УДК 004.522:004.934

ОПЫТ ПРИМЕНЕНИЯ ИНСТРУМЕНТАЛЬНОЙ СИСТЕМЫ SPHINX ДЛЯ РЕШЕНИЯ ЗАДАЧИ РАСПОЗНАВАНИЯ РЕЧЕВЫХ КОМАНД УПРАВЛЕНИЯ КОМПЬЮТЕРНЫМИ СИСТЕМАМИ

Бондаренко И.Ю., Савкова Д.Г.
Донецкий национальный технический университет
кафедра прикладной математики и информатики
E-mail: das.savkova@gmail.com

Аннотация

Бондаренко И.Ю., Савкова Д.Г. Опыт применения инструментальной системы Sphinx для решения задачи распознавания речевых команд управления компьютерными системами. *Данное исследование посвящено разработке системы распознавания речевых команд управления компьютерными системами на базе инструментальной среды Sphinx 4. Проанализирован математический аппарат скрытых марковских моделей, реализованный в Sphinx 4 для организации процесса распознавания речи и обучения такому распознаванию. Проведены экспериментальные исследования системы распознавания речевых команд на базе Sphinx 4, направленные на оценку точности работы этой системы как в однодикторном, так и в дикторонезависимом режимах. Сделаны практические рекомендации по эффективности применения инструментальной среды Sphinx 4 для организации речевого интерфейса компьютерных систем.*

Введение

Информационные технологии занимают уникальное положение в современном обществе. В отличие от других научно-технических достижений, средства вычислительной техники и информатики применяются практически во всех сферах интеллектуальной деятельности человека, способствуя научно-техническому прогрессу. В последнее время особое внимание как исследователей, так и конечных пользователей уделяется разработке и применению автоматизированных систем с речевым человеко-машинным интерфейсом. Устная речь является наиболее естественным и простым для человека способом общения, поэтому речевые технологии находят все большее распространение в роботехнике, управлении компьютерными устройствами, системах телекоммуникаций [1].

Особо актуальным речевой человеко-машинный интерфейс является в следующих двух случаях:

1. управление мобильными компьютерными устройствами (телефонами, смартфонами, нетбуками и т. д.);
2. управление персональным компьютером людьми с ограниченными физическими возможностями (прежде всего, с нарушениями опорно-двигательного и зрительного аппарата).

В первом случае использование традиционных средств тактильно-зрительного интерфейса (клавиатуры, мыши, дисплея) затруднено для пользователей из-за миниатюрности используемых компьютерных устройств, а во втором случае — и вовсе невозможно вследствие физических особенностей самих пользователей.

Ключевым элементом любого речевого интерфейса является система автоматического распознавания речи. Существует ряд методов построения таких систем. Наиболее популярными являются две группы методов:

1) методы распознавания речи, основанные на применении искусственных нейронных сетей [2];

2) методы распознавания речи, основанные на применении скрытых марковских моделей [3].

Авторы в течение многих лет ведут разработку систем автоматического распознавания как изолированных речевых команд, так и слитной речи, на базе нейросетевого подхода [4]. Естественным образом возникает задача сравнить эти системы распознавания с другими системами, функционирующими на базе скрытых марковских моделей.

Исходя из вышеописанного, в данной работе была поставлена следующая цель: разработать и исследовать систему автоматического распознавания речевых команд для управления персональным компьютером на базе математического аппарата скрытых марковских моделей.

В качестве объекта исследования была выбрана инструментальная система Sphinx 4, разработанная американскими исследователями [5]. Эта система предоставляет разработчикам удобный инструментарий для исследования скрытых марковских моделей, а после определённой доработки может использоваться как система автоматического распознавания речевых команд управления компьютерными устройствами. Преимуществами Sphinx по сравнению с аналогичной инструментальной системой HTK [6] являются:

1. открытая лицензия, по которой поставляется система Sphinx, что позволяет свободно использовать эту систему как в исследовательских, так и в коммерческих целях;
2. язык программирования Java, на котором написана система Sphinx, что идеально подходит для использования систем распознавания речи, построенных на базе Sphinx, в мобильных устройствах основанных на android от google, а также в сервлетах для последующего использования на устройствах типа CLDC.

Задачами данной работы являлись:

- исследовать математический аппарат скрытых марковских моделей для автоматического распознавания речи;
- на базе инструментальной системы Sphinx разработать и обучить программную систему распознавания речевых команд управления компьютером, работающую как в однодикторном, так и в дикторонезависимом режимах;
- на материалах собственного речевого корпуса длительностью свыше полутора часов оценить точность работы созданной системы распознавания речи.

Постановка задачи распознавания речевых команд

В качестве математического аппарата, применяемого для распознавания речи в Sphinx4, применяются скрытые марковские модели [7]. Марковская модель – это вероятностный автомат с конечным числом состояний, изменяющий своё состояние один раз в единицу времени. При этом наблюдателю известны состояния и вероятности переходов между состояниями (матрица переходов). Таким образом, марковская модель описывает некоторый вероятностный процесс. Каждому наблюдаемому событию этого процесса соответствует одно из состояний модели.

Скрытая марковская модель, в отличие от обычной, описывает два вероятностных процесса – ненаблюдаемый (основной) и наблюдаемый (вспомогательный). О ходе основного процесса (например, процесса произнесения фонем устной речи) мы пытаемся судить по наблюдаемым событиям вспомогательного процесса (например, по изменению кратковременных спектральных характеристик звукового сигнала). В скрытой марковской модели задаётся не только множество состояний, но и алфавит символов наблюдения. Наблюдение является вероятностной функцией состояния. Для определения скрытой марковской модели необходимо задать следующие элементы:

1) N – число состояний в модели. Хотя состояния и являются скрытыми от наблюдателя, им можно приписать физический смысл. Так, в распознавании речи под состояниями можно подразумевать фонемы слов или составные элементы фонем. Переход между состояниями осуществляется мгновенно.

2) M – число различных символов наблюдения, которые могут порождаться моделью, т.е. размер дискретного алфавита. В распознавании речи в качестве такого дискретного алфавита может использоваться набор классов или кластеров, на которые разбивается множество возможных значений кратковременных спектральных характеристик звукового сигнала.

3) распределение вероятностей переходов между состояниями (или матрица переходных вероятностей) $A = \{a_{ij}\},\ i = \overline{1..N},\ j = \overline{1..N}$.

4) распределение вероятностей появления символов наблюдения в состоянии j, $B = \{b_j(k)\},\ j = \overline{1..N},\ k = \overline{1..M}$.

5) начальное распределение вероятностей состояний $\pi = \{\pi_i\},\ i = \overline{1..N}$.

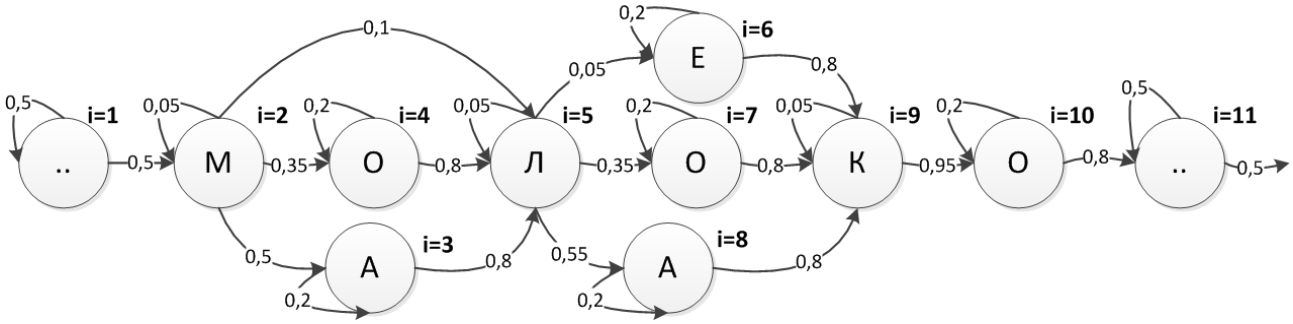

Рисунок 1 - Лево-правая скрытая марковская модель с 11 состояниями, описывающая вероятностный процесс произнесения слова «МОЛОКО»

В распознавании речи используются лево-правые скрытые марковские модели (см. рис.1). Их последовательность состояний обладает свойством, которое выражается в том, что с увеличением времени индекс состояния i также увеличивается или же остается неизменным. Т.е. состояния переходят слева направо, а наоборот сделать переход нельзя. Этот тип отлично подходит для описания процессов с прямым ходом времени, в частности, для распознавания речевых сигналов.

