(12) МЕЖДУНАРОДНАЯ ЗАЯВКА, ОПУБЛИКОВАННАЯ В СООТВЕТСТВИИ С ДОГОВОРОМ О ПАТЕНТНОЙ КООПЕРАЦИИ (PCT)

(19) Всемирная Организация Интеллектуальной Собственности
Международное бюро

(43) Дата международной публикации
20 мая 2021 (20.05.2021)

WIPO | PCT

(10) Номер международной публикации
WO 2021/096380 A1

(51) Международная патентная классификация:
G10L 15/16 (2006.01) *G10L 25/51* (2013.01)
G10L 15/20 (2006.01) *G10L 25/30* (2013.01)

(21) Номер международной заявки: PCT/RU2019/000818

(22) Дата международной подачи:
15 ноября 2019 (15.11.2019)

(25) Язык подачи: Русский

(26) Язык публикации: Русский

(71) Заявитель: **ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЦРТ-ИННОВАЦИИ" ("STC-INNOVATIONS LIMITED")** [RU/RU]; ул. Красуцкого, д. 4, лит. А Санкт-Петербург, 196084, Saint Petersburg (RU).

(72) Изобретатели: **МЕДЕННИКОВ, Иван Павлович (MEDENNIKOV, Ivan Pavlovich)**; ул. Красных Фортов, д. 15, кв. 10 Ленинградская область, г. Сосновый Бор, 188540, Leningradskaya oblast, g. Sosnovyj Bor (RU). **ПРИСЯЧ, Татьяна Николаевна (PRISYACH, Tatiana Nikolaevna)**; ул. Гаккелевская, д. 30, кв. 132 Санкт-Петербург, 197372, Saint-Petersburg (RU). **РОМАНЕНКО, Алексей Николаевич (ROMANENKO, Aleksei Nikolaevich)**; ул. Энгельса, д. 56, кв. 5 Тюменская область, г. Ханты-Мансийск, 628011, Tyumenskaya oblast', g. Hanty-Mansijsk (RU). **КОРЕНЕВСКАЯ, Мария Максимовна (KORENEVSKAYA, Mariya Maksimovna)**; ул. Полярников, д. 8, кв. 14 Санкт-Петербург, 192171, Saint-Petersburg (RU). **СОРОКИН, Иван Витальевич (SOROKIN, Ivan Vitalyevich)**; ул. Адмирала Черокова, д. 20В, кв. 1578 Санкт-Петербург, 198206, Saint-Petersburg (RU). **ХОХЛОВ, Юрий Юрьевич (HOHLOV, Yuriy Yurievich)**; ул. Адмирала Черокова, д. 22, кв. 481 Санкт-Петербург, 198206, Saint-Petersburg (RU).

(54) **Title:** METHOD FOR TRAINING A NEURAL NETWORK TO RECOGNIZE ACOUSTIC EVENTS IN AN ACOUSTIC SIGNAL

(54) **Название изобретения:** СПОСОБ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ РАСПОЗНАВАНИЮ ЗВУКОВЫХ СОБЫТИЙ В ЗВУКОВОМ СИГНАЛЕ

(57) **Abstract:** Proposed is a method for training a neural network to recognize acoustic events in an acoustic signal. A method for training an auxiliary neural network to identify a room impulse response includes obtaining a plurality of reverberating acoustic signals by the convolution of a plurality of non-reverberating acoustic signals with a plurality of room impulse responses; extracting training features that characterize each of the reverberating acoustic signals in the obtained plurality of reverberating acoustic signals; feeding to the input of a neural network said training features of a reverberating acoustic signal together with an identifier of the room impulse response corresponding to said reverberating acoustic signal, for each of the reverberating acoustic signals in the obtained plurality of reverberating acoustic signals. Furthermore, the method for training a neural network to recognize acoustic events in an acoustic signal includes calculating R vectors for a plurality of acoustic signals, which involves feeding an acoustic signal to the input of an auxiliary neural network and reading the R vector for the corresponding acoustic signal at the output of one of the hidden layers of the auxiliary neural network. The method also includes extracting training features that characterize each of the acoustic signals in the plurality of acoustic signals, and feeding to the input of a neural network the training features of an acoustic signal and label information with respect to the acoustic signal, as well as the corresponding R vector for the acoustic signal, for each of the acoustic signals in said plurality of acoustic signals. The technical result is more accurate speech recognition under reverberant conditions.

