Система речевого ввода информации для семантических баз знаний

Кузьмин А.А., *магистр физико-математических наук, аспирант кафедры радиофизики и цифровых медиа-технологий факультета радиофизики и компьютерных технологий БГУ*

• распознавание речи • речевой интерфейс баз знаний
• программный пакет HTK.

Данная работа является первым шагом на пути к созданию автономного интерфейсного компонента, который станет значимой частью целой технологии проектирования интеллектуальных систем, разрабатываемой в рамках открытого проекта OSTIS. Благодаря использованию такого относительно автономного компонента, у разработчиков, не являющихся специалистами в сфере анализа речевых сигналов, в перспективе появится возможность создания речевых интерфейсов для своих приложений. Все основные этапы реализации прототипа, включая создание системы Скрытых Марковских Моделей и непосредственно инструментарий распознавания речи, осуществлялись с помощью открытого программного пакета HTK.

This study is the first step in creating an autonomous front-end component, which will become an important part of the whole technology of design intelligent systems. This technology is being developed under an open source project OSTIS. Due to such relatively autonomous components, even the developers who are not specialists in the field of analysis of speech signals will be able to create speech interfaces for their own applications in the future. All the major stages of the prototype, including the establishment of Hidden Markov Models and tools directly to the speech recognition performed using open source software package HTK. Manual Speech Data Labeling, Monophones, Hidden Markov Model Tool Kit, Continious speech recognition, Semantic Knowledge Base.

Введение

Анализируя современные тенденции развития интерфейсов информационных и интеллектуальных систем, небезосновательным выглядит факт, что после интерфейса командной строки (1960–80-е гг.) и графического интерфейса (1980–2000-е гг.), будущее принадлежит комплексному пользовательскому интерфейсу, позволяющему задействовать, кроме зрения, разные органы

чувств человека, в первую очередь — слух. В пользу этого утверждения говорят следующие факты:

- речь — наиболее популярная форма коммуникации между людьми;
- нет необходимости в непосредственном контакте при взаимодействии, поскольку микрофон и динамики могут располагаться на расстоянии;
- руки и глаза остаются свободными, что делает голосовой интерфейс приоритетным в некоторых ситуациях, например, в процессе вождения транспортного средства или параллельном использовании нескольких приложений одновременно;
- в настоящее время в мире существует порядка 1,3 миллиарда мобильных телефонов, что в пять раз превышает количество компьютеров, подключённых к Интернету. Это обеспечивает огромный рынок для будущих систем автоматического диалога;
- благодаря достаточно высокому уровню интеллекта современных систем, появляется возможность увеличить точность распознавания и понимания устной речи.

В свете всех перечисленных тенденций, закономерной и актуальной выглядит разработка речевых способов управления для поддержки средств навигации и поиска в семантических сетях. Как правило, от таких приборов требуется, с одной стороны, обеспечивать обработку большого объёма запросов, а с другой — в процессе использования как можно меньше отвлекать пользователя от объекта поиска. При этом неизбежно растёт число элементов графического меню, что в свою очередь замедляет время поиска нужной опции и заставляет отвлекаться от искомого объекта [1].

Речевой способ взаимодействия — наиболее естественный интерфейс для общения человека с человеком. Это влечёт за собой простоту изучения и использования речевого интерфейса при взаимодействии с базами знаний. Обсуждаемый вид взаимодействия является альтернативным каналом обмена данными между оператором и системой, который позволяет освободить руки и глаза человека при подаче команд. Пользователь может осуществлять запросы, устно отдавая соответствующие команды, в процессе работы, передвижения или манипуляции другими объектами. Дополнительный комфорт интерфейса — следствие отсутствия необходимости в специальных устройствах, таких как: мышь, палочка или перчатки данных. Таким образом, широкому кругу пользователей, включая пожилых людей и инвалидов, будет удобно приспосабливаться к речевому интерфейсу.

Целью данной работы является выбор подходящей технологии и методики для создания модуля распознавания слитной русской речи, призванного обеспечить использование голосового ввода для осуществления запроса к базам знаний. При этом акцент ставится на разработку относительно автономного компонента, который может быть использован разработчиком, не имеющим высокой квалификации в сфере обработки сигналов или теории Скрытых Марковских Моделей (далее — СММ). Такая подсистема должна стать частью целой технологии проектирования интеллектуальных систем [7], разрабатываемой в рамках открытого проекта OSTIS.

