ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 004.934
DOI 10.21685/2072-3059-2019-2-1

И. В. Алексеев, М. А. Митрохин

СОВРЕМЕННЫЕ МЕТОДЫ РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ ПОСТРОЕНИЯ ГОЛОСОВОГО ИНТЕРФЕЙСА УПРАВЛЕНИЯ СИСТЕМАМИ СПЕЦИАЛЬНОГО НАЗНАЧЕНИЯ

Аннотация.

Актуальность и цели. Объектом исследования являются современные технологии распознавания речи. Предмет исследования – методы построения и обучения систем распознавания речи. Целью работы является анализ современных технологий распознавания речи на примере некоторых систем для определения возможности их применения в голосовом интерфейсе управления системами специального назначения.

Материалы и методы. Исследования выполнены с использованием методов теории вероятностей и методов распознавания образов.

Результаты. Проведен анализ требований и ограничений функционирования интерфейсов управления системами специального назначения. Рассмотрены основные аспекты реализации систем распознавания речи и некоторые особенности различных технологий определения структурных единиц речи.

Выводы. Рассмотренные технологии распознавания речи потенциально применимы в интерфейсах управления специальных систем, но требуются дополнительные исследования по оценке их эффективности.

Ключевые слова: пользовательский интерфейс, речевой интерфейс, скрытые марковские модели, нейронные сети, распознавание речи.

I. V. Alekseev, M. A. Mitrokhin

MODERN SPEECH RECOGNITION METHODS FOR CONSTRUCTING A VOICE-CONTROL INTERFACE FOR SPECIAL PURPOSE SYSTEMS

Abstract.

Background. The object of the research is modern technologies of speech recognition. The subject of the study is methods of constructing and teaching speech recognition systems. The purpose of the work is to analyze modern speech recogni-

tion technologies using the example of some systems to determine the possibility of their use in the voice interface of management of special purpose systems.

Materials and methods. Researches were conducted with the use of methods of probability theory and methods of pattern recognition.

Results. The analysis of the requirements and limitations of the operation of the management interfaces for special purpose systems is carried out. The main aspects of the implementation of speech recognition systems and some features of various technologies for determining the structural units of speech are considered.

Conclusions. The considered speech recognition technologies are potentially applicable in the management interfaces of special systems, but further research is required to evaluate their effectiveness.

Keywords: user interface, speech interface, hidden markov model, neural networks, speech recognition.

Введение

С момента появления электронно-вычислительных машин их пытались применять в различных областях, в том числе для использования в человеко-машинных системах специального назначения, к которым относятся автоматизированные системы управления, бортовые системы различных технических средств, а также системы массового обслуживания. Такие системы находят применение в управлении транспортными потоками, при обеспечении безопасности важных объектов, для управления технологическим и иным оборудованием. К этим системам всегда предъявлялись повышенные требования по надежности и безопасности [1, 2].

На сегодня при разработке систем специального назначения с большим числом задач порядка нескольких десятков или сотен и высоким уровнем автоматизации особое внимание уделяют пользовательскому интерфейсу в целях обеспечения максимального удобства и эффективности пользования им. С учетом особенностей функционирования систем специального назначения и требований, предъявляемых к их пользовательским интерфейсам, можно сделать вывод, что дальнейшее повышение эффективности взаимодействия оператора и вычислительной системы возможно при использовании наиболее естественных для человека способов обмена информацией – жестов, письма и речи. Так, уже обеспечивается поддержка ввода с сенсорных панелей. Однако наиболее привлекательным для использования в таких системах и наиболее перспективным является голосовой интерфейс, потому что это наиболее естественный и устойчивый способ общения. Крупнейшие мировые компании в своих разработках предлагают речевой интерфейс как альтернативу графическому. Более того, в системах специального назначения также давно применяются различные системы голосового оповещения оператора о различных событиях для повышения эффективности обратной связи. Ввиду этого для ускорения работы оператора в системе специального назначения, повышения надежности и эффективности его работы целесообразно применение прямой связи оператора с ЭВМ посредством голоса и получения в итоге полноценного речевого интерфейса.