(57) **Реферат:** Предложен способ обучения нейронной сети распознаванию звуковых событий в звуковом сигнале. Способ обучения вспомогательной нейронной сети определять импульсную характеристику помещения включает получение множества реверберированных звуковых сигналов путем применения ко множеству нереверберированных звуковых сигналов операции свертки с множеством импульсных характеристик помещения; выделение обучающих признаков, характеризующих каждый из полученного множества реверберированных звуковых сигналов; подачу на вход нейронной сети указанных обучающих признаков реверберированного звукового сигнала вместе с идентификатором импульсной характеристики помещения, соответствующей указанному реверберированному звуковому сигналу, для каждого из полученного множества реверберированных звуковых сигналов. При этом способ обучения нейронной сети распознаванию звуковых событий в звуковом сигнале включает вычисление R-векторов для множества звуковых сигналов, которое включает подачу звукового сигнала на вход вспомогательной нейронной сети и считывание R-вектора для соответствующего звукового сигнала на выходе одного из скрытых слоев вспомогательной нейронной сети. Способ также включает выделение обучающих признаков, характеризующих каждый из указанного множества звуковых сигналов; и подачу на вход нейронной сети указанных обучающих признаков звукового сигнала, информации о разметке в отношении звукового сигнала, а также соответствующего R-вектора для звукового сигнала для каждого из указанного множества звуковых сигналов. Техническим результатом является повышение точности распознавания речи в условиях реверберации.

WO 2021/096380 A1

[продолжение на следующей странице]

(74) **Агент: НИЛОВА, Мария Иннокентьевна (NILOVA, Maria Innokentievna)**; BOX 1125 ПАТЕНТИКА Санкт-Петербург, 190000, Saint Petersburg (RU).

(81) **Указанные государства** *(если не указано иначе, для каждого вида национальной охраны)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Указанные государства** *(если не указано иначе, для каждого вида региональной охраны)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), евразийский (AM, AZ, BY, KG, KZ, RU, TJ, TM), европейский патент (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Опубликована:

— *с отчётом о международном поиске (статья 21.3)*

СПОСОБ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ РАСПОЗНАВАНИЮ ЗВУКОВЫХ СОБЫТИЙ В ЗВУКОВОМ СИГНАЛЕ

Область техники

Настоящее изобретение относится к области распознавания речи, детектирования акустических событий и разделения дикторов, в частности к способам обучения нейронных сетей распознавать звуковое событие в звуковом сигнале для повышения точности распознавания речи, детектирования акустических событий и разделения дикторов в записях, выполненных в помещениях с помощью среднего и/или дальнего микрофона.

Уровень техники

Из уровня техники известны решения, позволяющие повысить точность распознавания речи за счёт определения i-векторов, характеризующих конкретных дикторов. Из статей «Robust i-vector based Adaptation of DNN Acoustic Model for Speech Recognition» под авторством Sri Garimella и др. и «Speaker adaptation of neural network acoustic models using i-vectors», IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013 под авторством G. Saon, H. Soltau и др. раскрыт способ обучения нейронной сети путём подачи на неё, помимо звукового сигнала или выделенных из него признаков, характеризующих звуковой сигнал, вычисленных i-векторов, характеризующих дикторов, что приводит к улучшению распознавания речи.

Кроме того, в других способах и системах при распознавании реверберированных звуковых сигналов используют различные способы деревербирации реверберированного звукового сигнала. Например, в US10170134B2, US10062392B2 и US201603667 реверберированный звуковой сигнал предварительно деревербируют.

Однако зачастую, вследствие того, что при распознавании речи не учитываются характеристики помещения, в котором был записан звуковой сигнал, точность распознавания речи может быть ниже требуемой.

Раскрытие сущности изобретения

Для повышения точности распознавания речи в настоящем изобретении предложен способ обучения нейронный сетей, позволяющий учитывать характеристики помещения, в котором был записан звуковой сигнал, и предназначенный для решения проблемы снижения точности распознавания речи, записываемой в условиях реверберации.

Согласно первому аспекту настоящего изобретения, предложен способ обучения нейронной сети определять импульсную характеристику помещения, включающий получение множества реверберированных звуковых сигналов путём применения ко множеству нереверберированных звуковых сигналов операции свёртки с множеством импульсных характеристик помещения; выделение обучающих признаков, характеризующих каждый из полученного множества реверберированных звуковых сигналов; подачу на вход нейронной сети указанных обучающих признаков реверберированного звукового сигнала вместе с идентификатором импульсной характеристики помещения, соответствующей реверберированному звуковому сигналу, для каждого из полученного множества реверберированных звуковых сигналов.