Теория СММ была выбрана как методологическая основа для создания модуля распознавания речевых запросов. В качестве набора инструментов, реализующего все основные функции и алгоритмы, был использован пакет HTK.

1. Некоторые аспекты основ теории Скрытых Марковских Моделей

Исторически данная методология появилась в рамках решения проблемы создания модели сигнала, как способа разработки и отладки систем обработки сигналов, а также извлечения информации об источнике. Впервые идея использовать математическую основу Марковских цепей для распознавания речи разрабатывается в классических для данной сферы работах Л.Е. Баума и его коллег [8, 9]. Интерпретация этих идей нахо-

дит своё отражение в статье Л.Р. Рабинера, пожалуй, одной из самых распространённых статей по СММ [3].

Основная идея заключается в том, что определённое значение физически наблюдаемых признаков, вычисленных на основе анализа акустического сигнала произнесённой фразы, является обусловленной нахождением в том или ином состоянии СММ, определённым образом связанной с этой моделью. Проиллюстрируем этот принцип, используя традиционный формализм [11].

Предположим, что каждая фраза представляется последовательностью векторов наблюдений O, определённом как:

$$O = o_1, o_2, o_3, \ldots, o_t, \quad (1)$$

где $\mathbf{o}_t$ — это вектор, наблюдаемый в момент времени t. Тогда проблема распознавания может быть рассмотрена как задача вычисление следующей величины:

$$\arg\max_i \{ P(w_i | O) \}, \quad (2)$$

где w_i — *i*-е слово в словаре. Рассчитывается эта величина с помощью правила Байеса:

$$P(w_i | O) = \frac{P(O | w_i) P(w_i)}{P(O)}. \quad (3)$$

Таким образом, при заданном наборе вероятностей $P(w_i)$ слово, которое вероятнее было произнесено, зависит только от значения $P(O | w_i)$. На практике в случае многомерности вектора наблюдений прямой расчёт совместной условной вероятности $P(o_1, o_2, \ldots | w_i)$ является весьма сложной задачей. Однако, если использовать Марковскую модель в качестве параметрической модели генерации слова, то задача расчёта $P(O | w_i)$ заменяется намного более простой проблемой нахождения параметров этой самой Марковской модели.

Может быть показано, что полное описание СММ включает [11]:

1. N — количество состояний модели.
2. M — количество различных символов для дискретного вектора наблюдений.
3. Матрица распределения вероятности перехода из состояния i в состояние j: $A = \{a_{ij}\}$.
4. Распределение вероятности значений вектора наблюдений для состояния j, $B = \{b_j(k)\}$, где $b_j(k)$ — эмиссионная вероятность наблюдать k-й вектор наблюдений в момент времени t, при условии, что модель находится в состоянии j.
5. Начальное распределение состояний $\pi = \pi_i$, т.е. вероятности того, что первым состоянием будет состояние i.

Для удобства можно ввести компактное обозначение для модели:

$$\lambda = (A, B, \pi). \quad (4)$$

Существуют три классические проблемы, которые необходимо решить для того, чтобы использовать СММ в реально работающих приложениях:

Проблема 1. Дана последовательность векторов наблюдений: $O = O_1, O_2, \ldots O_T$, и модель $\lambda = (A, B, \pi)$. Как эффективно рассчитать $P(O | \lambda)$, т.е. вероятность получить такую последовательность наблюдений при данной модели?

Проблема 2. Дана последовательность векторов наблюдений: $O = O_1, O_2, \ldots O_T$, и модель $\lambda=(A, B, \pi)$. Как выбрать соответствующую последовательность состояний: $Q = q_1, q_2, \ldots q_T$, которая оптимальна в определённом смысле, т.е. наилучшим образом «объясняет» такую последовательность наблюдений?

Проблема 3. Как настроить параметры модели $\lambda=(A, B, \pi)$, для того чтобы максимизировать вероятность $P(O|\lambda)$?

Для разработанной системы вместо эргодического типа СММ была использована т.н. лево-правая модель или модель Байеса, которая, во-первых, проще для расчётов, а во-вторых, больше подходит для приложений, связанных с распознаванием речи [10].