1. Требования к интерфейсам управления и обобщенная структура систем распознавания речи

Диалог человека-оператора с вычислительной техникой в любых системах представляет собой двусторонний обмен данными и командами (рис. 1).

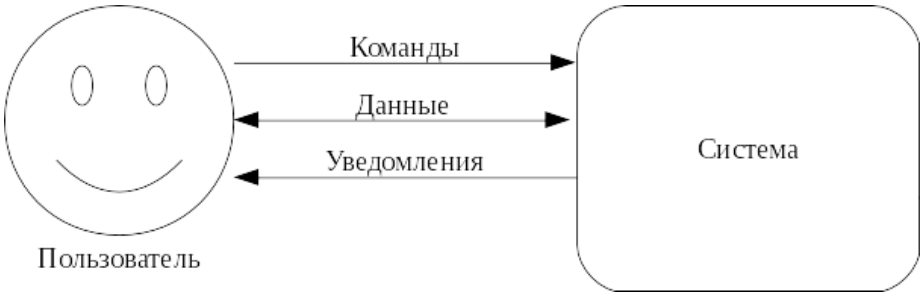

Рис. 1. Обмен информацией (диалог) оператора и ЭВМ

Для реализации речевого интерфейса в первую очередь необходима надежная эффективная подсистема (в составе рассматриваемой системы специального назначения) распознавания речи. С учетом повышенных требований к системам специального назначения эта подсистема должна задействовать лишь локальные вычислительные ресурсы и обладать высокой эффективностью.

Голосовой интерфейс программного обеспечения системы специального назначения, как и любой другой, должен удовлетворять всем требованиям к интерфейсам (естественность, интуитивность, дружественность, согласованность, адаптивность, непротиворечивость, неизбыточность), а также, ввиду своей специфики, некоторым дополнительным:

1) корректное распознавание. Система голосового управления должна верно распознавать все команды, в том числе при наличии различного рода помех;

2) высокая скорость распознавания. Система, которая распознает фразу слишком долго, может не только не повышать эффективность работы оператора, но и понижать производительность всей системы в целом;

3) оптимальное использование машинного времени и оперативной памяти. Поскольку распознавание речи – ресурсоемкая задача, то максимальные затраты локальных вычислительных ресурсов алгоритмом распознавания должны быть ограничены.

Все существующие системы распознавания речи имеют схожую структуру [3] (рис. 2).

Рис. 2. Структура систем распознавания речи

Технология распознавания речи следующая. Речевой сигнал читается из аудиофайла либо записывается с помощью аудиокарты. Полученный сигнал передается на внешний интерфейс (front-end) системы распознавания речи, реализующий функции получения входного сигнала, его предварительной обработки, разбиения на кадры и выделения признаков.

Обработка сигнала включает автоматическую регулировку уровней сигнала, фильтрацию шумов, выделение эха, обнаружение наличия речевого фрагмента (voice activity detection, VAD), определение конца фразы. Затем

подготовленный сигнал разбивается на короткие кадры (frames) продолжительностью от 10 до 300 мс в различных системах.

После раскадровки для каждого фрейма определяется вектор признаков, при этом используется, как правило, быстрое преобразование Фурье. Однако немногие системы распознавания речи используют такие параметры, как единственные. Обычно вместе с этими характеристиками используются другие, например кепстральные характеристики и их производные. Основным методом получения признаков для дальнейшего распознавания речи является метод выделения мел-кепстральных коэффициентов (Mel Frequency Cepstral Coefficients, MFCC) [2, 3], позволяющий по вычисленному кепстру использовать сигнал возбуждения после фильтра речевого тракта, что значительно повышает точность распознавания речи в дикторонезависимых системах.

Поток признаков поступает на вход акустического модуля системы, где происходит распознавание структурных единиц речи. Акустический модуль, реализующий конкретную технологию распознавания, – основной в системе. Каждый вектор признаков, полученный от внешнего интерфейса системы, сравнивается с имеющимися акустическо-фонетическими образцами, хранящимися в базе данных.

