УДК 004.522

А.Л. РОНЖИН

СПОСОБЫ ОЦЕНИВАНИЯ СИСТЕМ АУДИОЛОКАЛИЗАЦИИ ВЫСТУПАЮЩИХ В ЗАЛЕ СОВЕЩАНИЙ

Ронжин А.Л. **Способы оценивания систем аудиолокализации выступающих в зале совещаний**.

Аннотация. Применение методов аудиолокализации позволяет оценить положение и направление головы говорящего в помещении. Подобные системы в настоящее время популярны при разработке интеллектуальных систем сопровождения мероприятий в залах совещаний. В работе проанализирован ряд методик для оценивания производительности систем аудиолокализации, а также их интеграции с системами видеомониторинга. На примере разработанного интеллектуального зала совещаний проведена оценка точности аудиолокализации выступающих, находящихся в 32 креслах.

Ключевые слова: аудиолокализация, определение речевой активности, интеллектуальное пространство, автоматизация протоколирования мероприятий.

Ronzhin A.L. **Evaluation methods for sound source localization systems of speakers in a meeting room.**

Abstract. The employing of sound source localization methods allows to evaluate location and head orientation of a speaker in a room. At present such systems are popular at development of intelligent support systems for smart meeting rooms. In this paper, a set of the metrics for performance evaluation of sound source localization systems as well as their integration with video monitoring systems are analyzed. Accuracy estimation of speaker positions, located on 32 chairs was carried out in the developed smart meeting room.

Keywords: sound source localization, speech activity detection, smart space, automatic meeting summarization.

1. Введение. Применение «ненавязчивых» способов регистрации поведения пользователей особенно актуально в различного рода прототипах интеллектуального пространства [1], например, в умных домах, конференц-залах, помещениях с встроенными ассистивными технологиями для людей с ограниченными возможностями [2]. Анализ аудиоактивности пользователя в интеллектуальном пространстве позволяет оценить его положение, моменты речевой активности, фокус внимания и другие параметры, описывающие его текущее поведение.

Регистрация аудиосигнала может быть организована как с помощью микрофонов, установленных в помещении, так и закрепленных за самим пользователем. Использование последних или же проводных микрофонов отрицательно сказывается на естественности поведения диктора и вызывает определенные организационные сложности при установлении процесса коммуникации. В связи с чем при проектиро-

вании интеллектуальных помещений чаще всего применяются встроенные микрофоны, расположение которых выбирается таким образом, чтобы расстояние между ними и источниками полезного сигнала было минимальным.

В интеллектуальных залах совещаний обычно выделяется несколько зон, требующих регистрации аудиосигнала, за каждой из которых закрепляется свой набор сенсоров. При проектировании интеллектуального зала в университете южной Калифорнии аудиозаписывающее оборудование было сосредоточено рядом с конференц-столом [3]. В разработанном в проекте CHIL интеллектуальном зале кроме оборудованного конференц-стола так же осуществляется мониторинг зоны презентаций, где основной докладчик перемещается во время выступления [4]. В проекте DICIT основное внимание уделяется распознаванию голосовых команд пользователей системы интерактивного телевидения, которые могут располагаться в четырех заранее известных системе местах в помещении [5].

Методы определения положения источника звука в настоящее время активно используются в различных приложениях, в частности, в системах громкой голосовой связи, при поддержке распределённых совещаний, протоколировании дикторов, дистанционном распознавании речи [6, 7]. Оценивание положения диктора выполняется с помощью многоканальной обработки аудиосигналов, записанных несколькими микрофонами одновременно. В помещениях с высоким уровнем реверберации обычно используются три типа подходов SSL [8]: 1) управление диаграммой направленности массива микронов; 2) спектральный анализ с высоким разрешением; (3) оценивание разницы времени получения сигнала. Последняя группа методов относительно проста в реализации и чаще всего используется на практике [6].

В следующих разделах обсуждаются методики оценивания производительности систем аудиообработки для различных прикладных задач, в том числе определения речевой активности в многоканальном аудиопотоке, определения положения выступающего, а также направления его головы. В последнем разделе приведены экспериментальные результаты по оцениванию разработанной системы аудиолокализации в интеллектуальном зале совещаний.

