(19) Евразийское патентное ведомство

(11) 039495 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2022.02.03

(21) Номер заявки
202090169

(22) Дата подачи заявки
2020.01.28

(51) Int. Cl. ***G06T 13/40*** (2006.01)
G10L 21/10 (2006.01)

(54) СПОСОБ И СИСТЕМА ДЛЯ СОЗДАНИЯ МИМИКИ НА ОСНОВЕ ТЕКСТА

(31) **2019144357**
(32) **2019.12.27**
(33) **RU**
(43) **2021.06.30**

(71)(73) Заявитель и патентовладелец:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Ефимов Альберт Рувимович,
Гонноченко Алексей Сергеевич,
Владимиров Михаил Александрович
(RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-A1-20120130717
US-A1-20170308904
US-A1-20190147838
US-A1-20190164327
RU-C1-2708027

(57) Данное техническое решение, в общем, относится к области обработки данных изображения, а в частности, к способу и системе для создания мимики на основе текста. Техническим результатом, достигаемым при решении вышеуказанной технической задачи, является обеспечение возможности создания видеопотока с анимированным изображением 3D-модели головы с размещенной на ней динамической текстурой лицевой маски на основе данных речевого сигнала. Указанный технический результат достигается благодаря осуществлению способа обработки речевого сигнала для формирования видеопотока, выполняемого по меньшей мере одним вычислительным устройством, содержащего этапы, на которых получают данные по меньшей мере одного речевого сигнала; разделяют участки речевого сигнала, содержащие информацию о голосе, на временные окна; формируют для каждого временного окна изображение частотного спектра для получения последовательности изображений частотного спектра; на основе последовательности изображений частотного спектра определяют последовательность данных о множестве координат точек, образующих лицевую маску; размещают лицевую маску на 3D-модели головы для формирования последовательности кадров, содержащих изображение 3D-модели головы с размещенной на ней лицевой маской; на основе последовательности изображений частотного спектра формируют последовательность кадров динамической текстуры лицевой маски; формируют последовательность кадров, содержащих изображение результирующей 3D-модели головы с размещенной на ней динамической текстурой лицевой маски на основе последовательности кадров, содержащих изображение 3D-модели головы с размещенной на ней лицевой маской, и кадров динамической текстуры лицевой маски; формируют последовательность кадров с изображением результирующей 3D-модели головы на фоне сцены; объединяют полученную на предыдущем шаге последовательность кадров в видеопоток.

039495 B1

039495 B1

Область техники

Данное техническое решение, в общем, относится к области обработки данных изображения, а в частности, к способу и системе для создания мимики на основе текста. Представленное решение может быть использовано в процессе генерирования по меньшей мере одной анимированной модели - цифрового аватара на основе данных речи и движений.

Уровень техники

Из уровня техники известны различные решения, направленные на создание цифровых аватаров. Например, известен способ предоставления анимации в реальном времени для персонализированного анимированного аватара, раскрытый в заявке US 2012130717 (A1), опубл. 2012.05.24. В известном решении выполняют обучение одной или нескольких анимированных моделей для предоставления набора вероятностных движений для одной или нескольких верхних частей тела аватара, основываясь, по меньшей мере частично, на данных речи и движения; ассоциирование одной или нескольких заданных фраз эмоциональных состояний с одной или несколькими анимированными моделями; получение речевого ввода в реальном времени; идентификацию эмоционального состояния, которое должно быть выражено, основываясь, по меньшей мере частично, на одной или нескольких заранее определенных фразах, соответствующих, по меньшей мере частично, речевому вводу в реальном времени; и генерирование анимированной последовательности движений одной или нескольких верхних частей тела аватара путем применения одной или нескольких анимированных моделей в ответ на речевой ввод в реальном времени, причем анимированная последовательность движений выражает идентифицированное эмоциональное состояние.

Также известна система удаленного обслуживания клиентов, раскрытая в заявке US 2017308904 (A1), опубл. 2017.10.26. Известная система включает в себя блок интерактивного отображения в месте нахождения клиента, обеспечивающий двустороннюю аудиовизуальную связь с удаленным агентом по обслуживанию/продажам, при этом коммуникация обеспечивается посредством виртуального цифрового ведущего, отображаемого на дисплее. В известную систему также интегрирована виртуальная система "Digital Actor". "Digital Actor" используется в решении для электронного обучения, при производстве фильмов и в качестве ведущего на телевидении, в Интернете или других вещательных приложениях.