Любая система автоматического распознавания речи может функционировать в одном из двух режимов: режиме распознавания и режиме обучения. Использование скрытых марковских моделей для функционирования системы в режиме распознавания организовано следующим образом. В системе присутствует словарь распознаваемых элементов (например, речевых слов) размером V. Для каждого словарного элемента сформирована скрытая марковская модель $\lambda_v = \{A_v, B_v, \pi_v\},\ v = \overline{1..V}$. На вход поступает последовательность спектральных характеристик звукового сигнала, рассматриваемая как последовательность наблюдений $O = O_1 O_2 ... O_T$. Для каждого словарного элемента вычисляется $P(O \mid \lambda_v)$ – вероятность появления последовательности наблюдений O для скрытой марковской модели λ_v, или, попросту говоря, вероятность того, что во входном звуковом сигнале присутствует v-й элемент словаря распознавания. Результат распознавания – номер распознанного словарного элемента v_{res} – определяется следующим образом:

$$v_{res} = \arg\max_{1 \le v \le V}(P(O \mid \lambda_v)) \quad (1)$$

Использование скрытых марковских моделей для функционирования системы в режиме обучения организовано следующим образом. Существует множество последовательностей наблюдений $O_L = \{\{O_1 O_2 ... O_{T_1}\}, \{O_1 O_2 ... O_{T_2}\}, ... \{O_1 O_2 ... O_{T_L}\}\}$, сформированное на основе множества обучающих речевых сигналов v-го элемента словаря распознавания. Необходимо подстроить параметры скрытой марковской модели

$\lambda_v = \{A_v, B_v, \pi_v\}$ так, чтобы максимизировать вероятности $P(O_\ell \mid \lambda_v)$, $\ell = \overline{1..L}$. Выполняя подстройку параметров, получаем скрытую марковскую модель *v*-го словарного элемента, наилучшим образом соответствующую обучающим последовательностям наблюдений. Применяя вышеописанную процедуру для всех элементов словаря распознавания $v = \overline{1..V}$, мы обучаем нашу систему автоматического распознавания речи.

Эксперименты по распознаванию речевых команд управления компьютером

При проведении экспериментальных исследований использовалась инструментальная среда Sphinx 4, предназначенная для автоматизации проектирования систем распознавания речи на базе скрытых марковских моделей. С помощью Sphinx 4 была построена система распознавания речевых команд управления персональным компьютером со словарём объёмом 100 слов.

Для обучения данной системы дикторонезависимому распознаванию была использована свободная речевая база VoxForge [8]. Общая продолжительность аудиоматериала базы 9,5 часов. В записи принимали участие 18 дикторов.

Для обучения системы речевого управления однодикторному распознаванию команд была сформирована собственная речевая база. Продолжительность аудиоматериала в ней составила 1,31 часа. Все записи выполнял один диктор.

Для проверки эффективности распознавания было создано 4 словаря: на 10 слов, 25, 50 и 100. Чтобы слова из словаря объединить в различные словосочетания, с помощью которых оператору удобно было бы отдавать команды компьютеру, был создан набор контекстно-свободных грамматик (пример такой грамматики для словаря из 25 слов показан на рис.2).

```
grammar dic25;
public <openclose> = (открыть | закрыть) (файл | папку | меню | в новой вкладке | в
новом окне);
public <offon> = (включить | выключить) (звук | вайфай | блютуз | тачпад | экран);
public <abort> = (завершить) (работу | приложение);
public <commands> = (копировать | вставить | вырезать | отменить | повторить);
```

Рисунок 2 – Пример контекстно-свободной грамматики для организации речевого управления компьютером со словарём объёмом 25 слов

В качестве источника аудиоматериала, на котором оценивалась точность работы обученной системы распознавания, использовалась собственная шестидикторная речевая база данных. В её формировании приняли участие трое дикторов-мужчин и трое дикторов-женщин. В их задачи входило чтение всех предложений грамматики для 4 вышеописанных словарей. Запись производилась в несжатом формате raw с частотой 16000 Гц и разрядностью звука 16 бит в соответствии с требованиями Sphinx4.

Для численной оценки точности распознавания использовался стандартный критерий WER (Word Error Rate) [9]. Он вычисляется по следующей формуле:

$$WER = \frac{S + D + I}{N} \cdot 100\%, \qquad (2)$$

где S – количество замен (substitutions), D – количество удалений (deletions), I – количество вставок (insertions), которые необходимо применить к цепочке распознанных слов, чтобы она совпала с цепочкой правильных слов, а N – общее количество правильных слов.

Было проведено три эксперимента, направленных на оценивание точности распознавания речевых команд системой автоматического распознавания, разработанной с использованием инструментальной среды Sphinx 4 на базе математического аппарата скрытых марковских моделей.

Целью первого эксперимента было определение точности дикторонезависимого распознавания речевых команд. Система распознавания обучалась на материале речевой базы VoxForge, а тестировалась на материале собственной шестидикторной речевой базы. Результаты распознавания представлены на рис.3.

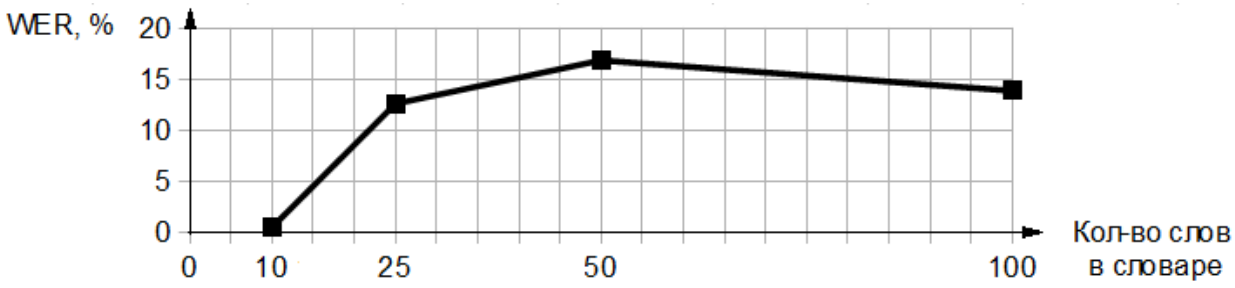

Рисунок 3 –. График зависимости точности дикторонезависимого распознавания речевых команд от объема словаря на материале тестовой шестидикторной речевой базы

Целью второго эксперимента было определение точности однодикторного распознавания речевых команд. Система распознавания обучалась на материале собственной однодикторной речевой базы, а тестировалась на дополнительном аудиоматериале длительностью около восьми минут, записанном тем же диктором. Результаты распознавания представлены на рис.4.

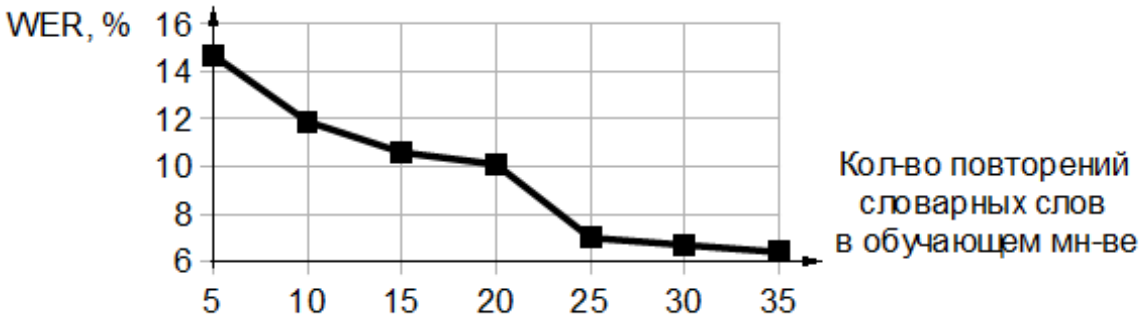

Рисунок 4 – График зависимости точности однодикторного распознавания речевых команд (на базе словаря 100 слов) от объема обучающего множества

Целью третьего эксперимента было определение того, насколько система распознавания снижает точность своей работы, если её обучать как однодикторную, а использовать в дикторонезависимом режиме. Система распознавания обучалась на материале однодикторной речевой базы, а тестировалась на материале шестидикторной речевой базы. Результаты распознавания представлены на рис.5.

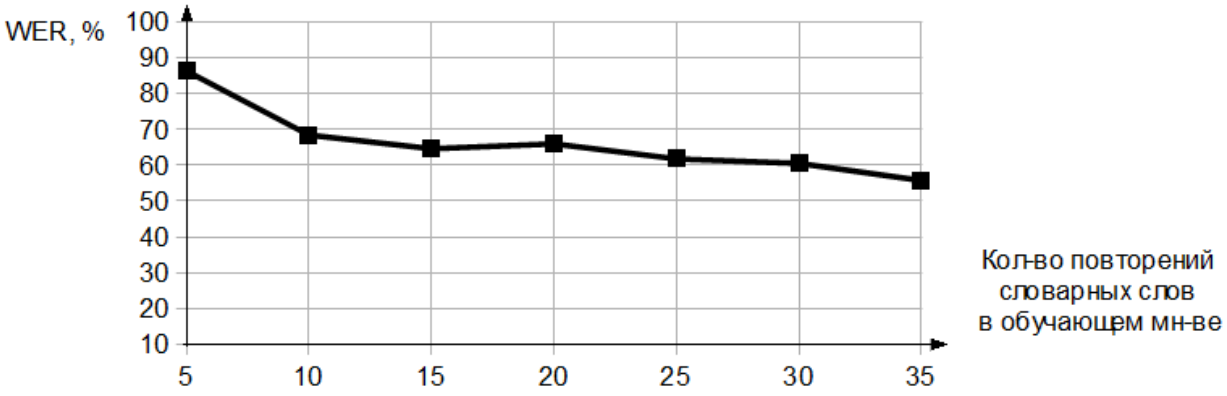

Рисунок 5 – График зависимости точности распознавания речевых команд от объема словаря на материале тестовой шестидикторной речевой базы при условии, что система распознавания обучалась в однодикторном режиме (на одного диктора)

Выводы

В данной работе была создана и исследована система автоматического распознавания речевых команд для управления персональным компьютером на базе математического аппарата скрытых марковских моделей. В ходе выполнения работы были получены следующие результаты:

1) исследован математический аппарат скрытых марковских моделей применительно к решению задачи автоматического распознавания речи;

2) на базе инструментальной системы Sphinx разработана и обучена программная система распознавания речевых команд управления компьютером, работающая в двух режимах: в однодикторном и дикторонезависимом;

3) на материалах собственного речевого корпуса длительностью свыше полутора часов (1,31 часа обучающего аудиоматериала и 0,2 часа тестового материала) оценена точность работы созданной системы распознавания речи.

Анализ результатов экспериментальных исследований показывает следующее.

1. Чем больше словарь системы распознавания, тем хуже точность работы этой системы в дикторонезависимом режиме. Известно, что речевой интерфейс становится удобен пользователям лишь в том случае, если ошибка распознавания речевых команд при использовании этого интерфейса не превышает 5%. Для системы дикторонезависимого распознавания речевых команд, разработанной на базе скрытых марковских моделей, этот порог превышается уже на словаре объёмом 25 слов. Таким образом, применение такой системы для полноценного речевого управления компьютером без предварительной подстройки под диктора не представляется оправданным: высокая точность достигается лишь при малом словаре, размер которого затрудняет речевой диалог, а на словаре более приемлемых размеров снижается точность распознавания.

2. При использовании системы речевого управления компьютером в однодикторном режиме оператору необходимо собрать достаточно большой объём аудиоматериала, чтобы достичь точности распознавания, близкой к 5 %. Кроме того, на результат обучения и, как следствие, на точность распознавания влияет не только объем материала, но и его качество. Чем понятнее будут произнесены слова, тем ниже будет ошибка. Также немаловажен порядок слов. Его обязательно нужно менять, иначе усилия, приложенные к записи большого объема материала, будут безрезультатными. Каждая запись должна быть дольше 5 секунд, но меньше 30. Между словами должны быть соблюдены паузы. Слова произносятся в соответствии с транскрипцией в словаре.

3. Категорически не рекомендуется использовать систему распознавания речи, настроенную на использование в однодикторном режиме, т.е. на одного оператора, для распознавания речевых команд других операторов.

Таким образом, можно сделать следующие выводы:

1) применение инструментария Sphinx для создания системы дикторонезависимого речевого управления компьютерными устройствами нецелесообразно, а для создания системы с подстройкой под диктора – затруднено ввиду громоздкой процедуры обучения, требующей со стороны оператора значительных усилий по формированию обучающего аудиоматериала;

2) аппарат скрытых марковских моделей неплохо подходит для описания процессов с прямым ходом времени, таких как порождение речи, но применение только этого аппарата для распознавания речевых команд без привлечения других мощных методов распознавания (нейронных сетей, нечёткой логики и т.п.) не всегда эффективно.

Дальнейшие исследования будут направлены на усовершенствование системы Sphinx в двух направлениях:

−разработка инвариантной системы фонетических признаков речевого сигнала, вычисляемой с помощью искусственных нейронных сетей и используемой при формировании входных сигналов для скрытых марковских моделей;

−исследование алгоритмов распознавания, альтернативных скрытым марковским моделям.

Целью такого усовершенствования будет являться повышение точности дикторонезависимого распознавания речи.

Список литературы

1. Ронжин А.Л., Карпов А.А., Ли И.В. Система автоматического распознавания русской речи SIRIUS. - Спб.: СПИИРАН, 2006. - 12 с.

2. Федяев О.И., Бондаренко И.Ю. Сегментация речевого сигнала на основе bagging-коллектива нейросетевых детекторов фонем // Материалы 8-й Международной научно-практической конференции «Математическое и программное обеспечение интеллектуальных систем» MPZIS-2010. – Днепропетровск: ДНУ. – 2010. – с. 238-239.

3. Аграновский А.В., Леднов Д.А. Теоретические аспекты алгоритмов образотки и классификации. - М.: Радио и связь, 2004. - 164 с.

4. Федяев О.И., Бондаренко И.Ю. Организация системы автоматического распознавания речи на основе коллектива распознающих автоматов // Материалы 4-й международной научно-технической конференции "Моделирование и компьютерная графика - 2011". Донецк. - 2011 . - с. 309-316.

5. CMU Sphinx Open Source Toolkit For Speech Recognition Evaluation [Electronic resourse] / Интернет-ресурс. - Режим доступа: http://cmusphinx.sourceforge.net/ [проверено 1.04.2012]. - Загл. с экрана.

6. What is HTK? [Electronic resourse] / Интернет-ресурс. - Режим доступа: http://htk.eng.cam.ac.uk/ [проверено 1.04.2012]. - Загл. с экрана.

7. Рабинер Л.Р. Скрытые марковские модели и их применение в избранных приложениях при распознавании речи// ТИИЭР. - 1984. - Т.72, № 2. - с. 86-120.

8. Welcome – Russian Evaluation [Electronic resourse] / Интернет-ресурс. - Режим доступа: http://www.voxforge.org/ru [проверено 1.04.2012]. - Загл. с экрана.

9. Word Error Rate [Electronic resourse] / Интернет-ресурс. – Режим доступа: http://en.wikipedia.org/wiki/Word_error_rate [проверено 1.04.2012]. - Загл. с экрана.