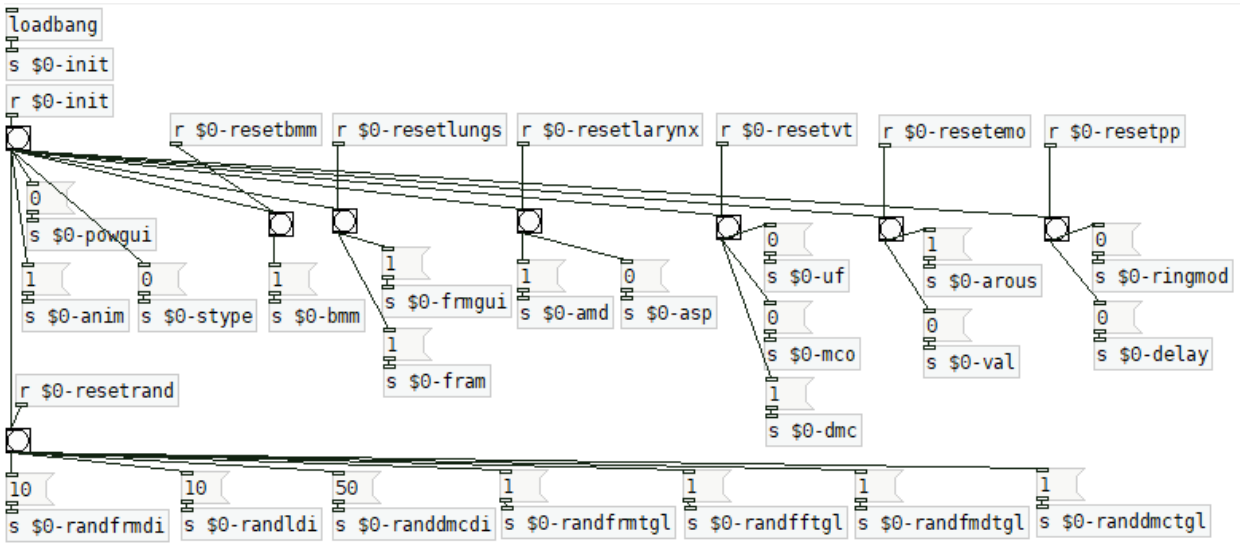
Частотная модуляция (ЧМ, FM) – вид аналоговой модуляции, при котором