Также известен AI-ведущий "Синьхуа", раскрытый в статье "Watch: China's Xinhua unveils the world's first AI news anchor", опубл. 09.11.2018 в Интернет по адресу: https://www.hindustantimes.com/tech/china-s-xinhua-unveils-the-world-s-first-ai-news-anchor/storv-ACUifqUzGSI8IEURbrlmBJ.html.

Представленные решения обладают рядом недостатков, таких как статичная поза головы, обусловленная ограничениями их версии технологии (двухмерный синтез мимики), из-за чего мимика ведущего выглядит скованно и менее естественно.

3D-синтез в свою очередь позволит развивать технологию в full-3D рендер, что существенно расширяет область применения технологии (использование виртуальных студий, "свободная камера" и пр.).

Сущность технического решения

Технической проблемой или задачей, поставленной в данном техническом решении, является создание простого и надежного способа и системы для создания мимики на основе текста.

Техническим результатом, достигаемым при решении вышеуказанной технической задачи, является обеспечение возможности создания видеопотока с анимированным изображением 3D-модели головы с размещенной на ней динамической текстурой лицевой маски на основе данных речевого сигнала. Данные речевого сигнала могут быть произвольно сгенерированы средствами синтеза речи (системы text-to-speech), эмитирующих человеческий голос, соответствующий по своим параметрам голосу диктора.

Указанный технический результат достигается благодаря осуществлению способа обработки речевого сигнала для формирования видеопотока, выполняемого по меньшей мере одним вычислительным устройством, содержащего этапы, на которых

получают данные по меньшей мере одного речевого сигнала;

разделяют участки речевого сигнала, содержащие информацию о голосе, на временные окна;

формируют для каждого временного окна изображение частотного спектра для получения последовательности изображений частотного спектра;

на основе последовательности изображений частотного спектра определяют последовательность данных о множестве координат точек, образующих лицевую маску;

размещают лицевую маску на 3D-модели головы для формирования последовательности кадров, содержащих изображение 3D-модели головы с размещенной на ней лицевой маской;

на основе последовательности изображений частотного спектра формируют последовательность кадров динамической текстуры лицевой маски;

формируют последовательность кадров, содержащих изображение результирующей 3D-модели головы с размещенной на ней динамической текстурой лицевой маски на основе последовательности кадров, содержащих изображение 3D-модели головы с размещенной на ней лицевой маской, и кадров динамической текстуры лицевой маски;

формируют последовательность кадров с изображением результирующей 3D-модели головы на фо-

не сцены;

объединяют полученную на предыдущем шаге последовательность кадров в видеопоток.

В одном из частных примеров осуществления способа дополнительно выполняют этапы, на которых при получении данных речевого сигнала определяют шум-голос; выделяют участки речевого сигнала, содержащие информацию о голосе; причем при определении шума-голоса учитываются данные фонетической разметки.

В другом частном примере осуществления способа дополнительно выполняют этап геометрической валидации данных о множестве координат точек, сформированных для временного окна, содержащихся в последовательности данных о множестве координат точек, образующих лицевую маску. В другом частном примере осуществления способа упомянутый этап геометрической валидации данных о множестве координат точек содержит этапы, на которых на основе данных фонетической разметки голоса определяют контрольные координаты точек лицевой маски, валидацию которых необходимо выполнить; определяют расстояние между контрольными точками лицевой маски, информация о которых содержится в данных о множестве координат точек, сформированных для временного окна, и сравнивают его с заданным значением расстояния для этих контрольных точек согласно данным о фонетических разметках голоса; причем если определенное расстояние между контрольными точками лицевой маски не превышает заданного значения расстояния между этими контрольными точками, то данные о множестве координат точек проходят валидацию.

В другом частном примере осуществления способа размещение лицевой маски на 3D-модели головы осуществляется посредством прикрепления лицевой маски по известным пограничным вершинам 3D-модели головы и соответствующим для этих вершин точкам лицевой маски.

В другом частном примере осуществления способа дополнительно выполняют этап, на котором размещают на изображениях 3D-модели головы текстуры зубов и языка согласно положению лицевой маски.

В другом частном примере осуществления способа дополнительно выполняют этап морфинга изображений результирующей 3D-модели головы.

В другом частном примере осуществления способа дополнительно выполняют этап цветокоррекции и совмещения динамических текстур для устранения пульсирующего изменения цвета лицевой маски посредством слияния по цвету упомянутых изображений 3D-модели головы, содержащихся на предыдущих и последующих кадрах.