2. Многоканальная оценка речевой активности. Для оценивания точности работы систем обработки аудиосигнала, а также определения наличия речи в аудиосигнале были предложены различного рода способы оценивания [9-14]. Точность сегментации границ речи в аудиопотоке чаще всего оценивают по ошибке первого и второго рода,

вычисляя число пропущенных (miss rate (MS)) и ложных сегментов речи (false alarm (FA)) соответственно. В работе [9] оценка точности сегментации речи в аудиосигналах производилась по корпусу NIST Rich Transcription на основе отношения длительности сегментов с ошибочным решением о наличии речи к длительности всех речевых сегментов в аудиосигнале.

Отметим, что оценка NIST направлена на тестирование методов по базам данных, в которых заранее определены речевые и неречевые сегменты. Например, система, которая допускает 5% ошибок в определении сегментов речи и 5% ошибок при определении неречевых сегментов в аудиосигнале, будет иметь совершенно другие показатели для тестовых баз данных с отличающейся пропорцией речевых и неречевых сегментов. В случае 90% речи и 10% шума оценка NIST составит 5.6%, а при равных пропорциях – оценка будет равна 10%. Учитывая данный факт, были введены три дополнительных оценки [10]: уровень несоответствия (MR); уровень ошибок определения речи SDER и уровень ошибок определения неречевых сегментов NDER, определяемых следующим образом.

Уровень несоответствия MR определяется как отношение длительности сегментов, классифицированных неверно, к длительности всех сегментов. Уровень ошибок определения речи SDER равен отношению длительности сегментов с ошибочным определением речи к длительности всех речевых сегментов. Наконец, уровень ошибок определения неречевых сегментов NDER определяется как отношение длительности ошибочно определенных неречевых сегментов к длительности всех неречевых сегментов.

На основе последних двух оценок в работе [9] добавлена еще одна дополнительная оценка: средний уровень ошибок классификации сегментов в аудиосигнале $ADER = (SDER + NDER)/2$. При оптимизации параметров метода определения речевой активности в этой же работе предложено использовать следующее контрольное условие: $|SDER - NDER|/(SDER + NDER) \le 0.1$.

При тестировании многоканальной системы определения речевой активности оценки суммируются по всем каналам M [11], например, при подсчете ошибки первого и второго рода:

$$MS = \frac{\sum_{k=1}^{M} T_k^{(MS)}}{\sum_{k=1}^{M} T_k^{(S)} + \sum_{k=1}^{M} T_k^{(MS)}}, \quad FA = \frac{\sum_{k=1}^{M} T_k^{(FA)}}{\sum_{k=1}^{M} T_k^{(S)} + \sum_{k=1}^{M} T_k^{(FA)}},$$

где $T_k^{(S)}$ - длительность сегментов речи в канале k, определенных системой как речь; $T_k^{(MS)}$ - длительность сегментов речи, пропущенных системой; $T_k^{(FA)}$ - длительность неречевых сегментов, определенных системой как речь. На рисунке 1 показан пример оценивания автоматической сегментации участка речи, выделенного предварительно вручную.

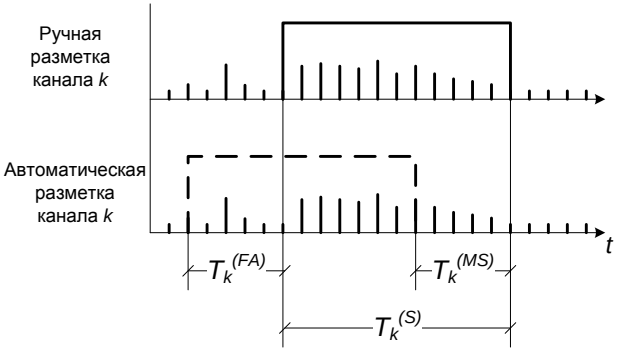

Рис. 1. Пример расчета ошибок автоматической сегментации речи в канале *k*.