Таким образом, нейронная сеть, обученная в соответствии с первым аспектом настоящего изобретения, позволяет определять импульсную характеристику помещения, близкую к той, которая соответствует реверберированному звуковому сигналу. При этом информация, полученная на выходе одного из скрытых слоёв нейронной сети, обученной в соответствии с первым аспектом, представляет собой высокоуровневое представление импульсной характеристики помещения, или R-вектор. R-вектор, таким образом, несёт в себе информацию о характеристиках помещения (например, о геометрии и коэффициентах поглощения звука поверхностями), а также информацию о положении источника речевого сигнала и микрофона.

В одном варианте реализации указанная в первом аспекте подача на вход нейронной сети также может включать одновременную с ней подачу на вход нейронной сети обучающих признаков, характеризующих ещё один реверберированный звуковой сигнал, вместе с идентификатором импульсной характеристики помещения, соответствующей этому ещё одному реверберированному звуковому сигналу. Кроме того, в данном варианте реализации способ также включает этап проверки того, совпадают ли импульсные характеристики указанных реверберированного звукового сигнала и ещё одного реверберированного звукового сигнала, и этап сообщения нейронной сети результата указанной проверки. Данный вариант реализации позволяет обучить

нейронную сеть также определять то, были ли записаны различные звуковые сигналы в одном помещении и при тех же положениях источника звука и микрофона.

Ещё в одном варианте реализации указанная подача на вход нейронной сети также может включать одновременную подачу на вход нейронной сети информации о дикторе, соответствующем указанным реверберированному звуковому сигналу и ещё одному реверберированному звуковому сигналу. В данном варианте реализации этап проверки, совпадают ли импульсные характеристики двух звуковых сигналов, или нет, также включает проверку того, совпадают ли идентификаторы дикторов для указанных реверберированного звукового сигнала и ещё одного реверберированного звукового сигнала. Данный вариант реализации позволяет повысить точность определения нейронной сетью того, что звуковые сигналы соответствуют одному и тому же диктору.

Согласно второму аспекту настоящего изобретения, предложен способ обучения нейронной сети распознавать звуковое событие, например, речь, из звукового сигнала, включающий вычисление R-векторов для множества звуковых сигналов, причём вычисление R-вектора для каждого звукового сигнала включает подачу звукового сигнала на вход вспомогательной нейронной сети, обученной при помощи способа в соответствии с первым аспектом настоящего изобретения, и считывание R-вектора для соответствующего звукового сигнала на выходе одного из скрытых слоёв вспомогательной нейронной сети; а указанный способ также включает выделение обучающих признаков, характеризующих каждый из указанного множества звуковых сигналов; и подачу на вход нейронной сети указанных обучающих признаков звукового сигнала, фонетической информации о звуковом сигнале, а также соответствующего R-вектора для звукового сигнала для каждого из указанного множества звуковых сигналов.

Обучение нейронной сети с использованием R-векторов, характеризующих помещение и условия, в которых была произведена запись звукового сигнала, позволяет существенно повысить точность распознавания речи обученной нейронной сетью.

Настоящее изобретение может быть в частности использовано для повышения точности распознавания речи, детектирования акустических событий и разделения дикторов, в системах автоматического распознавания речи, системах детектирования акустических событий, системах, использующих алгоритмы разделения дикторов (распознавание речи, идентификация и верификация дикторов), работающих с записями акустического сигнала, выполненными в условиях средней и сильной реверберации.

Осуществление изобретения

Согласно одному из вариантов осуществления настоящего изобретения, в способе обучения нейронных сетей, или акустических моделей, предназначенных для распознавания речевого сигнала, искаженного реверберацией, используют специальным образом подготовленные обучающие данные. Эти обучающие данные могут представлять собой речевой сигнал, записанный с помощью ближнего микрофона и искаженный путём применения к нему математической операции свёртки с импульсной характеристикой помещения (room impulse response, RIR), что позволяет смоделировать соответствующий речевой сигнал, записанный в условиях реверберации. При этом могут использоваться как реальные RIR (полученные для реальных помещений), так и RIR, смоделированные математически, например с использованием метода изображений, известного из уровня техники. Импульсная характеристика в общем представляет собой выходной сигнал динамической системы как реакцию на входной сигнал, который представляет собой простой импульс минимальной ширины и максимальной амплитуды. В терминах настоящего изобретения система представляет собой помещение, а простой импульс представляет собой идеальный звуковой дельта-импульс. Иными словами, импульсная характеристика помещения представляет собой переходную функцию между источником звука и микрофоном.