информационный сигнал управляет частотой несущего колебания.

Приложение

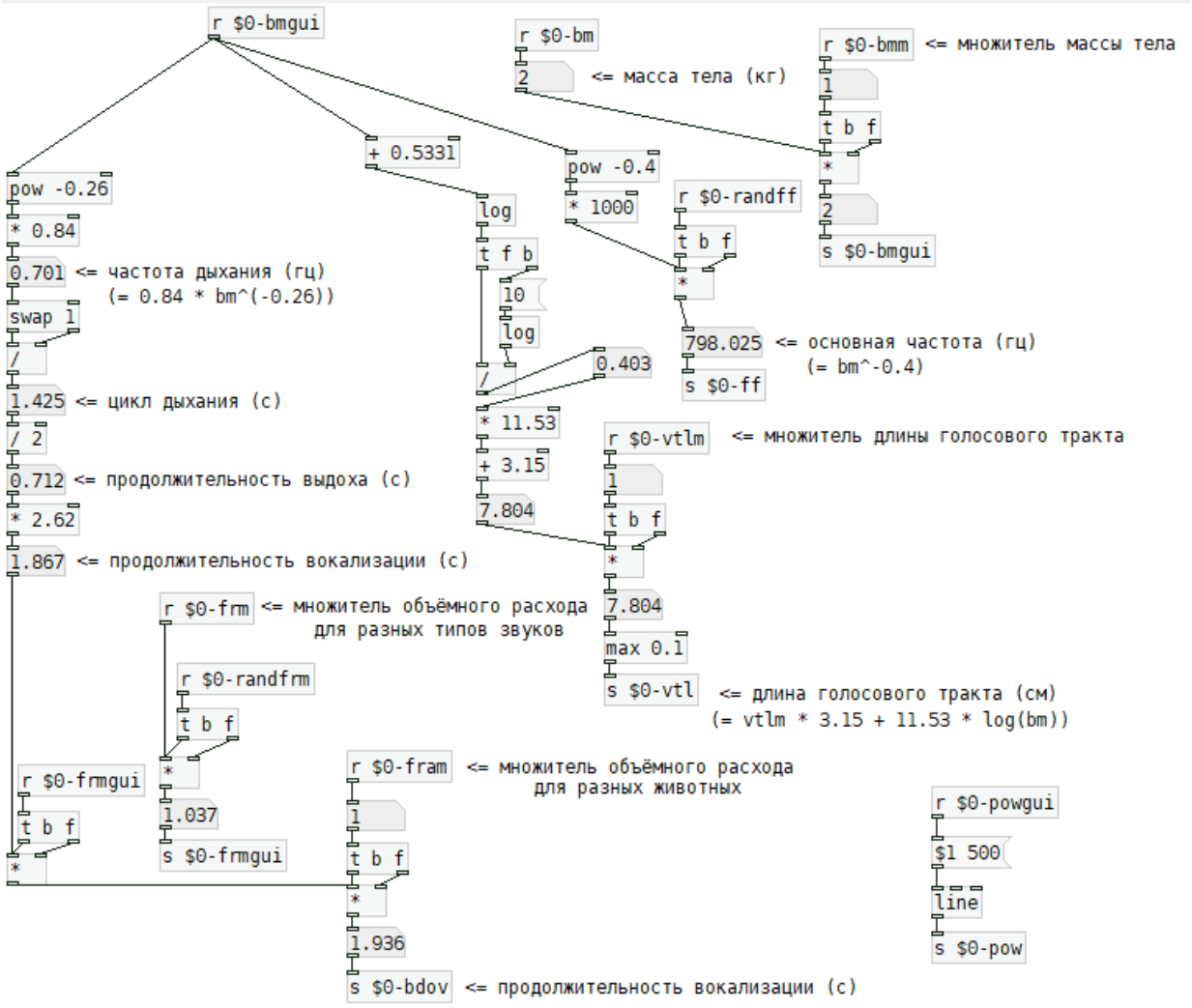