При настройке параметров алгоритма определения границ речи приходится выбирать некоторый компромисс между числом пропущенных и ложных сегментов. Для этой цели служит общепринятая оценка *DCF* (detection cost function), равная минимальному значению линейной комбинации оценок P_{MS} и P_{FA}:

$$DCF = \min(C_{MS} * P_{MS}(\theta) * P_{target} + C_{FA} * P_{FA}(\theta) * (1 \quad P_{target})),$$

где θ - номер эксперимента, C_{MS}, C_{FA} – веса для ошибок первого и второго рода соответственно, P_{target} – априорно задаваемая вероятность наличия речи в канале. Поскольку во время совещания паузы возникают достаточно редко, то при ложном переключении канала речь текущего диктора может быть отрезана, поэтому в нашем случае одинаково важно как не пропустить сегменты речи, так и не захватить ложные сегменты. Следовательно, на начальном этапе веса будут идентичны $C_{miss} = C_{false} = 1$. Значение P_{target} устанавливается заранее, если известно какой тип сегментов (речь/тишина) будет преобладать, поэтому на предварительном этапе эта оценка пока также не будет учитываться $P_{target} = 0.5$.

В ходе оптимизации подбираются такие значения параметров алгоритма, при которых оценка *DCF* будет иметь минимальную величину. При этом оценки P_{MS} и P_{FA} будут иметь разное значение. Другой известной мерой оценки алгоритмов является *EER* (Equal Error Rate), которая соответствует эксперименту с номером θ_{EER}, в котором оценки P_{MS} и P_{FA} имеют наиболее близкие значения:

$$\theta_{EER} = \arg\min_{\theta} \left| P_{miss}(\theta) - P_{false}(\theta) \right|,$$

В ходе оптимизации значения параметров изменяются с некоторым шагом, поэтому достичь равенства ошибок бывает затруднительно, поэтому при вычислении оценки *EER* используется усредненное значение оценок:

$$ERR(\theta) = \frac{P_{miss}(\theta) + P_{false}(\theta)}{2}$$

3. Оценивание систем аудиолокализации. Далее рассмотрим способы оценки точности работы алгоритмов определения положения источника звука. Довольно распространенной оценкой является ошибка локализации $e(t)$, рассчитываемая следующим образом: $e(t) = d(p(t), g(t))$, где d евклидово расстояние между координатами $p(t)$, определенными системой аудиолокализации в момент времени t, и действительным положением источника звука $g(t)$.

Ошибка локализации оценивается только для того момента времени, в котором определен источник звука в анализируемом пространстве и получен соответствующий набор координат. Уровень ошибок локализации принято разделять на два класса: (1) аномальные или грубые ошибки, если $e(t) > D$; (2) не аномальные или незначительные ошибки, если $e(t) \le D$, где $D-$ некоторая пороговая величина, значение которой определяется эмпирически в ходе тестирования системы локализации в зоне эксплуатации. Для аудиолокализации выступающего в интеллектуальном зале и определения других источников звука в нем приемлемой ошибкой локализации считается отклонение до 0.5 метра [4, 10, 14].

Такая классификация была введена для того, чтобы разделить ошибки локализации в зашумленной обстановке, когда система грубо, но правильно определяет положение источника, и ошибки, возникающие вследствие определения ложных несуществующих источников [15]. Кроме того, аномальные ошибки, возникающие в процессе

локализации в реальных условиях, могут получаться в случае некорректной разметки (т.е. реально существующий источник был определен с неправильными координатами) или если в процессе выступления основного диктора был зафиксирован еще один источник звука. Возникновение аномальных ошибок значительно понижает общую точность работы системы локализации. Поэтому кроме общей оценки ошибки локализации, необходим отдельный анализ незначительных ошибок. В частности в работе [10] предложен набор метрик для оценки точности локализации, включающий анализ как аномальных, так и незначительных ошибок:

1. Ошибка RMSE: среднеквадратическое отклонение, рассчитанное по всем результатам, полученным системой аудиолокализации. В некоторых случаях общая ошибка RMSE может быть значительно снижена за счет нескольких грубых ошибок локализации.
2. Ошибка Fine RMSE: среднеквадратическое отклонение, рассчитанное по всем выходным значениям локализации, классифицированным как незначительные ошибки.
3. Уровень точности локализации (P_{cor}): отношение незначительных ошибок ко всему числу произведенных оценок локализации. Данная метрика позволяет оценить достоверность метода локализации. В процентном соотношении уровень точности локализации вычисляется по формуле [14]: $P_{cor} = \frac{N_{fe}}{N_T} * 100$, где N_{fe} – число незначительных ошибок, N_T – число всех оценок локализации.
4. Смещение: средняя ошибка локализации, вычисленная отдельно для каждой координатной оси. Данная метрика позволяет определить на основе статистического анализа наличие стабильных отклонений по каждой оси и настроить параметры метода локализации с целью минимизации ошибки.

