УДК 004.522

Аналитический обзор методов распознавания речи в системах голосового управления

Гребнов С.В., асп.

Рассматриваются основные методы распознавания речи на основе скрытых марковских моделей, применяемые в системах голосового управления: метод скользящего окна и метод моделей заполнителей. Показаны ограничения применимости методов на практике и актуальность создания нового, более совершенного метода.

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

Analytical survey of speech recognition methods used in voce control systems

Grebnov S.V., postgraduate student

The article describes general methods of speech recognition based on hidden markov models (HMM) used in voice control systems: sliding window models and garbage/filler models. Also showed limitations of these methods and urgency of creation new one.

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

Введение. Построение системы голосового управления является в настоящее время актуальной задачей. Такие системы способны существенно облегчить взаимодействие пользователя с компьютерной системой. Особенно эта идея развита в концепции так называемых «умных домов». Более того, иногда голосовой интерфейс является необходимой компонентой, например, когда речь идет о людях с нарушениями опорно-двигательного аппарата.

Задачей таких систем является выделение и распознавание из потока звукового сигнала (как речевого, так и не речевого) заранее определенного набора речевых команд. Примером такой команды может служить фраза «Включить свет». При этом система не должна реагировать на другие участки речевого сигнала, включая и те, которые содержат отдельные слова предопределенных команд.

При создании подобной системы разработчик сталкивается с определенными проблемами. Во-первых, отсутствие математической модели семантики речевого сигнала, что выражается в том, что для определения семантики речевого сигнала могут применяться только вероятностные и эвристические методы, не дающие точного результата и точность которых обратно пропорциональна количеству смысловых единиц, на которые они рассчитаны. Во-вторых, индивидуальные характеристики говорящего: специфика произношения, акценты, ударения, хезитации[1]. В-третьих, работа со спонтанной речью и необходимость определения присутствия ключевого слова. В-четвертых, различия в акустической обстановке, шумы.

Ниже рассматриваются два основных метода голосового управления: метод скользящего окна и метод моделей заполнителей. Оба метода основаны на алгоритме распознавания речи с помощью скрытых марковских моделей (СММ). Приводится декомпозиция распространенного подхода построения систем голосового управления и принцип их работы; рассматриваются алгоритмы распознавания ключевого слова; в заключении приводятся ограничения применимости методов и делается вывод об актуальности создания нового, более совершенного метода.

Типичная архитектура систем голосового управления. Большинство современных систем автоматизированного распознавания используют модульную архитектура [3] с использованием блока шумоочистки (speech enhancement), детектора голоса (VAD), преобразователя сигнала в векторы особенностей[2] (front end) и главного модуля (search engine), включающего алгоритм распознавания ключевого слова. Цифровой сигнал сначала поступает в модуль шумоочистки, где повышается качество сигнала вследствие удаления шумов и внесенного каналом искажения. Затем детектор голоса выделяет участки сигнала, содержащие речь. Эти участки с помощью модуля преобразования сигнала в векторы особенностей превращаются в наборы коэффициентов, которые поступают в главный модуль, в котором происходит непосредственное определение наличия и распознавания команды. Таким образом, на выходе

[1] Хезитация – речевое колебание, связанное со спонтанностью речи: речевой сбой, заминка в речи, колебание в выборе слова или конструкции. Чаще всего Х выражается в виде паузы, нелексических вставных звуках, «словах-паразитах».

[2] Под вектором особенностей будем понимать фиксированный набор коэффициентов, характеризующий участок звукового сигнала.

главного модуля мы получаем информацию о наличии команды или ее отсутствии.

