(19) Евразийское патентное ведомство

(11) 042797 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.03.27

(21) Номер заявки
202293468

(22) Дата подачи заявки
2020.11.18

(51) Int. Cl. ***H04R 3/00*** (2006.01)
H04R 1/40 (2006.01)

(54) СПОСОБ ОПРЕДЕЛЕНИЯ МЕСТОПОЛОЖЕНИЯ ДИКТОРА С ИСПОЛЬЗОВАНИЕМ КОНФЕРЕНЦ-СИСТЕМЫ

(43) **2023.01.20**
(86) **PCT/RU2020/000615**
(87) **WO 2022/108470 2022.05.27**

(71)(73) Заявитель и патентовладелец:
ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЦРТ-ИННОВАЦИИ" (RU)

(72) Изобретатель:
Лаврентьев Александр Валерьевич, Попов Дмитрий Владимирович, Краев Станислав Сергеевич, Астапов Сергей Сергеевич (RU)

(74) Представитель:
Нилова М.И. (RU)

(56) IVAN HIMAWAN ET AL.: "Microphone Array Beamforming Approach to Blind Speech Separation", 28 June 2007 (2007-06-28), MACHINE LEARNING FOR MULTIMODAL INTERACTION; [LECTURE NOTES IN COMPUTER SCIENCE], SPRINGER BERLIN HEIDELBERG, BERLIN, HEIDELBERG, PAGE(S) 295-305, XP019086897, ISBN: 978-3-540-78154-7, pages 1-7; figure 1

US-A1-2020068297

ZHAO XIAOYAN ET AL.: "Accelerated steered response power method for sound source localization via clustering search", SCIENCE CHINA PHYSICS, MECHANICS & ASTRONOMY, vol. 56, no. 7, 1 July 2013 (2013-07-01), pages 1329-1338, XP055827260, Heidelberg ISSN: 1674-7348, DOI: 10.1007/s11433-013-5108-3, Retrieved from the Internet: URL:https://link.springer.com/content/pdf/10.1007/s11433-013-5108-3.pdf>, the whole document

PRANDI G. ET AL.: "Acoustic source localization by fusing distributed microphone arrays measurements", 2010 18TH EUROPEAN SIGNAL PROCESSING CONFERENCE, IEEE, 25 August 2008 (2008-08-25), pages 1-5, XP032760946, ISSN: 2219-5491 [retrieved on 2015-04-03], the whole document

(57) Предложен способ определения местоположения дикторов с использованием конференц-системы, согласно которому задают входные параметры конференц-системы и входные параметры для построения сетки поиска и кластеризации, принимают многоканальный звуковой сигнал, содержащий речь диктора, разбивают микрофоны конференц-системы на подмножества микрофонов, каждое из которых определяет отдельную пространственную область поиска, получают квазирегулярную сетку поиска с использованием входных параметров для построения сетки поиска; назначают каждую точку квазирегулярной сетки поиска одному из подмножества микрофонов с формированием множества дискретных точек для каждого из подмножеств микрофонов. Способ также включает осуществление параллельной акустической локализации в многоканальном звуковом сигнале, содержащем речь диктора, для каждой отдельной пространственной области поиска с использованием алгоритма SRP-PHAT (Steered Response Power Phase Transform) с получением SRP-значений для точек каждого множества дискретных точек; осуществляют агломеративную иерархическую кластеризацию SRP-значений дискретных точек с использованием входных параметров кластеризации с выделением кластера пространственных точек, задающих область локализации диктора, и определяют местоположение диктора в выделенном кластере. Предлагаемый способ сокращает количество вычислений по сравнению с существующими способами и увеличивает скорость выполнения локализации при поддерживании высокой точности поиска местоположения диктора, присущую методам поиска по всей сетке.

042797 B1

042797 B1

Область техники

Настоящее изобретение относится к области определения пространственного положения источника звука, а более конкретно - к определению местоположения дикторов в помещении для переговоров с использованием конференц-системы, расположенной в этом помещении.

Уровень техники

Конференц-системы широко используются для проведения деловых встреч или производственных совещаний, в которых может принимать участие большое количество участников (дикторов). Как правило, конференц-системы используются в специальных помещениях для переговоров и содержат распределенные микрофоны, которые размещены на общем столе и соединены с использованием аудиомикшера. Дикторы часто не ограничены в выборе места и могут располагаться относительно микрофонов конференц-системы произвольным образом. Для осуществления автоматического протоколирования переговоров в этих случаях требуется точно идентифицировать диктора, что особенно затруднительно, когда одним микрофоном конференц-системы пользуются два или более дикторов. В таких случаях для определения принадлежности каждой сказанной фразы определенному участнику переговоров при протоколировании переговоров полезно определять пространственное положение диктора в помещении для переговоров. Пространственное положение активного диктора может быть эффективно определено с помощью методов определения местоположения источника звука.

В качестве надежного способа определения положения источника звука в шумной среде с реверберацией известен алгоритм SRP-PHAT (Steered-Response Power Phase Transform).

KR 20140015894 (дата публикации 07.02.2014) раскрывает способ определения положения источника звука (Method for estimating location of sound source). Известный способ поиска по сетке основан на разделении направления источника звука с использованием конструкции с микрофонами и вычислении значений SRP-PHAT с использованием совокупностей пар микрофонов в конструкции. Однако для достижения достаточной точности определения положения способ требует образования большой поисковой сетки (сетки точек) и обработки большого количества сигналов, полученных от конструкции с микрофонами. Это приводит к существенному повышению вычислительной сложности, требуемой для осуществления способа.

CN 109709517 (дата публикации 03.05.2019) раскрывает способ поиска источника звука по направляющей сетке на основе алгоритма имитации отжига с использованием алгоритма SRP-PHAT (SRP-PHAT sound source positioning grid searching method based on simulated annealing algorithm), согласно которому осуществляют поиск глобального максимума в регулярной сетке поиска SRP с помощью алгоритма имитации отжига (simulated annealing algorithm). Согласно этому способу, для сокращения вычислительных ресурсов, требуемых для подсчета SRP значения для каждой точки сетки поиска, сокращают количество анализируемых точек в сетке поиска. Однако это повышает вероятность исключения локальных максимумов из анализа, что может привести к снижению точности определения положения источника звука.