Согласно одному из вариантов настоящего изобретения, проводят обучение акустической модели на реверберированных данных (смоделированных и/или натуральных) с использованием вспомогательных признаков, являющихся высокоуровневым представлением импульсной характеристики помещения, или R-векторами. Высокоуровневые представления импульсной характеристики помещения косвенно содержат информацию (зависящую от способа обучения нейронной сети, используемой для их извлечения) о характеристиках помещения (например, о геометрии и коэффициентах поглощения звука поверхностями), а также информацию о положении источника речевого сигнала и микрофона. Иными словами, R-векторы несут в себе информацию о свойствах помещения и об условиях записи, которые вносят искажения. Авторами настоящего изобретения было обнаружено, что использование R-векторов в обучении акустической модели, совместно с обучающими признаками, характеризующими звуковой сигнал, позволяет существенно улучшить точность распознавания речи или детектирования акустических событий. Кроме того, расстояния между R-векторами, вычисляемыми для различных фрагментов звука, могут использоваться для повышения точности разделения дикторов. В различных вариантах осуществления расстояния между

R-векторами могут быть вычислены, например, как евклидова метрика или расстояние Махалонобиса.

Согласно различным вариантам реализации предлагаемого способа, процесс обучения нейронных сетей, или акустических моделей, можно разделить на два основных этапа. Сначала проводят обучение вспомогательной нейронной сети для выделения вспомогательных признаков, являющихся высокоуровневым представлением импульсной характеристики помещения, т.е. R-векторов.

Согласно одному варианту реализации, для обучения вспомогательной нейронной сети используют базу данных фонограмм, или звуковых сигналов, записанных с помощью ближнего микрофона, т.е. нереверберированных фонограмм. К указанным фонограммам применяют операцию свёртки с множеством импульсных характеристик помещения, что позволяет смоделировать звуковые сигналы, записанные с применением среднего и дальнего микрофонов в помещениях с различными акустическими характеристиками, т.е. реверберированные звуковые сигналы. Однако в других вариантах реализации, вместо смоделированных реверберированных звуковых сигналов могут быть использованы реверберированные звуковые сигналы, полученный непосредственно путём записи в различных помещениях.

Размер выборки импульсных характеристик помещения может достигать десятков или сотен тысяч при использовании топологии вспомогательной нейронной сети с выходным слоем Softmax (т.е. слоем нейронной сети, в котором в качестве нелинейности используется функция Softmax). При использовании топологии AM без выходного Softmax слоя количество импульсных характеристик помещения может быть увеличено на несколько порядков.

После получения реверберированных звуковых сигналов из них выделяют обучающие признаки, характеризующие звуковой сигнал. В общем, выделение признаков — это разновидность абстрагирования, процесс снижения размерности, в котором исходный набор исходных переменных сокращается до более управляемых групп (признаков) для дальнейшей обработки, оставаясь при этом набором, достаточным для точного и полного описания исходного набора данных. Выделение признаков используется в машинном обучении. Выделение признаков начинают с исходного набора данных, в данном случае – полученного реверберированного звукового сигнала, и далее выводят вторичные значения (признаки), в отношении которых предполагается, что они должны быть достаточно информативными и не быть избыточными, что способствует

последующему процессу обучения нейронной сети и достижению лучшей обобщающей способности нейронной сети, а в некоторых случаях ведёт и к лучшей человеческой интерпретации данных. В настоящем изобретении обучающими признаками, характеризующими звуковой сигнал, могут быть любые признаки звукового сигнала, подходящие для обучения нейронной сети, как известно из уровня техники. Такие обучающие признаки могут включать в себя, например, MFCC, FBANK, PLP, фрагменты звука в исходном виде (waveform) и т.д. MFCC (Mel-frequency cepstral coefficient) представляет собой кратковременный энергетический спектр звукового сигнала. FBANK (filter bank) представляет собой массив полосовых фильтров, который разделяет входной звуковой сигнал на множество компонентов, каждый из которых содержит отдельный частотный поддиапазон первоначального звукового сигнала.

Затем к обучающим признакам, характеризующим звуковой сигнал, могут применять CMN (cepstral mean normalization), или нормализацию кепстров к среднему, которая представляет собой эффективный способ нормализации для надёжного распознавания речи. Применение CMN позволяет уменьшить искажения, обусловленные разными уровнями записи звука и уменьшить, или даже исключить, «канальные» искажения спектров, обусловленные, например, разными кодеками в телефонном канале, различными марками микрофонов, усилителей и т.п., для надёжного выделения обучающих признаков путём линейного преобразования кепстральных коэффициентов.

Затем обучающие признаки, характеризующие реверберированный звуковой сигнал, вместе с идентификатором импульсной характеристики помещения, соответствующей указанному реверберированному звуковому сигналу, для каждого из полученного множества реверберированных звуковых сигналов, подают на вход обучаемой вспомогательной нейронной сети. Таким образом, на вход вспомогательной нейронной сети подаётся множество комбинаций, включающих обучающие признаки, характеризующие смоделированный реверберированный звуковой сигнал, и идентификатор импульсной характеристики помещения, соответствующей реверберированному звуковому сигналу, так что вспомогательная нейронная сеть обучается определять импульсную характеристику помещения, в котором был записан или смоделирован звуковой сигнал.

Высокоуровневыми представлениями импульсных характеристик для подаваемых на вход обученной приведённым выше образом вспомогательной нейронной сети реверберированных звуковых сигналов являются выходы (или активации) нейронной сети, получаемые с одного из промежуточных слоёв нейронной сети. Это может быть любой

скрытый слой нейронной сети, но, как правило, этот слой имеет меньшую размерность, чем остальные, для того, чтобы R-векторы получались небольшой размерности (например, 40-512 нейронов). Это позволяет упаковать информацию, специфичную для идентификации импульсной характеристики помещения, в R-вектор небольшой размерности. Согласно одному варианту реализации, при создании нейронной сети обеспечивают наличие скрытого слоя небольшой размерности, предназначенного для извлечения R-векторов на его выходе.

Таким образом, на первом основном этапе обеспечено получение вспомогательной нейронной сети, обученной определять импульсную характеристику помещения, в котором был записан реверберированный звуковой сигнал. Кроме того, на основании данной нейронной сети можно создать экстрактор, т.е. нейронную сеть без одного или нескольких верхних (последних) слоёв, позволяющий получать выходы скрытого слоя, т.е. обеспечена возможность извлечения R-векторов на выходе одного из скрытых слоёв вспомогательной нейронной сети, как правило, но без ограничения, имеющего наименьшую размерность.

В одном из вариантов реализации при обучении вспомогательной нейронной сети на её вход вместе с подачей обучающих признаков, характеризующих реверберированный звуковой сигнал, и вместе с идентификатором импульсной характеристики помещения, соответствующей указанному реверберированному звуковому сигналу, могут также одновременно подавать обучающие признаки, характеризующие ещё один реверберированный звуковой сигнал, вместе с идентификатором импульсной характеристики помещения, соответствующей этому ещё одному реверберированному звуковому сигналу. При этом способ также включает этап проверки того, совпадают ли импульсные характеристики указанных реверберированного звукового сигнала и ещё одного реверберированного звукового сигнала, и этап сообщения нейронной сети результата указанной проверки. Таким образом, вспомогательную нейронную сеть обучают определять вероятность того, что два звуковых сигнала, информация о которых подана на вход вспомогательной нейронной сети, были записаны в одном помещении. В других вариантах реализации способ обучения нейронной сети может не включать подачу идентификаторов импульсных характеристик помещения на вход нейронной сети. В данном варианте реализации на вход вспомогательной нейронной сети подают обучающие признаки, характеризующие два реверберированных звуковых сигнала, а также информацию о том, были ли они записаны в одинаковых или разных условиях, т.е. соответствует ли им одна и та же импульсная характеристика помещения или разные

характеристики помещения. Например, на вход нейронной сети вместе с обучающими признаками, характеризующими звуковые сигналы, может подаваться также сигнал, соответствующий единице, в случае, когда оба сигнала соответствуют одной импульсной характеристики помещения, или ноль, в случае, когда сигналы соответствуют разным импульсным характеристикам помещения. В других вариантах реализации на вход нейронной сети могут подавать обучающие признаки, характеризующие более двух звуковых сигналов одновременно.

Кроме того, в одном варианте реализации подача на вход вспомогательной нейронной сети информации о двух звуковых сигналах, при её обучении, также включает подачу на её вход информации о дикторах, соответствующих указанным реверберированному звуковому сигналу и ещё одному реверберированному звуковому сигналу. В данном варианте реализации этап проверки того, совпадают ли импульсные характеристики указанных реверберированного звукового сигнала и ещё одного реверберированного звукового сигнала, также включает проверку того, совпадают ли идентификаторы дикторов для указанных реверберированного звукового сигнала и ещё одного реверберированного звукового сигнала. Таким образом нейронную сеть могут обучать определять не только вероятность того, что звуковые сигналы были записаны в одном помещении, но и вероятность того, что они относятся к одному диктору. Согласно одному из вариантов реализации, на вход вспомогательной нейронной сети могут подавать только обучающие признаки, характеризующие два звуковых сигнала, а также сигнал, соответствующий единице, если оба звуковых сигнала соответствуют одной импульсной характеристике помещения, а также одному диктору, и сигнал, соответствующий нулю, в противном случае.