2. Основные методы распознавания структурных единиц речи

В акустическом модуле в основном применяют глубокие рекуррентные нейронные сети или скрытые марковские модели на основе гауссовых смесей распределений.

Метод скрытых марковских моделей требует представления слов в виде последовательности состояний процесса, соответствующих акустическим фрагментам (единицам) речи и в пространстве признаков описываемых функциями плотности вероятности. Математический аппарат марковских цепей, используемый в применении к случайным процессам, известен с 1960-х гг. Первые применения его к распознаванию речи относятся к 1970-м гг. [2].

Цепь Маркова для представления речи выглядит как граф состояний S_i с переходами в дискретные моменты времени. При этом вероятность a_{ij} перехода в следующее состояние определяется только предыдущим состоянием процесса (рис. 3) [2].

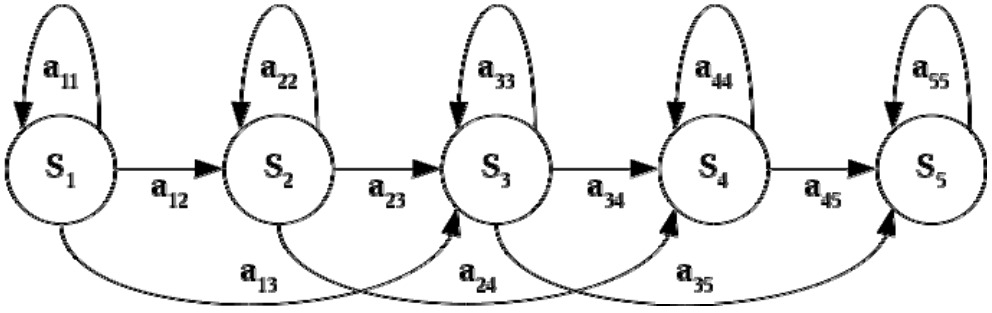

Рис. 3. Структура скрытой марковской модели в применении к речи

Обучение Марковской цепи заключается в определении функций плотности вероятности состояний и матрицы вероятности переходов по априорным данным. В качестве априорных используются аудиозаписи речевых сигна-

лов и соответствующие им тексты. Некоторые записи, используемые на первом этапе обучения, размечены специалистами в области фонетики и лингвистики на структурные единицы, для которых будут создаваться марковские модели. Эти записи разбиваются на короткие кадры и далее для каждой фонемы переводятся в последовательность векторов-признаков. По множеству признаков для фонем строятся функции плотности вероятности, которые можно аппроксимировать, например, конечным количеством гауссовых функций:

$$p(\mathbf{x} \mid \omega_i) = \frac{1}{(2\pi)^{n/2} \left| \sum_i \right|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^t \sum_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right),$$

где x – классифицируемый элемент; ω_i – один из n возможных классов; $\sum_i$ – матрица ковариации данного класса; $\boldsymbol{\mu}_i$ – математические ожидания данного класса.

Для обучения может использоваться EM-алгоритм либо его частные реализации. Реализация EM-алгоритма представляет собой итерации двух шагов [3].

На **E-шаге** оценивается вектор ожидаемых значений (**e**xpectation) скрытых переменных G по вектору текущих значений параметров θ.

Для этого вычисляется условная вероятность $p(x, \theta_j)$ того, что объект x получен из j-й компоненты смеси: $p(x, \theta_j) = p(x) P(\theta_j \mid x) = \omega_j p_j(x)$.

Из формулы Байеса по выражению

$$P(\theta_j \mid x_i) = g_{ij} = \frac{\omega_j p_j(x_i)}{\sum_{s=1}^{k} \omega_s p_s(x_i)}$$

определяется апостериорная вероятность того, что обучающий объект x_i получен из j-й компоненты смеси. При этом

$$\sum_{j=1}^{k} g_{ij} = 1 \text{ для любого } i = 1, \ldots, m$$

есть полная вероятность принадлежности объекта x_i одной из k компонент смеси.

Величины $p(x, \theta_j)$ и g_{ij} используют в качестве скрытых переменных.