В другом предпочтительном варианте осуществления заявленного решения представлено устройство обработки речевого сигнала, содержащее по меньшей мере одно вычислительное устройство и по меньшей мере одно устройство памяти, содержащее машиночитаемые инструкции, которые при их исполнении по меньшей мере одним вычислительным устройством выполняют указанный выше способ.

Краткое описание чертежей

Признаки и преимущества настоящего изобретения станут очевидными из приводимого ниже подробного описания изобретения и прилагаемых чертежей, на которых

на фиг. 1 представлена принципиальная архитектура технологии;

на фиг. 2 - пример системы для создания мимики на основе текста;

на фиг. 3 - пример фонетической разметки фразы "Добрый день";

на фиг. 4 - пример общего вида вычислительного устройства.

Подробное описание изобретения

Ниже будут описаны понятия и термины, необходимые для понимания данного технического решения.

В данном техническом решении под системой подразумевается, в том числе компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, четко определенную последовательность операций (действий, инструкций).

Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы).

Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройств хранения данных. В роли устройства хранения данных могут выступать, но не ограничиваясь, жесткие диски (HDD), флеш-память, ПЗУ (постоянное запоминающее устройство), твердотельные накопители (SSD), оптические приводы.

Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

Видео фрейм (video frame) - видеокадр, кадр изображения, телевизионный кадр.

Риг - термин в компьютерной анимации, который описывает набор зависимостей между управляющими и управляемыми элементами, созданный таким образом, чтобы управляющих элементов было меньше, чем управляемых.

Программно-аппаратный комплекс "Цифровой аватар" базируется на 2 технологических комплек-

сах (см. фиг. 1):

синтез речи на основе произвольного текста [1], в котором текстовую информацию (TXT) направляют на вход нейронной сети (DNN) для его обработки посредством технологии Text-to-speech (TTS) и получения сгенерированного нейронной сетью речевого сигнала (VOICE);

синтез изображения на основе сгенерированной речи [2], в котором речевой сигнал (VOICE) направляют на вход нейронной сети (DNN) для формирования мимики (Mimic) и изображения цифрового аватара (VIDEO).

В данном технологическом комплексе также используются несколько технологий: известные технологии постобработки композитинга (объединения изображений в кадре) и новой технологии "Генерации мимики цифрового аватара на основе голоса".

В соответствии со схемой, приведенной на фиг. 2, система 100 для создания мимики на основе текста, выполняющая описанный выше синтез изображения на основе сгенерированной речи, содержит устройство 10 памяти и устройство 20 обработки речевого сигнала.

Устройство 10 памяти может представлять собой оперативную или постоянную память, предназначенную для хранения речевых сигналов, синтезированных на основе текста описанным выше методом, например, в виде звуковых дорожек. Для генерации упомянутого речевого сигнала могут использоваться решения, предлагаемые компаниями ЦРТ, АБК и Google. Дополнительно вместе с речевым сигналом могут быть сохранены данные фонетической разметки голоса, сгенерированные при синтезировании текста. Фонетическая разметка голоса может быть сформирована известными из уровня техники методами, например, посредством решений, предлагаемых компанией ЦРТ. Данные фонетической разметки содержат указатели времени начала и окончания произнесения звука в рамках речевого сигнала - звуковой дорожки. Пример данных фонетической разметки представлен на фиг. 3, на котором изображен речевой сигнал фразы "Добрый день", а также в нижней части чертежа - данные фонетической разметки.

Устройство 20 обработки речевого сигнала может быть выполнено на базе по меньшей мере одного вычислительного устройства и содержать модуль 21 выделения сигнала/шума, модуль 22 артикуляционного синтеза, модуль 23 позиционирования лицевой маски, модуль 24 формирования кадра (фрейма) динамической текстуры и модуль 25 обработки кадров. Модуль 21 выделения сигнала/шума и модуль 25 обработки кадров могут быть выполнены на базе вычислительного устройства, оснащенного специализированным программным обеспечением для выполнения приписанных им в настоящей заявке функций. Модуль 22 артикуляционного синтеза, модуль 23 позиционирования лицевой маски и модуль 24 формирования кадра динамической текстуры могут быть выполнены на базе свёрточных нейронных сетей, заранее обученных на выборке данных, отобранной разработчиком данных модулей.

В соответствии с заложенным программным алгоритмом устройство 20 обработки речевого сигнала обращается к устройству 10 памяти для извлечения данных по меньшей мере одного речевого сигнала, которые могут быть сохранены, например, при частоте дискретизации 22 кГц моноканал 16 бит, которые поступают в модуль 21 выделения сигнала/шума. Дополнительно вместе с данными речевого сигнала из упомянутого устройства 10 могут быть извлечены данные фонетической разметки голоса для данного речевого сигнала.