К основным особенностям созданной системы также можно отнести и то, что вектор наблюдений имеет непрерывный диапазон значений. Как следствие — в качестве описания эмиссионной вероятности выступает сумма плотностей вероятностей Гауссовых случайных величин. Кроме того, пакет HTK позволяет каждый вектор наблюдений в момент времени t разделить на S независимых потоков данных o_{st}. Это сделано для возможности смоделировать сразу несколько источников информации. Тогда формула для расчёта $b_j(o_t)$ выглядит следующим образом:

$$b_j(o_t) = \prod_{s=1}^{S}\left[\sum_{m=1}^{M_s} c_{jsm} N(o_{st}, \mu_{jsm}; \Sigma_{jsm}\right]^{\gamma_n}, \quad (5)$$

где M_s — количество компонент в смеси потока S, c_{jsm} — вес m-ой компоненты, а $N(o_{st}; \mu_{jsm}; \Sigma_{jsm}$ — многомерный гауссиан с вектором средних значений μ и ковариационной матрицей Σ:

$$N(O, \mu, \Sigma) = \frac{1}{\sqrt{2\pi^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1}(o-\mu)}, \quad (6)$$

где n — размерность вектора наблюдений o.

2. Задача обработки сигналов

Для успешного решения проблемы распознавания речевых сигналов большую роль играет выделение их информативных признаков. Они должны удовлетворять нескольким критериям [2, 3]:

- быть легко вычисляемыми;
- сохранять всю необходимую информацию, которая содержится в сигнале;
- в режиме обучения на основании обучающего множества векторов признаков создавать модель, которая будет более общей, чем обучающий материал.

В качестве таких признаков используются кепстральные параметры, так как они наилучшим образом подходят для работы с речевыми сигналами и являются широко распространённым представлением спектральных характеристик в задачах обработки подобных сигналов [4]. В данной работе использовались следующие шаги расчёта мел-частотных кепстральных коэффициентов [5]:

Предыскажение. На первом этапе речевой сигнал обрабатывался цифровым фильтром первого порядка. Цель этого фильтра — усиление энергии на высоких частотах, которая обычно уменьшается при генерировании речевого сигнала.

Сегментация. На втором этапе сигнал сегментировался на статистически однородные блоки. Во время формирования вектора признаков предполагалось, что сигнал можно рассма-

тривать как квазистационарный процесс, состоящий из последовательности фреймов, которые могут обрабатываться независимо.

Взвешивание. Выигрышным является умножение каждого кадра на весовую функцию для минимизации нежелательных концевых эффектов. В данной работе в качестве такой функции использовалось окно Хемминга.

Спектральный анализ. После разбиения сигнала на фреймы, вычислялось БПФ [6]. После получения спектра по причине его симметричности используется только половина получившихся данных.

Мел-шкала. Слуховая система человека не является линейной относительно спектра аудиосигнала. Практика показывает, что использование такого нелинейного подхода способно увеличить качество распознавания. Наибольший интерес представляет не детальное представление сигнала, а форманты — спектральные максимумы. Именно они дают основную (необходимую) информацию для распознавания речи и верификации дикторов. Для выделения формант производится сглаживание спектра путём применения набора полосовых фильтров. Для локализации фильтров использовалась мел-шкала.

Операция логарифмирования. Поскольку воспринимаемая громкость сигнала приблизительно логарифмическая, то на выходе каждого рассчитывается усреднённое значение E_j и выполняется операция логарифмирования: $m_j = logE_j$.

Дискретное косинусное преобразование (ДКП). Полученные коэффициенты достаточно сильно кореллированы, таким образом, очень важным является использование кепстрального преобразования для последующего использования в СММ с диагональными матрицами ковариации.

3. Тестовые диалоги для интерфейса семантических баз знаний

При создании прототипа речевого интерфейса для осуществления различного рода запросов в семантических базах знаний был выбран ряд тестовых диалогов пользователя с системой с целью выбора необходимых в общении слов и фраз для последующего создания СММ с соответствующими параметрами. Примером такого рода диалогов могут служить следующие ситуации:

Вопрос: Что это такое (в окне геометрического редактора выделяется некоторая фигура)?

Ответ: Это треугольник со сторонами a, b и углом C, равным 45 градусам, между ними.

Вопрос: Как они связаны (в окне системы выделяются понятия треугольника и тригонометрии)?

Ответ: Эти два понятия связанны в теореме синусов и теореме косинусов.

Вопрос: Что из этого следует?

Ответ: Это даёт возможность расчёта численных характеристик конкретного треугольника.

Вопрос: Приведите пример.

Ответ: Зная длины двух сторон и значение угла между ними, можно рассчитать длину третьей стороны.

Вопрос: Приведите примеры другой теоремы.

Ответ: Сумма углов треугольника равна 180 градусам.

И т. д.

Из существующего ограниченного набора запросов, во-первых, был сформирован набор базовых слов, сгруппированный по назначению:

Язык вопросов

Ключевые узлы:

Вопрос

Ответ

Что это такое?