WO 2017093554 (дата публикации 08.06.2017) раскрывает конференц-систему с микрофонным массивом и способ захвата речи в конференц-системе. Согласно этому способу осуществляют "грубое" определение положения диктора по разряженной сетке поиска, где используется сравнительно небольшое количество точек поиска, а затем осуществляют поиск по "локальной" более плотной (насыщенной точками поиска) сетке для точного определения положения диктора. Известный способ снижает влияние шумов при определении положения диктора путем установления пороговых значений SRP, а также сокращает количество вычислений при реализации алгоритма SRP-PHAT за счет исключения некоторых областей из сетки поиска, однако такое исключение снижает точность определения положения диктора при захвате речи конференц-системой.

Раскрытие сущности изобретения

Задачей настоящего изобретения является создание способа определения местоположения активных дикторов, для выполнения которого требуется меньшее количество вычислений при сохранении высокой точности определения местоположения диктора, присущей методам поиска по всей сетке поиска.

Еще одной задачей настоящего изобретения является сокращение времени определения местоположения диктора и обеспечение возможности определения местоположения в большем пространстве и в режиме реального времени.

Еще одной задачей настоящего изобретения является снижение потребления энергии при определении местоположения диктора с использованием конференц-системы.

Указанные задачи решены способом определения местоположения дикторов с использованием конференц-системы, как указано в данной заявке. Способ включает задание входных параметров конференц-системы и входных параметров для построения сетки поиска и кластеризации; прием многоканального звукового сигнала, содержащего речь дикторов, разбивку множества микрофонов конференц-системы на подмножества микрофонов, каждое из которых определяет отдельную пространственную область поиска дикторов и получение квазирегулярной сетки поиска с использованием входных параметров для построения сетки поиска. Способ также включает присвоение каждой точке полученной ква-

зирегулярной сетки поиска одному из подмножества микрофонов с формированием множества дискретных точек для каждого из подмножеств микрофонов; осуществление параллельного акустического определения положения в многоканальном звуковом сигнале для каждой отдельной пространственной области поиска с использованием алгоритма SRP-PHAT (Steered Response Power Phase Transform) с получением метрики когерентности звуковых волн (SRP-значения) для каждой точки каждого множества дискретных точек; и осуществление агломеративной иерархической кластеризации SRP-значений дискретных точек с использованием входных параметров кластеризации с получением кластера пространственных точек, задающих местоположение диктора.

Заявленный способ сокращает количество вычислений по сравнению с существующими методами, увеличивает скорость выполнения поиска местоположения диктора и обеспечивает высокую точность поиска, присущую поиску по всей сетке поиска.

Согласно предпочтительному варианту реализации способа входными параметрами конференц-системы являются координаты каждого микрофона в системе координат помещения, входными параметрами для построения сетки поиска являются номера пар микрофонов, дискретный шаг и шаг расширения конуса, параметр отсеивания точек сетки, а входными параметрами кластеризации являются число точек L и порог расстояния.

Также, согласно предпочтительному варианту реализации, квазирегулярную сетку поиска получают путем определения множества дискретных точек в общей пространственной области поиска и удаления таких точек, которые находятся от других точек множества на расстоянии, меньшем, чем установленное (минимальное) расстояние.

Согласно одному из вариантов реализации изобретения, многоканальный звуковой сигнал поступает непосредственно с микрофонов конференц-системы, либо считывается из файла с записью звукового сигнала, сделанную во время переговоров.

Согласно одному из вариантов реализации, в полученном кластере пространственных точек, задающем местоположение диктора, определяют взвешенное значение центра кластера, если последующий потребитель данных о местоположении диктора не требует информации о дисперсии кластера и форме кластера.

Согласно одному из вариантов реализации, звуковой сигнал, содержащий речь диктора, дополнительно обрабатывают с определением временных промежутков, соответствующих речевым сегментам звукового сигнала, а параллельную акустическую локализацию и агломеративную иерархическую кластеризацию осуществляют только для определенных временных промежутков. Такую обработку могут осуществлять с применением детектора активности речи (VAD, voice activity detector). Определение временных интервалов, соответствующих речевым сегментам звукового сигнала, и выполнение кластеризации только для таких интервалов повышает точность определения местоположения диктора, когда большое количество сторонних шумов искажает речь дикторов, путем исключения из анализа звуковых акустических событий, не принадлежащих дикторам, а также дополнительно снижает вычислительную сложность, требуемую для осуществления способа.

Краткое описание чертежей

Настоящее изобретение далее будет описано подробно со ссылками на сопроводительные чертежи, на которых:

фиг. 1 показывает блок-схему способа согласно предпочтительному варианту реализации настоящего изобретения;

фиг. 2 показывает пример расположения микрофонов конференц-системы за столом для совещаний и пример расположения дикторов относительно микрофонов;

фиг. 3 иллюстрирует построение локальной сетки поиска (усеченный конус) для одной пары микрофонов конференц-системы;

фиг. 4 иллюстрирует результат построения локальных сеток поиска (усеченные конусы) для каждой пары микрофонов выбранных n-подмассивов конференц-системы, размещенных на прямоугольном столе для совещаний, показанном на фиг. 2, до удаления точек, которые находятся ближе к соседним точкам, чем установленный нижний предел расстояния;

фиг. 5 иллюстрирует этапы построения сетки поиска и определения кластера пространственных точек, задающих пространственную область определения положения диктора.

Осуществление изобретения

Заявленный способ реализуется с использованием компьютерного устройства, связанного с конференц-системой с возможностью обмена данными. Конференц-система содержит распределенные микрофоны (фиг. 2), соединенные с использованием аудиомикшера. Микрофоны предпочтительно имеют кардиоидную диаграмму направленности.