Метод скрытых марковских моделей. В качестве метода распознавания большинство современных систем используют метод скрытых марковских моделей [3, 4, 5]. Анализ применимости СММ для распознавания речи приводится в [6, 7]. Использование СММ для распознавания речи базируется на следующих предположениях: речь может быть разбита на сегменты (состояния), внутри которых речевой сигнал может рассматриваться как стационарный, переход между этими состояниями осуществляется мгновенно; вероятность символа наблюдения, порождаемого моделью, зависит только от текущего состояния модели и не зависит от предыдущих. Чаще всего используются СММ с тремя состояниями (рис. 1).

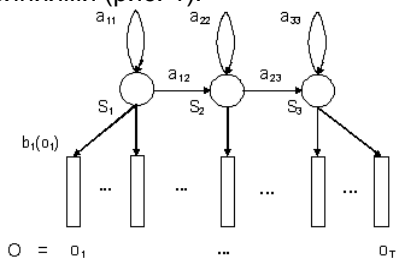

Рис. 1. СММ с тремя состояниями

СММ представляет собой конечный автомат, изменяющий свое состояние в каждый дискретный момент времени t. Переход из состояния s_i в состояние s_j осуществляется случайным образом с вероятностью a_{ij}. В каждый дискретный момент времени модель порождает вектор наблюдений o_t (который в конкретной задаче является вектором особенностей, полученным в преобразователе сигнала) с вероятностью $b_j(o_t)$. Распределение плотности вероятности наблюдений моделируется конечной гаусовской смесью с четырьмя компонентами. Каждая такая модель обозначает один из звуков русского языка или отсутствие звука (одна из моделей).

Алгоритмы распознавания ключевого слова [8, 9, 10] используют эти модели для определения команд в потоке речи. Наиболее часто эта задача решается с помощью метода скользящего окна (sliding window) [11] и метода моделей-заполнителей (filler models) [12].

Метод скользящего окна. Суть метода скользящего окна [11] заключается в определении вхождения ключевого слова с помощью алгоритма Витерби(Viterbi) [5], который широко применяется для распознавания слитной речи (CSR). Этот алгоритм решает следующую задачу: дан вектор наблюдений (o), требуется определить наиболее подходящую последовательность СММ (s) и переходов между их состояниями для этого вектора наблюдений (рис. 2). Далее будем называть такую последовательность *путем*[3]. Так, на рис. 2 изображены все возможные пути для данного участка сигнала и определенной последовательности СММ; утолщенной линией обозначен наиболее вероятный путь. Так как ключевое слово может начинаться и заканчиваться в любом месте сигнала, то этот метод перебирает все возможные пары начала и конца вхождения ключевого слова и находит самый вероятный путь для ключевого слова и этого отрезка, как если бы ключевое слово присутствовало в нем. Для каждого найденного вероятного пути ключевого слова применяется функция правдоподобия, основанная на срабатывании, если значение пути, рассчитанное в соответствии с применяемым методом оценки пути, больше предопределенного значения. Часто для оценки пути используется значение вероятности, полученное с помощью алгоритма Витерби.

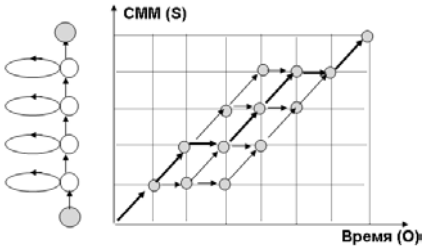

Рис. 2. Пример работы алгоритма Витерби (утолщенная линия соответствует наиболее вероятной последовательности СММ)

Главным недостатком такого подхода является то, что он перебирает все возможные варианты вхождения ключевого слова, что создает большую вычислительную сложность. Кроме этого, метод распознавания команды на основе этого алгоритма заключается в применении его 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 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 же из-за неправильного процесса обучения.

Если второе ограничение можно устранить за счет правильного выбора ключевого слова и качественной РБД, то вычислительную

[3] Под путем будем понимать возможную последовательность СММ и их состояний для определенного участка сигнала

сложность изменить не удастся. Тем самым метод может применяться только в системах голосового управления с небольшим словарем команд, которые не требуют работы в режиме реального времени или в системах, которые имеют значительные вычислительные ресурсы (суперкомпьютеры и др.).