При получении данных речевого сигнала упомянутый модуль 21 определяет на основе амплитудно-фазовой частотной характеристики (АФЧХ) звуковой волны (аудиопотока речи) шум-голос, выделяет участки, содержащие информацию о голосе и размечает остальные фрагменты речевого сигнала, не содержащие информацию о голосе, соответствующими тайм-кодами (временным кодом), которые маркируются как молчание. В случае поступления данных фонетической разметки голоса, в частности данных о времени начала и окончания произнесения звука, эти данные также учитываются упомянутым модулем 21 при определении шум-голос, в связи с чем улучшается качество его отделения. Далее данные речевого сигнала с выделенными участками, содержащими информацию о голосе, поступают в модуль 22 артикуляционного синтеза, который разделяет участки, содержащие информацию о голосе, на временные окна, (например, равные 200 мс), причем одно временное окно соответствует одному кадру (фрейму) видео. Дополнительно в упомянутый модуль 22 могут быть направлены данные фонетической разметки голоса.

После того как участки речевого сигнала, содержащие информацию о голосе, были разделены на временные окна, модуль 22 артикуляционного синтеза преобразует АФЧХ звуковой волны для каждого временного окна, содержащего информацию о голосе, в изображение частотного спектра, например двумерное изображение графика зависимости относительной энергии звуковых колебаний от частоты. Упомянутое преобразование может быть выполнено посредством преобразования Фурье - семейства математических методов, основанных на разложении исходной непрерывной функции от времени на совокупность базисных гармонических функций (в качестве которых выступают синусоидальные функции) различной частоты, амплитуды и фазы. Из определения видно, что основная идея преобразования заключается в том, что любую функцию можно представить в виде бесконечной суммы синусоид, каждая из которых будет характеризоваться своей амплитудой, частотой и начальной фазой. Таким образом, в упомянутом модуле 22 формируется последовательность изображений частотного спектра. Далее модуль 22 артикуляционного синтеза на основе последовательности изображений частотного спектра определяет последовательность данных о множестве координат точек, образующих лицевую маску. Для этого каж-

дое изображение частотного спектра подается на вход свёрточным нейронным сетям, которыми упомянутый модуль 22 может быть оснащен, в которых полученное изображение преобразуется в матрицу изображения, содержащую значения пикселей изображения, после чего полученная матрица изображения умножается на матрицу (ядро) свертки поэлементно, где матрица свёртки - это матрица коэффициентов, заданная разработчиком для каждого слоя свёрточной нейронной сети в зависимости он требуемой предварительной обработки поступившего на вход изображения частотного спектра. Методы подбора коэффициентов матрицы свертки широко известны из уровня техники и далее не будут раскрыты в настоящей заявке (см., например, статью "Матричные фильтры обработки изображений", опубл. 25.04.2012 в Интернет по адресу: https://habr.com/ru/post/142818/).

Далее на основе результатов умножения матрицы изображения на матрицу свертки, полученных на каждом слое свёрточной нейронной сети, формируется выходное изображение частотного спектра, которое подается на вход специализированной нейронной сети, заранее обученной на различных изображениях частотных спектров и соответствующих этим изображениям множестве координат точек (landmarks), образующих лицевую маску, на выходе которой упомянутый модуль 22 получает данные о множестве координат точек, образующих лицевую маску, для каждого временного окна. Таким образом, в модуле 22 формируется последовательность данных о множестве координат точек, образующих лицевую маску. Методы определения множества координат точек, образующих лицевую маску, на основе изображения лица широко известны из уровня техники (см, например, статью "Распознавание лиц. Создаем и примеряем маски", опубл. 12.12.2017 в Интернет по адресу: https://habr.com/ru/company/ epam systems/blog/343514/) и более подробно не будут раскрыты в материалах настоящей заявки. Изображения лица, на основе которых выполняют определение множества координат точек, могут быть получены посредством обработки данных видеоизображений лица, например, диктора, разделенное на временные окна, соответствующие временным окнам для речевого сигнала.

