УДК 004.522

АВТОМАТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ЧЕЛОВЕКА ПО ГОЛОСУ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА НА ОСНОВЕ МОДЕЛИ ГАУССОВЫХ СМЕСЕЙ

Д. И. Лукьянов, доцент кафедры РУС РГРТУ, к.т.н.; d.i.lukyanov@outlook.com
А. С. Михайлова, магистрант РГРТУ; alenkamihailova@yandex.ru

Исследованы методы автоматической идентификации человека по голосу. ***Целью работы*** *является обоснование алгоритма идентификации человека, на основе речевого сигнала, обеспечивающего высокую вероятность идентификации. Проведены экспериментальные исследования алгоритма на основе метода среднеквадратического отклонения, минимального расстояния, гауссовых смесей. Идентификация человека с использованием алгоритма на основе метода среднеквадратического отклонения не является полностью инвариантной к исходным речевым сигналам (РС), так как является эффективной для коротких речевых сообщений, содержащих одно слово. Идентификация человека с использованием алгоритма на основе метода минимального расстояния позволяет проводить идентификацию на уровне мужчина/женщина. Одним из наиболее эффективных является метод на основе моделей гауссовых смесей. Оптимальное число кластеров для распознавания равно 5. Реализация алгоритма на основе модели гауссовых смесей позволяет проводить корректную идентификацию с вероятностью до 75 %.*

Ключевые слова: *распознавание речевой информации, системы распознавания, идентификация человека по голосу, метод среднеквадратического отклонения, метод минимального расстояния, метод мел-кепстрального преобразования спектра РС, кепстр, модель гауссовых смесей, оптимальное число кластеров.*

DOI: 10.21667/1995-4565-2017-61-3-19-24

Введение

Речь является формой общения людей с помощью фонетических конструкций, которая сложилась в ходе исторического развития человека. С её помощью человек познает окружающий мир, передает свои знания и умения другим людям. Устная составляющая речи проявляется в виде высказываний в звуковой форме, которые возможны благодаря голосовому аппарату человека.

У каждого человека свои индивидуальные голосовые характеристики [1]. Во время общения люди способны на подсознательном уровне различать голоса других людей, однако вычислительная техника не может решить данную задачу с помощью простых методов.

Обработка речи была одной из самых захватывающих областей исследования. Сигналы обычно обрабатываются в цифровом виде, поэтому обработку речи можно рассматривать как частный случай цифровой обработки сигнала.

Более 40 лет назад была поставлена задача определения личности человека по его индивидуальным голосовым характеристикам. Стоит отметить, что исследования в этой области продолжаются и в настоящее время [1] и в последние годы наблюдается тенденция к повышению качественных характеристик систем распознавания речевой информации, однако основная проблема автоматической идентификации диктора по голосу еще не решена. Поэтому сейчас актуален поиск новых решений в данной области, но не следует забывать про существующие алгоритмы и их оптимизацию.

Определение голосовых характеристик человека находит применение во многих сферах человеческой деятельности, таких как: криминалистика, судебная экспертиза, обеспечение безопасности, банковские технологии, электронная коммерция, телематика.

Современные методы идентификации речи в реальном времени выдвигают высокие требования к вычислительным ресурсам, но часто их объем ограничен. Так, в мобильных устройствах невозможно применять многие из существующих алгоритмов, что заставляет находить более эффективные методы.

В решении задачи идентификации диктора заинтересованы государственные учреждения, бизнес-структуры и другие категории различных

пользователей информационных услуг. В настоящее время ведутся интенсивные научные исследования идентификации человека по голосу, однако реальное применение таких систем на практике ограничено вычислительными ресурсами и сложностью различных алгоритмов, что подтверждается регулярными годовыми отчетами логистической компании Gartner Group [2]. По данным компании, лишь малое количество пользователей (до 1% от общего числа) удовлетворено эффективностью систем распознавания голосовых характеристик диктора.

В развитие направления идентификации человека по голосу внесли вклад такие ученые, как Л. Рабинер, заложивший научные основы распознавания речи статистическими методами, F. Wilpon, P. Lee, A. Higgins, Т. К. Винцюк, А. А. Карпов, А. Л. Ронжин. Профессором В. В. Савченко была создана информационная теория восприятия речи (ИТВР) [2], фундаментом которой служит критерий минимума информационных рассогласований (МИР) [3] и кластерная модель речевых единиц.

