РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР им.А.А.ДОРОДНИЦЫНА

МОДЕЛИ, МЕТОДЫ, АЛГОРИТМЫ И АРХИТЕКТУРЫ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР
им.А.А.ДОРОДНИЦЫНА
МОСКВА 2006

УДК 621.391;518.5

Ответственный редактор
доктор физ.-матем. наук В.В. Рязанов

В представленном сборнике работ рассматриваются организационные, теоретические и практические аспекты проблемы создания систем распознавания речи. В работе "Информационная модель системы документационного обеспечения НИОКР"авторы предлагают простую, интуитивно понятную систему документационного обеспечения (СДО) НИОКР в области инженерии наукоемких программных систем. В работе "Гибридные модели: скрытые марковские модели и нейронные сети, их применение в системах распознавания речи"представлен обзор современных моделей и методов распознавания речи. В работе "Выделение незнакомых слов и акустических событий при распознавании речи"анализируются существующие методы обнаружения незнакомых слов на основе мер сходства. В работе "Обнаружение новых слов и невербальных событий при распознавании речи"рассмотрен алгоритм GdAlg формирования оценок достоверности результатов распознавания, основанных на отношении правдоподобия. Сборник работ может быть полезен руководителям и разработчикам сложных наукоемких программных систем, в частности систем искусственного интеллекта и распознавания речи.

Рецензенты: В.В.Стрижов,
В.Ф.Огарышев

Научное издание

Предисловие

Для коллективов научно-технической интеллигенции переход экономики на инновационный путь развития фактически означает переход от стабильного государственного обеспечения к конкурентной борьбе за различные источники финансирования, главным из которых должен стать коммерчески значимый результат работы коллектива, т.е. инновация.

В широком смысле под инновациями понимается прибыльное использование новшеств в виде новых технологий, видов продукции или услуг, организационно-технических или социально-экономических решений в различных областях деятельности. В более узком аспекте, научно-технические инновации представляют материальное воплощение новых знаний, открытий, изобретений и научно-технических разработок в производстве с целью их коммерческой реализации для удовлетворения определенных запросов потребителей.

Современная глобализация экономической деятельности и обострение конкуренции делают необходимым использование специальных знаний методологии и методов творческой работы для решения задач научной, технической и предпринимательской деятельности. Проблема повышения продуктивности творческого мышления непосредственно связана с проблемой повышения уровня организованности знаний и является одной из ключевых проблем научно-исследовательской, опытно-конструкторской или инновационной работы.

В представленном сборнике работ рассматриваются организа-

ционные, теоретические и практические аспекты проблемы создания систем распознавания речи. В работе "Информационная модель системы документационного обеспечения НИОКР"авторы предлагают простую, интуитивно понятную систему документационного обеспечения (СДО) НИОКР в области инженерии наукоемких программных систем. В работе рассматривается информационная модель процесса разработки наукоемких программных систем, её отображение в модель системы документационного обеспечения НИОКР, описываются концептуальная и логическая модели, архитектура и реализация СДО НИОКР. Целью предложенной информационной технологии является повышение уровня организации представления знаний, получаемых в ходе выполнения НИОКР.

В работе "Гибридные модели: скрытые марковские модели и нейронные сети, их применение в системах распознавания речи"представлен обзор современных моделей и методов распознавания речи. Рассматриваемая в статье гибридная модель реализована во многих системах распознавания слитной речи с большими словарями и продемонстрировала лучшие результаты по сравнению с системами построенными на основе каждой из моделей, составляющих гибрид.

В работе "Выделение незнакомых слов и акустических событий при распознавании речи"анализируются существующие методы обнаружения незнакомых слов на основе мер сходства, приводится численное исследование эффективности предложенного алгоритма выявления незнакомых слов на основе оценок правдоподобия для наблюдаемого речевого сигнала при заданном

множестве акустико-фонетических моделей. В работе предложен метод оценки априорных дисперсий счетов, выбора априорного порога, а также процедура адаптации среднего и порога, в соответствии с наблюдаемыми значениями интегральных счетов. Показано, что эффективность предложенных априорных оценок счетов и метода адаптации порога практически соответствует использованию апостериорных оценок пороговых счетов на настроечной выборке.

В работе "Обнаружение новых слов и невербальных событий при распознавании речи"приведен обзор современных методов проверки корректности распознавания речи. На основе этого обзора реализован алгоритм GdAlg формирования оценок достоверности результатов распознавания, основанных на отношении правдоподобия. Приведено подробное описание алгоритма и численных результатов его применения на корпусе данных Favor.

В информационном обществе уже давно существует потребность в автоматическое распознавание речи. Однако несмотря на многочисленные прогнозы известных специалистов, широкого использования технологий распознавания речи в повседневной деятельности пока не наблюдается. Одной из основных причин этого является недостаточная адекватность и робастность существующих методов распознавания речи. Надеемся, что предлагаемый сборник работ окажется полезным для руководителей и разработчиков наукоемких программных систем, в частности систем искусственного интеллекта или распознавания речи.

Информационная модель системы документационного обеспечения НИОКР

А.В.Чичагов, В.Я.Чучупал, К.А.Маковкин

Аннотация

В работе рассматривается модель процесса разработки наукоемких программных систем, ее отображение в модель системы документационного обеспечения (СДО) НИОКР. Описываются концепция, архитектура и реализация СДО НИОКР. Ключевые слова: методология, методика, НИОКР, НИР, модель, ПС, НПС, артефакт, документация, обеспечение.

Введение

В повседневной практике словосочетание «научно-исследовательская опытно-конструкторская работа» (НИОКР), как и слово «работа», обычно используют в нескольких различных значениях, среди которых отметим следующие:

- *процесс*, т. е. интеллектуальная трудовая деятельность коллектива авторов (исполнителей),
- *результат*, т. е. некоторый интеллектуальный объект или продукт, который можно использовать независимо от авторов работы.

Интеллектуальный продукт может иметь форму материального изделия («hardware»), программного обеспечения («software»), документации («docware»), услуги (консультации, обучения, т. е. полезной информации, представленной в коммуникативном виде) или их сочетания. При этом термин "программное обеспечение"охватывает широкий спектр программных изделий различного функционального назначения и качества исполнения от прототипов с низким уровнем надежности до высокотехнологичных изделий, готовых к длительной эксплуатации внешними пользователями.

Заказчиками и/или потребителями НИОКР обычно являются государственные организации или коммерческие предприятия, цель которых наладить серийный выпуск, внедрение/сбыт и обслуживание наукоемкого продукта, поэтому основную часть стоимости НИОКР составляет информация, т. е. специальные знания («docware» и «software.applogics»), а не стандартная или физическая/материальная часть («software.interface» и «hardware»). Выражаясь более просто, можно сказать, что результатом НИОКР является описание способа разработки/изготовления наукоемкого технического (программного) изделия и, возможно, экземпляр (прототипа) изделия.

Началом любой работы является мысль или идея, которая иногда материализуется в виде контракта или гранта. Ответ на вопрос, как возникает идея и каким образом ее легко можно материализовать, авторам в настоящее время детально неизвестен, поэтому в данной работе не рассматривается. Однако известно, если работу рассматривать в методологическом аспекте

(в отличие от финансово-экономического аспекта), то «результат работы» соответствует «процессу работы». Из этого утверждения следует вывод, что одним из путей повышения качества *результата* работы является рационализация *процесса* работы.

В настоящей работе рассматривается проблема создания наукоемких программных или программно-аппаратных систем (НПС/НПАС), представителем которых является, например, класс систем цифровой обработки, анализа, синтеза и распознавания речевых сигналов. Данный класс систем отличается от класса сложных программных или программно-аппаратных систем (СПС/СПАС) тем, что в научно-техническом задании на разработку НПС неопределенными артефактами являются не только архитектура и алгоритмы функционирования проектируемой системы, но также методы и модели естественных процессов, лежащих в основе этих алгоритмов.