Дополнительно модуль 22 артикуляционного синтеза может быть выполнен с возможностью геометрической валидации упомянутых данных о множестве координат точек, сформированных для временного окна, содержащихся в последовательности данных о множестве координат точек, образующих лицевую маску. Для этого модуль 22 артикуляционного синтеза на основе данных фонетической разметки голоса определяет контрольные координаты точек лицевой маски, валидацию которых необходимо выполнить. Например, данные о фонетической разметки голоса могут указывать на наличие звука "б", "м" или "п" во временном окне, в связи с чем на основе данной информации модулем 22 артикуляционного синтеза могут быть определены контрольные точки лицевой маски, характеризующие расположение центров поверхности верхней и нижней губ, уголков губ и пр. Далее упомянутый модуль 22 определяет расстояние между контрольными точками лицевой маски, информация о которых содержится в данных о множестве координат точек, сформированных для временного окна, и сравнивает его с заданным значением расстояния для этих контрольных точек согласно данным о фонетической разметки голоса для данного временного окна. Например, для данных о фонетической разметке голоса, указывающих на наличие звука "б", "м" или "п" во временном окне, значение расстояния между контрольными точками, характеризующими расположение центров поверхности верхней и нижней губ, может равняться минимальному расстоянию, близкому к 0 (нулю), для обеспечения визуализации смыкания губ на 3D-модели головы. Если определенное расстояние между контрольными точками лицевой маски больше заданного значения расстояния между этими контрольными точками, т.е. больше минимального расстояния, то данные о множестве контрольных точек не проходят валидацию, а модуль 22 артикуляционного синтеза осуществляет возврат на предыдущий этап, и изображение частотного спектра, содержащееся в упомянутой последовательности изображений, на основе которого были определены данные о множестве координат точек, не прошедшие валидацию, снова подается описанным выше способом на вход свёрточной нейронной сети до получения данных о множестве координат точек, проходящей проверку (валидацию).

Если определенное расстояние между контрольными точками лицевой маски не превышает заданного значения расстояния между этими контрольными точками, то данные о множестве координат точек проходят валидацию. Информация о контрольных точках для данных фонетической разметки и значения расстояния между контрольными точками может быть заранее задана разработчиком модуля 22 артикуляционного синтеза.

Далее последовательность данных о множестве координат точек, прошедших валидацию, направляются в модуль 23 позиционирования лицевой маски, который извлекает из памяти, которой он дополнительно может быть оснащен, 3D-модель головы, например головы диктора, и выполняет размещение лицевой маски на 3D-модели головы, после чего формирует кадр (фрейм) с изображением 3D-модели с размещенной на ней лицевой маской. Позиционирование лицевой маски на 3D-модель головы может осуществляться посредством ее прикрепления по известным пограничным вершинам 3D-модели головы и соответствующим для этих вершин точкам лицевой маски, заданными разработчиком упомянутого модуля 23. Таким образом, на выходе упомянутого модуля 23 формируется последовательность кадров, содержащих изображение 3D-модели головы с размещенной на ней лицевой маской.

В альтернативном варианте реализации заявленного решения пограничные вершины 3D-модели го-

ловы и соответствующие для этих вершин точки лицевой маски могут быть определены известными из уровня техники методами, например посредством библиотеки dlib с помощью последовательного выполнения нахождения лица в кадре, нахождения на нем неподвижных точек и получения их координат.

Дополнительно в памяти упомянутого модуля 23 может быть сохранены текстуры зубов и языка для их размещения на изображении 3D-модели головы согласно положению лицевой маски. Положение текстур зубов и языка может быть заданы разработчиком модуля 23 исходя из анатомических свойств человека. Привязка зубов и языка осуществляется в рамках проектирования 3D-модели. Такие элементы обладают ригами с жесткой логикой работы и относятся к стандартной практике создания "скелета" 3D-модели.