Для повышения точности слежения за объектами в интеллектуальном пространстве методы аудиолокализации и видеомониторинга объединяются в одной системе. Так как оценки о положении объектов принимаются с различной частотой, то необходим некоторый алгоритм синхронизации аудиовизуальной обработки сигналов. Поэтому в работе [10] был введен дополнительный набор вторичных метрик, описывающих возможности определения наличия речи в аудиосигнале на базе оценок алгоритма аудиолокализации:

1. Уровень вывода значений: число оцененных положений источника звука за одну секунду.
2. Уровень ложных срабатываний: процентное соотношение числа выходных значений локализации в случае тишины к общему числу сегментов, помеченных как тишина в аудиосигнале.
3. Уровень пропусков сигнала: рассчитывается для моментов времени, когда модуль аудиолокализации не выдает никаких оценок, даже если кто-то выступает. Данная метрика равна отношению числа пропущенных аудиосегментов к общему числу сегментов с выступлениями дикторов.

Введение уровней пропуска и ложного срабатывания обусловлено тем фактом, что системы локализации дикторов включают некоторый неявный процесс определения акустических событий [10]. Существующие алгоритмы аудиолокализации предполагают наличие дикторов в анализируемом помещении и выполняют оценку положения источника звука с учетом координат дикторов. При проектировании приложений для реальных условий эксплуатации необходимо учитывать как высокую точность локализации, так и сбалансированный уровень пропусков и ложных срабатываний.

4. Определение ориентации головы диктора. Также в области обработки аудиосигнала применяются метрики, разработанные для оценивания направления (ориентации) головы участников мероприятий. Подобное тестирование проводилось в рамках проекта CLEAR по базе данных, которая была накоплена в ходе проекта CHIL – Computers In the Human Interaction Loop в процессе проведения семинаров [16]. Для оценки системы определения направления головы были введены три основные метрики [17]:

1. Средняя ошибка угла в горизонтальной плоскости (PMAE), измеряемая в градусах, позволяет оценить погрешность определения направления (ориентации) головы участника мероприятия.
2. Уровень правильной классификации ориентации головы в горизонтальной плоскости (PCC), измеряемый в процентах, позволяет оценить способность системы корректно классифицировать положение головы относительно восьми классов, каждый из которых покрывает сектор равный 45 градусам.
3. Уровень правильной классификация ориентации головы в горизонтальной плоскости внутри сектора (PCCR), измеряемый в процентах, позволяет оценить способность системы кор-

ректно классифицировать положение головы с допустимой ошибкой ±1 сектор.

Последние оценки используются для решения задачи не только определения направления головы, но и выявления фокуса внимания участников мероприятия в интеллектуальном зале, что может применяться как для управления презентационным оборудованием, так и для оценки поведения участников и их вклада в решение проблем, обсуждаемых на мероприятии.

5. Эксперименты. Оценка работы системы аудиолокализации производилась в интеллектуальном зале СПИИРАН с размерами 8.85x7.15x4.80м и со временем реверберации аудиосигнала 0.62 с [18]. Многоканальная аудиоплата Presonus FirePod была использована для записи сигналов со всех микрофонов, обслуживающих конференц-стол, зону выступлений, а также ряды из 32 кресел.

В первую очередь была проверена система многоканальной записи речи участников, сидящих за конференц-столом. Для экспериментальной проверки алгоритма определения речевой активности была подготовлена тестовая база данных, содержащая пятиканальную аудиозапись с частотой дискретизации отсчетов 16кГц. Длина сегмента речи равнялась 1600 отсчетам. Длина скользящего окна составляла 10 сегментов. Окно сдвигалось с шагом равным одному сегменту. Общая длительность речевого сигнала в базе данных составила 28 минут. В ходе эксперимента пять участников последовательно читали предложения различной длины из одного текста. Распечатанные листы бумаги с текстом лежали на столе перед каждым участником. Таким образом, в данном эксперименте была создана несколько искусственная ситуация: участники не перебивали друг друга, а читали предложения последовательно; между микрофоном и участником не возникали помехи (руки, бумаги, другие предметы), лицо диктора было направлено преимущественно в сторону микрофона на протяжении всей записи.