Также нейронная сеть, обученная в соответствии с предшествующим вариантом реализации, может позволять находить точки (моменты времени) смены дикторов, так как разным дикторам соответствуют разные импульсные характеристики помещений (при условии, что дикторы находятся в разных точках пространства). Иными словами, подавая соседние звуковые сигналы на вход нейронной сети, можно найти моменты времени, когда один диктор сменяет другого. В этих случаях первый и второй звуковые сигналы будут принадлежать разным дикторам и поэтому будут соответствовать разным импульсным характеристикам помещений. Точки (моменты времени) смены диктора разобьют фонограмму на сегменты, принадлежащие одному диктору. После этого для каждого сегмента, принадлежащего одному диктору, можно вычислить R-вектор. Затем могут

выполнить кластеризацию речевых фрагментов с использованием расстояния между R-векторами в качестве метрики.

В качестве неограничивающего примера ниже приведена архитектура вспомогательной нейронной сети, которая может быть использована в настоящем изобретении. Однако специалисту в данной области техники очевидно, что могут быть использованы и другие подходящие топологии.

Архитектура вспомогательной нейронной сети, на основании которой можно создать экстрактор, является следующей. Первые три слоя содержат по 512 нейронов с активациями ReLU (rectified linear unit, линейный выпрямляющий блок) и групповой нормализацией (Batch normalization). Данные слои могут иметь архитектуру с временной задержкой: первый слой склеивает вместе входящие временные фрэймы ft; t ±3; t±2; t±1g (где t представляет собой текущий временной фрэйм); второй слой склеивает активации предыдущего слоя для значений времени ft; t ± 2g, а третий слой - для ft; t ± 3g. Слои 4 и 5 имеют 512 и 1500 нейронов соответственно с активацией ReLU и групповой нормализацией без какой-либо склейки. Следующий слой представляет собой слой статистического объединения. Данный слой собирает первые 10 тысяч фрэймов входного сегмента и вычисляет их среднее и нормальное отклонение в виде единого выхода для всего сегмента. Далее выполнены слои 7 и 8 размерностью 512 нейронов и, наконец, выходной слой Softmax. Нейронную сеть обучают классифицировать N RIR в тренировочных данных. Выходные данные седьмого слоя используют для выделения R-векторов.

На втором основном этапе выполняют обучение основной нейронной сети для распознавания звуковых событий из звукового сигнала, где звуковые события могут представлять собой, например, речь. Для этого берут множество звуковых сигналов, или произнесений, обучающей выборки, которая может содержать реверберированные (смоделированные или натуральные) и нереверберированные звуковые сигналы, и для каждого из них вычисляют R-вектор. Для этого каждый из указанных звуковых сигналов подают на вход вспомогательной нейронной сети, описанной выше, и считывают соответствующие им R-векторы на выходе одного из скрытых слоёв вспомогательной нейронной сети, как указано выше.

Затем для каждого из указанных звуковых сигналов выделяют обучающие признаки, характеризующие звуковой сигнал. Указанные обучающие признаки могут быть как такими же, как были выделены при обучении вспомогательной нейронной сети, так и

отличающимися от них обучающими признаками. Однако выделение обучающих признаков происходит в целом, как описано выше в отношении вспомогательной нейронной сети.

Затем выделенные обучающие признаки, характеризующие звуковой сигнал, вместе с соответствующими R-векторами, а также информацию о разметке звукового сигнала, для каждого из указанного множества звуковых сигналов, подают на вход основной нейронной сети для того, чтобы обучить её распознавать звуковые события. Согласно одному из вариантов реализации, R-векторы могут быть конкатенированы к обучающим признакам характеризующим звуковой сигнал. В других вариантах реализации R-векторы могут подаваться на отдельный вход нейронной сети или конкатенироваться к выходам одного или нескольких скрытых слоёв. В случае использования отдельного входа, перед одним из скрытых слоёв можно выполнять конкатенацию или суммирование (если сделать размерности одинаковыми) выходов предыдущих слоёв, соответствующих разным входам сети — основным обучающим признакам и R-векторам. Например, если первый слой нейронной сети является свёрточным, то R-векторы можно добавлять как отдельный канал. В других вариантах реализации R-векторы могут быть присоединены к указанным обучающим признакам другим подходящим образом. При этом информация о разметке включает в себя данные о том, в какой момент времени происходило то или иное звуковое событие.

Согласно одному неограничивающему примеру, информацию о разметке можно описать следующим образом. Приведённый в качестве примера звуковой сигнал имеет длительность D, в различных примерах длительность звукового сигнала может значительно отличаться, и она не является ограниченной каким-либо предельным значением. В данном примере звуковой сигнал разбивается на сегменты с шагом 0,01 секунды. В данном примере первые 50 сегментов, например, могут соответствовать тишине, или паузе, затем 20 сегментов могут соответствовать фонеме A, затем 15 сегментов могут соответствовать фонеме B, следующие 25 сегментов могут соответствовать фонеме C, затем 50 сегментов могут соответствовать звонку телефона и так далее для всех сегментов данного звукового сигнала. В других вариантах реализации звуковой сигнал может быть разбит на временные фрэймы следующим образом. Сначала задают размер окна и смещение. Как правило, смещение меньше размера окна, поэтому окна идут с перекрытием. Количество фреймов для сигнала длительностью D вычисляется как $N=(D-L)/S+1$, где L - размер окна, а S — смещение. D, L и S задаются в отсчётах звука, а деление целочисленное.

Затем информацию о разметке подают на вход нейронной сети для каждого звукового сигнала для того, чтобы при обучении нейронная сеть имела всю информацию о звуковом файле. Таким образом, нейронная сеть обучается на основании подаваемых на её вход обучающих признаков, характеризующих звуковые сигналы, информации о разметке и соответствующих R-векторов распознавать звуковые события в звуковом сигнале. При этом специалисту в данной области техники понятно, что количество сегментов или временных фрэймов, которое умещается в секунде звукового файла, может отличаться как в большую, так и в меньшую сторону, и не ограничено какими-либо предельными значениями. Кроме того, специалисту в данной области техники будет очевидно, что информация о разметке может быть представлена в любом другом виде, пригодном для описания звукового сигнала. Согласно одному из вариантов реализации, в случае, когда звуковой сигнал представляет собой речь, информация о разметке содержит фонетическую информацию.

Таким образом, на втором основном этапе обеспечено получение нейронной сети, обученной распознавать звуковые события в звуковом сигнале и адаптированной к помещению/положению источника звука/микрофона путём привнесения через R-векторы дополнительной информации об акустических свойствах помещения, в котором был записан подаваемый на вход нейронной сети звуковой сигнал, и о положении источника звука/микрофона в указанном помещении.

При этом следует отметить, что специалисту в данной области техники очевидно, что топология как основной, так и вспомогательной нейронной сети может быть различной. Однако, в других вариантах реализации топология основной нейронной сети может быть такой же, как представлена выше для вспомогательной нейронной сети. При этом целевые классы основной нейронной сети будут отличаться от целевых классов вспомогательной нейронной сети.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ обучения нейронной сети определять импульсную характеристику помещения, включающий

получение множества реверберированных звуковых сигналов путём применения ко множеству нереверберированных звуковых сигналов операции свёртки с множеством импульсных характеристик помещения;

выделение обучающих признаков, характеризующих каждый из полученного множества реверберированных звуковых сигналов;

подачу на вход нейронной сети указанных обучающих признаков реверберированного звукового сигнала вместе с идентификатором импульсной характеристики помещения, соответствующей указанному реверберированному звуковому сигналу, для каждого из полученного множества реверберированных звуковых сигналов.

2. Способ по п. 1, в котором указанная подача также включает одновременную подачу на вход нейронной сети обучающих признаков, характеризующих ещё один реверберированный звуковой сигнал, вместе с идентификатором импульсной характеристики помещения, соответствующей этому ещё одному реверберированному звуковому сигналу, а способ также включает этап проверки того, совпадают ли импульсные характеристики указанных реверберированного звукового сигнала и ещё одного реверберированного звукового сигнала, и этап сообщения нейронной сети результата указанной проверки.

3. Способ по п. 2, в котором указанная подача также включает одновременную подачу на вход нейронной сети информации о дикторе, соответствующем указанным реверберированному звуковому сигналу и ещё одному реверберированному звуковому сигналу, а указанный этап проверки также включает проверку того, совпадают ли идентификаторы дикторов для указанных реверберированного звукового сигнала и ещё одного реверберированного звукового сигнала.

4. Способ обучения нейронной сети распознавать звуковое событие в звуковом сигнале, включающий

вычисление R-векторов для множества звуковых сигналов, причём вычисление R-вектора для каждого звукового сигнала включает

подачу звукового сигнала на вход вспомогательной нейронной сети, обученной при помощи способа по любому из пп. 1-3, и

считывание R-вектора для соответствующего звукового сигнала на выходе одного из скрытых слоёв вспомогательной нейронной сети; а указанный способ также включает

выделение обучающих признаков, характеризующих каждый из указанного множества звуковых сигналов; и

подачу на вход нейронной сети указанных обучающих признаков звукового сигнала, информации о разметке в отношении звукового сигнала, а также соответствующего R-вектора для звукового сигнала для каждого из указанного множества звуковых сигналов.