Теоретическая часть

Существует три основных подхода [4] к решению задачи идентификация диктора (ИД) по голосу.

Первый подход основывается на том, что из речевого сигнала (РС) выделяют вектор характерных свойств, который учитывает особенности восприятия звука человеком [5]. Этот подход основан на анализе несущих частот сигнала и его выравнивании по уровню (громкости). Среди наиболее применимых технологий, использующих такой подход, можно выделить метод оценки кепстральных коэффициентов частоты основного тона (ОТ) и метод оценки коэффициентов линейного предсказания. В дальнейшем может быть смоделировано сопоставление с шаблоном, характерным для человеческого восприятия, обеспечивающее большую устойчивость к воздействию шумов [6].

Второй подход основан на выделении и анализе РС из смеси зашумленного сигнала на входе системы распознавания. Основной причиной нестабильной работы систем ИД являются различия поступающих в систему зашумленных РС от эталонов, полученных в результате обучения незашумленными РС. Главной задачей этого подхода является уменьшение различия. Предполагается, что шум во входном сигнале аддитивный и стационарный. Полученные таким образом оценки среднего значения усредненного шума вычитаются из кепстра, вычисленного по зашумленным данным. Ключевым недостатком такого подхода является необходимость точной оценки характеристик шума, которую практически провести довольно тяжело.

Третий подход основывается на использовании многомерных пространств [7]. Ключевая идея этого подхода заключается в определении линейного отображения, которое сводит к минимуму значение функции стоимости. Примером этого подхода может служить как основной компонентный анализ, так и независимый компонентный анализ [8].

Основная идея метода мел-кепстрального преобразования спектра РС заключается в том, что на коротких интервалах времени (до 20 мс) вычисляется значение мгновенного спектра мощности, к которому затем применяется инверсное преобразование Фурье от логарифма этого спектра (кепстр), далее находятся коэффициенты кепстра [9]. В системах голосовой идентификации обычно используют от 10 до 30 коэффициентов [10].

После решения вопроса выбора характеристик необходимо выбрать наиболее эффективный метод, позволяющий произвести идентификацию человека по голосу с высокой вероятностью.

Основные методы автоматической идентификации человека по голосу [11]: динамическая трансформация временной шкалы (Dynamic Time Warping), скрытая марковская модель (Hidden Markov Modeling), векторное квантование (Vector Quantization), метод опорных векторов, модель гауссовых смесей (ГС).

Динамическая трансформация временной шкалы представляет собой алгоритм для измерения подобия между двумя временными последовательностями. Этот метод вычисляет оптимальное соответствие между двумя последовательностями.

Скрытая марковская модель (СММ) – стохастическая модель пространства состояний со случайными переходами между состояниями, где вероятность перехода зависит только от текущего состояния. СММ может использоваться для решения задач классификации скрытых параметров на основе наблюдаемых. Каждому скрытому состоянию с определенной вероятностью соответствует наблюдаемое состояние. Текущее значение вероятности зависит только от конечного числа предыдущих состояний. Закон смены состояния не меняется во времени [12].

Векторное квантование – разбиение пространства возможных значений векторной величины на конечное число областей. Этот метод обработки сигнала позволяет моделировать плотность вероятности, функции распределения векторов. Первоначально этот метод использо-

вался для сжатия данных. Он работает путем разделения большого набора векторов на группы, имеющие приблизительно одинаковые значения. В методе векторного квантования выборка из обучающих векторов преобразуется в фиксированное множество кодовых векторов. Одним из распространённых методов формирования подобного множества, именуемого также кодовой книгой, является алгоритм К – средних [13].

В основе метода СКО лежит перевод исходных данных в частотную область. В начале алгоритма проводится оконная обработка исходных данных, затем с помощью быстрого преобразования Фурье переводится в частотную область, где вычисляется спектральная плотность мощности полученного спектра. Затем СКО сравнивается с априорной информацией о параметрах речи диктора. Если полученное СКО меньше некоторого порога, то происходит подтверждение идентификации.

В методе минимального расстояния в качестве суммарной оценки можно определить такую, сумма расстояний до которой от оценок всех дикторов будет минимальной.

На данный момент существует большое количество методов, позволяющих решать задачу идентификации человека по голосу. Тем не менее, одним из наиболее эффективных методов является модель ГС, который взят за основу в рамках данной работы.

Гауссова модель представляет собой параметрическую функцию плотности вероятности. Данная модель являются удачной вариацией стохастической модели для построения систем распознавания [14]. Они удобны для моделирования характеристик голоса диктора, канала звукозаписи, окружающей среды. Каждая из компонент модели отражает некоторые общие, но индивидуальные для каждого диктора особенности голоса. Именно поэтому данный подход можно успешно применять для решения задачи идентификации диктора.

Обычные системы распознавания речи используют модели ГС на основе СММ. Модель ГС описывается выражением (1). Она представлена в виде взвешенной суммы M компонент [15]:

$$P(\overline{x}|\lambda)=\sum_{i=1}^{M} p_i b_i(\overline{x}), \qquad (1)$$

где $\overline{x}$ – D-мерный вектор случайных величин, λ – модель диктора, $p_i, 1\le i\le M$ – веса компонентов модели, $b_i, 1\le i\le M$ – функции плотности распределения компонент модели:

$$b_i(\overline{x})=\frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}}\times$$

$$\times\exp\left\{-\frac{1}{2}(\overline{x}-\overline{\mu_i})^T\sum_i^{-1}(\overline{x}-\overline{\mu_i})\right\}, \qquad (2)$$

где $\overline{\mu_i}$ – вектор математического ожидания и $\sum_i$ – ковариационная матрица. При этом веса смеси удовлетворяют условию:

$$\sum_{i=1}^{M} p_i = 1.$$

Векторы математического ожидания, ковариационные матрицы и веса смесей полностью определяют модель гауссовой смеси для каждого компонента модели:

$$\lambda=\left\{p_i,\overline{\mu_i},\sum_i\right\}, i=1,...M. \qquad (3)$$

При использовании данного метода каждого из дикторов можно представить в качестве модели гауссовых смесей λ.

Для того чтобы построить модель диктора, необходимо точно оценить её параметры, которые наиболее точно соответствуют распределению векторов признаков обучающего высказывания. Существует определенный ряд методов для оценки параметров модели. Одним из наиболее популярных и хорошо себя зарекомендовавшим является метод оценки максимального правдоподобия [16]. Цель оценки заключается в определении параметров модели, которые максимально повысят вероятность правдоподобия этой модели при заданных данных для обучения.

Для детерминированной последовательности векторов $X=\left\{\overline{x_1},...,\overline{x_T}\right\}$ правдоподобие модели ГС может быть записано в виде:

$$P(X|\lambda)=\prod_{t=1}^{T} P(\overline{x_t}|\lambda). \qquad (4)$$

Выражение (4) представляет нелинейную функцию от набора параметров λ. Используя эту функцию невозможно непосредственное вычисление правдоподобия ГС [17].

Пусть $S=\left\{S_1,S_2,...,S_N\right\}$ – группа дикторов, которые представлены набором моделей ГС $\lambda_1,\lambda_2,...,\lambda_N$.

Необходимо найти такую модель, которая при идентификации максимизирует значение апостериорной вероятности для заданного образца:

$$S=\arg\max_{1\le k\le N} P(\lambda_k|X)=\arg\max_{1\le k\le N} P(X|\lambda_k). \qquad (5)$$

Логарифмируя полученное выражение и учитывая независимость между наблюдениями, получаем систему идентификации дикторов:

$$S=\arg\max_{1\le k\le N}\sum_{t=1}^{T}\log P(\overline{x_t}|\lambda_k) \qquad (6)$$

Модели ГС представляют собой эффективный алгоритм, который позволяет проводить

идентификацию с высокой точностью распознавания [18]. Однако возникает ряд проблем, связанных с выбором числа компонентов модели и инициализацией её начальных параметров.

На рисунке 1 отображена схема работы системы идентификации диктора в режиме обучения.