Структура программного обеспечения:



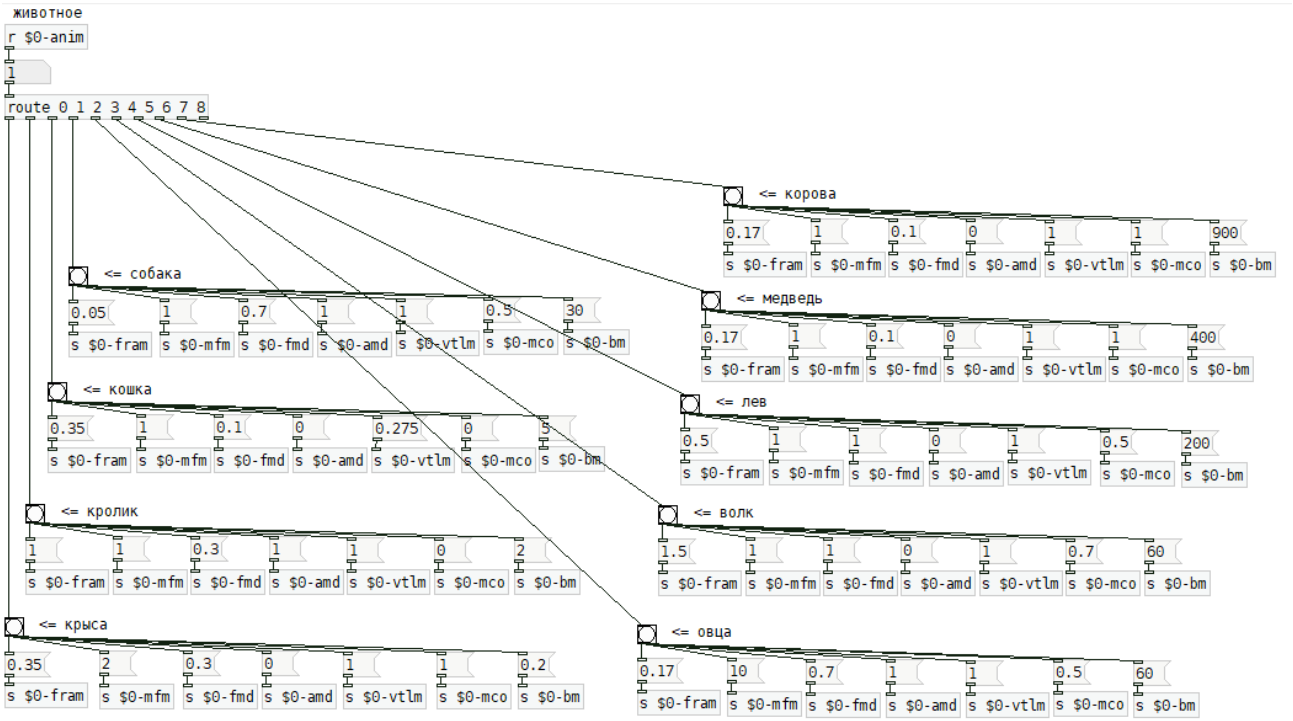
Подпатч initialise:



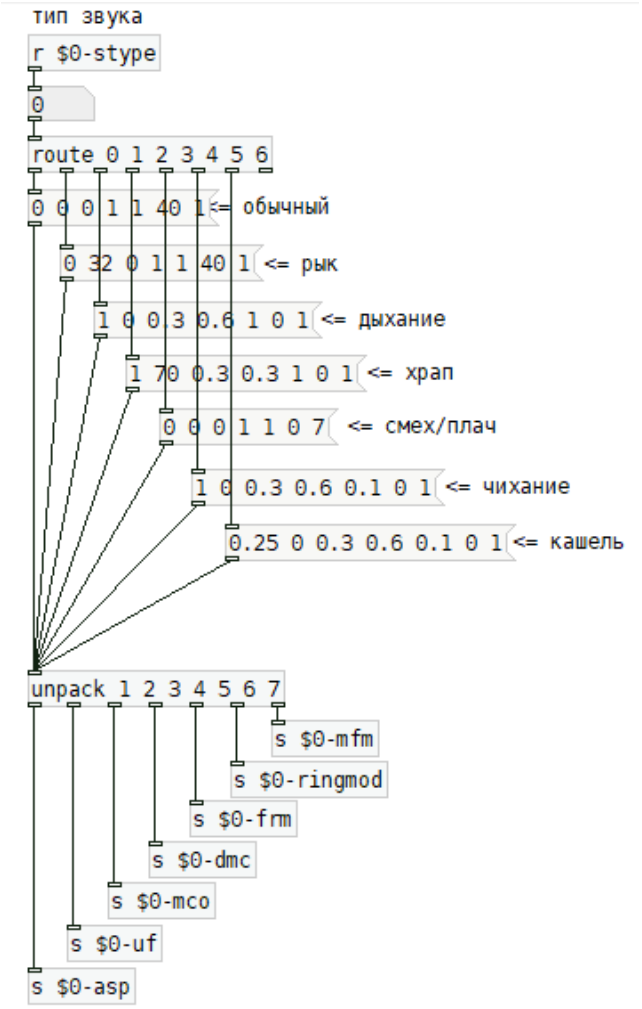
Подпатч body:



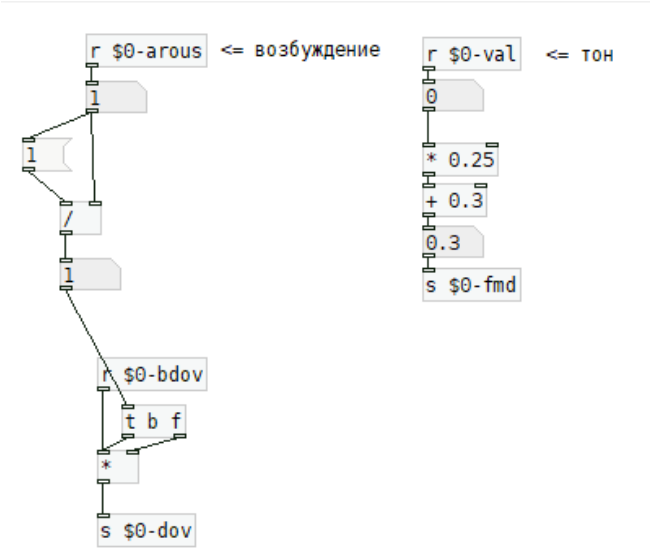
Подпатч animal:



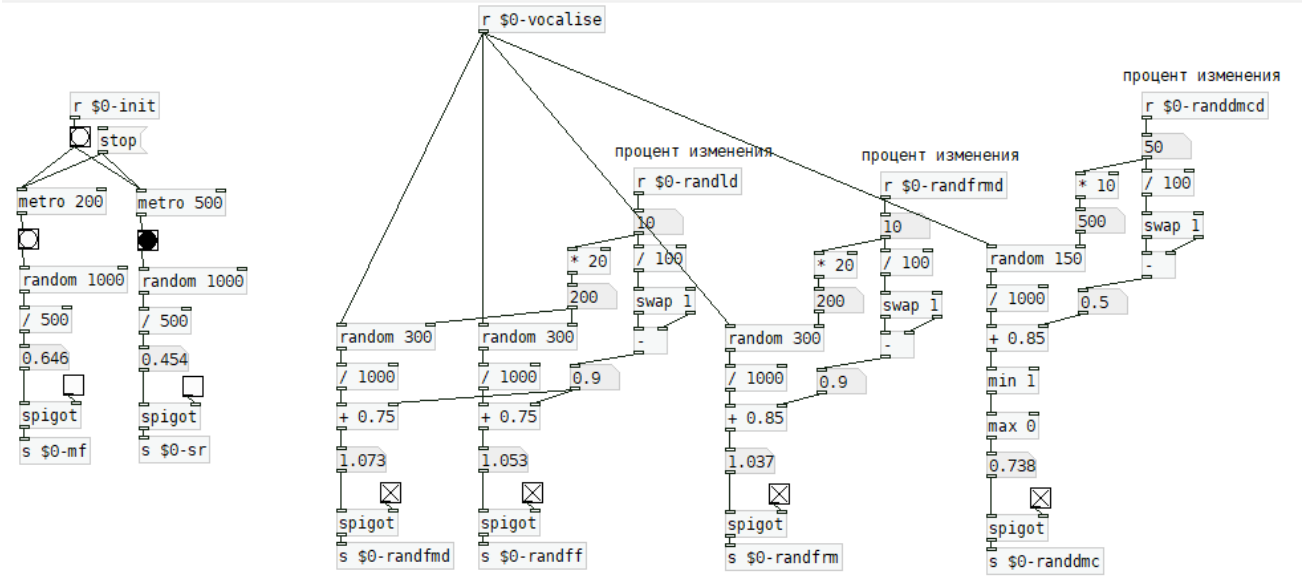
Подпатч sound-type:



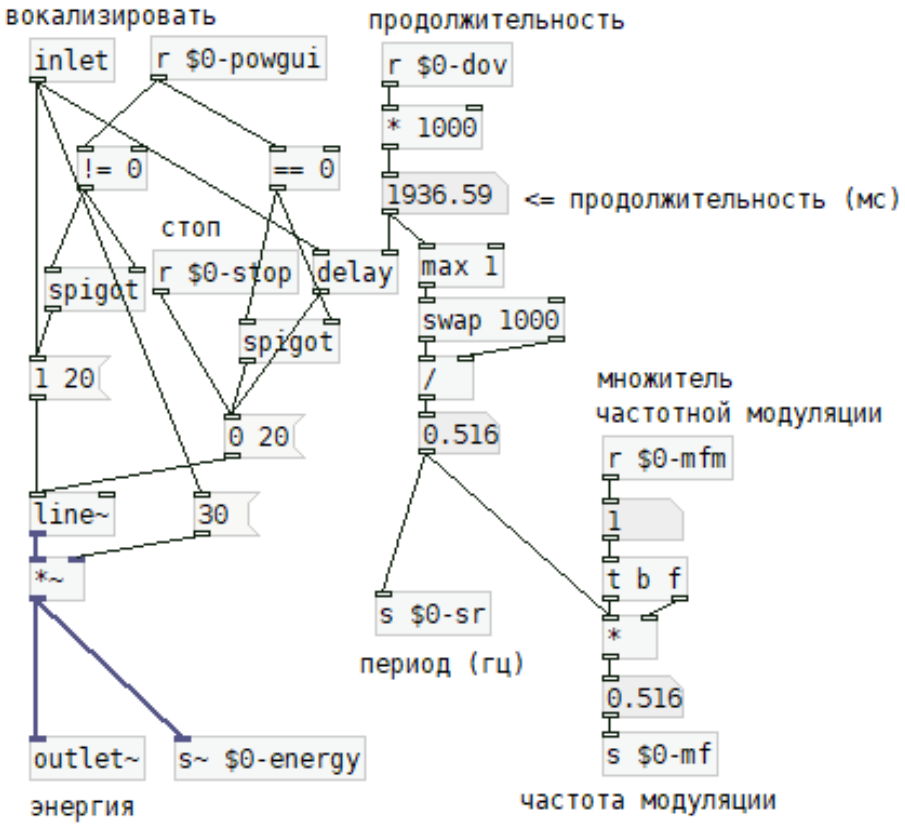
Подпатч emotion:



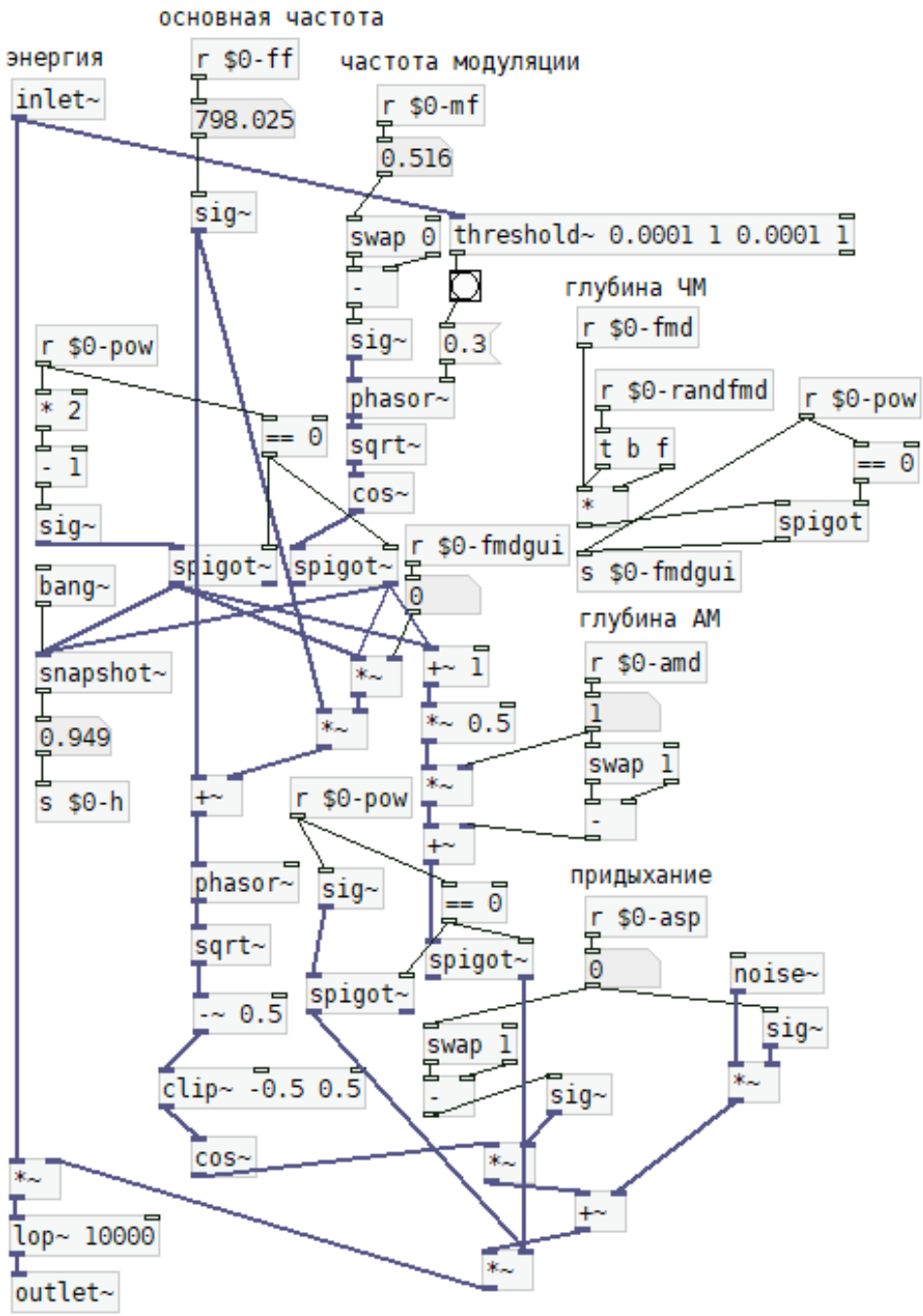
Подпатч random:



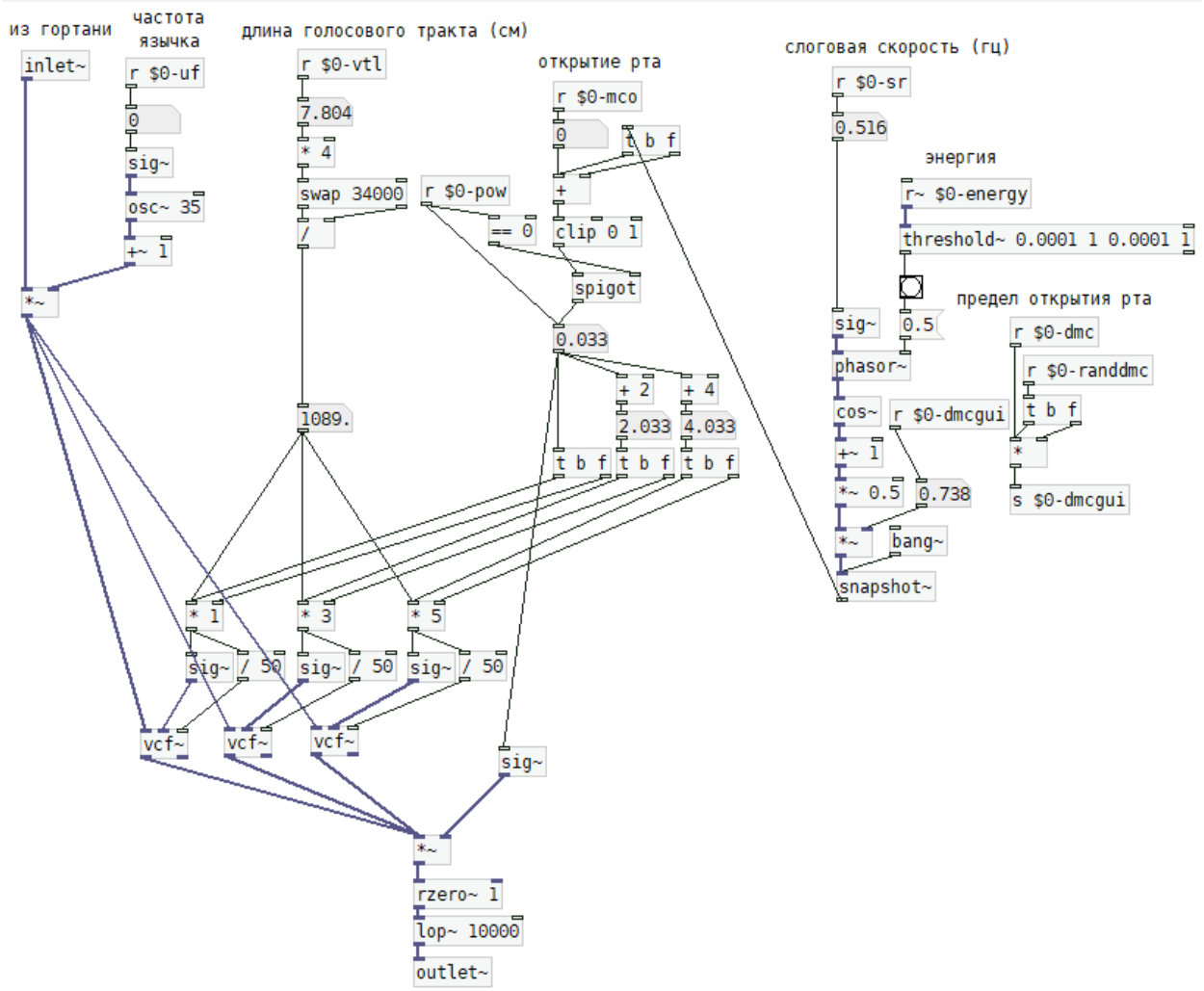