…

Предметная область

Ключевые узлы:

Треугольник

Произвольный треугольник

Прямоугольный треугольник

Вершина

…

Во-вторых, сформированы стандартные синтаксические последовательности этих слов для перехода от распознавания отдельных слов к распознаванию слитной речи.

4. Сравнение характеристик распознавания созданных систем

В качестве вариантов рассматривались системы со следующими характеристиками: монофонные без разметки обучающих данных, монофонные с разметкой обучающих данных и системы на основе связанных трифонов, где обучение проводилось без разметки обучающих данных. Критерием для сравнения выступали среднее время распознавания на один фрейм, точность распознавания по фразам, а также точность распознавания по отдельным словам. Обучение осуществлялось на одинаковом ограниченном наборе из 50 фраз, содержащих в сумме 217 слов. Результаты представлены в таблице 1.

Табл. 1. Оценка характеристик систем распознавания

	Среднее время распознавания на один фрейм, с/фрейм	Точность распознавания фраз, %	Точность распознавания слов, %
Система на основе монофонов. Обучение без разметки данных	0.011396	74.00	94.97
Система на основе связанных трифонов. Обучение без разметки данных	0.0118026	72.00	93.53
Система на основе монофонов. Обучение с разметкой данных	0.0116263	98.00	99.57

На основе вышеприведённой информации, выбор был сделан в пользу системы на основе монофонов, для создания которой использовались данные, размеченные вручную. Решающим фактором стал высокий процент точности распознавания по фразам при приемлемом времени распознавания. Такие результаты можно объяснить следующим образом: исходя из условия ограниченности словаря и последовательностей слов, целе-

сообразным выглядит разметить ограниченное число обучающих файлов вручную, с другой стороны, эта же ограниченность препятствует созданию качественных трифонов, тем более связанных.

5. Алгоритм создания системы СММ для распознавания

Полный цикл создания хорошо обученных СММ включает два основных этапа: подготовку данных для обучения и непосредственно само обучение (рис. 1). Во многом успех распознавания зависит от качества обучающей информации. Поэтому этой части работы уделялось особенно много внимания.

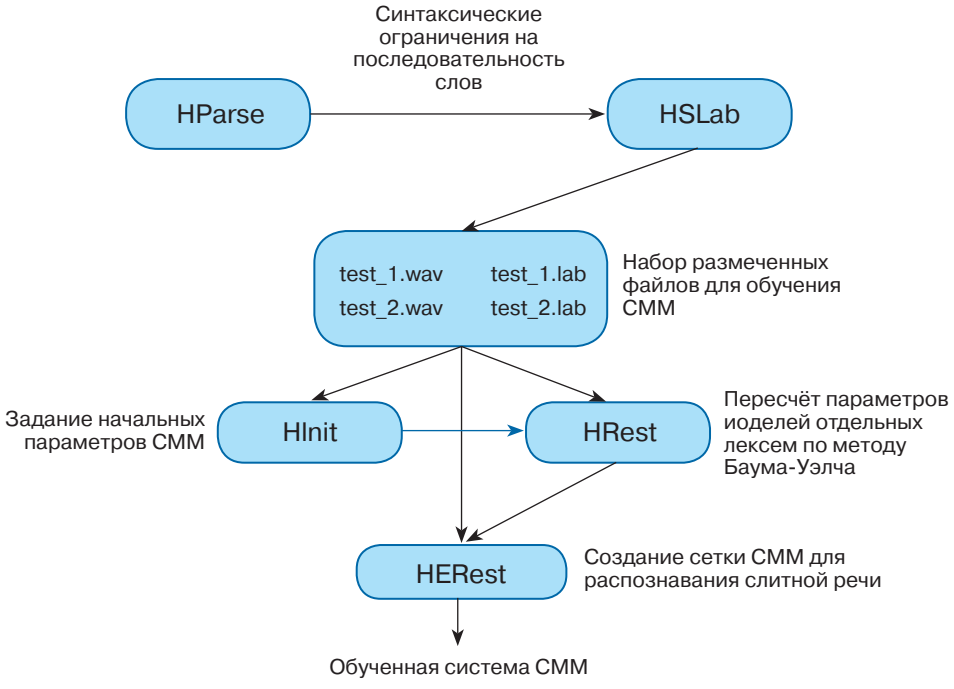

Рис. 1. Алгоритм создания набора СММ для распознавания

Подготовка данных

Шаг 1 — грамматика. Первым шагом было создание грамматики каждого из запросов, которая в данном случае представляла сетку, которая включала в себя строгие последовательности слов, допустимых для распознавания. Такие последовательности реализовывали ряд однозначно трактуемых запросов к базе знаний. Ограничения на возможный порядок слов играют важную роль в распознавании и стабильной работе всей системы, так как априори исключает последовательности, которые противоречат правилам синтаксиса или не могут быть идентифицированы как запросы.