Рисунок 1 – Общая схема работы системы в режиме обучения

Экспериментальные исследования

В эксперименте принимали участие 9 дикторов, каждый из которых произносил по 9 фраз. Общее число реализаций 81. Записи производились в одно и то же время дикторами, находящимися в спокойном состоянии, что позволяет свести к минимуму ошибки, возникающие из-за изменения голоса диктора под действием внешних факторов [19]. В качестве тестовых фраз выступали акустически взвешенные записи, рекомендованные ГОСТ Р 50840-95. Частота дискретизации составляла 8 кГц, разрядность квантования – 16 бит.

В работе проведены экспериментальные исследования метода среднеквадратического отклонения (СКО), метода минимального расстояния и модели ГС.

Идентификация человека с использованием алгоритма на основе метода СКО не является полностью инвариантной к исходным РС, так как является эффективным для коротких речевых сообщений, содержащих одно слово. ИД с использованием алгоритма на основе метода минимального расстояния позволяет проводить идентификацию на уровне мужчина/женщина.

Проведены экспериментальные исследования эффективности идентификации человека с использованием алгоритма на основе метода ГС и по определению количества элементов, входящих в состав модели. На рисунке 2 отображена схема работы системы идентификации диктора в режиме идентификации.

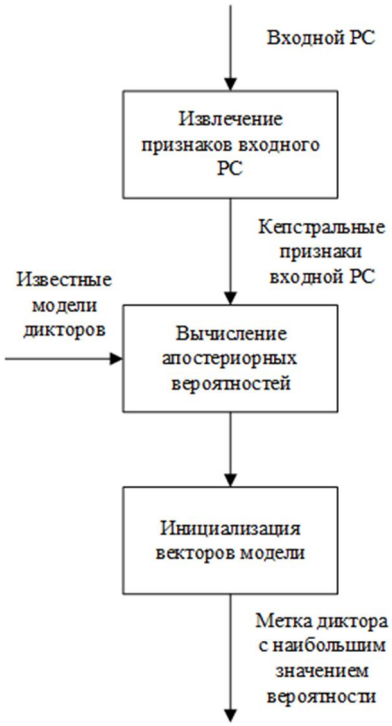

Рисунок 2 – Общая схема работы системы в режиме идентификации

Экспериментальный набор голосов дикторов состоял из 9 элементов. Продолжительность обучающего и тестового РС составляла 3…5 с. В качестве тестового высказывания выступала 1 фраза, остальные использовались для обучения. С целью определения эффективных параметров системы необходимо определить наиболее значимую полосу частот, в которой сосредоточена основная информация, требуемая для идентификации. Для этого записи пропускают через гребенку фильтров с переменной частотой среза. На выходе получают фильтрованный сигнал, который оценивался шестью аудиторами. В результате было установлено, что основная полоса частот, необходимая для идентификации диктора, сосредоточена в диапазоне 500…1500 Гц.

С целью определения числа компонент, входящих в состав модели, рассматривались модели

с числом компонент от 1 до 10. На рисунке 3 изображена зависимость вероятности правильной идентификации от числа компонент модели ГС. Погрешность результатов составила 3 %.

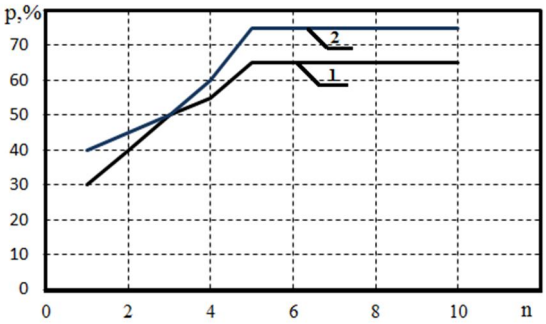

Рисунок 3 – Зависимость вероятности идентификации от числа компонент модели гауссовых смесей

Анализ зависимости, представленной на рисунке 3 (кривая 1), показывает, что наиболее эффективное число компонент равно 5, так как с дальнейшим увеличением числа компонент увеличение вероятности правильной идентификации человека по РС пренебрежимо мало. Точность идентификации при этом достигает 65 %.

Дополнительно проводился эксперимент с увеличением длительности сигнала в обучающей выборке до 60 секунд и увеличением длительности тестового высказывания до 10 секунд. На рисунке 3 (кривая 2) изображена зависимость вероятности идентификации от числа компонент модели ГС.