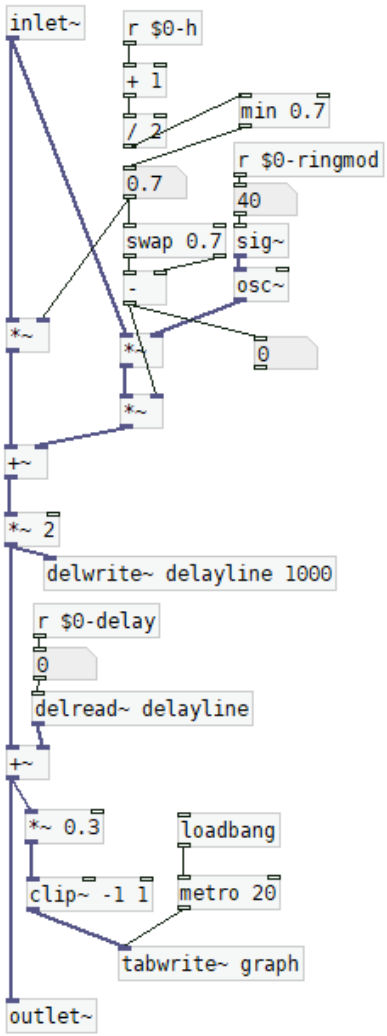
Подпатч lungs:



Подпатч larynx:



Подпатч vocal-tract:



Подпатч postproc: