УДК 004.93

ОСОБЕННОСТИ ДИСТАНЦИОННОЙ ЗАПИСИ И ОБРАБОТКИ РЕЧИ В АВТОМАТАХ САМООБСЛУЖИВАНИЯ

А. Л. Ронжин,
канд. техн. наук, доцент
А. А. Карпов,
канд. техн. наук, старший научный сотрудник
И. А. Кагиров,
младший научный сотрудник
Санкт-Петербургский институт информатики и автоматизации РАН

Рассматривается ряд проблем, возникающих при дистанционной записи речи в зашумленных условиях. Повысить точность выделения границ полезного речевого сигнала предлагается за счет применения спектрально-пространственного анализа многоканального звукового сигнала.

Ключевые слова *— дистанционное распознавание речи, пассивная локализация, корреляция взаимного спектра, речевой пользовательский интерфейс.*

Введение

С развитием технологий «повсеместных вычислений» (ubiquitous computing) и разработкой так называемых «окружающих интеллектуальных пространств» (ambient intelligence space) к речевым технологиям предъявляются все более жесткие требования, в частности, система должна воспринимать речь диктора, свободно перемещающегося в помещении, т. е. самостоятельно определять местонахождение источника полезного сигнала. Подавляющее большинство существующих систем распознавания речи способно обрабатывать только речь диктора, записанную с помощью микрофона-гарнитуры, расположенного непосредственно перед ртом диктора, саму же запись рекомендуется проводить в тихом, звукоизолированном помещении. Однако очевидно, что далеко не каждый пользователь готов к таким ограничениям. Для развития и внедрения речевых технологий необходимо сделать процесс записи речи максимально удобным для пользователя, прежде всего, обеспечив дистанционную запись речи в условиях фонового шума и параллельных разговоров в помещении.

Следует признать, что технологии распознавания речи все еще на таком уровне, что основными областями, в которых они действительно востребованы, остаются индустрия развлечений и сфера обслуживания. Причиной тому является их недостаточная устойчивость по отношению к вариативности речи, окружающей акустической обстановке и другим неблагоприятным факторам. Так, по сравнению с автоматической системой распознавания профессиональная стенографистка в несколько раз быстрее набирает текст, а в условиях перегрузок и сильных аудиошумов вместо голосовых команд по-прежнему проще использовать манипуляторы и джойстики, расположенные непосредственно в руках оператора (пилота самолета).

Таким образом, принципиальными для функционирования системы становятся, с одной стороны, фильтрация шумов и с другой — автоматическая локализация пользователя (объекта-источника сигнала).

Именно поэтому одним из объектов самого пристального внимания в области автоматического распознавания речи стала проблема записи речи при помощи микрофонов, расположенных на значительном расстоянии от диктора, а не установленных в непосредственной близости ото рта.

При обработке речевого сигнала посредством двух и более микрофонов (т. е. массива датчиков) используется способность бинаурального слуха оценивать пространственную акустическую обстановку [1]. Кроме непосредственной функции приема звуковых сигналов, массив микрофонов используется для пространственной локализа-

ции источников звука и фильтрации полезного сигнала за счет управления диаграммой направленности массива микрофонов. Благодаря этому система воспринимает и анализирует звуки, исходящие из узкой области рабочего пространства, и значительно ослабляет или даже совсем отсекает звуки, приходящие со всех остальных направлений.

В настоящей статье дается краткий обзор существующих методов пространственной обработки сигнала при помощи массива микрофонов и представлены результаты экспериментов, проведенных при использовании нескольких модификаций метода обобщенной функции взаимной корреляции и ряда пространственных моделей массива микрофонов.

Специфика дистанционной записи и распознавания речи

При переносе системы распознавания речи из лабораторных условий в обычные мы сталкиваемся с рядом новых проблем и особенностей речевого взаимодействия. Пожалуй, самым сложным случаем (и в то же время одним из самых распространенных) для автоматической системы будет ситуация cocktail party, когда в помещении находится большое число людей, свободно перемещающихся и разговаривающих. В такой обстановке система записывает многомерный звуковой сигнал, содержащий все звуки источников, находящихся в помещении, в котором проходит запись. При использовании методов спектрально-пространственной фильтрации необходимо разделить все звуковые сигналы, произвести идентификацию дикторов, определить их положение и, наконец, распознать их речь. Рассмотрим специфику этой задачи более подробно.

Из-за недостатка обучающих данных по спонтанной русской речи сегодня лишь немногие системы распознавания являются дикторонезависимыми, поэтому при обработке речевого сигнала, в котором содержится речь нескольких дикторов, требуется предварительная настройка на каждого диктора персонально (записывается некоторый обучающий набор фраз) или же производится адаптация моделей фонем по первым фразам конкретного диктора. Таким образом, для обработки речевого сигнала, содержащего фразы нескольких дикторов, либо требуется дикторонезависимый модуль обработки речи, модели фонем которого обучены на достаточно представительном корпусе фраз, наговоренных значительным числом дикторов, либо модуль идентификации речи диктора должен сегментировать входной сигнал на участки, принадлежащие разным дикторам, и адаптировать модели фонем персонально для каждого диктора. Первый способ является более эффективным, поскольку в этом случае проводится обучение с «учителем».

Эксперименты показывают, что длительность «перекрывающейся» речи в условиях деловой встречи, конференции или совещания может достигать 70 % [1]. Эта проблема настолько серьезна, что вызывает сбои даже у систем распознавания, использующих персональные радиомикрофоны, расположенные непосредственно у рта диктора. Тем не менее, поскольку в одной точке пространства может находиться только один диктор, то именно пространственно-спектральная фильтрация многомерного аудиосигнала, записанного через массив микрофонов, позволит разделить сигналы, исходящие из разных точек и разных источников.

В процессе изложения своих идей докладчик иногда передвигается, сопровождая речь жестами. При перемещении диктора расстояние между ним и массивом микрофонов постоянно меняется, соответственно, будет меняться и уровень речевого сигнала (громкость). Для одновременной записи как удаленных, так и близкорасположенных источников необходимо применение аналого-цифровых преобразователей, обеспечивающих запись аудиосигнала в широком динамическом диапазоне.

Проблема вариативности уровня сигнала проявляет себя в первую очередь на этапе определения границ речевого сигнала. Большинство алгоритмов определения активности голоса выделяют речь в аудиосигнале, анализируя энергию сигнала или его спектра. Если энергия сегмента превышает некоторый заданный порог, то система считает, что начался речевой сигнал. В результате при фиксированном пороге получится, что границы одного и того же сигнала, проигранного с разного расстояния до микрофона, будут отличаться между собой. Фраза, произнесенная с близкого расстояния, будет записана целиком, а при увеличении расстояния часть фразы может пропасть, так как энергия сигнала уменьшится, и некоторые сегменты по энергии не будут превышать заданный порог. Наличие взрывных согласных в сигнале приводит к тому, что сигнал после смычки не записывается совсем. Дело в том, что при снижении уровня сигнала громкость участков в районе смычки становится незначительной, в результате чего длительность паузы значительно увеличивается и часто воспринимается как окончание фразы. Инвариантность данных алгоритмов по отношению к амплитуде сигнала можно повысить, приняв во внимание расстояние до источника сигнала.

При использовании массива микрофонов необходимо обеспечить синхронную запись сигна-

лов по каждому из датчиков. К сожалению, это не всегда удается реализовать технически. Наш собственный опыт показывает, что если подключать отдельные USB-микрофоны, то их синхронизация представляет собой сложную задачу, поскольку операционная система накладывает свои ограничения и выставляет различные приоритеты устройствам. Использование же специализированных аудиоплат, осуществляющих синхронизацию и управление потоками самостоятельно, хотя и позволяет частично решить данную проблему, но сопряжено с дополнительными трудностями технического (и финансового) характера.

В ряде экспериментов [1] для верификации результата распознавания используется комбинация массива микрофонов и близкостоящего микрофона. Массив микрофонов служит для определения положения диктора, а когда говорящий определен, его речь также записывается с помощью ближайшего к нему микрофона. Для распознавания используются сигналы, записанные как при помощи массива микрофонов, так и близкостоящего микрофона, и полученные результаты затем сопоставляются.

Методы спектрально-пространственной обработки звуковых сигналов

При разработке систем, анализирующих распространяющиеся в пространстве сигналы, в первую очередь возникают проблемы интерференции различных сигналов. Если полезный сигнал и помехи имеют одни и те же частотно-временные характеристики, то временная фильтрация сигнала не эффективна. Однако источники полезного сигнала и помех часто располагаются в различных точках пространства, в этом случае использование пространственного фильтра для ослабления уровня помех невозможно.

Применение временного фильтра требует обработки данных, собранных в некотором временном интервале (окне). Аналогично, пространственная фильтрация производится над данными, полученными в некотором пространственном окне [2]. Пространственная разрешающая способность зависит от размера окна — чем больше окно, тем лучше разрешение. Однако абсолютный размер окна не так важен, как его отношение к длинам принимаемых волн. На практике чаще используются высокочастотные сигналы, в том числе и потому, что для их приема требуются антенны с меньшими габаритными размерами.

Рассмотрим наиболее общие подходы к решению задачи локализации пользователя в пространстве. В настоящий момент до конца не определено, какой класс методов лучше всего использовать для решения задачи локализации источника речевого сигнала [3–5]. Прежде всего, это связано с тем, что методы были изначально ориентированы на локализацию узкополосных сигналов, а речь имеет относительно широкий частотный диапазон, и для ее обработки требуется некоторая модификация данных методов.

В большинстве подходов используются спектральные методы обработки звукового сигнала [6–8], а также параметрические методы для статистической оценки положения источника звука и фильтрации полезного сигнала **[9]. Спектральные** методы довольно привлекательны с точки зрения простоты вычислений, поэтому они применяются чаще. В связи с этим в данной работе за основу был выбран также спектральный подход.

Среди спектральных подходов следует выделить метод формирования луча (beam forming) — обработку сигнала, которая обеспечивает системе пространственную селективность [4]. Его назначение состоит в выделении компонент сигнала, распространяющегося из определенной точки пространства. Задача формирования луча или диаграммы направленности, т. е. избирательного приема излучения в пространстве с разных направлений, подобна полосовой фильтрации: сигналы, принадлежащие лучу, пропускаются, а сигналы, ему не принадлежащие, ослабляются.

Одна из простейших систем формирования луча основана на взвешенном суммировании сигналов с задержкой. Суммарный сигнал образуется усреднением взвешенных и задержанных копий сигнала, поступающих с микрофонов. Задержки выбираются таким образом, что центр полосы пропускания располагается вдоль определенного направления в пространстве. Например, если все микрофоны расположены в одной плоскости и мы хотим направить луч перпендикулярно этой плоскости, то задержки для всех микрофонов должны равняться нулю. Тогда плоские волны, идущие перпендикулярно, будут складываться по фазе, а волны, поступающие из других направлений, будут складываться с различными фазами и преимущественно ослабляться.

Для обработки узкополосных сигналов применяется метод формирования луча, который производит дискретизацию сигнала в пространстве. На выходе фильтра в момент времени t появляется некоторый сигнал $y(t)$, являющийся линейной комбинацией данных, поступивших с J датчиков в момент времени t:

$$y(t) = \sum_{l=1}^{J} \mathbf{w}_l^* \mathbf{x}_l(t),$$

где $\mathbf{w}$ — вектор весов; $\mathbf{x}(t)$ — вектор входных данных, а * представляет комплексное сопряжение,

что упрощенно соответствует перемножению данных с некоторыми весами. Тем не менее, предполагается, что данные и веса являются комплексными числами, поскольку во многих приложениях используются энергия и фаза сигналов.

Для широкополосных сигналов важно учитывать дискретизацию как в пространстве, так и во времени. Выход фильтра в этом случае может быть описан следующим выражением:

$$y(t) = \sum_{l=1}^{J} \sum_{p=0}^{t-1} \mathbf{w}_{l,p}^{*} \mathbf{x}_l(t-p),$$

где $t - 1$ – число задержек на каждом из J каналов датчиков. Если сигнал на каждом датчике рассматривать как вход, то формирователь луча представляет собой систему с несколькими входами и одним выходом. Тогда для обоих методов формирования луча можно записать

$$y(t) = \mathbf{w}^{H} \mathbf{x}(t),$$

где H — эрмитов оператор.

В зависимости от способа настройки весового вектора **w** существуют различные методы формирования луча [10]. Наибольшее распространение получили формирователь луча Бартлетта [9] и формирователь луча Кэпона [11]. Первый метод использует постоянный набор весов и задержек для объединения сигналов, поступивших с микрофонов, преимущественно учитывая только информацию о расположении микрофонов в пространстве и направление волны. Второй же метод, наоборот, анализирует эту информацию вместе со свойствами принятого сигнала, что позволяет уменьшить количество помех.

В задачах, где требуется только определение положения источника звука без пространственной фильтрации полезного сигнала, применяется измерение времени задержки между сигналами, записанными двумя или более микрофонами. В большинстве приложений используются методы обобщенной функции взаимной корреляции (General Cross Correlation — GCC) [6] или обработки фазы сигналов [7, 12] для оценивания задержки прихода звуковой волны. В таких методиках положение диктора определяется с помощью набора оценок задержек, вычисленных по разным микрофонам. Основной недостаток всех предложенных методов наблюдается в условиях высокой реверберации, когда происходит множественное отражение звуковых волн от стен помещения и основной сигнал перемешивается с его отраженными копиями. Также следует отметить, что проблема локализации нескольких источников звука стала исследоваться совсем недавно, хотя такая ситуация наиболее свойственна для реальных условий. Для повышения точности локализации также эффективна интеграция методов аудио- и видеообработки [4]. За последние десятилетия в этой сфере были достигнуты большие успехи, однако интенсивность исследований по-прежнему высока. Наиболее актуальной проблемой все еще остается недостаточная устойчивость систем по отношению к шумовым помехам и реверберации.

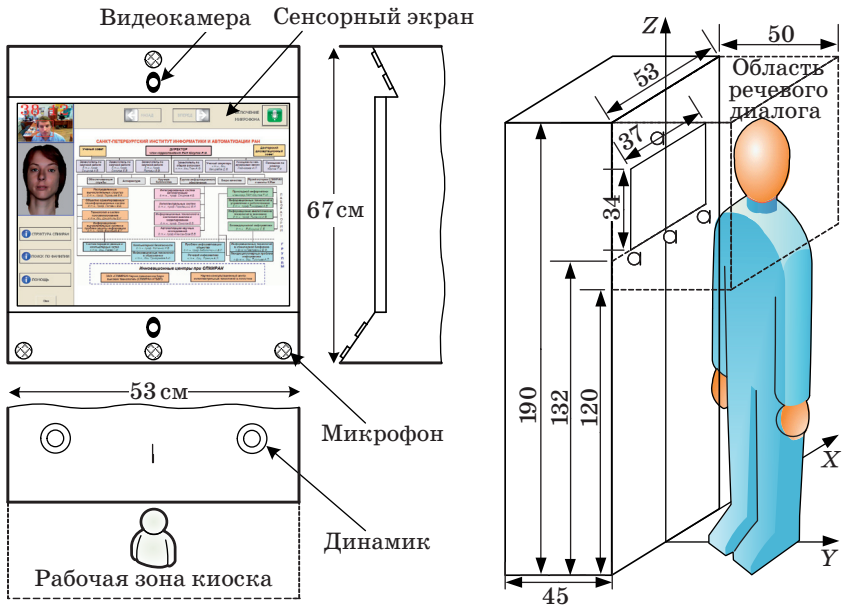
Тестирование разработанного метода локализации источника звуковых сигналов

При разработке системы для локализации диктора в пространстве перед киоском (автоматом) были применены три алгоритма: GCC-SCOT, GCC-PHAT и LMS, основанные на оценке времени задержки прихода сигнала к двум разным микрофонам [12]. При проведении экспериментов изменялись следующие параметры: 1) расстояние между микрофонами; 2) расстояние от источника звука до микрофона; 3) отклонение источника звука от линии массива микрофонов [13]. Звуковой сигнал записывался синхронно двумя микрофонами с частотой дискретизации 16 кГц. Комплексное преобразование Фурье вычислялось для сегмента сигнала размером 512 отсчетов с шагом 128 отсчетов.

При сравнении сигналов, записанных двумя разными микрофонами, определялся угол отклонения источника звука от линии массива микрофонов. Оценка данного угла производилась только для сегментов, максимальное значение функции взаимной корреляции которых превышало заданный порог. Длительность тестовой фразы составляла 2,3 с. Усредненная оценка угла вычислялась для сегментов записанного сигнала в диапазоне 0,5–1,8 с. Алгоритмы показали примерно одинаковые результаты, поэтому за основу был взят алгоритм GCC-PHAT, поскольку он требовал меньших вычислительных ресурсов.

С 2007 г. разрабатывается система МИДАС (многомодальный интерактивно-диалоговый автомат самообслуживания), распознающая присутствие клиента и вербально взаимодействующая с ним на естественном языке. Все эксперименты по дистанционной записи и распознаванию речи проводились с помощью разработанного многомодального киоска. В состав аппаратной части массива входят четыре микрофона «Октава» МК-012 и звуковая плата PreSonus Firepod.

Точность локализации источника полезного звука (речи клиента) и качество принимаемого аудиосигнала в сильной степени зависят от пространственного расположения микрофонов в массиве и самого массива в корпусе киоска. В киоске (рис. 1) используется массив микрофонов, три из которых расположены на одной линии под экра-



■ ***Рис. 1.*** *Расположение устройств ввода-вывода интерфейсной части и общий вид киоска*

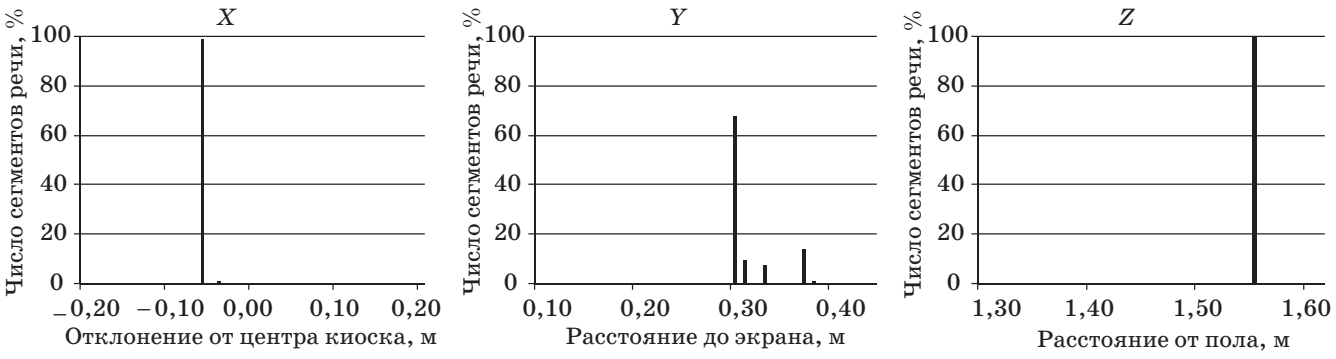
ном, а один находится над ним. Такое положение позволяет определять координаты источников звука в трехмерном пространстве. При этом один микрофон располагается по центру под монитором максимально близко ко рту пользователя, что обеспечивает уверенный прием аудиосигнала, кроме того, пользователь частично закрывает его от внешних шумов. Допустимая область речевого диалога на рисунке ограничена пунктирными линиями. В ней может находиться лицо (и рот) потенциального клиента во время общения с киоском.

В нашем киоске для определения положения источника звука в пространстве используется оценка времени задержки прихода сигнала по четырем парам микрофонов. Затем методом триангуляции рассчитываются координаты источника, а по длительности сигнала и энергии спектра принимается решение о наличии речи в записанном звуковом потоке.

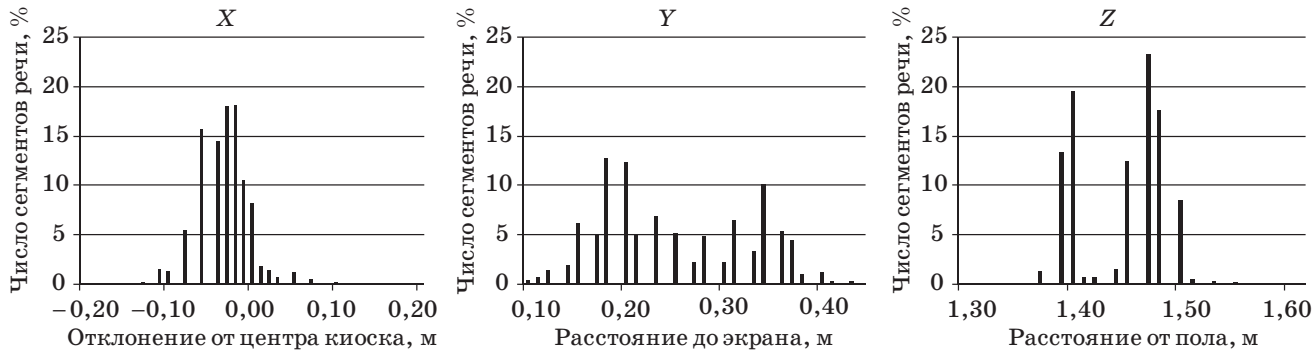
В первом эксперименте для определения точности работы алгоритма трехмерной локализации набор из 2500 фраз был проигран через динамик, располагавшийся перед киоском в течение всего процесса проигрывания. Разброс значений координат центра динамика, которые были определены в ходе записи речевого сигнала через массив микрофонов, показан на рис. 2.

Как можно видеть, большинство речевых сегментов было локализовано с одинаковыми значениями координат (– 0,06; 0,32; 1,55). Некоторый разброс параметров присутствует по оси *Y*. Тем не менее, используя методы сглаживания, можно добиться более устойчивого определения координат источника.

В ходе следующего эксперимента фразы говорили пользователи киоска (5 пользователей разного роста и пола, каждый из которых произнес по 500 фраз). Так как параметры системы локализации остались прежними, то увеличившийся



■ ***Рис. 2.*** *Значения координат центра динамика, определенные автоматически*



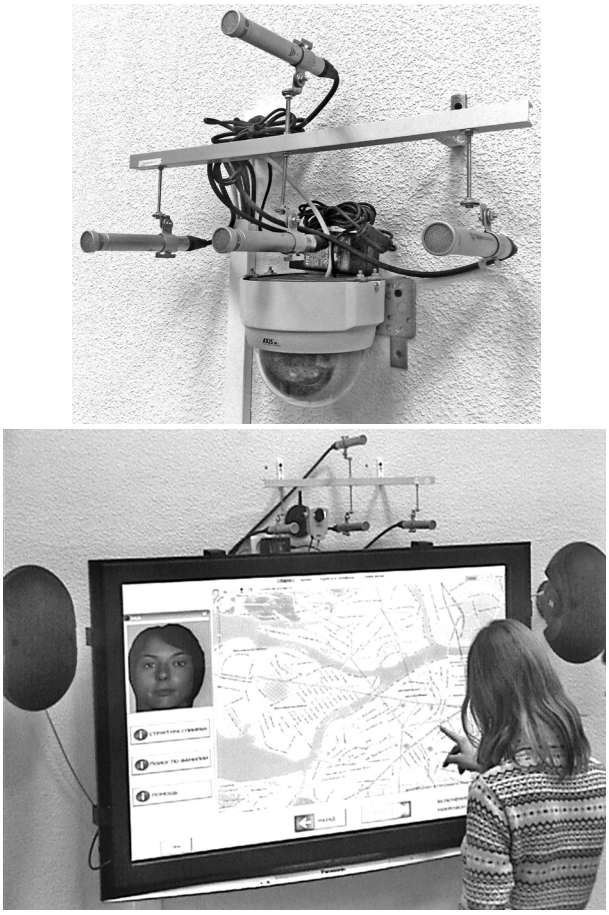
■ ***Рис. 3.*** *Значения координат центра рта пользователя, определенные автоматически по голосу*

разброс значений координат связан с естественными небольшими передвижениями пользователей в процессе взаимодействия с киоском. Разброс значений координат центра рта пользователей, которые были определены в ходе записи речевого сигнала через массив микрофонов, показан на рис. 3.

Для данной системы ось X проходила вдоль нижней границы передней панели киоска, ось Y — через центр киоска в сторону клиента, ось Z — вертикально по центру передней панели киоска. Таким образом, центр координат находился посередине нижней границы передней панели киоска, как показано на рис. 1. Приведенные графики показывают, что рабочая зона взаимодействия с киоском может быть ограничена по оси X шириной киоска (53 см), а по оси Y — 50 см, что достаточно для удобного управления сенсорным монитором, и, наконец, по оси Z допустимая область речевого диалога может быть ограничена от 120 до 180 см. Также было замечено, что при произнесении команды пользователь иногда наклонялся ближе к киоску, интуитивно стремясь повысить громкость речевого сигнала. В проведенном эксперименте не участвовали дети, однако при разработке многомодальных игровых/обучающих приложений следует учесть, что нижняя граница рабочей зоны по оси Z должна быть существенно снижена.

Разработанная модель дистанционной локализации источника звука и алгоритм спектрально-пространственного определения границ речи были успешно внедрены в многомодальные справочные системы СПИИРАН и карту Санкт-Петербурга.

Массив микрофонов, собранный по схеме «перевернутая Т», представлен на рис. 4. Соседние элементы массива расположены в 20 см друг от друга. При этом массив совмещен с видеокамерой, вращающейся в горизонтальной и вертикальной плоскостях, что позволяет использовать его для бимодальной локализации и слежения за



■ ***Рис. 4.*** *Прототип массива микрофонов для «интеллектуального зала» СПИИРАН*

диктором. Эта модель массива микрофонов была использована при оснащении «интеллектуального зала» института.

В 2009 г. киоск был размещен в коридоре здания института и апробирован в реальных условиях эксплуатации. В конце работ планируется улучшенный прототип киоска выставить, например, в музеях или вокзалах для массового обслуживания клиентов.