В ходе прослушивания всех записей вручную были выставлены границы фраз в каждом канале с точностью до одного аудиосегмента. Полученная разметка использовалась в качестве эталонной, по которой оценивалось качество автоматической сегментации. В данном эксперименте с помощью предложенного алгоритма границы фраз участников были определены с точностью *EER=9.16%* [19].

В следующем эксперименте основное внимание уделялось определению активных участников в зоне кресел, содержащей места для 32 сидящих участников. Описание разработанного метода определения положения источника звука, основанного на подходе GCC-PHAT, и

конфигурация четырех массивов микрофонов, расположенных в зале, приведены в [20, 21].

Аудиосигналы записывались в условиях реверберации и фоновых шумов в зале. Отношение сигнал/шум (SNR) изменялось при увеличении расстояния между креслом, в котором сидит участник, и массивом микрофонов. При измерении SNR один и тот же полезный сигнал проигрывался с 32 установленных в зале кресел. Результаты представлены в таблице 2.

Таблица 2. **Оценка SNR для 32 кресел в зале**.

Массив микрофонов	Расстояние от массива до центра кресла			
	Минимальное расстояние, м	SNR, дБ	Максимальное расстояние, м	SNR, дБ
MA_1	3.9	18	8.7	16
MA_2	2.9	19	9.0	16
MA_3	1.9	20	6.8	16
MA_4	2.9	19	9.0	16

На предварительном этапе оценки координаты центров кресел были использованы как координаты положения выступающих участников. Для определения реальных координат 32 потенциальных позиций участников один тестирующий последовательно произнес один и тот же набор фраз из каждого кресла. При этом точность определения положения источника звука (отношение оценок со среднеквадратическим отклонением менее 0.5 м к общему числу оценок) составила всего 53%. Среднее отклонение вычисленных координат выступающего от координат центра кресла, в котором он сидел, составило около 0.45 м. Экспериментальная оценка 32-х потенциальных позиций участников была использована в следующем эксперименте. На основе полученных оценок была выполнена корректировка координат 32-х положений участников, сидящих в креслах.

В серии следующих экспериментов принимало участие от одного до трёх человек. Каждый участник эксперимента выполнял последовательность действий по заранее разработанному сценарию: 1) сесть в кресло, расположенное в комнате; 2) дождаться визуального подтверждения на «Умной доске» о регистрации участника в кресле; 3) проговорить фразу, состоящую из цифр от одного до десяти; 4) затем пересесть в другое свободное кресло. Длительность каждой произнесенной фразы равнялась приблизительно 6 секундам.

В ходе этой серии экспериментов система аудиолокализации использовалась для определения положения диктора 2875 раз. Точность определения положения сидящих выступающих составила 97%. В таблице 3 представлено распределение значений среднеквадратического отклонения RMSE для 32-х позиций участников, расположенных в четырёх рядах по восемь кресел в каждом.

Таблица 3. **Значения оценки Fine RMSE для участников, выступавших с 32 кресел в зале**.

	Колонки кресел							
Ряды кресел	0.19	0.12	0.12	0.15	0.12	**0.09**	**0.09**	0.16
	0.23	0.17	**0.05**	0.18	**0.06**	**0.04**	**0.06**	**0.09**
	0.11	0.11	**0.09**	0.14	**0.06**	0.11	0.13	0.11
	0.14	0.18	0.11	**0.08**	**0.05**	0.17	0.14	0.15

Результаты экспериментов показали, что рассчитанные значения оценки Fine RMSE для всех 32-х кресел меньше, чем 0.23м. Позиции участников, в которых ошибка в определении составила меньше, чем 10см, выделены жирным шрифтом в таблице 9. Из полученных экспериментальных данных можно сделать вывод, что точность определения положения источника звука зависит от трех факторов: 1) отклонение источника звука от нормали массива микрофонов; 2) расстояние от источника звука до массива микрофонов; 3) число массивов, использованных при определении положения источника звука. Небольшое значение ошибки в определении положения источника звука в креслах, указанных в правой верхней части таблицы, связано с первым и вторым факторами, а в креслах, указанных в средней части таблицы, – с первым и третьим факторами. В последующих экспериментах будет оцениваться не только положение выступающего, но и направление его головы с помощью методик, описанных выше.