5. Способ по п. 4, в котором звуковое событие представляет собой речь.

ОТЧЁТ О МЕЖДУНАРОДНОМ ПОИСКЕ

Международная заявка № PCT/RU 2019/000818

A. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ:
G10L 15/16 G10L 15/20 G10L 25/51 G10L 25/30

Согласно международной патентной классификации (МПК-8)

B. ОБЛАСТИ ПОИСКА:

Проверенный минимум документации (система классификации и индексы) МПК-8:
G10L

Другая проверенная документация в той мере, в какой она включена в поисковые подборки:

Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, поисковые термины):
EPO-Internal

C. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ:

Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
X	YURI KHOKHLOV ET AL: "R-Vectors: New Technique for Adaptation to Room Acoustics" , INTERSPEECH 2019, 15 сентября 2019 (2019-09-15) , стр. 1243-1247, XP055718478, ISCA DOI : 10.21437/ Inter speech.2019 -2645	1, 4, 5
A	реферат раздел 1, 2 под-раздел 3.1, 3.2	2, 3

[X] последующие документы указаны в продолжении графы C.

[ ] данные о патентах-аналогах указаны в приложении.

* Особые категории ссылочных документов:

A документ, определяющий общий уровень техники

E более ранний документ, но опубликованный на дату международной подачи или после нее

O документ, относящийся к устному раскрытию, экспониро-ванию и т.д.

P документ, опубликованный до даты международной подачи, но после даты испрашиваемого приоритета и т.д.

"P" документ, опубликованный до даты международной подачи, но после даты испрашиваемого приоритета.

T более поздний документ, опубликованный после даты приоритета и приведенный для понимания изобретения

X документ, имеющий наиболее близкое отношение к предмету поиска, порочащий новизну и изобретательский уровень

Y документ, порочащий изобретательский уровень в сочетании с одним или несколькими документами той же категории

& документ, являющийся патентом-аналогом

"&" документ, являющийся патентом-аналогом

Дата действительного завершения международного поиска:	Дата отправки настоящего отчёта о международном поиске:
29 июля 2020 (29.07.2020)	17 августа 2020 (17.08.2020)
Наименование и адрес Международного поискового органа: EP	Уполномоченное лицо: Телефон №

ОТЧЁТ О МЕЖДУНАРОДНОМ ПОИСКЕ

Международная заявка № PCT/RU 2019/000818

C. (Продолжение), ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ

Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
X	IVAN MEDENNIKOV ET AL: "The STC ASR System for the VOICES from a Distance Challenge 2019", INTERSPEECH 2019, 15 сентября 2019 (2019-09-15), стр. 2453-2457, XP055718489, ISCA D0I : 10.21437/Interspeech.2019-1574	4,5
A	фигура 1 раздел 3 под-раздел 4.1, 4.2	1-3
A	SNYDER DAVID ET AL: "X-Vectors: Robust DNN Embeddings for Speaker Recognition", 2018 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP) , IEEE, 15 апреля 2018 (2018-04-15), стр. 5329-5333, XP033403941, D0I : 10.1109/ICASSP.2018.8461375 [найдено 2018-09-10] под-раздел 2.3	1-5

INTERNATIONAL SEARCH REPORT

International application No
PCT/RU2019/000818

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L15/16 G10L15/20 G10L25/51 G10L25/30
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	YURI KHOKHLOV ET AL: "R-Vectors: New Technique for Adaptation to Room Acoustics", INTERSPEECH 2019, 15 September 2019 (2019-09-15), pages 1243-1247, XP055718478, ISCA DOI: 10.21437/Interspeech.2019-2645	1,4,5
A	abstract sections 1, 2 sub-sections 3.1, 3.2	2,3

-/--

[X] Further documents are listed in the continuation of Box C.

[ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 July 2020	17/08/2020

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Tilp, Jan

INTERNATIONAL SEARCH REPORT

International application No
PCT/RU2019/000818

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	IVAN MEDENNIKOV ET AL: "The STC ASR System for the VOiCES from a Distance Challenge 2019", INTERSPEECH 2019, 15 September 2019 (2019-09-15), pages 2453-2457, XP055718489, ISCA DOI: 10.21437/Interspeech.2019-1574	4,5
A	figure 1 section 3 sub-sections 4.1, 4.2 -----	1-3
A	SNYDER DAVID ET AL: "X-Vectors: Robust DNN Embeddings for Speaker Recognition", 2018 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), IEEE, 15 April 2018 (2018-04-15), pages 5329-5333, XP033403941, DOI: 10.1109/ICASSP.2018.8461375 [retrieved on 2018-09-10] sub-section 2.3 -----	1-5