Вычислительное устройство содержит вычислительный модуль, пользовательский интерфейс (блок ввода-вывода), запоминающее устройство, оперативную память, а также, при необходимости, периферийные устройства и графическую карту.

Согласно одному из возможных вариантов реализации, показанному на фиг. 2, микрофоны конференц-системы расположены на прямоугольном столе для совещаний, который имеет размеры 8,8 на 2,8

м, оснащен 30 микрофонами (мелкие круги с нумерацией снаружи), объединенными в единый распределенный микрофонный массив. За столом находятся 27 дикторов (пунктирные круги с нумерацией внутри) обоих полов; мужчины и женщины обозначаются соответственно частым и редким пунктиром.

Микрофоны конференц-системы расположены на расстоянии приблизительно 0,4-1 м от края стола и направлены к ближайшему краю стола. Расстояние между микрофонами не менее 0,3-0,4 м и варьируется в пределах 0,3-1,0 м. Общая длина микрофонного массива по длинной стороне продолговатого стола до 12-15 м. Подразумевается, что дикторы находятся в естественном сидячем положении за столом, однако, возможен вариант случайной рассадки за столом, т.е. не требуется, чтобы каждый диктор сидел непосредственно перед одним микрофоном. Количество дикторов может быть больше, меньше, или равно количеству микрофонов; расстояние между дикторами может быть меньше расстояния между микрофонами, стоящими в их окрестности. Каждый отдельный микрофон включает в себя один сенсор. Все микрофоны подключены к единому аудиомикшеру, обеспечивающему синхронную запись и одинаковое усиление сигнала. Многоканальная запись подается с аудиомикшера в вычислительное устройство, которое производит вычисления согласно заявленному способу определения местоположения дикторов.

Предпочтительный вариант реализации способа определения местоположения дикторов показан на фиг. 1, где слева от вертикальной пунктирной линии приведены этапы способа, которые осуществляются перед этапом акустической локализации и могут выполняться однократно при настройке конференц-системы к работе в определенном помещении. Данные по результатам выполнения этих этапов могут сохраняться в запоминающем устройстве вычислительного устройства и при необходимости повторного определения местоположения дикторов с использованием конференц-системы (при условии неизменности введенных начальных условий, таких как пространственные координаты микрофонов, номера пар микрофонов для построения локальных сеток поиска, номера микрофонов в n-подмассивах, параметры сетки поиска, параметры отсеивания точек поиска) могут быть повторно использованы для выполнения этапов способа по акустической локализации и агломеративной кластеризации (показаны справа от пунктирной линии на фиг. 1). Таким образом, могут быть исключены этапы построения квазирегулярной сетки и присвоения множества точек каждому подмассиву микрофонов, что снижает вычислительную нагрузку на вычислительное устройство и позволяет выполнять способ определения местоположения диктора в реальном времени.

На первом этапе способа согласно настоящему изобретению задают входные параметры геометрии конференц-системы, к которым относятся координаты каждого микрофона относительно стола. Также на первом этапе задают входные параметры для создания сетки поиска, к которым относятся номера пар микрофонов, параметры сетки (дискретный шаг, шаг расширения конуса), параметр отсеивания точек (минимальное значение). Также, на первом этапе задают входные параметры кластеризации, к которым относятся число точек L и порог расстояния. Указанные параметры задаются пользователем (инженером) через пользовательский интерфейс вычислительного устройства.

На следующем этапе принимают многоканальный звуковой сигнал, содержащий речь диктора. Согласно предпочтительному варианту реализации звуковой сигнал, содержащий речь диктора, обрабатывают детектором активности речи (VAD), посредством которого определяют временные интервалы, соответствующие речевым сегментам звукового сигнала. Такую обработку предпочтительно осуществляют при наличии большого количества сторонних шумов для исключения из последующего анализа источников акустических событий, не представляющих собой дикторов.

На следующем этапе осуществляют разбиение общей пространственной зоны поиска на отдельные участки, каждый из которых обрабатывается отдельным подмножеством микрофонов из множества микрофонов (микрофонная решетка, микрофонный массив). Указанное разбиение также осуществляется пользователем через пользовательский интерфейс вычислительного устройства путем выбора номеров микрофонов, входящих в каждое подмножество. Общее число микрофонов в конференц-системе на столе для совещаний обозначается как M. Для каждого из этих M микрофонов определяется подмножество микрофонов, которые участвуют в процедуре акустической локализации в области поиска в окрестности данного микрофона. Это подмножество микрофонов обозначается как n-подмассив. Значение n выбирается относительно геометрии стола и расположения микрофонов в этой зоне поиска. Например, если выбирается n=3, микрофон 1 (фиг. 2), который находится на углу стола и направлен в сторону длинной стороны стола, включается в 3-подмассив микрофонов {1, 2, 3}; микрофон 2, микрофоны по соседству от которого направлены в ту же сторону, включается в 3-подмассив {1, 2, 3}; микрофон 5 включается в 3-подмассив {4, 5, 6}; микрофон 13 включается в 3-подмассив {13, 14, 15} и т.д. Данное разбиение производится предварительно при настройке конференц-системы под конкретные стол и помещение; параметры разбиения заносятся в базу параметров предложенного способа (запоминающее устройство вычислительного устройства).

На следующем этапе способа посредством вычислительного устройства определяют множество дискретных точек в пространстве поиска дикторов. Множество точек определяется в горизонтальной плоскости на уровне высоты микрофонов. Такой поиск дискретных точек может быть осуществлен в условиях стола прямоугольной формы (квадратной, четырехугольной) или круглой, а также сложных форм (овал, трапеции, Г-образные, Т-образные, П-образные и т.п.). Процесс определения множества точек для

одной пары микрофонов проиллюстрирован на фиг. 3. Для определения множества дискретных точек выбирается пара последовательных микрофонов (m_1, m_2). Для этой пары выбирается множество линий, параллельных линии (ряду точек), соединяющей данные микрофоны. Каждая линия из множества линий удаляется от линии, соединяющей два микрофона, с инкрементируемым дискретным шагом. При этом длина каждой линии увеличивается с увеличением её отдаления от линии, соединяющей два микрофона. Таким образом, получается усеченный конус покрытия локальной зоны поиска для этих двух микрофонов. На каждую линию наносится множество точек с дискретным пространственным шагом. Таким образом, для двух микрофонов получается дискретная сетка точек поиска диктора в окрестности этих микрофонов. Например, при выборе инкремента шага между линиями, равного 0,1 м, и дискретного шага между накладываемыми точками, равного 0,1 м, получается регулярная сетка поиска с дискретным шагом 0,1 м. При этом распределение линий относительно линии, соединяющей два микрофона, производится под тем же углом, под которым установлены микрофоны относительно поверхности стола.

Таким образом, если микрофоны установлены на прямоугольном столе (фиг. 4) в направлении края стола, каждый усеченный конус будет направлен на часть края стола в направлении потенциального расположения дикторов относительно расположения микрофонов. Также возможно ограничивать зону поиска путем включения или изъятия пар микрофонов из установленных параметров с использованием пользовательского интерфейса вычислительного устройства. Например, на фиг. 2 для микрофонов {1, 30}, {12, 13}, {15, 16}, {27, 28}, возможно включение данных пар для построения сетки поиска на углах стола в случае, если дикторы потенциально могут располагаться на углах стола, и возможно изъятие этих пар при построении сетки поиска в случае, если дикторы в этих местах располагаться не будут. Пары микрофонов, относительно которых требуется производить поиск дикторов (создавать регулярную сетку поиска), заносятся в базу параметров заявляемого способа, хранящуюся в запоминающем устройстве вычислительного устройства. При неизменной конфигурации конференц-системы, геометрии стола и помещения, параметры пар микрофонов не требуют изменений в ходе последующей эксплуатации настоящего изобретения.

Далее, множество дискретных точек, построенных для каждой из указанных пар микрофонов, проходит отсеивание с критерием минимального взаимного расстояния между точками (предел расстояния). Поскольку локальные сетки поиска (для каждой пары микрофонов) могут накладываться друг на друга, могут возникать точки, чрезмерно близкие друг к другу. Расчет метрики акустической локализации для таких чрезмерно близких точек не повышает точности локализации, но использует дополнительные вычислительные ресурсы. Поэтому точки сетки, расположенные на расстоянии от соседних точек, которое меньше, чем установленный предел расстояния (равен дискретному шагу между точками, установленному в качестве параметров построения усеченного конуса), удаляются. Практический опыт применения показал, что такое отсеивание позволяет исключить из анализа до половины точек без ущерба точности акустической локализации, что практически вдвое сокращает используемые вычислительные ресурсы. На этом этапе образуется квазирегулярная сетка поиска по всей площади поиска потенциальных положений дикторов в окрестности кромки стола.

На следующем этапе заявляемого способа каждая точка из общей сетки поиска назначается конкретному n-подмассиву микрофонов, данные с которого впоследствии будут обрабатываться вычислительным устройством с целью оценки метрики акустической локализации. Для этого находится микрофон, расстояние между которым и выбранной точкой будет минимальным, и который направлен в сторону такой выбранной точки. Стоит отметить, что каждой точке из сетки поиска назначается только один n-подмассив.

Таким образом, каждому n-подмассиву назначается множество точек, которые этот n-подмассив использует в решении задачи акустической локализации. Вычисления на данном этапе являются независимыми, а, следовательно, могут быть реализованы с помощью параллельного вычисления на разных ядрах CPU многопроцессорного вычислительного устройства. Таким образом, выполнение первого и второго этапов способа позволяет применять параллельные вычисления и достигнуть таким образом значительного прироста скорости обработки многоканального сигнала.

На следующем этапе способа каждый n-подмассив микрофонов выполняет задачу акустической локализации звуковых событий из многоканального сигнала, поступающего в микрофоны конференц-системы (обработка в режиме реального времени) или подаваемого через пользовательский интерфейс вычислительного устройства (офлайн-обработка записи), на назначенном данному n-подмассиву множестве точек из общей квазирегулярной сетки поиска. Акустическая локализация по каждому n-подмассиву производится параллельно другим. Акустическая локализация производится согласно алгоритму SRP-PHAT, который оценивает т.н. SRP значение совокупной энергии сигнала в каждой конкретной точке пространства. Для этого алгоритм SRP-PHAT использует пространственные координаты микрофонов (задаются на первом этапе способа) в n-подмассиве и пространственные координаты точек сетки поиска этого n-подмассива. Значение SRP в точке рассчитывается как сумма значений кросс-корреляции между всеми парами сигналов n-подмассива со сдвигом кросс-корреляции, соответствующим теоретическим временным задержкам прихода акустических волн из данной точки до каждого из микрофонов n-подмассива. Коэффициент PHAT нормирует это суммарное значение кросс-корреляции на межканаль-

ную мощность сигнала. Таким образом, для каждой точки в сетке поиска рассчитывается нормированная на мощность сигнала метрика акустической локализации или метрика когерентности звуковых волн, т.н. SRP значение. Это значение увеличивается при уменьшении расстояния от точки до диктора, и имеет максимум в том месте, где положение точки совпадает с положением диктора. Алгоритм SRP-PHAT осуществляется в частотной области оконного преобразования Фурье (STFT). Многоканальный сигнал во временной области преобразуется в STFT массив с неким временным шагом кратковременного преобразования (кадр STFT); значение SRP рассчитывается для каждого сигнала (кадра) STFT с этим дискретным шагом. Алгоритм SRP-PHAT терпим к условиям средней реверберации и применим в конференц-системах применяемых для реализации способа согласно настоящему изобретению.

На следующем этапе способа согласно настоящему изобретению агрегируют данные со всех n-подмассивов и оценивают координаты активного диктора с помощью агломеративной иерархической кластеризации. Согласно предлагаемому способу, в каждый дискретный момент времени кадра STFT значения SRP со всех n-подмассивов объединяются в множество значений по всей квазирегулярной сетке поиска. Существует однозначное соответствие: для каждой дискретной точки поиска определено одно SRP значение, при этом все SRP значения сортируются по убыванию. Для дальнейшего анализа выбирается определенное число L наибольших значений SRP (например, 100 значений); выбирается L подмножество точек сетки поиска (из числа точек L на первом этапе способа), которым соответствуют эти L значений SRP. По данным L точкам производится агломеративная кластеризация, которая определяет кластеры со взаимным евклидовым расстоянием между точками в кластере не больше определенного верхнего порога расстояния (заданы на первом этапе способа) и получением кластеров по метрике евклидового расстояния между центрами кластеров. Верхний порог указанного расстояния определен максимальной площадью желаемого кластера, которая должна быть сопоставима с окрестностью зоны головы активного диктора, во избежание попадания в эту зону другого диктора и посторонних шумов. Результатом агломеративной кластеризации является некое количество кластеров, которым назначаются L точек поиска. Кластеры могут быть двумерными или трехмерными. Для каждого из выбранных таким образом кластеров определяются две метрики качества кластера: количество точек сетки поиска в кластере и суммарное значение SRP в кластере. Если существует такой кластер, для которого обе метрики качества достигают максимальных значений, данный кластер точек определяется как кластер, содержащий координаты предполагаемого диктора в данный момент времени STFT кадра. Такой кластер пространственных точек образует пространственную область, в которой локализуется голова активного диктора и которая примерно соответствует по размерам голове диктора (фиг. 5). Если же максимальное суммарное значение SRP принадлежит не самому большому кластеру по метрике количества точек, данный кластер проверяется на достаточный размер, и если проверка успешна, то данный кластер определяется как кластер, содержащий координаты предполагаемого диктора в данный момент времени STFT кадра. Если подходящий кластер не найден по метрикам качества, в данный момент времени STFT кадра результат кластеризации определяется как пустое множество. Это происходит в моменты тишины (отсутствия активных дикторов) или в моменты сложных шумовых условий, когда из-за присутствия множества источников точечных шумов или нестационарного шума высокой мощности акустическая локализация не способна выделить полезный источник (диктора) среди общего фона нецелевых источников.

Если в данный момент времени STFT кадра определен пространственный кластер, содержащий предполагаемое положение диктора, финальная оценка местоположения (координат) диктора вычисляется как взвешенное значение центра кластера, т.е. среднее по пространственным точкам кластера, взвешенное на значения SRP, соответствующие этим точкам. Таким образом, формируется временная последовательность оценок координат активных дикторов с частотой дискретизации, равной длине кадра STFT. Например, если длина кадра STFT равна 100 мс, частота временной дискретизации оценок координат активных дикторов будет равна 100 мс, или 10 оценок в секунду.

При использовании детектора активности речи (VAD), создающего временную сегментацию активности речи в подаваемом звуковом сигнале, может быть осуществлено определение координат активного диктора в период времени, соответствующему одной фразе диктора. Каждый временной сегмент из этой временной сегментации соответствует одной слитной фразе. Для определения координат диктора, соответствующих фразе, определяется покадровая последовательность оценок координат, соответствующая интервалу времени этого сегмента сигнала. Данная последовательность координат проходит этап кластеризации, в ходе которого определяется кластер точек, со взаимным евклидовым расстоянием между точками в кластере не больше определенного верхнего порога. Таким образом из расчета удаляются единичные точки, значительно удаленные от центра кластера. Эти точки обусловлены неточностью локализации в моменты микро-пауз в речи одной слитной фразы и влиянием сторонних шумов. Координаты диктора, соответствующие одной фразе, вычисляются как среднее значение точек данного кластера (центр формы кластера). В данном случае подразумевается, что диктор находится в квазистационарном положении при произнесении одной короткой связной фразы.

Результатом работы настоящего изобретения может являться как определение местоположения активного диктора в пространстве в виде покадровой оценки координат активных дикторов, так и локализация каждой слитной фразы в виде оценки координат, соответствующих каждой слитной фразе. Даль-

нейшее применение пространственной информации о положении дикторов не входит в область применения настоящего изобретения. Сама по себе оценка положения активного диктора в каждый момент времени является индикатором нахождения диктора в том или ином месте за столом совещаний. Резкая смена оцененного положения является индикатором смены диктора, что способствует повышению качества разделения и диаризации дикторов. Пространственная информация, примененная в совокупности с биометрической моделью, потенциально способствует повышению качества идентификации и диаризации дикторов. Пространственная информация потенциально применима в бимодальных и мульти-модальных конференц-системах, где пространственная информация, полученная с помощью анализа аудиосигнала, объединяется с пространственной информацией о дикторах, полученной с помощью анализа изображений видео сигнала. При подготовке протокола совещаний пространственная информация может заноситься в протокол либо в сыром виде, т.е. каждая детектированная и распознанная фраза снабжается метаинформацией о положении говорящего, либо как результат идентификации и диаризации говорящего (в случае объединения с биометрическим, мульти-модальным или каким-либо другим методом идентификации), т.е. каждая детектированная и распознанная фраза снабжается ID меткой идентифицированного говорящего.

Настоящее изобретение представляет собой программный комплекс, работающий на базе вычислительного устройства с использованием микрофонов конференц-системы, соединенной с вычислительным устройством с возможностью обмена данными. Вычислительное устройство через пользовательский интерфейс на первом этапе способа в качестве входных параметров принимает данные о геометрии конференц-системы: координаты микрофонов на столе для совещаний (устанавливаются при установке конференц-системы), а также номера микрофонов для разбиения всего микрофонного массива на n-подмассивы, параметры расстояний между точками и номера пар микрофонов для создания квазирегулярной сетки поиска; параметры STFT преобразования (длина кратковременного окна (кадра), количество пересечений кратковременных кадров); параметры кластеризации для оценки координат дикторов. Вычислительное устройство производит построение квазирегулярной сетки поиска, т.е. определение координат множества точек поиска. При неизменных параметрах: координат микрофонов на столе для совещаний, номеров микрофонов для разбиения всего микрофонного массива на n-подмассивы, параметров расстояний между точками и номеров пар микрофонов, процедура построения квазирегулярной сетки поиска осуществляется один раз при первом выполнении способа согласно настоящему изобретению. При изменении указанных выше параметров процедура построения квазирегулярной сетки поиска производится заново при инициализации программного комплекса; при этом структура квазирегулярной сетки поиска обновляется и остается неизменной до следующего обновления указанных выше параметров.

Программный комплекс, реализующий настоящее изобретение, может взаимодействовать программными модулями, не входящими в состав настоящего изобретения: модуль детектирования речи и модуль потребления результатов работы программного комплекса. При работе в онлайн-режиме программный комплекс считывает блоки из буфера многоканального потока аудиоданных; в офлайн-режиме программный комплекс использует сохраненную в файл запись совещания. Модуль детектирования речи синхронизирован с программным комплексом по временным отсчетам аудиоданных и, таким образом, обрабатывает тот же блок поточных данных или сохраненную запись, что и программный комплекс. Модуль потребления в данном контексте подразумевает любую потребитель, использующий результаты оценки координат дикторов, например, система идентификации и диаризации дикторов, протоколирующее приложение.

Программный комплекс обрабатывает либо онлайн-поток многоканальных аудиоданных, либо офлайн-запись конференции или совещания. Функционал программного модуля при этом остается неизменным, функции выполняются одинаково в онлайн- и офлайн-режимах. В онлайн-режиме на программный модуль накладывается ограничение по обработке в реальном времени RT. Реальным временем в данном контексте является временная продолжительность самой обрабатываемой записи. Если обработка записи занимает ровно столько же времени, сколько длится сама запись, значение RT=1. Если обработка занимает меньше времени, значение RT<1. При достаточных вычислительных мощностях настоящий программный комплекс выполняет условие работы в реальном времени RT<1.

Практическое применение программного комплекса было произведено на вычислительном устройстве со следующими значимыми параметрами: CPU типа Xeon E7 v4, 48 виртуальных ядер; 252 ГБ оперативной памяти. Многопоточность вычислений программного комплекса осуществляется на CPU, графическая карта не задействуется. Далее приводятся данные испытаний, произведенных на записи конференц-системой, приведенной на фиг. 2. Запись продолжительностью 83 мин, включающая речь 27 дикторов, обрабатывалась программным комплексом в офлайн-режиме (подается целиком) и в симуляции онлайн-режима (подается блоками по 5 мин каждый). Запись представляет собой 30-канальный WAV PCM 16 файл, записанный в несжатом виде с частотой дискретизации равной 16000 Гц. Запись была получена в типовых условиях совещания, содержит спонтанную речь, присутствуют перебивания и одновременная речь.

Шумовая обстановка типовая, присутствуют точечные источники шумов офисных предметов, по-

суды под кофе/воду. Фоновым шумом является шум кондиционера и видео проектора. Реверберационная картина помещения описывается временем реверберации T60=600 мс. Во время записи положения дикторов мягко фиксировались; не разрешалось вставать, пересаживаться. Обычные движения и смещения в рамках посадочного места не запрещались. Для подсчета точности определения местоположения дикторов координаты каждого диктора были измерены в координатной системе помещения до начала совещания. Координаты измерялись при ровной посадке каждого диктора на его посадочном месте.

При реализации способа согласно настоящему изобретению, общий микрофонный массив поделен на 3-подмассивы согласно таблице. Для каждой из пар последовательных микрофонов создана дискретная сетка локальной зоны поиска. Исключения составляют пары микрофонов, стоящих на углах стола и, соответственно, направленные перпендикулярно друг другу: {1, 30}, {12, 13}, {15, 16}, {27, 28}. Дискретный шаг между точками выбран 0,1 м. Результирующая общая зона поиска покрывает весь периметр стола, начинаясь от 0,3 м вглубь от края стола, заканчиваясь на 1,2 м от края стола. Общее число точек в сетке поиска составило 3600 точек; это число точек поделено для обработки 3-подмассивами, каждому 3-подмассиву выделено 120-220 точек, в зависимости от расположения сетки поиска относительно микрофонов. Каждый 3-подмассив осуществлял задачу акустической локализации в параллельном режиме. Каждому 3-подмассиву выделено одно виртуальное ядро CPU. Таким образом, загруженность ядер на задачу локализации составляет 30 из 48 ядер. Шаг временной дискретизации STFT составляет 100 мс. Результат локализации по общей сетке поиска осуществлялся с параметром количества выбираемых точек L=50.

Присвоение 3-подмассива каждому микрофону конференц-системы

Номер микрофона	3-подмассив
1	{1, 2, 3}
2	{1, 2, 3}
3	{2, 3, 4}
4	{3, 4, 5}
5	{4, 5, 6}
6	{5, 6, 7}
7	{6, 7, 8}
8	{7, 8, 9}
9	{8, 9, 10}
10	{9, 10, 11}
11	{10, 11, 12}
12	{10, 11, 12}
13	{13, 14, 15}
14	{13, 14, 15}
15	{13, 14, 15}
16	{16, 17, 18}
17	{16, 17, 18}
18	{17, 18, 19}
19	{18, 19, 20}
20	{19, 20, 21}
21	{20, 21, 22}
22	{21, 22, 23}
23	{22, 23, 24}
24	{23, 24, 25}
25	{24, 25, 26}
26	{25, 26, 27}
27	{25, 26, 27}
28	{28, 29, 30}
29	{28, 29, 30}
30	{28, 29, 30}

Для исключения влияния ошибки VAD на точность определения местоположения, в условиях испытаний используется эталонная разметка фраз каждого диктора, созданная вручную. Точность работы

настоящего изобретения вычисляется следующим образом. Каждой зафиксированной фразе назначается метка координат в системе координат помещения. Оценка считается верной, если координаты, назначенные фразе, находятся ближе всего к эталонным измеренным координатам диктора, которому данная фраза принадлежит, но не более чем 0,5 м от этих эталонных координат. Если же оценка координат находится ближе к координатам диктора, не произносившего данную фразу, или более чем 0,5 м от координат верного диктора, оценка считается неверной. Отношение количества фраз с неверной оценкой координат к общему количеству фраз является критерием ошибки определения местоположения в процентах.

На данной записи продолжительностью в 83 мин ошибка определения координат составила 3,01%. При загруженности 30 виртуальных ядер CPU из 48 показатель RT составил RT=0,18 при офлайн-обработке записи и RT=0,23 при онлайн-обработке потока записи. Пиковая загруженность оперативной памяти при этом составила 23 ГБ. Средняя загруженность оперативной памяти составила 12 ГБ.

Выше описан предпочтительный, но не единственный, вариант реализации заявляемого способа, и объем правовой охраны, испрашиваемой в настоящей заявке, ограничен только нижеследующей формулой изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ определения местоположения дикторов с использованием конференц-системы, согласно которому:

a) задают входные параметры конференц-системы, входные параметры для построения квазирегулярной сетки поиска, которыми являются номера пар микрофонов, дискретный шаг, шаг расширения конуса и предел расстояния, и входные параметры кластеризации, содержащие порог расстояния;

b) принимают многоканальный звуковой сигнал, содержащий речь диктора;

c) разбивают множество микрофонов конференц-системы на подмножества микрофонов, каждое из которых определяет отдельную пространственную область поиска, которая выделена из общей пространственной области поиска;

d) получают квазирегулярную сетку поиска с использованием входных параметров для построения сетки поиска путем определения множества дискретных точек в общей пространственной области поиска и исключают такие точки, которые находятся от других точек на расстоянии меньшем, чем установленный предел расстояния;

e) присваивают каждую точку квазирегулярной сетки поиска только одному из подмножества микрофонов с формированием множества дискретных точек для каждого из подмножеств микрофонов;

f) осуществляют параллельную акустическую локализацию в многоканальном звуковом сигнале для каждой отдельной пространственной области поиска с использованием алгоритма SRP-PHAT (Steered Response Power Phase Transform) с получением SRP-значений для каждой точки каждого множества дискретных точек;

g) осуществляют агломеративную иерархическую кластеризацию SRP-значений дискретных точек с использованием порога расстояния с выделением кластера пространственных точек, задающих местоположение диктора.

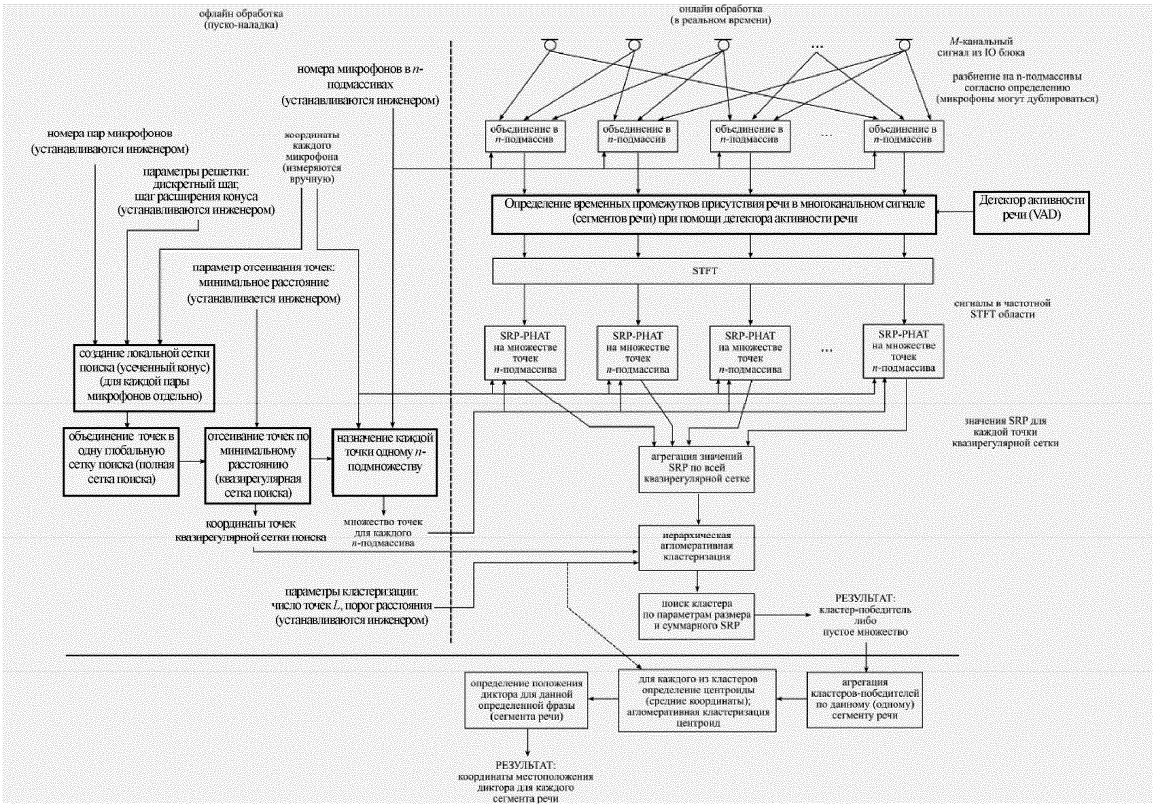
2. Способ по п.1, согласно которому входными параметрами конференц-системы являются координаты каждого микрофона в системе координат помещения, а входным параметром кластеризации дополнительно является число точек L.