5. Заключение. Описанные в статье способы количественного оценивания систем автоматической обработки аудиосигналов могут быть применены для локализации источников звука, многоканальной оценки речевой активности и комплексной оценки поведения пользователей на основе анализа аудио, видео, радиочастотных и других данных, используемых для регистрации текущей ситуации в зоне мониторинга. Разработанные методы аудиолокализации и определения речевой активности были протестированы по рассмотренным методикам. В результате точность определения положения дикторов для 32-х кресел,

расположенных в интеллектуальном зале составила 97%, а границы фраз участников были определены с точностью 9.16% .

Литература

1. *Ан.Л. Ронжин, А.А. Карпов.* Проектирование интерактивных приложений с многомодальным интерфейсом // Доклады ТУСУРа. –2010. –№ 1 (21), часть 1. – С. 124-127.
2. *А.А. Карпов, И.А. Кагиров.* Формализация лексикона системы компьютерного синтеза языка жестов // Труды СПИИРАН. СПб.: Наука, Вып. 16, 2011, С. 123-140.
3. *Rozgic, V., Busso, C., Georgiou, P.G., and Narayanan, S.S.* Multimodal meeting monitoring: Improvements on speaker tracking and segmentation through a modified mixture particle filter // IEEE International Workshop on Multimedia Signal Processing (MMSP), 2007, pp.60-65.
4. *Omologo, M., Svaizer, P.G, Brutti, A., Cristoforetti, L.* Speaker Localization in CHIL Lectures: Evaluation Criteria and Results // Machine Learning for Multimodal Interaction. Berlin: Springer, 2006, pp.476–487.
5. *Brutti, A., Omologo, M. and Svaizer, P.* Comparison between different sound source localization techniques based on a real data collection // Hands-Free Speech Communication and Microphone Arrays (HSCMA), Trento, Italy, May 2008.
6. *Benesty, J., Sondhi, M., and Huang, Y.* Handbook of Speech Processing // Springer, 2008.
7. *Zhang, C., Yin, P., Rui, Y., Cutler, R., Viola, P., Sun, X., Pinto, N., and Zhang, Z.* Boosting-Based Multimodal Speaker Detection for Distributed Meeting Videos // IEEE Transactions on Multimedia, Vol.10, No.8, 2008, pp.1541-1552.
8. *DiBiase, J., Silverman, H., and Brandstein, M.* Robust Localization in Reverberant Rooms // ser. Microphone arrays-Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag, ch. 8, 2001, pp.157–180.
9. *Macho D., Nadeu C., Temko A.* Robust Speech Activity Detection in Interactive Smart-Room Environments // Springer Berlin/Heidelberg, S. Renals, S. Bengio, J. Fiscus (Eds.): MLMI 2006, LNCS 4299, 2006, pp. 236 – 247.
10. *Omologo M., Svaizer P., Brutti A., Cristoforetti L.* Speaker localization in CHIL lectures: Evaluation criteria and results. Springer Berlin/Heidelberg, Steve Renals and Samy Bengio (Eds.): MLMI 2005: Revised and selected papers, Edinburgh, UK, July 11-13 2005, pp. 476-487.
11. *Laskowski K., Schultz T.* Simultaneous multispeaker segmentation for automatic meeting recognition. In Proc. of EUSIPCO, Poznan, Poland, September 2007, pp. 1294–1298.
12. *Мещеряков Р.В.* Система оценки качества передаваемой речи // Доклады ТУСУР, 2010. - N2(22) - С.324-329.
13. *Косарев, Ю.А., Ли, И.В., Ронжин, А.Л., Скиданов, Е.А., Savage J.* Обзор методов понимания речи и текста, Труды СПИИРАН / Под ред. Р.М. Юсупова вып. 1 т. 2 – СПб.: «Анатолия», 2002, С. 157-195.
14. *Nishiura T., Yamada T., Nakamura S., Shikano K.* Localization of multiple sound sources based on a CSP analysis with a microphone array. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 2, Instanbul, Turkey, June 5-9 2000, pp. 1053-1056.
15. *Champagne B., Bedard S., Stephenne A.* Performance of time-delay estimation in the presence of room reverberation // IEEE Transactions on Speech and Audio Processing, March 1996, Vol. 4, Issue 2, pp. 148-152.