Отличительной особенностью этого класса систем является также то, что для описания НПС требуется более тщательная проработка вопроса «представления знаний», т. е. представления огромного массива информации состоящей из артефактов (документов) различных видов. Другими словами, представления математических моделей, методов, алгоритмов, архитектур, программ, тестов и артефактов других видов, без организации представления которых, включая сохранение семантических связей между артефактами и актуальную информацию о выполненных модификациях артефактов, реализовать высококачественные и надежные НПС представляется проблематичным.

Модель процесса разработки НПС можно представить в виде модели временного «рискового предприятия», т. е. человеко-машинной среды, в рамках которой выполняется интеллектуальный трудовой процесс (НИОКР). Несмотря на широкое применение средств вычислительной техники в трудовых процессах этого вида, рациональная организация (унификация) представления знаний, полученных в ходе выполнения НИОКР, в настоящее время практически отсутствует. Это обстоятельство, по мнению авторов, является тормозом на пути развития инновационной экономики.

Действительно, существующая документация результатов интеллектуальной деятельности по сути является бумажно-ориентированной и обычно представляет «дерево папок файлов». «Бумажность» состоит в том, что документация этого вида поддерживает логическую структурированность, но не поддерживает семантическую связность («дальние логические связи») артефактов НИОКР. Многие артефакты НИОКР (точнее, практически все артефакты) связаны между собой таким образом, что изменение одного артефакта требует модификации семантически связанной цепочки артефактов. Например, изменение артефакта «модель процесса» требует модификации артефактов «метод решения», «алгоритм», «программа», а также ряда других артефактов («тест», «инструкция», «план работ» и пр.).

Сложный целостный документ/описание сложного целостного объекта/системы можно представить в виде системы, состоящей из семантически связанных разделов документа. Причем каждый раздел также может состоять из подразделов или

представлять некоторый артефакт, иначе говоря, целесообразное унифицированное описание, основанное на специальных знаниях. Рациональная структуризация (декомпозиция) документа может существенно сократить колличество «дальних логических связей» между разделами, но не ликвидировать их полностью, т. к. в этом случае разрушается целостность описания и, следовательно, адекватность восприятия/представления или распознавания субъектом реального объекта по заданному описанию.

Поддержка «папка-ориентированной» документации НИОКР выше «среднего» объема как при частой, так и при редкой модификации артефактов в целостном (непротиворечивом) состоянии является серьезной проблемой, которую обычно пытаются решить, используя неформальные или ненадежные способы. В настоящей работе предлагается надежный способ решения обозначенной проблемы, а именно, создание простой, интуитивно понятной *системы поддержки* или системы документационного обеспечения (СДО) НИОКР.

В работе рассматривается информационная модель процесса разработки НПС, ее отображение в модель системы документационного обеспечения НИОКР, описываются концепция, архитектура и опытная реализация (прототип) СДО НИОКР. Целью предлагаемой информационной технологии является повышение уровня организации представления знаний, получаемых в ходе выполнения НИОКР. Вопросы, касающиеся обработки и использования знаний, в настоящей работе не рассматриваются.

Модель процесса разработки НПС

Методология разработки программных систем (ПС) достаточно хорошо известна и включает следующие виды работ: постановка задачи, анализ (изучение) проблемы, проектирование, конструирование/кодирование и испытание разработанного изделия. В качестве основного системообразующего признака здесь используется вид или «характер работы», т. е. специализация или язык, который используется и на котором представляется результат работы. При этом термин «специализация» отражает как формальные, так и неформальные синтаксическую, семантическую и прагматическую части используемого языка (точнее, терминологически связного подмножества естественного/русского языка).

Если указанные виды работ выполняются последовательно, то их называют этапами, а соответствующая методика разработки носит название каскадного процесса или водопадной модели процесса разработки ПС. Эта модель представляет линейный процесс без возвращений к пройденным этапам. Основная слабость указанной модели - отсутствие механизма исправления ошибок, которые могут допустить авторы/разработчики ПС. В настоящее время данная модель разработки используется в основном в учебно-образовательных целях, а также «вне» области инженерии ПС.

На практике разработку «сложных программных систем» (СПС) или «наукоемких программных систем» (НПС) проводят за несколько (десятков или сотен) циклов. В начале

разрабатывается система взаимодействия компонент (система управления), а на последующих циклах добавляется требуемая функциональность в соответствии с определенным приоритетом. Такая методика разработки ПС получила название спиральной или инкрементной модели процесса разработки [1]. При этом важно отметить, знания, полученные на предыдущих циклах процесса, используются в текущем цикле процесса разработки программных систем.

Указанную последовательность работ также называют «жизненным циклом ПС», подчеркивая наличие обратной связи от этапа испытания к этапу анализа или исследования объекта (рис.1). В принципе, полная модель процесса разработки ПС должна допускать все возможные пути обмена информацией между участниками НИОКР, поэтому граф, представленный на рис.1, более корректно назвать симплексом, однако цикл уже описывает основную идею. Подробное описание характера работ, соответствующих этапам разработки СПС, можно найти в любом учебнике по инженерии ПС, например [2] и поэтому, здесь не приводится.

Следует отметить, что термин «этап» здесь используется в силу исторически сложившейся терминологии и, вообще говоря, означает «функциональный специализированный интеллектуальный (трудовой) процесс определения/формулировки задачи, разработки метода (проекта) решения и конструктивного решения задачи, а также апробирования полученного решения» при отсутствии 100% гарантии достижения намеченного результата в установленный срок.

В инженерии программных систем различают «тяжеловесные» (архитектурно- или проектно-ориентированные) и «легковесные» (тест-ориентированные) процессы разработки ПС. Примером «тяжеловесного» процесса является унифицированный процесс (UP), в котором предлагается документировать артефакты всех этапов указанных на рис.1 работ, а для представления артефактов этапов анализа и проектирования систем предложен специальный унифицированный язык моделирования UML [3].

Примером «легковесного» процесса является экстремальный процесс (XP), в котором предлагается документировать код программы и тесты этого кода с краткими комментариями указанных артефактов [4]. Для представления артефактов этапа тестирования предложена специальная унифицированная нотация описания модульных тестов на выбранном языке программирования целевой системы. Подключаемый к среде программирования специализированный модуль (jUnit/xUnit) позволяет автоматизировать выполнение тестов разрабатываемых модулей. Идеологию экстремального процесса (XP) дополняет список полезных «практик» разработки ПС [4] , [5].

По мнению авторов, основное различие между указанными подходами сводится к различию зафиксированных в каждом из подходов форм представления артефактов этапов работ. Действительно, хорошо известно, что информация может существовать в двух видах, а именно, в виде коммуникации и в виде документации. Проще говоря, в виде устной и письменной речи и их телекоммуникационных расширений. Артефакты указанных на рис.1 этапов работ можно представлять в различных видах

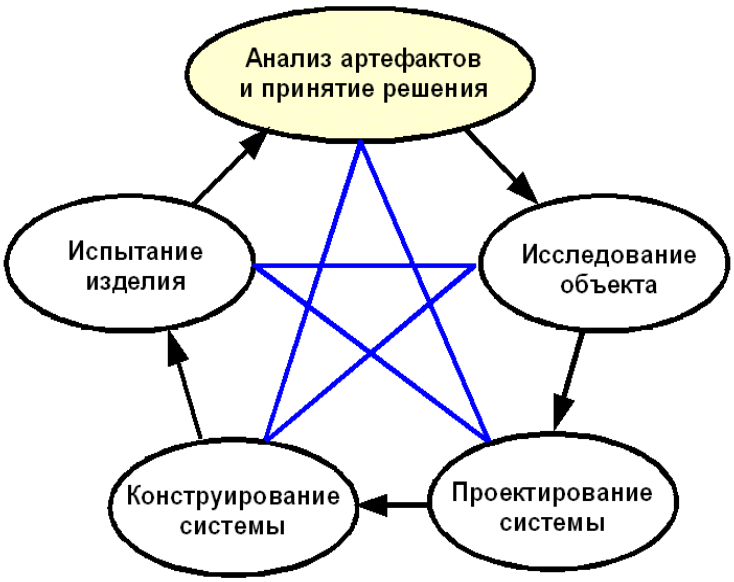

Рис. 1. Модель НИОКР

т. е. в устном виде, в виде «эскиза» (неунифицированной схемы), «чертежа» (унифицированной схемы), «текста» и пр. Поэтому различие между обоими подходами сводится к различию форм представления артефактов, которые, вообще говоря, совсем не обязательно представлять в том виде, в котором их предлагают представлять в указанных подходах, а необходимо, представлять в виде, который наиболее подходит для их выражения.