Также последовательность изображений частотного спектра направляется в модуль 24 формирования кадров динамической текстуры, который для каждого временного окна на основе изображения частотного спектра формирует кадр динамической текстуры лицевой маски, которые далее объединяются в последовательность кадров динамической текстуры лицевой маски, таким образом, обеспечивая формирование дискретного изображения текстуры лицевой маски. Динамическая текстура лицевой маски может содержать информацию о участках лица, которые следует отобразить с покраснением, пульсацией, изменением фактуры лица и пр. Для формирования кадров динамической текстуры лицевой маски упомянутый модуль 24 может быть оснащен сверточной нейронной сетью, заранее обученной на различных изображениях частотных спектров и соответствующих этим изображениям кадрах динамической текстуры лицевой маски. Изображения лица, на основе которых формируют кадры динамической текстуры лицевой маски, также могут быть получены посредством обработки данных видеоизображений лица, например, диктора, разделенное на временные окна, соответствующие временным окнам для речевого сигнала. Соответственно, изображения частотных спектров подаются на вход упомянутой нейронной сети, а на ее выходе формируются кадры динамической текстуры лицевой маски. Последовательности кадров, сформированные модулем 23 позиционирования лицевой маски и модулем 24 формирования кадров динамической текстуры, далее направляются в модуль 25 обработки кадров, который извлекает из кадра о динамической текстуре лицевой маски, например, первого кадра в последовательности кадров динамической текстуры лицевой маски, информацию о динамической текстуре лицевой маски, и размещает динамическую текстуру лицевой маски на кадре, в частности на первом кадре в последовательности кадров, содержащем изображение 3D-модели головы с размещенной на ней лицевой маской. Аналогичным образом динамическая текстура лицевой маски размещается на всех кадрах в последовательности кадров, содержащих изображение 3D-модели головы с размещенной на ней лицевой маской, в соответствии с номерами этих кадров и номерами тех же кадров, содержащих информацию о динамической текстуре лицевой маски. Таким образом, в модуле 25 обработки кадров формируется последовательность кадров, содержащих изображение результирующей 3D-модели головы с размещенной на ней динамической текстурой лицевой маски.

Дополнительно для придания монолитного вида модуль 25 обработки кадров может быть оснащен матричными фильтрами обработки изображений, широко используемыми в уровне техники для обработки изображений. Посредством данных фильтров упомянутый модуль 25 может выполнить анализ на сформированных кадрах, содержащих изображение результирующей 3D-модели головы, границ перехода между лицевой маской и 3D-модели головы, заполнить разрывы между изображениями 3D-модели головы и динамической текстурой лицевой маски, а также выполнить размытие цветовых границ между текстурами.

Также дополнительно модуль 25 обработки кадров может быть оснащен соответствующей 3D-моделью для формирования характеристик поверхности 3D-модели головы (рельеф), которые могут быть совмещены с такими параметрами как "свойства материала" (базовый оттенок, фактура).

После этого модуль 25 обработки кадров формирует последовательность кадров с изображением результирующей 3D-модели головы на фоне сцены и текстурой зубов и языка на основе данных последовательности кадров, содержащих изображение результирующей 3D-модели головы. Для этого упомянутый модуль 25 извлекает изображение заранее заготовленной сцены, например изображение какого-либо помещения, добавляет в изображение сцены изображение результирующей 3D-модели головы, содержащееся в первом кадре упомянутой выше последовательности кадров, текстуры зубов и языка для данного изображения 3D-модели головы согласно положению лицевой маски, после чего формирует кадр с изображением результирующей 3D-модели головы на фоне сцены, т.е. с динамической текстурой лицевой маски и текстурой зубов и языка. Аналогичным образом модуль 25 обработки кадров формирует кадры для второго и последующих кадров, содержащих изображение результирующей 3D-модели головы, вследствие чего в модуле 25 обработки кадров формируется последовательность кадров с изображением результирующей 3D-модели головы на фоне сцены с текстурой зубов и языка на основе данных последовательности кадров, содержащих изображение результирующей 3D-модели головы. Далее модуль 25 обработки кадров осуществляет объединение полученной последовательности кадров в видеопоток, причем на этапе объединения последовательности кадров упомянутый модуль 25 может быть выполнен с возможностью морфинга изображений результирующей 3D-модели головы с размещенной на ней динамической текстурой лицевой маски для обеспечения плавности переходов между упомянутыми

изображениями 3D-модели головы и устранения резкого изменения формы лицевой маски посредством слияния по геометрии упомянутых изображений 3D-модели головы, содержащихся на предыдущих и последующих кадрах, а также с возможностью цветокоррекции и совмещения динамических текстур для устранения пульсирующего изменения цвета лицевой маски посредством слияния по цвету упомянутых изображений 3D-модели головы, содержащихся на предыдущих и последующих кадрах. Технологии морфинга изображений, цветокоррекции и совмещения динамических текстур широко известны из уровня техники и более подробно не будут раскрыты в настоящей заявке.

Полученный видеопоток устройством 20 обработки речевого сигнала известными из уровня техники методами объединяется с данными речевого сигнала, после чего видеопоток сохраняется в устройстве 10 памяти и может быть выведен на устройства вывода информации по соответствующему запросу. При необходимости полученный видеопоток может быть сжат с целью уменьшения занимаемого объёма данных.

В общем виде (см. фиг. 4) вычислительное устройство содержит объединенные общей шиной информационного обмена один или несколько процессоров (201), средства памяти, такие как ОЗУ (202) и ПЗУ (203), интерфейсы ввода/вывода (204), устройства ввода/вывода (205) и устройство для сетевого взаимодействия (206).