Анализ зависимости, представленной на рисунке 3 (кривая 2), показывает, что наиболее эффективное число компонент равно 5, так как с дальнейшим увеличением числа компонент увеличение вероятности идентификации человека по РС пренебрежимо мало. Точность идентификации при этом достигает 75 %.

Из представленных зависимостей можно сделать вывод, что с увеличением длительности обучающего высказывания увеличивается вероятность правильной идентификации диктора.

Аналогичные исследования проводил И.А. Рахманенко под руководством профессора, д.т.н. Р.В. Мещерякова. Созданная ими реализация модели гауссовой смеси позволяет производить корректную идентификацию с точностью более 95 % с оптимальным числом компонент, равным 5.

Заключение

Проведены экспериментальные исследования основных методов идентификации человека по голосу. Показана целесообразность использования метода ГС. Отмечено, что идентификации человека с использованием алгоритма на основе метода СКО не является полностью инвариантной к исходным РС, так как является эффективной для коротких речевых сообщений, содержащих одно слово. ИД с использованием алгоритма на основе метода минимального расстояния позволяет проводить идентификацию на уровне мужчина/женщина. Одним из наиболее эффективных методов является метод на основе модели ГС. Особенность этого метода заключается в том, что отдельные гауссовы компоненты хорошо моделируют характеристики голоса диктора, необходимые для идентификации. При данном наборе дикторов оптимальное число кластеров для распознавания методом ГС равно 5. Предложенная реализация алгоритма на основе модели гауссовой смеси позволяет производить корректную идентификацию с вероятностью до 75 %.

Библиографический список

1. **Doddington George.** Speaker recognition: Identifying people by voice ТИЕR. – 1985. – № 11 (73), С. 129–146.

2. **Запрягаев С. А., Коновалов А. Ю.** Распознавание речевых сигналов. Вестник ВГУ, 2009. № 2. С. 39–48.

3. **Сорокин В. Н., Вьюгин В. В., Тананыкин А. А.** Распознавание личности по голосу: аналитический обзор // Информационные процессы. –2012. – Т. 12. –№ 1. С. 1–30.

4. **Rabiner L. A.** tutorial on hidden markov models and selected applications in speech recognition. 1989. P. 257–286.

5. **T. Sakai and S. Doshita,** «The phonetic typewriter, information processing 1962», Proc. IFIP Congress 1962.

6. **J. K. Baker,** «Stochastic modeling for automatic speech understanding», Academic Press 1975.

7. **X. Aubert, R. Haeb-Umbach and H Ney,** «Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models», Proc. of ICASSP, 1993 Vol. II, Pp. 648-651.

8. **Dempster A., Laird N., Rubin D.** Maximum likelihood from incomplete data via the EM algorithm J. Royal Stat. Society. 1977. Vol. 39. p. 138.

9. **Campbell J. P.** Speaker Recognition: A Tutorial Proceedings of the IEEE. 1997. Vol. 85, no 9. pp. 1437-1462.

10. **Bilmes, Jeff A. A.** Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models // Berkley, CA: International Computer Science Institute. 1998. Pp. 7–13.

11. **Fletcher H.** Auditory patterns Reviews of modern physics. 1940. Vol. 12, no. 1. p. 47

12. **Furui S.** Digital Speech Processing, Synthesis and Recognition Marcel Dekker, New York, 1989.

13. **Atal B.** Automatic recognition of speakers from their voices Proc. IEEE. 1976. Vol. 64. Pp. 460-475.

14. Методы автоматического распознавания речи. Под редакцией У.Ли. М.: МИР, 1983.

15. **Садыхов Р. Х., Ракуш В. В.** Модели гауссовых смесей для верификации диктора по произвольной речи. Доклады БГУИР. Минск, 2003. № 4. С. 95–103.

16. **Кульбак С.** Теория информации и статистика. М.: Наука, 1967. 408 с.

17. **J. K. Baker.** «Stochastic modeling for automatic speech understanding», Academic Press 1975.

18. **X. Huang, A. Acero, H. Hon.** Spoken languageprocessing: a guide to theory, algorithm, and systemdevelopment. – Prentice Hall PTR, 2001. P. 936.