Например, семантика модели процесса разработки НПС обычно несколько шире той семантики, которая может быть выражена средствами языка UML. Так для описания сложных математических моделей и вычислительных методов, которые используются в системах цифровой обработки, анализа, синтеза и распознавания речевых сигналов язык UML, как и существующие языки программирования, не являются подходящими, так как они не выражают непосредственно семантику предметной области (которую, как нетрудно догадаться, выражают соответствующие разделы математики и физики).

Из-за того, что формы представления артефактов в унифицированном (UP) и экстремальном (XP) подходах выбраны разные, различаются «языки» описания процессов или терминологии, используемые в указанных подходах. Действительно, ЖЦ ПС в XP-интерпретации можно представить в виде цикла (модификация/рефакторизация, интеграция, тестирование), где «тестирование» является синонимом «испытания», «интеграция» соответствует «сборке» или программированию системы, а модное слово «рефакторинг» - объединению этапов анализа, проектирования (выполняемых в «легковесном» (коммуникационном)

виде, т. е. в «устной» и/или «эскизной» формах) и модульного программирования. Таким образом, в методологическом аспекте оба подхода соответствуют инкрементальной модели процесса разработки программных систем.

Различие между XP- и UP-подходами состоит в том, что в XP-подходе предлагается упростить документирование ПС посредством кардинального уменьшения объема самой документации, т. е. с помощью уничтожения «излишней» информации/артефактов, которые, как полагают, могут быть «легко» восстановлены из сохраненных артефактов, точнее говоря, источного кода программы. При разработке функционально не сложных ПС данное предложение вполне уместно и практически целесообразно. Однако в случае создания НПС, восстановление (например, для проведения небольшой модификации ПС) математической модели системы из источного кода программы может потребовать несколько больше времени и ресурсов, чем имеется, поэтому для этого класса систем XP-подход может оказаться не достаточно эффективным.

Важным отличием процесса разработки НПС от процесса разработки СПС является то, что на момент разработки ещё не сложился аппарат адекватного решения выбранного класса задач и который, возможно, сложится в результате разработки НПС удовлетворяющей соответствующим критериям качества. В терминологии, используемой в инженерии программных систем, такие классы задач можно назвать развивающимися предметными областями знаний, в отличие от развитых предметных областей, в контексте которых разрабатываются СПС.

В англоязычной литературе по инженерии ПС этап изучения или исследования естественного объекта/субъекта (предмета исследования) называют анализом предметной области, обозначая этим термином сбор и анализ необходимой информации, определение требований потенциальных пользователей целевой системы, разработку понятийной/аналитической модели. В русскоязычной литературе данный этап обычно определяют как научно-исследовательскую работу (НИР), или, точнее, прикладную НИР (ПНИР), которая несколько отличается от поисковой (фундаментальной) НИР (ФНИР).

Фундаментальной научно-исследовательской работой называют этап изучения естественного объекта/субъекта (предмета исследования) используя научные методы (и инструменты) исследования. Под этим термином обычно понимают сбор и анализ необходимой информации, проведение опытов/экспериментов, а также разработку теории (понятийной модели) изучаемого класса явлений, систем или процессов.

В случае, когда НИР является составной частью НИОКР или инновационной работы, данную НИР называют прикладной научно-исследовательской работой. ПНИР обычно имеет более конкретную (в практическом аспекте) цель работы и, кроме сбора и анализа информации, проведения опытов/экспериментов и разработки аналитической модели конкретного явления, системы или процесса (модели предметной области), обычно включает в свой состав определение концепции целевой системы, т.е. технических и, возможно, маркетинговых требований потенциальных пользователей изделия или продукта.

Обычно ФНИР финансирует государство или общественные фонды, тогда как ПНИР - государство или частные организации (при наличии государственной поддержки), поэтому в экономико-юридическом аспекте результаты ФНИР представляют общественное достояние, а результаты ПНИР - коммерческую/государственную тайну или интеллектуальную собственность.

На этапе исследования определяются:

- представление и структура объекта исследования (модель предметной области),
- «предметные» модели компонент и характеристики связей между компонентами модели объекта исследования (аналитические/математические модели компонент и связей).

Производится апробирование разработанных артефактов, т. е. экспертиза и/или испытание предложенных моделей. В частности, используя средства иммитационного моделирования, вычислительного и/или натурного экспериментов анализируется логическая непротиворечивость, математическая (формальная) корректность и физическая (фактическая) адекватность предложенных моделей и методов решения.

Модель жизненного цикла (ЖЦ) НИР приведена на рис.2. Данная модель может также служить в качестве абстрактной модели ЖЦ других этапов (фаз) НИОКР или инновационных работ.

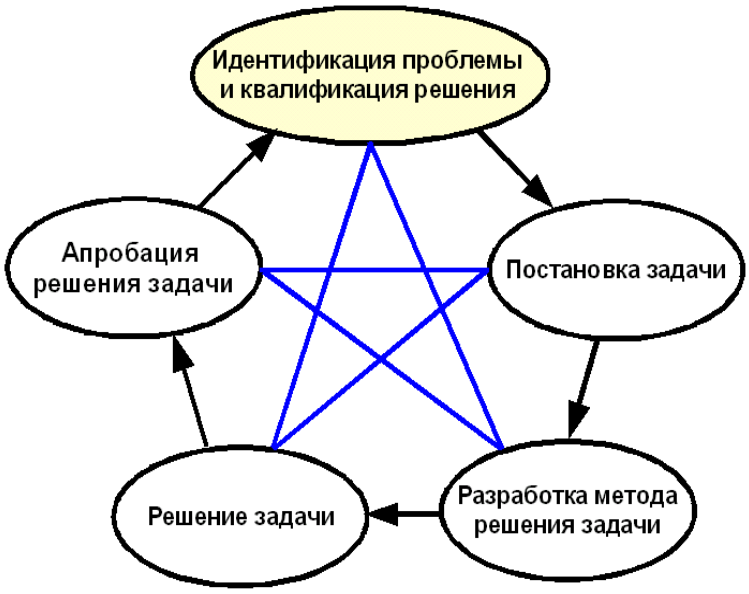

Рис. 2. Модель НИР

Для описания математических моделей и вычислительных методов принято использовать специальные редакторы математических текстов, например, TeX/LaTex. Для выполнения математических вычислений и численных экспериментов применяют математические пакеты MATLAB, MathCAD и др. Для выполнения опытных испытаний или иммитационного моделирования исследуемых явлений, систем или процессов используются специализированные пакеты программ или соответствующие интерактивные редакторы. Использование совокупности указанных средств позволяет создавать качественные и надежные артефакты (документы), которые в дальнейшем можно использовать для создания наукоемких технических/программных систем.

В случае отсутствия/недостаточности имеющихся средств для проведения необходимых исследований объекта, формулируются научно-технические задания на их разработку. При этом в роли потенциальных пользователей требуемых средств оказываются сами исследователи (аналитики), заинтересованные в их разработке. Аналогичные технические задания, в случае отсутствия/недостаточности имеющихся средств, могут также формулироваться архитекторами, конструкторами и испытателями (тестерами) целевой системы. Заметим, что исходное задание на разработку целевой системы также является следствием отсутствия/недостаточности существующих средств решения некоторой важной проблемы и формулируется лицом, принимающим решения, т. е. заказчиком или руководителем НИОКР.

Разрабатываемые артефакты можно разделить на два вида:

- целевые артефакты, т. е. артефакты, предназначенные для разработки НПС,
- инструментальные артефакты, т. е. артефакты, предназначенные для разработки целевых артефактов.

Иными словами, структуру *разрабатываемой* НПС можно представить состоящей из двух частей: *разрабатываемого* целевого ядра и *разрабатываемой* инструментальной оболочки. В документации НИОКР «целевые» и «инструментальные» артефакты, вообще говоря, следует различать.

Этапы сегмента ОКР цикла НИОКР, а именно, этапы проектирования, конструирования и испытания изделия представляют рациональное распределение работ в соответствии с сложившейся специализацией. На этапе проектирования разрабатываются:

- архитектура и алгоритмы компонент/модулей системы,
- спецификации интерфейсов компонент/модулей системы (API), включая интерфейс пользователя (UI) и, если необходимо, интерфейс с хранилищем/базой данных (xDBC).

Наиболее развитым языком проектирования систем является унифицированный язык моделирования UML, хотя на практике часто используют и другие нотации системного моделирования. Язык UML [3] предназначен для визуального определения, проектирования и документирования артефактов программных систем. Разработанный в 1994г. язык UML фактически является стандартным языком объектно-ориентированного проектирования и обычно используется в контексте унифицированного

процесса разработки программных систем (UP). В настоящее время наиболее развитыми средствами поддержки разработки СПС является номенклатура инструментальных сред под общим названием «IBM Rational» или «рациональный унифицированный процесс (RUP)» [6].

На этапе конструирования (кодирования) выполняется реализация:

- механизма взаимодействия компонент системы (системы управления),
- компонентов/модулей системы и утилит.

В настоящее время наиболее распространенными языками программирования («софт-конструирования») ПС являются языки C/C++, java, C# , а также ряд других языков, поддерживающих объектно-ориентированный подход (ООП). Для каждого из перечисленных языков существуют обширные библиотеки классов и технической документации. Для повышения производительности работы программистов разработаны соответствующие инструментальные средства или интегрированные среды разработки (редактирования, кодирования, отладки и тестирования), в частности, Eclipse IDE [8], NetBeans, MS Visual Studio. Указанные языки программирования, включая базовую парадигму программирования, интенсивно развиваются. В качестве примера развития парадигмы ООП можно привести аспектно-ориентированный подход [7].

На этапе испытания производятся:

- комплексные испытания изделия (β-тестирование),
- написание комплекта эксплуатационной/пользовательской документации.

Производится экспертиза результатов опытных испытаний и эксплуатационной документации. В настоящее время фактическим стандартом эксплуатационного/пользовательского документационного обеспечения является электронная документация, состоящая из связанных HTML-документов, которая вытеснила устаревшие форматы файлов документов или бумажно-ориентированную txt-документацию [9].

Язык разметки гипертекстов HTML представляет достаточно простой для освоения человеком машинно-ориентированный язык, основной особенностью которого является возможность определения ссылок или, другими словами, средств выражения связей между HTML-документами (или различными частями одного HTML-документа). Существующие программы-визуализаторы или HTML-редакторы обеспечивают комфортное (удобное для человека) изучение/просмотр и менее комфортное создание/модификацию связанного набора HTML-документов. При этом средства всемирной паутины WWW реализуют связанность HTML-документов в «мировом масштабе».

На этапе анализа артефактов и принятия решения выполняется:

- анализ полученных результатов,
- коррекция (плана) НИОКР.

Апробация (проектов) принимаемых решений обычно называется «обсуждением» с исполнителями НИОКР или «совещанием» с другими заинтересованными лицами. Язык, на котором проводятся эти мероприятия, обычно является «русским». Великое могущество русского языка состоит в наличии элегантных выразительных средств описания содержания этих мероприятий.

В целом управление НИОКР представляет комплекс мероприятий, направленных на обеспечение выполнения проекта, и предполагает следующие виды деятельности:

- планирование работ, т. е. расчет количества необходимых ресурсов, разработка графика работ и пр.,
- организация работ, т. е. распределение и перераспределение ролей и обязанностей, контроль за количеством израсходованных средств и пр.,
- управление работами, т. е. контроль выполнения запланированных работ, решение текущих проблем, обмен информацией с заинтересованными лицами и пр.

Практика показывает, что полное и оперативное согласование и решение текущих вопросов проекта повышает вероятность успешного выполнения НИОКР, тогда как отсутствие плана работ либо схемы организации или управления НИОКР, а также отсутствие необходимой и достоверной информации о ходе выполнения проекта как у руководителей и заинтересованных лиц, так и у исполнителей НИОКР существенно уменьшает вероятность успешного выполнения НИОКР.

Принципы разработки СДО НИОКР

Ниже приводится набор основополагающих принципов построения СДО НИОКР и приводится их краткая интерпретация.

Научность (понятность) – при определении нового понятия (модели) объекта исследования необходимо использовать достоверные, т. е. научно обоснованные понятия. Введение нового понятия не должно нарушать целостность (непротиворечивость) базовой системы понятий (знаний). Достоверность нового понятия необходимо обосновывать теоретически (логически), а адекватность понятия объекту - обосновывать экспериментально (физически), т. е. проверять на практике. Объект исследования можно изучать теоретически – с различных точек зрения (систем понятий) или экспериментально – в различных системах измерений. Отношение соответствия «точки зрения» и «системы измерений» называется представлением. Использование научных представлений (т. е. представлений, лежащих в основе определенной области науки/знания) означает научную (предметную) обоснованность или понятность.

Системность – при определении нового понятия (модели) объект исследования следует рассматривать, с одной стороны, как автономную часть более крупной системы (среды), которая взаимодействует с рассматриваемым объектом, а с другой стороны, как систему, состоящую из автономных более мелких объектов (компонентов), которые взаимодействуют друг с другом. При этом компоненты системы также можно рассматривать

как системы и т.д. Этим способом определяются сложные многоуровневые иерархические модели объектов. Системообразующей или идентификационной характеристикой объекта/системы, в соответствии с которой определяют и обозначают часть среды как объект/систему, является цель, функция или предназначение системы.

Целостность (полнота) – при определении нового понятия (модели) объект исследования следует представлять иерархически, начиная с высшего целевого уровня. При этом необходимо соблюдать правило полноты декомпозиции целей, иначе говоря, каждая цель верхнего уровня должна быть представлена в виде подцелей нижнего уровня исчерпывающим образом, т. е. так, чтобы объединение подцелей полностью определяло исходную цель. Тоже самое, очевидно, относится и к определению структуры системы. Заметим, что целостность системы, как упорядоченной совокупности висимых компонент, проявляется также в том, что свойства системы не сводятся к свойствам отдельных компонент системы, а представляют новое качество, которое называют синергетическим эффектом.

Рациональность (простота) – при определении нового понятия (модели) объекта исследования необходимо руководствоваться методологическим принципом сформулированным английским средневековым философом и логиком У. Оккамом и который получил название «бритва Оккама». Этот принцип требует устранения из модели всех понятий, не являющимися интуитивно очевидными или не поддающимися проверке опытом: «Сущности не следует порождать без необходимости». Этот принцип также

называют принципом простоты. Заметим, что введенные понятия должны быть достоверны. Формальным выражением принципа рациональности является принцип оптимальности, т. е. выбор наилучшего (оптимального) варианта из множества допустимых по заданному критерию оценки.

Открытость – при определении нового понятия (модели) объекта исследования необходимо учитывать как влияние на систему внешней среды, так и влияние системы на внешнюю среду, в частности инфраструктуру системы. Создаваемая система должна быть согласована/совместима с внешней средой или, выражаясь аллегорически, воспринята, интегрирована в исторически сложившуюся инфраструктуру.

Эволюционность – при определении нового понятия (модели) объекта исследования следует учитывать возможное в дальнейшем развитие или расширение модели системы. Понятие развития, изменяемости, т.е. возможности количественных изменений характеристик системы при сохранении качественных особенностей следует закладывать в основу создания модели системы.

Гармоничность – при определении нового понятия (модели) объекта исследования следует обеспечить согласованность, соразмерность, слаженность взаимодействия компонент системы при условии ее целостности. Обычно гармоничность выражается в виде «красоты», совершенства или симметрии структуры или распределения функций компонент системы.

Формализуемость – при определении нового понятия (модели) объекта исследования следует различать «разнообразные» и

«однообразные» компоненты системы. Разнообразные компоненты системы соответствуют различным видам, типам или классам объектов. Однообразные компоненты системы соответствуют одному виду, типу или классу объектов. Операция определения вида, типа или класса объектов может осуществляться как на методологическом (понятийном, логическом), так и на методическом (прагматическом, технологическом) уровнях описания набора объектов и называется операцией унификации. Операция унификации представляет «логическую декомпозицию» описания набора объектов на форму и содержание (тип и значение). Процедура рекурсивной унификации называется классификацией. Результатом классификации является иерархия классов, т. е. рекурсивная форма представления некоторого семейства объектов («дерево» типов).

Концепция СДО НИОКР

В рассмотренной выше модели процесса разработки НПС были определены артефакты, которые обычно создаются в процессе НИОКР, однако какие-либо соображения о том, каким образом следует организовать их хранение, обработку и использование осталось за рамками модели. Обычно предполагается, что стандартных средств взаимодействия с файловой системой, которые встроены практически в любое приложение, вполне достаточно для решения этих задач. Это предположение вполне справедливо для отдельно взятого файла или небольшого набора

файлов.

Однако при выполнении НИОКР, обычно требуется управлять уже достаточно большим набором файлов или документов, которые представляют семантически связанные артефакты. Со временем управление таким набором документов превращается в проблему, ввиду того, что, во-первых, с ростом количества документов в наборе, количество семантических связей между элементами набора возрастает приблизительно квадратично и, во-вторых, не задокументированные семантические связи между занесенными в набор документами со временем забываются. Теоретическое решение проблемы эффективного управления большим набором семантически связанных документов известно и состоит в «повышении уровня организации информации».

Традиционными средствами организации работы с наборами документов являются хорошо известные архиваторы файлов и «подшивщики» тематически связанных документов, такие, например, как MS Office Binder. Подшивщики документов, как и архиваторы, позволяют собирать тематически связанные документы в одном файле, который называют подшивкой. «Подшитые» документы могут быть различных «офисных» типов (текст, диаграмма, таблица, презентация). Подшивка позволяет использовать единый стиль оформления содержащихся в ней документов: нумеровать страницы, использовать общие колонтитулы и пр. Кроме этого, файл подшивки можно целиком вывести на печать. Подшивщики, как и обычные архиваторы файлов, не ограничивают возможности работы с каждым документом подшивки по отдельности и в целом, можно сказать, представ-

ляют электронные «сборники» документов с дружественным интерфейсом.

Использование указанных инструментов для представления документации НИОКР имеет, однако, один существенный недостаток, а именно отсутствие механизма представления семантических связей между документами. Чтобы подчеркнуть важность данной информации, приведем следующие рассуждения. Предположим, что «информационная ценность», содержащаяся в наборе документов, может быть представлена в виде суммы двух слагаемых, а именно: первое слагаемое – информационная ценность, содержащаяся в наборе терминальных или первичных документов (артефактах), и второе слагаемое – информационная ценность, содержащаяся в семантических связях между первичными документами. Если принять это предположение, то по мере роста объема набора документов всё более существенную долю в общей информационной ценности, содержащейся в наборе документов, будет играть второе слагаемое ввиду более быстрого роста. Из этого рассуждения следует вывод: кроме сохранения «ценных» первичных документов необходимо также сохранять ценные «вторичные» документы, представляющие описание семантических связей между первичными документами. Широко известным способом представления семантических связей между документами является аппарат библиографических ссылок, который в эпоху всеобщей компьютеризации трансформировался в аппарат гипертекстовых ссылок.

Определим «документацию НИОКР» как рационально организованный набор семантически связанных научно-технических

документов обладающий свойствами *полноты*, т. е. отсутствия семантических пробелов/изъятий, *простоты*, т. е. отсутствия семантических излишеств/повторов и *понятности*, т. е. наличия семантической корректности/обоснованности (более подробное описание этих понятий было приведено выше).

Приведенное определение соответствует успешно завершенной НИОКР, тогда как в начале НИОКР, вообще говоря, значительное количество конкретных артефактов необходимо «оформить» документально. Проще говоря, их требуется найти, определить, измерить, вычислить, разработать, написать или нарисовать, т. е. создать. Всё это – задача коллектива НИОКР, т. е. научно-техническое («документационное») *произведение*.

Документационное *обеспечение*, о котором идет речь в этой работе, представляет «организационные» знания, т.е. рациональную схему (методологию) интеллектуального трудового процесса, способ/процедуру (методику) оформления/представления набора научно-технических документов и систему поддержки рационального представления *набора* документов, создаваемых участниками НИОКР. Указанную совокупность знаний (методологию, методику и реализацию/конструкцию) принято называть информационной технологией.

Архитектура СДО НИОКР

В процессе выполнения НИОКР документация отражает текущее состояние научно-технического проекта. В конце процесса

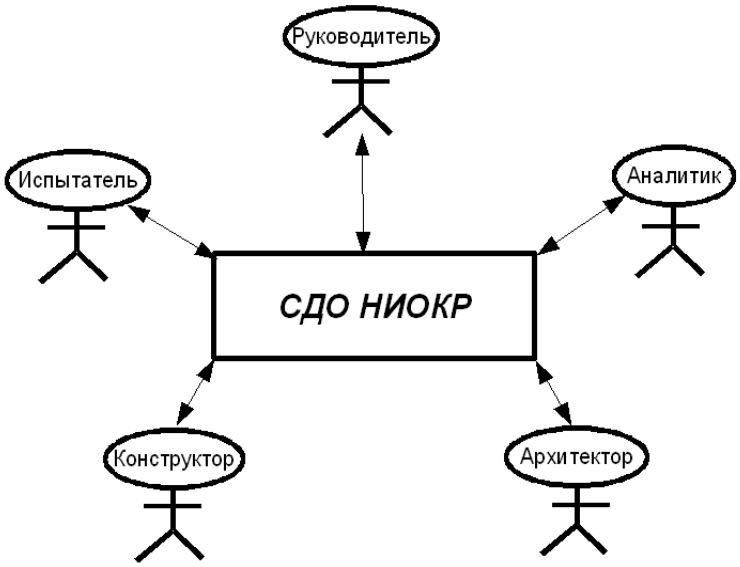

Рис. 3. Схема пользователей СДО НИОКР

выполнения НИОКР документация содержит формальный результат НИОКР. В узком смысле результатом НИОКР является разработанная целевая система или «продукт», который был определен в научно-техническом задании. В широком смысле результатом НИОКР является не только разработанная целевая система или «продукт», но и новые инструментальные (технологические) артефакты, которые потребовались для разработки целевой системы, а также знания и опыт, которые остались у участников НИОКР.