Процессор (201) (или несколько процессоров, многоядерный процессор и т.п.) может выбираться из ассортимента устройств, широко применяемых в настоящее время, например, таких производителей, как Intel™, AMD™, Apple™, Samsung Exynos™, MediaTEK™, Qualcomm Snapdragon™ и т.п. Под процессором или одним из используемых процессоров в устройстве (200) также необходимо учитывать графический процессор, например GPU NVIDIA или Graphcore, тип которых также является пригодным для полного или частичного выполнения способа, а также может применяться для обучения и применения моделей машинного обучения в различных информационных системах.

ОЗУ (202) представляет собой оперативную память и предназначено для хранения исполняемых процессором (201) машиночитаемых инструкций для выполнения необходимых операций по логической обработке данных. ОЗУ (202), как правило, содержит исполняемые инструкции операционной системы и соответствующих программных компонент (приложения, программные модули и т.п.). При этом в качестве ОЗУ (202) может выступать доступный объем памяти графической карты или графического процессора.

ПЗУ (203) представляет собой одно или более устройств постоянного хранения данных, например жесткий диск (HDD), твердотельный накопитель данных (SSD), флэш-память (EEPROM, NAND и т.п.), оптические носители информации (CD-R/RW, DVD-R/RW, BlueRay Disc, MD) и др.

Для организации работы компонентов устройства (200) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (204). Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь, PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п.

Для обеспечения взаимодействия пользователя с устройством (200) применяются различные средства (205) В/В информации, например клавиатура, дисплей (монитор), сенсорный дисплей, тач-пад, джойстик, манипулятор мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п. Средство сетевого взаимодействия (206) обеспечивает передачу данных посредством внутренней или внешней вычислительной сети, например Интранет, Интернет, ЛВС и т.п. В качестве одного или более средств (206) может использоваться, но не ограничиваться, Ethernet карта, GSM модем, GPRS модем, LTE модем, 5G модем, модуль спутниковой связи, NFC модуль, Bluetooth и/или BLE модуль, Wi-Fi модуль и др.

Дополнительно могут применяться также средства спутниковой навигации в составе устройства (200), например GPS, ГЛОНАСС, BeiDou, Galileo. Конкретный выбор элементов устройства (200) для реализации различных программно-аппаратных архитектурных решений может варьироваться с сохранением обеспечиваемого требуемого функционала.

Модификации и улучшения вышеописанных вариантов осуществления настоящего технического решения будут ясны специалистам в данной области техники. Предшествующее описание представлено только в качестве примера и не несет никаких ограничений. Таким образом, объем настоящего технического решения ограничен только объемом прилагаемой формулы изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ обработки речевого сигнала для формирования видеопотока, выполняемый по меньшей мере одним вычислительным устройством, содержащий этапы, на которых

получают данные по меньшей мере одного речевого сигнала;

разделяют участки речевого сигнала, содержащие информацию о голосе, на временные окна;

формируют для каждого временного окна изображение частотного спектра для получения последовательности изображений частотного спектра;

на основе последовательности изображений частотного спектра определяют последовательность данных о множестве координат точек, образующих лицевую маску;

размещают лицевую маску на 3D-модели головы для формирования последовательности кадров, содержащих изображение 3D-модели головы с размещенной на ней лицевой маской;

на основе последовательности изображений частотного спектра формируют последовательность кадров динамической текстуры лицевой маски;

формируют последовательность кадров, содержащих изображение результирующей 3D-модели головы с размещенной на ней динамической текстурой лицевой маски на основе последовательности кадров, содержащих изображение 3D-модели головы с размещенной на ней лицевой маской, и кадров динамической текстуры лицевой маски;

формируют последовательность кадров с изображением результирующей 3D-модели головы на фоне сцены;

объединяют полученную на предыдущем шаге последовательность кадров в видеопоток.

2. Способ по п.1, характеризующийся тем, что дополнительно выполняют этапы, на которых

при получении данных речевого сигнала определяют шум-голос;

выделяют участки речевого сигнала, содержащие информацию о голосе;

причем при определении шума-голоса учитываются данные фонетической разметки.

3. Способ по п.1, характеризующийся тем, что дополнительно выполняют этап геометрической валидации данных о множестве координат точек, сформированных для временного окна, содержащихся в последовательности данных о множестве координат точек, образующих лицевую маску.

