УДК 004.522

Двухуровневый метод распознавания голосовой команды

Гребнов С.В., асп.

Предложен двухуровневый метод распознавания голосовой команды, основанный на работе со скрытыми марковскими моделями и являющийся комбинацией двух алгоритмов распознавания речи: распознавания слитной речи и распознавания ключевого слова.

Ключевые слова: скрытые марковские модели, распознавание речи, голосовое управление, метод скользящего окна.

Two-level voice control approach

Grebnov S.V., post graduate student

The article describes two-level voice control approach based on hidden markov models (HMM) and combination of two speech recognition algorithms: continuous speech recognition and keyword spotting. Speed and accuracy of proposed approach are good enough to use it in real time practice.

Keywords: hidden markov models, speech recognition, keyword spotting, sliding window models.

Введение. В связи с успехами развития вычислительной техники и новых информационных технологий в последнее десятилетие определилась тенденция к нарастанию сложности систем управления, а также всех прочих видов человеко-машинных систем. Важной является возможность взаимодействия человека с машиной на языке, максимально приближенном к естественному языку человека, поскольку общение на естественном языке позволяет организовать эффективное и удобное взаимодействие оператора с системой. В настоящее время сфера внедрения систем распознавания речи существенно расширяется, захватывая различные отрасли производственной, административной и даже бытовой деятельности.

Основные применения системы распознавания речи и голосового управления относятся к сфере управления роботами и техническими устройствами различного назначения, а также автоматизации ввода различной информации в вычислительные информационные системы. В процессе распознавания голосовых команд производится выделение фрагментов речи, их последующая обработка и анализ в целях определения эталонной команды из словаря, соответствующей произнесенной.

В качестве метода распознавания большинство современных систем используют алгоритмы распознавания ключевого слова (Keyword spotting) [6, 7, 8], построенные на основе скрытых марковских моделей (СММ) [1, 2, 3]. Анализ применимости СММ для распознавания речи приводится в [4, 5]. Типичный метод распознавания голосовой команды заключается в применении алгоритма распознавания ключевого слова (Keyword spotting) ко всему речевому участку для каждой возможной команды из словаря команд. Такой подход имеет два существенных недостатка:

1) большая вычислительная сложность;

2) команды могут включать слова, которые плохо распознаются с помощью алгоритма распознавания ключевого слова.

Первая проблема возникает из-за необходимости применения алгоритма распознавания ключевого слова для каждой возможной команды из словаря, вторая – по следующим двум причинам:

- составные части команды содержат сложные для распознавания фонемы языка;
- существуют дефекты в некоторых моделях фонем, полученные в силу несбалансированности речевой базы данных (РБД), на которой производилось обучение, или неправильного процесса обучения.

Ниже рассматривается новый метод голосового управления, который позволяет решить эти проблемы за счет комбинации метода распознавания ключевого слова с методом распознавания слитной речи, приводится описание нового принципа распознавания голосовой команды, рассматривается разработанный однопроходный алгоритм распознавания ключевого слова, приводятся результаты тестирования, а также сравнение показателей разработанных модификаций и методов с существующими аналогами.

Двухуровневый метод распознавания команды. Для решения проблем, описанных выше, нами разработан следующий подход: добавление в команду заранее предопределенного ключевого слова и применение алгоритма распознавания ключевого слова только для этого вводного (ключевого) слова, а для оставшейся части команды использование стандартного алгоритма распознавания слитной речи (CSR).

Таким образом, использование предложенной системы заключается в произнесении команды, которая состоит из двух частей: ключево-

го слова (с помощью него система понимает, что обращаются к ней) и собственно команды (название действия, которое хочет выполнить пользователь). Например: «Агент, включить свет». В данном случае «Агент» – ключевое слово, а «включить свет» – непосредственно команда.

Выделение двух частей направлено на увеличение скорости работы и качества распознавания.

Увеличение скорости достигается за счет сокращения количества распознаваемых ключевых слов до одного, заранее предопределенного, и использования для распознавания непосредственно команды алгоритма распознавания слитной речи (CSR), который на порядок быстрее алгоритма распознавания ключевого слова и временем работы которого в данном случае можно пренебречь.