Метод моделей заполнителей. Для алгоритмов распознавания ключевого слова слово для распознавания представляется встроенным в инородную речь. На этом основании методы моделей заполнителей [12] обрабатывают эту инородную речь с помощью явного моделирования инородной речи за счет второстепенных моделей. Для этого в словарь системы распознавания добавляются «обобщенные» слова. Роль этих слов в том, чтобы любой сегмент сигнала незнакомого слова или неречевого акустического события был распознан системой как одно слово или цепочка из обобщенных слов. Для каждого обобщенного слова создается и обучается акустическая модель на корпусе данных с соответствующими размеченными сегментами сигнала.

На выходе из декодера выдается цепочка, состоящая из слов словаря (ключевых слов) и обобщенных слов. Обобщенные слова затем отбрасываются, и оставшаяся часть цепочки считается результатом распознавания.

Недостатком подхода с использованием слов-заполнителей является высокая вероятность ошибки, когда ключевые слова распознаются как обобщенные. Кроме этого, встает и вопрос об оптимальном выборе алфавита обобщенных слов. Это объясняется тем, что пространство акустических событий, моделируемое альтернативными моделями, очень большое и сложное, поэтому обучение целевых и альтернативных моделей играет важную роль в повышении эффективности метода. В итоге подготовка моделей заполнителей становится нетривиальным процессом, нацеленным на определенный набор команд. Это не дает возможности динамически изменять словарь ключевых слов с сохранением прежних показателей распознавания.

Заключение

Рассмотренные основные методы распознавания речи на основе скрытых марковских моделей: метод скользящего окна [11] и метод моделей заполнителей [12], – применяются в системах голосового управления и имеют определенные недостатки: первый метод – большую вычислительную сложность; второй – требует подробного дополнительного моделирования посторонней речи.

Эти недостатки создают неудобства и мешают применению систем голосового управления на практике. Таким образом, разработка нового алгоритма распознавания речи для систем голосового управления является актуальной задачей в настоящее время. Новый метод не должен требовать трудоемкого дополнительного моделирования посторонней речи и должен иметь низкую вычислительную сложность, которая бы позволяла применять его в режиме реального времени.

Список литературы

1. **Rose R.** Robust speech recognition techniques applied to a speech in noise task: European Conference on Speech Communication and Technology, Aarlborg, Denmark, 3–7 Sept, 2001.
2. **Ahadi S.** An Efficient front-end for automatic speech recognition // IEEE Trans. on Speech and Audio Processing. – 2003.
3. **Demuynck K.** Extracting, modeling and combining information in speech recognition // PhD thesis, ESAT. – 2001.
4. **Rosti I.** Linear gaussian models for speech recognition // PhD thesis, University of Cambridge. – 2004.
5. **Couvreur Chr.** Hidden Markov Models and Their Mixtures // DEA Thesis, Department of Mathematics, Catholic University of Louvain. – 1996.
6. **R. Rabiner L.** A tutorial on Hidden Markov Models and selected applications in speech recognition // Proceedings of the IEEE. – 1989.
7. **Morgan N.** Neural Network for Statistical Recognition of Continuous Speech // Proceedings of the IEEE. – 1995.
8. **Xhenyu X.** Comparison and combination of confidence measures in IWR // ISCSLP. – 2002.
9. **Hazen T.** Recognition confidence scoring and its use in speech understanding systems // Computer Speech and Language. – 2002.
10. **Mengusoglu E.** Use of acoustic prior information for confidence measure in ASR: European Conference on Speech Communication Technology. – 2005.
11. **Bridle J.** An efficient elastic template method for detecting given words in running speech // British Acoustical Society Meeting, Apr. – 1973.
12. **Higgins A.** Keyword recognition using template concatenation. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP, 1985.

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