19. **Сычев В. В.** Обработка речевых сигналов, ИМПБ РАС, Пущино, 2001-2005 гг.

UDC 004.522

AUTOMATIC IDENTIFICATION OF A PERSON ACCORDING TO VOICE USING THE ALGORITHM BASED ON GAUSSIAN MIXES MODEL

D. I. Lukyanov, associate professor, department of radiocontrol and communication, RSREU, PhD; d.i.lukyanov@outlook.com

A. S. Mikhailova, master student, RSREU; alenkamihailova@yandex.ru

Methods of automatic identification of a person according to voice are researched. ***The aim of the work is*** *determination of the most effective algorithm of identification of a person on the basis of speech signal providing high probability of correct identification. Pilot studies of the algorithm on the basis of method of mean square deviation, minimum distance, Gaussian mixes are conducted. Identification of a person using the algorithm on the basis of method of mean square deviation isn't completely invariant to initial RS as is effective for short speech messages containing one word. Identification of a person using the algorithm on the basis of method of minimum distance allows to carry out identification at the level of man/woman. One of the most effective methods is the method based on Gaussian mixes model. The optimum number of clusters for recognition is equal to 5. Implementation of the algorithm on the basis of model of Gaussian mixes allows to make correct identification with probability to 75 %.*

Key words*: recognitions of voice information, recognition systems, identification of a person according to voice, method of mean squared deviation, method of minimum distance, method swept-cepstral conversions of RS range, cepstrum, model of Gaussian compounds, optimum number of clusters.*

DOI: 10.21667/1995-4565-2017-61-3-19-24

References

1. **Doddington George.** Speaker recognition: Identifying people by voice TIIER. 1985. no 11 (73). 129‑146.

2. **Zapryagayev S. A., Konovalov A. Yu**. Recognition of speech signals. Bulletin of VSU, 2009, no. 2, pp. 39–48.

3. **Sorokin V. N., Vjugin V. V., Tananykin A. A**. Recognition of the personality on a voice: state-of-the-art review Information processes. 2012. Vol. 12. no 1. pp. 1-30.

4. **Rabiner L. A.** tutorial on hidden markov models and selected applications in speech recognition. 1989. pp. 257–286.

5. **T. Sakai and S. Doshita,** «The phonetic typewriter, information processing 1962», Proc. IFIP Congress 1962.

6. **J. K. Baker,** «Stochastic modeling for automatic speech understanding», Academic Press 1975.

7. **X. Aubert, R. Haeb-Umbach and H. Ney,** «Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models», Proc. of ICASSP, Vol. II, pp. 648-651 1993.

8. **Dempster A., Laird N., Rubin D.** Maximum likelihood from incomplete data via the EM algorithm J. Royal Stat. Society. 1977. Vol. 39. p. 138.

9. **Campbell J. P.** Speaker Recognition: A Tutorial Proceedings of the IEEE. 1997. Vol. 85, no 9. pp. 1437-1462.

10. **Bilmes, Jeff A. A.** Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models Berkley, CA: International Computer Science Institute. 1998. pp. 7–13.

11. **Fletcher H.** Auditory patterns Reviews of modern physics. 1940. Vol. 12, no. 1. p. 47

12. **Furui S**. Digital Speech Processing, Synthesis and Recognition Marcel Dekker, New York, 1989.

13. **Atal B.** Automatic recognition of speakers from their voices Proc. IEEE. 1976. Vol. 64. pp. 460-475.

14. Methods of automatic speech recognition. Edited U.Li. M.: Mir, 1983.

15. **Savchenko V. V.** Information theory of speech perception Proceedings of the universities. Electronics. 2007. Vol .6. pp. 10-14.

16. **Kullback S.** Information theory and statistics. M.: Nauka, 1967. p. 408.

17. **Baker. J. K.** «Stochastic modeling for automatic speech understanding», Academic Press 1975.

18. **Huang, X. Acero A., Hon H.** Spoken languageprocessing: a guide to theory, algorithm, and systemdevelopment. – Prentice Hall PTR, 2001. p. 936.

19. **Sychev V. V.** Processing speech signals, IBMP RAS, Pushchino.