16. *Stiefelhagen R., Garofolo J.* (Eds.) Multimodal Technologies for Perception of Humans. First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006 // Lecture Notes in Computer Science, Vol. 4122, Springer-Verlag, 2007.
17. *Abad A., Segura C., Nadeu C., Hernando J.* Audio-based approaches to head orientation estimation in a smart-room // In Proc. of Interspeech'2007, August 27-31, Antwerp, Belgium, pp. 590-593.
18. Ганбат, Д., Ронжин, А.Л., Найдандорж, Р., Будков, В.Ю., Прищепа, М.В. Разработка веб-системы для предоставления обучающих сервисов удаленным мобильным пользователям // Труды СПИИРАН. Вып. 13, СПб.: Наука, 2010, С. 21-34.
19. *Ронжин Ан.Л., Будков В.Ю.* Технологии поддержки гибридных е-совещаний на основе методов аудиовизуальной обработки // Вестник компьютерных и информационных технологий, № 4, 2011.
20. *Ронжин Ан.Л., Будков В.Ю., Ронжин Ал.Л.* Технологии формирования аудиовизуального интерфейса системы телеконференций // Автоматизация и современные технологии. № 5. 2011, С. 20-26.
21. *Ронжин Ал.Л., Ронжин Ан.Л.* Система аудиовизуального мониторинга участников совещания в интеллектуальном зале // Доклады ТУСУРа, № 1 (22), часть 1, 2011, С. 147-151.

Ронжин Александр Леонидович — младший научный сотрудник лаборатории речевых и многомодальных интерфейсов Учреждения Российской академии наук Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: технологии интеллектуального пространства, аудиолокализации, техническое зрение. Число научных публикаций — 17. ronzhinal@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Ronzhin Alexander Leonidovich — junior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: smart space, sound source localization, computer vision. The number of publications — 17. ronzhinal@iias.spb.su; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Поддержка исследований. Данное исследование поддержано Советом по грантам Президента РФ (проект № MD-501.2011.8); Министерством образования и науки РФ в рамках ФЦП «Научные и научно-педагогические кадры инновационной России на 2010-2012 гг.» (госконтракт № 14.740.11.0357).

Рекомендовано лабораторией автоматизации научных исследований, заведующий лабораторией Александров В.В., д-р техн. наук, проф.
Статья поступила в редакцию 04.07.2011.

РЕФЕРАТ

Ронжин А.Л. **Способы оценивания систем аудиолокализации выступающих в зале совещаний.**

Методы определения положения источника звука в настоящее время активно используются в различных приложениях, в частности, в системах громкой голосовой связи, поддержке распределённых совещаний, протоколировании дикторов, дистанционном распознавании речи. Для оценивания точности работы систем обработки аудиосигнала и локализации источника звука, а также определения наличия речи в аудиосигнале были предложены различного рода метрики. Точность сегментации границ речи в аудиопотоке чаще всего оценивают по ошибке первого и второго рода, вычисляя число пропущенных и ложных сегментов речи соответственно. Наиболее распространенная оценка систем аудиолокализации основывается на Евклидовом расстоянии между координатами, определенными системой аудиолокализации в момент времени, и действительным положением источника звука. Уровень ошибок локализации принято разделять на два класса: грубые и незначительные ошибки. Для аудиолокализации выступающего в интеллектуальном зале приемлемой ошибкой локализации считается отклонение до 0.5 метра.

Разработанные методы аудиолокализации и определения речевой активности были протестированы по рассмотренным методикам. В результате точность определения положения дикторов для 32 кресел, расположенных в интеллектуальном зале составила 97%, а границы фраз участников были определены с точностью 9.16%.

SUMMARY

Ronzhin A.L. **Evaluation Techniques of Speaker Localization Systems in a Meeting Room.**

At present sound source localization methods are actively applied in different spheres, in particular in hand-free voice communication systems, supporting of distributed meetings, speaker diarization, and distant speech recognition. Different types of metrics for evaluation of sound source localization and audio signal processing were proposed. An accuracy of speech boundary segmentation in audio stream is often estimated by first and second types of errors with calculation of value of missed and false segments, respectively. The most popular estimation of sound source localization systems is based on Euclidian distance between coordinates determined by sound source localization system and real sound source position at the same time moment. Localization error rate may be divided into two classes: rough errors and subtle errors. For sound source localization of a speaker in a smart room the acceptable error rate is less than 0.5m.

The developed sound source localization and speech activity detection methods were tested by described techniques. The experimental results showed that the sound source localization accuracy of speaker sitting in 32 chairs, which were located in the smart room, was 97% and the participants' phrase boundaries were detected with accuracy of 9.16%.