Движущей силой НИОКР является персонал «предприятия», схема распределения которого по специальностям в соответствии

с моделью процесса разработки НПС показана на рис.3. Архитектура СДО НИОКР показана на рис.4. Аппаратура, операционная система и WWW подробно описываются в соответствующих руководствах, а инструментарий составляют редакторы файлов различных типов данных, которые используются для создания, модификации, тестирования и использования документов (содержащих артефакты НИОКР). В отличие от инструментария, предоставляющего средства создания отдельных документов, основное назначение СДО НИОКР состоит в предоставлении удобных средств описания первичных документов и семантических связей между первичными документами, т. е. средств создания *системы* документации НИОКР.

Ядром СДО НИОКР является «docware», или интеллектуальное обеспечение процесса разработки НПС, однако с более утилитарной точки зрения в СДО НИОКР можно также включить «software», например, систему управления проектом, систему управления версиями, систему управления содержанием/контентом и др. Указанные средства предоставляют более удобный сервис и доступ к документации НИОКР по сравнению с стандартными средствами операционной системы.

В частности, система управления версиями (VCS) представляет систему поддержки текущего набора артефактов НИОКР в актуальном состоянии. Система позволяет вести историю изменений исходных текстов, блокировать одновременные модификации файла разными разработчиками, оперативно информировать разработчиков об изменении текущих версий исходных файлов в долгосрочном проекте НПС.

Рис. 4. Архитектура СДО НИОКР

Система управления содержанием/контентом (CMS) представляет каркас WWW сайта, обычно с открытым доступом к информационным ресурсам, расположенным на этом сайте, и поддержкой средств коллективного редактирования документов. В целом, системы VCS или CMS предоставляют набор удобных средств выполнения совместной работы над проектом удаленными друг от друга разработчиками.

Рассматриваемая в настоящей работе СДО НИОКР представляет унифицированный информационный пакет (контейнер или документационный модуль), из клонов которого формально конструируется фактическая документация НИОКР. Модульная структура документации имеет топологию дерева, где каждый унифицированный информационный пакет может представлять элемент декомпозиции произвольного уровня иерархии.

Документационный модуль или унифицированный информационный пакет, который будем обозначать SXP (сокращение англоязычного словосочетания «simple extensible package» или «simplex package»), состоит из двух частей, которые будем называть «заголовком» и «телом». Заголовок пакета (SXP.head) содержит краткое описание или информацию о содержании содержимого тела информационного пакета. Тело информационного пакета (SXP.body) может содержать артефакты НИОКР в виде первичных/терминальных документов и/или вложенные информационные пакеты (рис.5).

Заголовок или, с точки зрения пользователя-непрограммиста, интерфейс информационного пакета можно структурировать и представить в виде трех HTML-файлов:

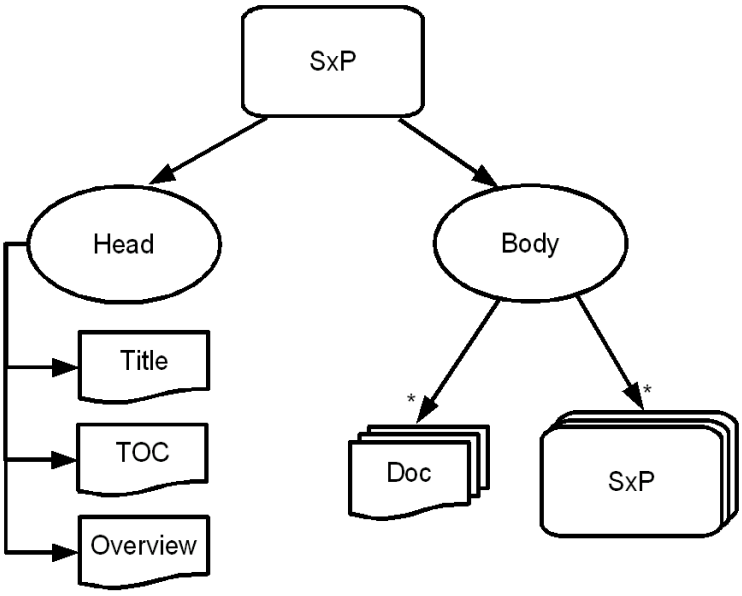

Рис. 5. Структура унифицированного информационного пакета

- **Title.html** содержит название раздела документации НИОКР, которое должно выражать цель/назначение или суть того, что заключено в теле информационного пакета,

- **TOC.html** содержит описание состава информационного пакета с точки зрения пользователя-непрограммиста, т. е. на семантическом уровне или, проще говоря, содержание раздела документации НИОКР,

- **Overview.html** содержит аннотацию раздела документации НИОКР, т. е. важную с точки зрения пользователя-непрограммиста более детальную (дополнительную) семантическую информацию об артефактах НИОКР, в частности ссылки на другие информационные пакеты, содержащие семантически связанные артефакты НИОКР.

Информационные пакеты верхних уровней иерархической структуры документации НИОКР могут содержать принятые проектные решения, пакеты средних уровней – детализации принятых решений, пакеты нижних уровней – «терминальные» артефакты НИОКР. Семантическая связь артефактов НИОКР, хранящихся в различных информационных пакетах, легко выражается с помощью механизма ссылок между HTML-документами, составляющими интерфейс информационного пакета.

Модульная организация СДО НИОКР предоставляет возможность организовывать пакеты документов в различные структуры или представления, например, создавать представления ана-

литика, архитектора, конструктора, испытателя или руководителя НИОКР (см. рис.3), иначе говоря, создавать презентационные представления различных специалистов (внешние модели) без дублирования терминальных артефактов НИОКР.

Литература

1. Якобсон А., Буч Г., Рамбо Дж. Унифицированный процесс разработки программного обеспечения. СПб.: Питер, 2003.

2. Брауде Э. Технология разработки программного обеспечения. СПб.: Питер, 2004.

3. Фаулер М. UML. Основы, 3-е издание. СПб.: Символ-Плюс, 2004.

4. Бек К. Экстремальное программирование: разработка через тестирование. СПб.: Питер, 2003.

5. Ауэр К., Миллер Р. Экстремальное программирование: постановка процесса. С первых шагов и до победного конца. СПб.: Питер, 2004.

6. Крачтен Ф. Введение в Rational Unified Process, 2-е издание, М.: Издательский дом «Вильямс», 2002.

7. Чарнецки К., Айзенекер У. Порождающее программирование: методы, инструменты, применение. - СПб.: Питер, 2005.

8. Гамма Э., Бек К. Расширения Eclipse: принципы, шаблоны и подключаемые модули. М.: КУДИЦ-ОБРАЗ, 2005.

9. Гультяев А.К. Help. Разработка справочных систем: Учебный курс. СПб.: Питер, 2004.

Гибридные модели: скрытые марковские модели и нейронные сети, их применение в системах распознавания речи

К.А.Маковкин

Аннотация

В статье сделан обзор существующих гибридных моделей – скрытых марковских моделей и нейронных сетей. Рассмотрены модели различных нейронных сетей и принципы их комбинирования со скрытыми марковскими моделями. Приведены краткие сравнительные характеристики систем распознавания речи, которые используют такую архитектуру. Ключевые слова: распознавание речи, скрытые марковские модели, нейронных сети, гибридные модели.

Введение

Применение методов статистической теории распознавания образов стало важным этапом в развитии автоматического распознавания речи (АРР). Это позволило исследователям использовать мощный аппарат математической статистики и теории вероятностей, что в свою очередь привело к существенному повышению качества распознавания. На сегодняшний день практически все известные системы распознавания речи основаны на статистических методах.

В рамках такого подхода речевой сигнал представляется как случайный образ, который необходимо распознать или, другими словами, преобразовать в некоторую последовательность слов W, и тогда задача распознавания речевого сигнала формулируется как классическая задача классификации образов по критерию максимума апостериорной вероятности, т.е. необходимо максимизировать апостериорную вероятность $P(W|X)$, где X – это наблюдаемая последовательность акустических векторов параметров речевого сигнала, а W – последовательность слов. Согласно формуле Байеса апостериорную вероятность можно переписать в виде:

$$\underset{W \in \Gamma}{\boldsymbol{argmax}}\, P(W|X) = \underset{W \in \Gamma}{\boldsymbol{argmax}}\, P(X|W) \cdot P(W) \tag{1}$$

где, Γ – множество всех возможных последовательностей слов, $P(W|X)$ – условная вероятность появления последовательности акустических векторов X для заданной последовательности слов W, а $P(W)$ – априорная вероятность появления последовательности слов W. Выражение $P(X|W)$ обычно называют акустико-фонетической моделью, а $P(W)$ – моделью языка [55].

Наиболее популярными технологиями акустико-фонетического моделирования речевого сигнала на сегодняшний день по праву являются технологии, основанные на скрытых марковских моделях (СММ) [56]. Использование СММ позволяет достичь довольно высокой точности распознавания, обеспечивает хорошее представление речевого сигнала и предоставляет мощный и гибкий инструмент для разработки систем распознавания. К сожалению, при неоспоримых преимуществах

СММ обладают целым рядом ограничений, например слабой дискриминантной мощностью, т.е. способностью разделять классы образов. Особенно это проявляется при обучении с использованием критерия максимума правдоподобия (МП) [55]. При использовании других критериев, например критерия максимума взаимной информации (МВИ), можно достичь большей разрешающей возможности, но этот алгоритм математически более сложный и требует большого числа ограничивающих предположений. Кроме того, использование акустической и фонетической контекстуальной информации требует значительного усложнения СММ, большого объема памяти для хранения параметров модели и большого количества обучающих данных.

Другим классом моделей, обеспечивающих акустико-фонетическое моделирование, являются модели искусственных нейронных сетей (ИНС). С середины 1980-х г. ИНС стали активно использоваться в системах распознавания речи. Исследователями было предложено довольно много различных архитектур нейронных сетей [40], которые показывали неплохие результаты по классификации речевых образов. Основным преимуществом, обеспечившим ИНС такое бурное использование, являются мощные дискриминантные способности, а также возможность обучаться и представлять неявные знания. Однако, несмотря на потенциальные возможности по классификации кратковременных акустико-фонетических единиц, таких как, например, фонемы, ИНС не стали основой моделью для создания систем АРР. Причиной тому послужил недостаток

ИНС, связанный со сложностью моделирования длительных последовательностей наблюдений, таких как, например, слова и целые высказывания, так как эти последовательности обычно обладают сильной временной изменчивостью. Эту проблему не решило даже использование рекуррентных архитектур сети. Другими словами, ИНС хорошо работают только со статическими образами, и их эффективность сильно снижается, когда на входе появляется некоторая динамика, т.е. образы подвержены, например нелинейным изменениям во времени.

В начале 90-х г. факт существования двух взаимодополняющих подходов привел исследователей к идее комбинировать СММ и ИНС в рамках одной, новой модели – гибридной СММ/ИНС модели [20,24,27,37,45,48]. Такая гибридная модель позволяет эффективно объединить преимущества марковских моделей и нейронной сети, т.е. СММ обеспечивает возможность моделирования долговременных зависимостей, а ИНС обеспечивает непараметрическую универсальную аппроксимацию, оценку вероятности, алгоритмы дискриминантного обучения, уменьшение числа параметров для оценки, которые обычно требуются в стандартных СММ.

Скрытые марковские модели

Практически все наиболее известные системы распознавания речи, созданные за последние двадцать пять лет основаны на статистических принципах и используют аппарат СММ. Основные положения теории СММ были сформулированы и опубликованы на рубеже 60-х – 70-х гг. в серии статей Баума и др. исследователей [12,13,14], а первые практические результаты использования СММ в системах АРР описаны Бейкером [11] и Елинеком с коллегами из IBM [1,10,33,34]. Позднее были написаны несколько обзорных статей, которые позволили использовать теорию СММ в практических приложениях [3,4,38].

Для простоты рассмотрим пример марковской модели для звука, которая изображена на рис. 1. Эта модель состоит из последовательности состояний, обозначенных $S_1, S_2, ..., S_5$, которые связаны мгновенными вероятностными переходами, изображенные стрелками и имеющие вероятность a_{ij}, т.е. вероятность перехода из i-го состояния в j-е. Возможны переходы только в следующее состояние и зацикливание. В каждый момент времени модель осуществляет вероятностный переход из одного состояния в другое или в то же самое состояние, при этом происходит излучение выходного акустического вектора y_k с выходным вероятностным распределением $b_n(y_k)$, соответствующим этому состоянию. Эти вероятности называют *эмиссионными вероятностями*. Тогда некоторое высказывание, описываемое последовательностью акустических векторов параметров $X =$

$\{x_1, x_2, ..., x_N\}$, можно промоделировать последовательностью дискретных стационарных состояний $Q = \{q_1, q_2, ..., q_K\}, K < N$, с мгновенными переходами между этими состояниями и последовательностью излученных при этом акустических векторов $Y = \{y_1, y_2, ..., y_N\}$.

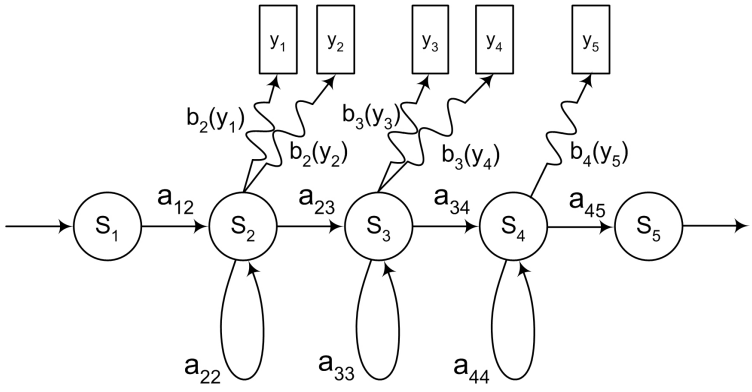

Рис. 1. Пример скрытой марковской модели

Таким образом, скрытая марковская модель состоит из марковской цепи с конечным числом состояний S_N и матрицей переходных (транзитивных) вероятностей a_{ij}, которые определяют длительность пребывания системы в данном состоянии, т.е. марковская цепь моделирует временные изменения речевого сигнала, а также конечного множества эмиссионных вероятностей $b_n(y_k)$, которые позволяют моделировать спектральные вариации сигнала. Этот подход определяет два одновременных стохастических процесса, один из которых является основным и ненаблюдаемым (т.е. скрытым) – это последовательность СММ-

состояний, и мы можем судить о нем только с помощью другого случайного процесса, т.е. по последовательности наблюдений (собственно, поэтому такая модель называется “скрытой” марковской моделью).

Для определения СММ необходимо задать следующие элементы:

1. Множество состояний модели $S = \{s_1, s_2, ...s_N\}$, где N – число состояний в модели. Состояние модели в момент времени t обозначается q_t.

2. Множество различных символов наблюдения, которые могут порождаться моделью $Y = \{y_1, y_2, ..., y_K\}$, где K – число символов наблюдения модели. Символы наблюдения соответствуют физическому выходу моделируемой системы.