Улучшение же качества распознавания достигается за счет введения заранее предопределенного, единственного ключевого слова, что дает возможность в качестве такого слова выбрать легкораспознаваемое слово или провести дополнительные меры для улучшения распознавания этого слова: подготовка дополнительного речевого материала, калибровка параметров алгоритма распознавания ключевого слова.

Кроме этого, можно использовать название системы или же имя в качестве ключевого слова, что дает следующие преимущества:

- является для человека более интуитивно привычным и удобным способом произношения команд;
- добавляет одинаковую для всех пользователей командную интонацию, что уменьшает количество интонационных вариаций произношения команды, тем самым упрощает процесс распознавания;
- уменьшается количество ложных срабатываний за счет сокращения количества ключевых слов до одного.

Недостатком предложенного подхода распознавания команды являются некорректные срабатывания системы, если после ключевого слова следуют команды или фразы не из словаря команд. Например, если пользователь скомандует «Агент, включить телевизор» и такой команды нет в словаре команд, то алгоритм распознает эту команду как команду из словаря, наиболее созвучную с произнесенной, например «Агент, включить свет». Это происходит по причине применения алгоритма распознавания слитной речи для распознавания непосредственно команды, который лишь распознает голосовую фразу в соответствии с заранее определенным словарем и не может определить, является ли данная фраза не из словаря(OOV, Out Of Vocabulary). Такую задачу решает алгоритм распознавания ключевых слов. Таким образом, предлагаемый подход вносит следующее ограничение на использование системы: пользователь должен знать набор поддерживаемых системой команд и запрещать их угадывание.

Архитектурной отличительной особенностью подхода является объединение двух различных алгоритмов распознавания речи в единый блок (search engine). В этом блоке (модуле) сначала делается вывод о наличии вводного слова и затем, если это слово присутствует, происходит распознавание самой команды. Таким образом, на выходе главного модуля мы получаем информацию о команде или ее отсутствии.

Алгоритм распознавание ключевого слова. Одним из популярных алгоритмов распознавания ключевого слова является метода скользящего окна (sliding window) [9]. Анализ метода, а также экспериментальные данные показывают, что метод содержит существенный недостаток – большую вычислительную сложность, что создает неудобства и мешает применению системы на практике.

Для устранения данного недостатка был разработан и реализован новый однопроходный алгоритм распознавания ключевого слова, включающий новый алгоритм распространения и новую функцию правдоподобия. Псевдокод алгоритма изображен на рис. 1.

```
для каждого участка входного сигнала o_t ∈ o
    создать новый путь для состояния s_1 ключевого слова
    для всех текущих путей p ∈ P
        для всех возможных переходов s_new из текущего стостояния пути s_current
        если (Фонема[s_new] == Фонема[s_current]) // переход внутри одной фонемы
            осуществить переход p в s_new
            увеличить длину пути p
            пересчитать правдоподобие p
        иначе
            Score = правдоподобие(p)
            если (Score > порог [Фонема[s_current]])
                если (Фонема[s_current] == последняя фонема ключевого слова)
                    вернуть true
                создать новый путь p_new для состояния s_new ключевого слова
                добавить p_new в P

    сокращение путей
вернуть false
```

Рис. 1. Псевдокод алгоритма распознавания ключевого слова

Используемый алгоритм определения ключевого слова использует те же модели, что и алгоритм распознавания команды, но не учитывает вероятности перехода. В каждый дискретный момент (t) этого алгоритма происходит следующее:

1. Стартует новый путь из текущей позиции сигнала (o_t) в начальное состояние ключевого слова (в начальное состояние СММ первой фонемы ключевого слова, s_1).

2. Каждый существующий путь дублируется: один остается в текущем состоянии, второй переходит в следующее состояние ключевого слова.

3. Происходит сокращение путей на основе функции правдоподобия.

Как показатель соответствия разработанный метод использует не только сами вероятности наблюдения $b_j(o_t)$, но и отношение вероятности наиболее подходящего состояния из всего набора моделей и текущего состояния ключевого слова $b_j(o_t) / b_{best}(o_t)$. Функция же правдоподобия основана на показателях худшей модели в ключевом слове, а не на общих показателях всего слова. Экспериментальные данные, приведенные далее, показали, что такой подход уменьшает ошибку, связанную с неправильным срабатыванием на слове, близком по звучанию. Кроме этого, экспериментальные данные показывают, что некоторые фонемы распознаются лучше, чем другие. Поэтому для каждой фонемы ключевого слова экспериментально был определен свой собственный порог срабатывания. Использование более высокого порога для хорошо распознающихся фонем и более низкого для плохораспознающихся позволило существенно улучшить качество распознавания. Кроме этого, для сокращения заведомо ложных путей функция правдоподобия использует специальный общий порог на каждое состояние пути. Путь удаляется из списка текущих путей, если не выполняется хотя бы одно из следующих условий:

1) показатель соответствия всех состояний пути больше общего порога;

2) среднее значение показателя соответствия текущий фонемы больше порога для этой фонемы.

Экспериментальные данные. *Речевая база данных.* Численные эксперименты выполнялись на речевом корпусе slSpeechCorp, разработанном в НИИ СпецЛаб и ООО «Спецлаборатория» [14]. Эта РБД содержит 10 ч речевого материала, записанного 40 людьми возрастом от 18 до 50 лет и транскрибированного вручную. Для транскрибирования использовался алфавит Russian SAMPA (Speech Assessment Methods Phonetic Alphabet) и разделение гласных фонем на ударные (stressed) и безударные (unstressed). Всего 50 фонем.

Для проведения тестирования вся РБД была разбита на два блока: один блок содержал 90 % речевого материала и использовался для обучения (использовался алгоритм Баума-Уелша (Baum-Welch) [5]), другой – 10 % и использовался для тестирования. Блоки содержали различных дикторов.

Конфигурация системы. Для тестирования применялась следующая конфигурация системы.

Детектор голоса состоит из двух частей, которые в сумме дают представление о наличии речевой составляющей в сигнале. Первая часть основана на изменении энергии сигнала, вторая – на периодичности, которая рассчитывается с помощью алгоритма, основанного на методе наименьших квадратов [10].

Для преобразования сигнала [13] в векторы особенностей за основу взят алгоритм, предложенный европейским институтом стандартов телекоммуникации(ETSI) [11], блок-схема которого изображена на рис. 2.

Для сбора особенностей алгоритм использует свойства человеческого восприятия. Кроме того, согласно [12], в него были внесены следующие модификации, направленные на улучшение качества работы: лифтрация (Liftering); вычет среднего кепстрального значения (Cepstral Mean Subtraction (CMS)); нормализация энергии (Energy Normalization). На выходе данного блока для каждого участка сигнала в 25 мс формируется вектор из 39 параметров. Первыми 13-ю из них являются кепстральные коэффициенты (12 mel-frequency cepstral coefficient) и логарифм энергии (logE), а остальными – производные 1-го и 2-го порядка этих коэффициентов (они показывают динамику изменения).

Результаты тестирования. Тестирование проводилось для распознавания слитной речи для словаря в 20 слов. В качестве материала для тестирования работы модуля шумоочистки использовался шум из специального набора шумов NOISEX92 [14]. На первом этапе было проведено сравнение работы алгоритма распознавания слитной речи для трех различных конфигураций модуля шумоочистки:

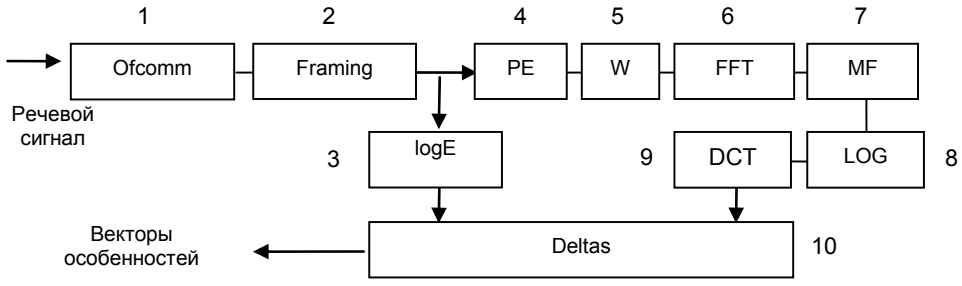

Рис. 2. Блок диаграмма преобразователя сигнала: 1 – компенсация смещения; 2 – разбиение на блоки; 3 – вычисление энергии; 4 – предварительная коррекция; 5 – применение оконной функции; 6- разложение на частотные составляющие; 7 – применение фильтров; 8 – нелинейное преобразование; 9 – косинус преобразования; 10 – вычисление 1-й и 2-й производной по времени

- без шумоочистки;
- с использованием алгоритма шумоочистки OM-LSA;
- с использованием алгоритма шумоочистки OM-LSA и с применением разработанных эвристик.