На **M-шаге** (**m**aximization) осуществляется максимизация логарифма полного правдоподобия:

$$Q(\Theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} \omega_j p_j(x_i) \to \max_{\Theta}.$$

В результате решения оптимизационной задачи Лагранжа находим

$$\omega_j = \frac{1}{m} \sum_{i=1}^{m} g_{ij}, \; j = 1, \ldots k.$$

$$\theta_j = \arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i, \theta),\ j = 1,...k\ .$$

Таким образом, на М-шаге определяются веса компонент смеси ω_j и оцениваются их параметры θ_j путем решения *k* независимых оптимизационных задач.

Для смесей нормальных распределений, обычно используемых в распознавании речи, результаты Е- и М-шага алгоритма запишутся следующим образом:

$\theta = (\omega_1,...,\omega_k;\ \mu_1,...,\mu_k;\ \sigma_1,...\sigma_k)$ – вектор параметров,

$$p_j(x) = N(x, \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2} \right)$$ – плотность распределения.

Е-шаг:

$$g_{ij} = \frac{\omega_j N(x_i, \mu_j, \sigma_j)}{\sum_{s=1}^{k} \omega_s N(x_i, \mu_s, \sigma_s)}\ .$$

М-шаг:

$$\omega_j = \frac{1}{m} \sum_{i=1}^{m} g_{ij}\ ,\ \mu_j = \frac{1}{m\omega_j} \sum_{i=1}^{m} g_{ij} x_i\ ,\ \sigma_j^2 = \frac{1}{m\omega_j} \sum_{i=1}^{m} g_{ij} (x_i - \mu_j)^2, j - 1,...,k\ .$$

На этом же этапе обучения скрытых марковских моделей после выполнения ЕМ-алгоритма программа анализирует длительности фрагментов и строит их гистограммы для вычисления вероятностей выхода из состояний, соответствующих данным структурным единицам.

Далее, зная оценки функций плотности вероятности и вероятностей перехода, используют алгоритм Баума – Уэлша или Витерби для максимизации вероятностей порождения априорных последовательностей признаков цепочками состояний [2].

На последнем этапе привлекаются неразмеченные речевые сигналы. Полученные параметры состояний итеративно применяют к фрагментам сигналов для автоматического разбиения на структурные единицы. Такой метод называется «принудительным выравниванием» (forced alignment). Далее выровненные фрагменты используются для дальнейшего уточнения параметров Марковской цепи.

Марковская модель на основе гауссовских смесей является одним из наиболее распространенных и оптимизированных аппаратов распознавания речи [4]. Поэтому эта технология распознавания остается одной из наиболее эффективных.

В последние десятилетия с развитием вычислительной техники и систем параллельных вычислений активно разрабатывалась и совершенствовалась нейросетевая технология распознавания образов, позволив повысить точность распознавания речи по сравнению с классическими моделями.

Дальнейшие успехи в распознавании речи связаны с использованием глубоких нейронных сетей, состоящих из многих слоев. Одним из методов обучения многослойных нейронных сетей является послойное обучение [1], в том числе с использованием автоэнкодерных сетей. Процесс обучения выглядит следующим образом. В качестве целевой функции для первого скрытого слоя рассматривается входной вектор признаков. Исходный вектор может содержать несколько последовательных мел-спектральных или мел-кепстральных векторов-признаков. Следующий слой нейронной сети обучают таким же образом воспроизводить выходные сигналы предыдущего слоя. Так обучают до 5–7 слоев. После того как инициализация первых слоев проведена, применяют известный алгоритм обратного распространения ошибки для всей сети с целевой функцией, отражающей принадлежность входного сигнала к соответствующему трифону.

Данный подход многократно превосходит классический подход на основе гауссовых смесей по эффективности и точности: многослойная сеть, обученная на речевом материале в 309 ч речи, показала лучшие результаты, чем метод с гауссовыми смесями, обученный на 2000 ч речи [2].