3. Способ по любому из пп.1, 2, согласно которому звуковой сигнал принимают либо с микрофонов конференц-системы, либо из файла с записью звукового сигнала.

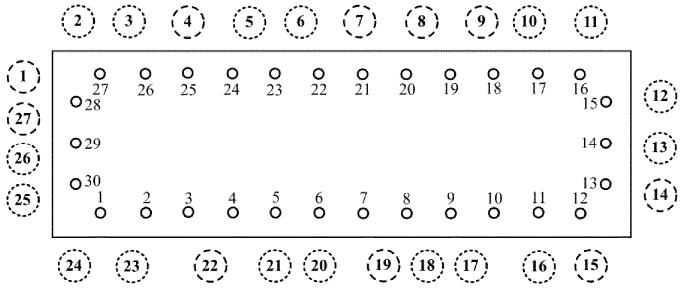
4. Способ по любому из пп.1-3, согласно которому в выделенном кластере пространственных точек, задающих местоположение диктора, определяют взвешенное значение центра кластера.

5. Способ по любому из пп.1-4, согласно которому звуковой сигнал, содержащий речь диктора, дополнительно обрабатывают с выделением временных интервалов, соответствующих речевым сегментам звукового сигнала, а параллельную акустическую локализацию и агломеративную иерархическую кластеризацию осуществляют только для выделенных временных интервалов.

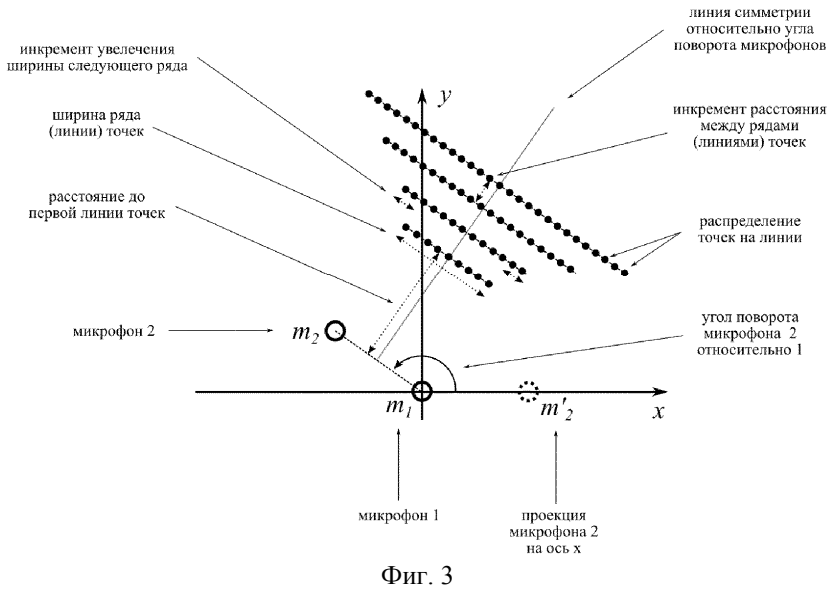
6. Способ по п.5, согласно которому выделение речевых сегментов осуществляют с применением детектора активности речи (VAD).



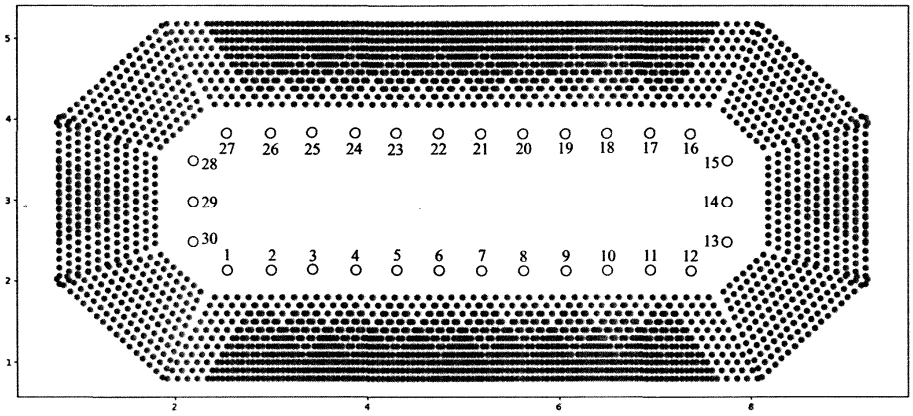
Фиг. 1



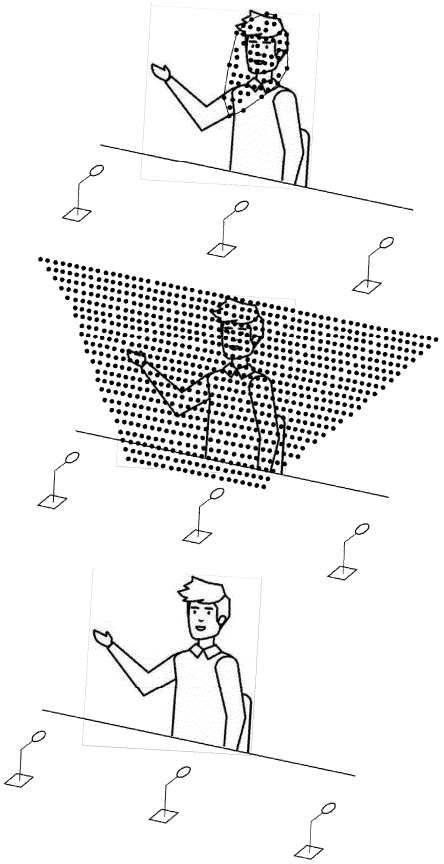
Фиг. 2



Фиг. 3


27 26 25 24 23 22 21 20 19 18 17 16
28
29
30
15
14
13
1 2 3 4 5 6 7 8 9 10 11 12

Фиг. 4



Фиг. 5