4. Способ по п.3, характеризующийся тем, что упомянутый этап геометрической валидации данных о множестве координат точек содержит этапы, на которых

на основе данных фонетической разметки голоса определяют контрольные координаты точек лицевой маски, валидацию которых необходимо выполнить;

определяют расстояние между контрольными точками лицевой маски, информация о которых содержится в данных о множестве координат точек, сформированных для временного окна, и сравнивают его с заданным значением расстояния для этих контрольных точек согласно данным о фонетических разметках голоса;

причем если определенное расстояние между контрольными точками лицевой маски не превышает заданного значения расстояния между этими контрольными точками, то данные о множестве координат точек проходят валидацию.

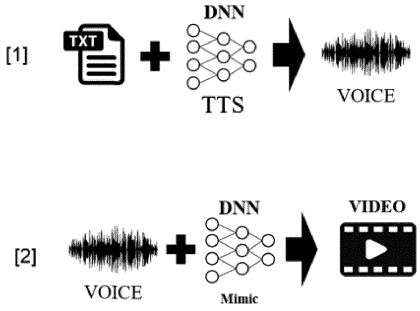
5. Способ по п.1, характеризующийся тем, что размещение лицевой маски на 3D-модели головы осуществляется посредством прикрепления лицевой маски по известным пограничным вершинам 3D-модели головы и соответствующим для этих вершин точкам лицевой маски.

6. Способ по п.1, характеризующийся тем, что дополнительно выполняют этап, на котором размещают на изображениях 3D-модели головы текстуры зубов и языка согласно положению лицевой маски.

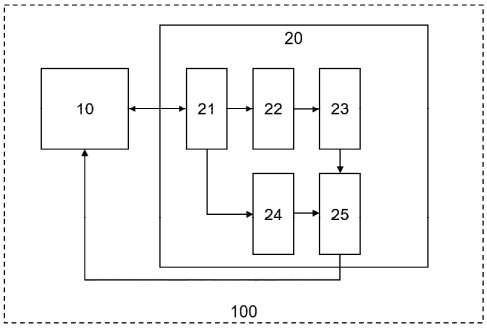
7. Способ по п.1, характеризующийся тем, что дополнительно выполняют этап морфинга изображений результирующей 3D-модели головы.

8. Способ по п.1, характеризующийся тем, что дополнительно выполняют этап цветокоррекции и совмещения динамических текстур для устранения пульсирующего изменения цвета лицевой маски посредством слияния по цвету упомянутых изображений 3D-модели головы, содержащихся на предыдущих и последующих кадрах.

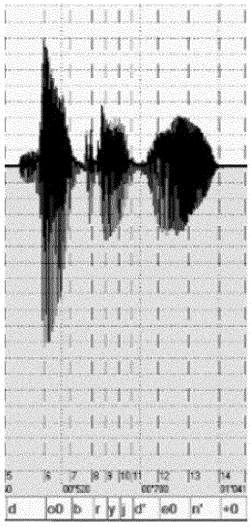
9. Устройство обработки речевого сигнала, содержащее по меньшей мере одно вычислительное устройство и по меньшей мере одно устройство памяти, содержащее машиночитаемые инструкции, которые при их исполнении по меньшей мере одним вычислительным устройством выполняют операции способа по любому из пп.1-8.



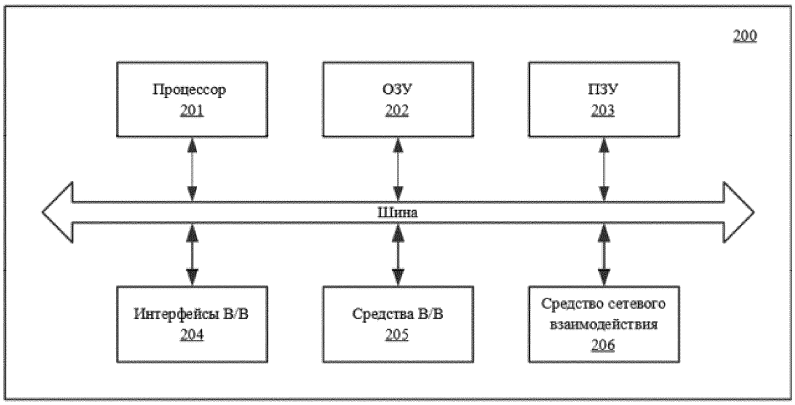
Фиг. 1



Фиг. 2



Фиг. 3



Фиг. 4

Евразийская патентная организация, ЕАПВ
Россия, 109012, Москва, Малый Черкасский пер., 2