На выходе акустического блока вне зависимости от его реализации получают поток распознанных структурных единиц речи с вероятностями правдоподобия, который поступает далее на вход лингвистического модуля. Задача этой подсистемы – найти наилучшее соответствие входного потока фонем и заданных в словаре слов и затем соответствие потока слов употребимым в языке фразам. В зависимости от объема используемого словаря и действующих синтаксических правил применяются различные стратегии поиска и отсева. При этом также часто используется вероятностная система сравнения результатов [3].

Как правило, данный блок содержит в себе реализации словаря и лингвистической модели. На основе последней часто строят конечный автомат, наиболее вероятная цепочка переходов между состояниями которого и есть произнесенная пользователем фраза. В настоящее время также находят применение другие лингвистические модели, например, на основе нейронных сетей.

Заключение

Таким образом, в настоящее время существует несколько типовых реализаций систем распознавания речи. Основные их отличия заключаются в используемой технологии распознавания структурных единиц речи. Наиболее проработанной в настоящее время является технология, основанная на скрытых марковских моделях. Однако она имеет потенциальные ограничения на объем данных для описания модели при использовании в системах специального назначения. Перспективным направлением является использование нейронных сетей, однако данные технологии пока находят применение лишь при наличии больших объемов размеченных записей речи и требуют высокопроизводительных вычислительных ресурсов. Обе технологии потенциально могут применяться в голосовом интерфейсе управления систем специального назначения, но для окончательных выводов требуется проведение исследований эффективности распознавания при ограничениях на вычислительные ресурсы и особенностей функционирования в условиях работы операторов систем специального назначения.

Библиографический список

1. **Хайкин, С.** Нейронные сети: полный курс : пер. с англ. / С. Хайкин. – 2-е изд., испр. – Москва : Вильямс, 2006. – 1104 с.
2. **Тампель, И. Б.** Автоматическое распознавание речи – основные этапы за 50 лет / И. Б. Тампель // Научно-технический вестник информационных технологий, механики и оптики. – 2015. – Т. 15, № 6. – С. 957–968.
3. **Huang, X.** Spoken language processing: a guide to theory, algorithm, and system development / X. Huang, A. Acero. – Prentice Hall, 2001. – 1008 p.
4. **Bourlard, H.** Towards increasing speech recognition error rates / H. Bourlard, H. Hermansky, N. Morgan // Speech Communication. – 1996. – Vol. 18, № 3. – P. 205–231. – DOI 10.1016/0167-6393(96)00003-9

References

1. Khaykin S. *Neyronnye seti: polnyy kurs: per. s angl.* [Neural networks: translation from English]. 2nd ed., corr. Moscow: Vil'yams, 2006, 1104 p. [In Russian]
2. Tampel' I. B. *Nauchno-tekhnicheskiy vestnik informatsionnykh tekhnologiy, mekhaniki i optiki* [Scientific and technical bulletin of informational technologies, mechanics and optics]. 2015, vol. 15, no. 6, pp. 957–968. [In Russian]
3. Huang X., Acero A. *Spoken language processing: a guide to theory, algorithm, and system development*. Prentice Hall, 2001, 1008 p.
4. Bourlard H., Hermansky H., Morgan N. *Speech Communication*. 1996, vol. 18, no. 3, pp. 205–231. DOI 10.1016/0167-6393(96)00003-9

Алексеев Илья Владимирович
аспирант, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

E-mail: aius@pnzgu.ru

Alekseev Il'ya Vladimirovich
Postgraduate student, Penza State University (40 Krasnaya street, Penza, Russia)

Митрохин Максим Александрович
доктор технических наук, заведующий кафедрой вычислительной техники, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

E-mail: vt@pnzgu.ru

Mitrokhin Maksim Aleksandrovich
Doctor of engineering sciences, head of sub-department of computer engineering, Penza State University (40 Krasnaya street, Penza, Russia)

Образец цитирования:

Алексеев, И. В. Современные методы распознавания речи для построения голосового интерфейса управления системами специального назначения / И. В. Алексеев, М. А. Митрохин // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2019. – № 2 (50). – С. 3–10. – DOI 10.21685/2072-3059-2019-2-1.