Шаг 2 — словарь. Упорядоченный словарь каждого из слов, входящий в запросы, создавался, следуя иерархии от фраз к отдельным словам. Закрытый словарь предоставлял возможность создания уникальной фонетической транскрипции, адаптированной для конкретного диктора и учитывающей региональные особенности произношения некоторых слов русского языка (например, как мягкий [ц] вместо мягкого [т]). Это было сделано с целью

демонстрации достаточно высокой точности распознавания на прототипе, созданном под конкретного диктора.

Шаг 3 — запись данных. На данном этапе осуществлялось создание набора файлов формата .wav, содержащих не менее трёх вариантов записей базовых слов, произнесённых диктором. В качестве инструмента выступала функция HSLab из пакета HTK, которая позволила не только записать данные, но и разметить их по содержащимся фонемам. Разметка производилась вручную. Такая работа является весьма трудоёмкой и кропотливой, однако, как было отмечено выше, именно качественное выделение фонем вручную стало залогом достаточно высокой точности распознавания даже на уровне монофонов. Следует отметить также, что чем короче слово, тем большее количество вариантов требуется для более качественного распознавания. Всего для обучения использовалось 113 фраз, состоящих в сумме из 613 слов.

Шаг 4 — кодирование данных. Финальным шагом в подготовке данных является обработка речевых сигналов и преобразование их в последовательности векторов признаков. Как и было упомянуто выше, в данной работе в качестве таких векторов были использованы кепстральные коэффициенты шкалы мел-частот. Обработка сигнала производилась согласно алгоритму, описанному в разделе «Задача обработки сигналов».

Помимо 13 кептсральных коэффициентов набор параметров содержал 13 дельта-коэффициентов и 13 коэффициентов ускорения. Таким образом, размерность вектора признаков составила величину, равную 39. Для хранения векторов признаков использовался специальный формат .MFCC (от англ. Mel Frequency Cepstral Coefficients —Кепстральные коэффициенты шкалы мел-чатсот).

Обучение

С этого момента начинается создание набора хорошо обученных CMM, эмиссионные вероятности в которых описываются суммой плотностей вероятностей гауссовых случайных величин.

Шаг 5 — создание начальных монофонов. Первым шагом в создании системы CMM является определение модели прототипа. Для системы, основанной на фонемах, хорошей топологией является лево-правая с тремя состояниями. Начальные значения для модели каждой фонемы рассчитывались на основании обучающих данных с помощью алгоритма Витерби. Программным средством здесь выступила функция HInit.

Шаг 7 — пересчёт коэффициентов монофонов. Параметры модели каждой фонемы пересчитывались по методу Баума-Уэлча (алгоритм прямого-обратного хода или алгоритм максимизации правдоподобия). Пересчёт производился с помощью функции HRest.

Шаг 8 — создание системы CMM для распознавания фраз. Последним шагом стала корректировка параметров моделей отдельных фонем, но уже связанных в рамках, как отдельных слов, так и целых предложений. Такая интегрированная система создавалась функцией HERest.

6. Распознавание

Своего рода «субстратом» для распознавания выступает многоуровневая сеть, которая образуется следующим образом: динамически в момент запуска распознавателя каждое слово в сетке запросов меняется на совокупность соответствующих фонем из словаря. Тут же каждая фонема заменяется соответствующей CMM. Таким образом формируется распознающая сеть.

Для каждой неизвестной фразы с Т-рассчитанными фреймами каждый путь через сетку от первого до последнего узла является потенциальной гипотезой распознавания.

Каждый из этих путей имеет логарифмированную вероятность, которая рассчитывается суммированием логарифмов вероятностей каждого индивидуального перехода в рамках путей, а также логарифмов эмиссионных вероятностей каждого состояния относительно соответствующих наблюдений.