В табл. 1 приведены результаты экспериментов, анализ которых показывает, что лучшие показатели распознавания показала система, использующая алгоритм OM-LSA с применением разработанных эвристик.

Таблица 1. **Результаты работы алгоритма распознавания слитной речи для разных вариантов модуля шумоочистки**

Вариант модуля шумоочистки	Правильность распознавания, %	
	Уровень шума 25дБ	Уровень шума 5дБ
Шумоочистка отсутствует	99%	25%
OM-LSA	96%	94%
OM-LSA + разработанные эвристики	99,5%	99%

Второй этап тестирования заключался в экспериментальном сравнении разработанного алгоритма распознавания ключевого слова с методом скользящего окна. В качестве показателей для сравнения были выбраны скорость работы и качество распознавания. При этом качество распознавания измерялось двумя величинами:

- правильностью срабатывания и распознавания голосовой команды;
- неправильным срабатыванием на участке сигнала, не содержащем голосовой команды.

Результаты экспериментов (табл. 2) показывают превосходство разработанного метода определения ключевого слова как в плане скорости, так и в плане качества распознавания.

Таблица 2. **Сравнение показателей разработанного алгоритма распознавания ключевого слова с методом скользящего окна**

Показатель	Значение	
	Метод скользящего окна	Разработанный алгоритм
Время работы алгоритма (полученное для входного участка продолжительностью 7, 4с)	20 с	0,5 с
Качество распознавания	86,1% / 5%	98% / 0,01%

Заключение

Отличительной чертой разработанного двухуровневого метода распознавания голосовой команды является разделение речевой команды на две части: ключевого слова и непосредственно команды. Экспериментальные данные подтверждают эффективность такого подхода: срабатывание на ключевом слове – 98%, не на ключевом слове – 0,01%, распознавание команды – 99,5%. Недостатком этого подхода является следующее ограничение: пользователь должен знать набор поддерживаемых системой команд и запрещать их угадывание. Предложенный новый однопроходный алгоритм распознавания ключевого слова и проведенное экспериментальное сравнение разработанного алгоритма с методом скользящего окна [12] показывают превосходство разработанного метода как в плане скорости, так и в плане качества.

Список литературы

1. **Demuynck K.** Extracting, modeling and combining information in speech recognition: PhD thesis, ESAT, 2001.
2. **Rosti I.** Linear gaussian models for speech recognition: PhD thesis, University of Cambridge, 2004.
3. **Couvreur Chr.** Hidden Markov Models and Their Mixtures // DEA Thesis, Department of Mathematics, Catholic University of Louvain. – 1996.
4. **R. Rabiner L.** A tutorial on Hidden Markov Models and selected applications in speech recognition // Proceedings of the IEEE. – 1989.
5. **Morgan N.** Neural Network for Statistical Recognition of Continuous Speech // Proceedings of the IEEE. – 1995.
6. **Xhenyu X.** Comparison and combination of confidence measures in IWR: ISCSLP, 2002.
7. **Hazen T.** Recognition confidence scoring and its use in speech understanding systems // Computer Speech and Language. – 2002.
8. **Mengusoglu E.** Use of acoustic prior information for confidence measure in ASR: European Conference on Speech Communication Technology. – 2005.
9. **Bridle J.** An efficient elastic template method for detecting given words in running speech: British Acoustical Society Meeting. – Apr. 1973.
10. **Tucker R.** Voice activity detection using a periodicity measure // Proceedings of the IEEE. – Vol. 139. – 1992.
11. **European** Telecommunications Standards Institute. ES 201 108 Distributed Speech Recognition Encoding. Proceedings of ETSI, 2003.
12. **Parihar N.** Performance analysis of advances front ends on the Aurora LV evaluation. M.S. Dissertation, Mississippi State University. – 2003.
13. **Koc A.** Acoustic feature analysis for robust speech recognition. M.S. Thesis, Bilkent University. – 2002.
14. **ООО «Спецлаборатория»,** http://www.goal.ru.

Гребнов Сергей Викторович,
Ивановский государственный энергетический университет,
аспирант кафедры программного обеспечения компьютерных систем,
e-mail: Sergei.Grebnov@gmail.com