Заключение

Предложенный метод предназначен для определения положения диктора в пространстве перед массивом микрофонов и дистанционной записи голосовых команд без использования гарнитуры-микрофона. Способность бинаурального слуха оценивать пространственную акустическую обстановку использована при разработке антропоморфных моделей записи и обработки речевого сигнала посредством двух или нескольких микрофонов. Кроме непосредственной функции приема звуковых сигналов, массив микрофонов и разработанный программный модуль используются для пространственной локализации источников звука и фильтрации полезного речевого сигнала. Благодаря этому массив микрофонов воспринимает и анализирует звуки, исходящие из определенной рабочей области пространства, и значительно ослабляет звуки, приходящие со всех остальных направлений. Модуль позволяет существенно повысить качество дистанционного распознавания речи в справочно-информационных системах, располагающихся в зашумленных местах массового использования.

Разработанная модель дистанционного распознавания русской речи проходит тестовую эксплуатацию в СПИИРАН в многомодальном справочном киоске, предлагающем в интерактивном режиме информацию о сотрудниках института, научных подразделениях и текущих мероприятиях. Также в киоске реализована многомодальная карта Санкт-Петербурга, где посредством голосового дистанционного запроса производится поиск улицы и вывод на экран сенсорного монитора интересующего участка карты. Кроме того, в «интеллектуальном зале» модуль дистанционного распознавания речи используется для голосового управления светом, шторами, телевизором и другими исполнительными устройствами. Внедрение систем дистанционного распознавания речи в системы массового обслуживания, банкоматы и справочные автоматы позволит обеспечить пользователю возможность интуитивного взаимодействия с системой за счет использования речевого интерфейса.

Данные исследования проводятся при финансовой поддержке Правительства Санкт-Петербурга и гранта Российского фонда фундаментальных исследований № 07-07-00073-а.

Литература

1. **Stanford V., Rochet C., Michel M., Garofolo J.** Beyond Close-talk — Issues in Distant Speech Acquisition, Conditioning Classification, and Recognition// Proc. ICASSP 2004 Meeting Recognition Workshop. Montreal, Canada, 2004. P. 123–127.
2. **The Digital** Signal Processing Handbook/Ed. V. Madisetti, D. Williams. — N. Y.: CRC Press, 1999. — 1776 p.
3. **Johnson D., Dugeon D.** Array Signal Processing: Concepts and Techniques. — NJ: Prentice Hall, Inc. Englewood Cliffs, 1993. — 512 p.
4. **Brandstein M., Ward D.** Microphone Arrays Signal Processing Techniques and Applications. — Berlin; Heidelberg; N. Y.: Springer-Verlag, 2001. — 398 p.
5. **Trees H.** Optimum Array Processing. — N. Y.: John Wiley & Sons, 2002. — 1456 p.
6. **Knapp C. H., Carter G. C.** The generalized correlation method for estimation of time delay//IEEE Trans. Acoustics Speech Signal Proc. 1979. Vol. 24. P. 320–327.
7. **Lathoud G., McCowan I. A.** A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays//Proc. of SAPA-2004. Korea, 2004. P. 93–105.
8. **Даджион Д., Мерсеро Р.** Цифровая обработка многомерных сигналов: Пер. с англ. — М.: Мир, 1988. — 488 с.
9. **Krim H, Viberg M.** Two decades of array signal processing research: the parametric approach//Signal Proc. Magazine. Cambridge, MA. Jul. 1996. Vol. 13. N 4. P. 67–94.
10. **Van Veen B. D., Buckley K. M.** Beamforming: A Versatile Approach to Spatial Filtering//IEEE ASSP Magazine. April 1988. P. 4–24.
11. **Capon J.** High-Resolution Frequency-Wavenumber Spectrum Analysis//Proc. IEEE. Aug. 1969. Vol. 57. N 8. P. 2408–2418.
12. **Omologo M., Svaizer P.** Acoustic event localization using a crosspower-spectrum phase based technique// Proc. of ICASSP. 1994. Vol. 2. P. 273–276.
13. **Ронжин А. Л., Карпов А. А.** Сравнение методов локализации пользователя многомодальной системы по его речи//Изв. вузов. Приборостроение. 2008. Т. 51. № 11. С. 41–47.