Работа распознавателя заключается в том, чтобы найти среди сетки путь с наибольшей логарифмированной вероятностью. Такой путь находится, благодаря использованию т.н. алгоритма прохода меток (Token Passing). Каждая метка представляет собой частичный путь, который дополняется на протяжении времени. В самом начале метка располагается в каждом из возможных начальных узлов. За каждый временной шаг метки распространяются вдоль состояний, пока не достигнут состояния, имеющего эмиссионную вероятность. Когда метка попадает в разветвляющийся участок, она копируется и исследование во всех возможных направлениях продолжается. Проходя вдоль переходов и через состояния, метка увеличивает логарифм своей текущей вероятности за счёт вероятностей перехода и эмиссионных вероятностей. На каждом временном шаге отбираются N лучших меток, и продолжается исследование.

Каждая метка сохраняет историю своего движения, которая включает в себя исключительно слова, расположенные в пройденных узлах, для последующего восстановления искомой фразы.

Основные особенности созданной системы:

1. Созданная система является дикторозависимой, так как действует пока на правах прототипа. Однако в дальнейшем уже созданные модели станут начальным приближением для создания более универсальной системы распознавания.

2. Словарь является закрытым (включает множество слов из строго определённых запросов). Это ограничение также помогает увеличить точность распознавания.

3. Плотность эмиссионной вероятности является суммой плотностей вероятностей Гауссовых случайных величин.

4. Для обучения использовались данные, размеченные вручную. Как отмечалось, это ключевой момент для качества работы системы.

Заключение

Результатом работы стало создание модуля распознавания речи для осуществления голосовых запросов к интеллектуальной базе знаний. На пути к успешному решению поставленной задачи были, в свою очередь, решены следующие подзадачи:

1. Выбран способ реализации системы с учётом требований к точности и времени распознавания.

2. Создан прототип модуля, который способен с высокой точностью распознавать широкий спектр запросов к базе знаний по геометрии. Все запросы являются русскоязычными.

3. Выработана методика создания модуля распознавания для русскоязычных фраз, которая может быть применена человеком, не являющимся специалистом в сфере речевых сигналов, для создания модуля распознавания запросов к базам знаний в других сферах помимо геометрии.

Результаты работы должны стать важным шагом на пути создания автономного IP-компонента интерфейса семантических баз знаний в рамках разработки технологии проектирования интеллектуальных систем в целом.

Литература

1. *Landauer T.K*. Selection from alphabetic and numeric menu trees using a touch screen: Breadth, depth, and width / T.K. Landauer, D.W. Nachbar. New York: ACM, 1985.

2. *Бовбель Е.И*. Статистические методы распознавания речи: скрытые марковские модели / Е.И. Бовбель, И.Э. Хейдоров // Зарубежная радиоэлектроника. Успехи современной электроники. 1998. № 3. С. 3654.

3. *Rabiner, L. R*. Fundamentals of Speach Recognition / L. R. Rabiner, B. N. Juang. New Jersey, 1993.

4. *Oppenhejm, A. V*. From Frequency to Quefrancy: A History of the Cepstrum / A.V. Oppenhejm, R. W. Schafer // IEEE Signal Processing Magazine. 2004. Vol. 21. P. 95–106.

5. *Бовбель Е.И*. Скрытые марковские модели и машины на опорных векторах от теории к практике / Е.И. Бовбель, И.Э. Хейдоров, Ю.В. Пачковский. Минск: БГУ, 2008.

6. *Опенгейм А.В*. Цифровая обработка сигналов / А.В. Опенгейм, Р.В. Шафер. М., 1979.

7. Open Semantic Technology for Intelligent Systems. [Электронный ресурс]. 2010. Режим доступа: http://www.ostis.net /. Дата доступа: 30.11.2010.

8. *Baum L. E*. Statistical inference for probabilistic functions of finite state Markov chains / Baum L. E., Petrie T., Ann. Math. Stat., vol. 37, pp. 1554–1563, 1960.

9. *Baum L. E*. A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov Chains / Baum L. E., Petrie T., G. Soules, N. Weiss, Ann. Math. Stat., vol. 41, no 1, pp. 164–171, 1970.

10. *Jelinek F*. Continuous speech recognition by statistical methods / F. Jelinek, Proc. IEEE, vol. 64, pp. 532–536, Apr. 1976.

11. *Rabiner L*. A. Tutorial on Hidden Markov Models and Selected Applications in Speech / L. A. Rabiner, Recognition. pp. 257–286. IEEE Press, 1988.

Сведения об авторе

Кузьмин Алексей Александрович —

магистр физико-математических наук, аспирант кафедры радиофизики и цифровых медиа-технологий факультета радиофизики и компьютерных технологий БГУ