3. Распределение вероятностей переходов между состояниями (или матрица переходных вероятностей) $A = \{a_{ij}\}$, где

$$a_{ij} = P\left[q_{t+1} = s_j | q_t = s_i\right], \qquad 1 \leq i, j \leq N, \tag{2}$$

при этом предполагается, a_{ij} не зависят от времени.

4. Множество распределений вероятностей появления символов наблюдения (их называют эмиссионными или выходными вероятностями) в состоянии j, $B = \{b_j(k)\}$, где

$$b_j(k) = P[y_k \text{ в момент } t | q_t = s_j], 1 \leq j \leq N, 1 \leq k \leq K \tag{3}$$

5. Начальное распределение вероятностей состояний $\Pi = \{\pi_i\}$

$$\pi_i = P[q_1 = s_i], \qquad 1 \leq i \leq N \tag{4}$$

Использование СММ в системах распознавания речи

Чтобы использовать СММ в системе АРР необходимо сделать несколько упрощающих, но очень важных предположений о речевом сигнале:

- последовательные наблюдения являются статистически независимыми и, следовательно, вероятность последовательности наблюдений есть просто произведение вероятности отдельных наблюдений;

- хотя речь представляет собой нестационарный процесс, он моделируется последовательностью векторов наблюдений, которые представляют собой кусочно-стационарный процесс;

- собственно марковское допущение, т.е. допущение о том, что вероятность пребывания в некотором состоянии в момент времени t зависит только от состояния, в котором процесс находился в момент времени $t-1$.

Теперь рассмотрим простую систему распознавания. Идеально было бы иметь СММ для каждого из возможных высказываний. Однако, очевидно, что это выполнимо только для очень ограниченных задач, например распознавание изолированных команд из небольшого словаря. Поэтому используют более мелкие речевые единицы, например фоны, которые с лингвистической точки зрения соответствуют фонемам. Для каждого фона необходимо

создать свою отдельную СММ, т.е. $M = \{m_1, m_2, ..., m_U\}$ – множество марковских моделей для всех возможных фонов, а $\Theta = \{\lambda_1, \lambda_2, ..., \lambda_U\}$ – множество связанных с ними параметров. Тогда M_i будет представлять марковскую модель некоторого слова, полученную конкатенацией элементарных моделей из множества M, при этом M_i состоит из L_i состояний $q_l \in S$ и $l = 1, 2, ..., L_i$, а множество параметров этой модели будет Λ_i, которое является подмножеством Θ. Произнесение каждого фона описывается последовательностью векторов спектральных характеристик сигнала. На этапе обучения для каждого слова (например, M_i) имеется последовательность, состоящая из множества повторений последовательностей векторов параметров X_{M_i}, соответствующих произнесению этого слова одним или несколькими дикторами и необходимо выбрать такое множество параметров Θ, которое максимизировало вероятность $P(M_i|X_{M_i}, \Theta)$ для всех обучающих высказываний X_{M_i}, связанных с M_i, т.е.

$$\underset{\Theta}{argmax} \prod_{i=1}^{I} P(M_i|X_{M_i}, \Theta) \tag{5}$$

Таким образом, обучение состоит в подборе параметров модели Θ в соответствии с некоторым критерием оптимальности. К сожалению, не существует известного аналитического выражения для этих параметров. Кроме того, на практике, располагая некоторой последовательностью наблюдений в качестве обучающих данных, нельзя указать оптимальный способ оценки параметров. Однако, используя итеративные процедуры, например алгоритм Баума-Уэлча или, что эквивалентно, EM-метод (метод математического ожидания-модификации) [22], или

градиентные методы [38], можно подобрать параметры модели таким образом, чтобы локально максимизировать вероятность $P(M|X,\Theta^*)$. Следует отметить, что эти алгоритмы принадлежат классу алгоритмов обучения "без учителя", так как они производят ненаблюдаемую оценку параметров распределения вероятностей, не требуя предварительной разметки. На этапе распознавания неизвестного высказывания X необходимо найти наиболее подходящую модель M_i, которая максимизировала $P(M|X,\Theta)$ при уже фиксированном множестве параметров Θ и наблюдаемой в данный момент последовательности X. Таким образом, результатом распознавания высказывания X будет слово, связанное с моделью M_i такое, что

$$i = \underset{\forall j}{argmax}\, P(M_j|X,\Theta) \tag{6}$$

Метод нахождения наилучшей модели основан на динамическом программировани и называется алгоритмом Витерби [56].

Обучение и распознавание связано с выбором некоторого критерия оптимальности. Таких критериев существует несколько. Все они имеют физический смысл и используются на практике. Выбранный критерий оптимальности (например, максимум правдоподобия или максимум апостериорной вероятности) оказывает влияние на такие параметры модели, как объем данных для обучения и требования к вычислительным ресурсам, точность распознавания, способность к обобщению данных из обучающей выборки. Одним из наилучших критериев может считаться Байесовский классификатор, основанный на апостериорной вероятности $P(M_i|X,\Theta)$ (или классификатор по макси-

муму апостериорной вероятности, MAP-оцениватель) того, что последовательность акустических векторов X была порождена M_i моделью с множеством параметров Θ. Используя правило Байеса $P(M_i|X,\Theta)$, можно записать в виде выражения

$$P(M_i|X,\Theta) = \frac{P(X|M_i,\Theta)P(M_i|\Theta)}{P(X|\Theta)}, \tag{7}$$

которое разделяет процесс оценки вероятности на две части: задачу акустического

$$\frac{P(X|M_i,\Theta)}{P(X|\Theta)} \tag{8}$$

и языкового $P(M_i|\Theta)$ моделирования. Целью языкового моделирования является оценка априорных вероятностей моделей высказываний $P(M_i|\Theta)$. Эта языковая модель обычно полагается независимой от акустических моделей и описывается в терминах независимого множества параметров Θ^*. Параметры языковой модели обычно оцениваются на больших текстовых базах данных [49].

Задачей акустического моделирования является оценка плотностей вероятностей (8), как правило, независимо от других моделей. Так как вероятность $P(X|M_i,\Theta)$ обусловлена только M_i, то она зависит только от параметров M_i модели и, опуская $P(X|\Theta)$, как в [19], выражение (8) можно переписать как $P(X|M_i,\Lambda_i)$, где Λ_i – множество параметров, связанных с моделью M_i. Таким образом, и обучение, и распознавание требует оценки вероятности $P(X|M_i,\Lambda_i)$, которая называется глобальным правдоподобием последовательности векторов параметров X при заданной M_i.

Далее $P(X|M_i, \Lambda_i)$ можно оценить как сумму

$$P(X|M_i, \Lambda_i) = \sum_{\{\Gamma_i\}} P(X, \Gamma_i|M_i, \Lambda_i), \tag{9}$$

где $\{\Gamma_i\}$ представляет собой множество всех возможных путей (последовательности состояний) длины L в модели M_i. При этом для каждой последовательности состояний вероятность появления последовательности наблюдений $X_1^L = \{x_1, x_2, ..., x_L\}$ определяется выражением

$$P(X_1^L|q_1^L, M_i, \Lambda_i) = \prod_{l=1}^{L} P(x_l|q_1^l, X_1^{l-1}, M_i, \Lambda_i). \tag{10}$$

Можно показать [17], что (9) можно вычислить с помошью алгоритма прямого-обратного хода [56], для которого необходимо рекурсивно вычислять так называемую прямую переменную

$$\mathtt{P}(\mathtt{q}_1^\mathtt{n}, \mathtt{X}_1^\mathtt{n}|\mathtt{M}_\mathtt{i}, \Lambda_\mathtt{i}) = \sum_{\mathtt{k}=1}^{\mathtt{L}} \mathtt{P}(\mathtt{q}_\mathtt{k}^{\mathtt{n}-1}, \mathtt{X}_1^{\mathtt{n}-1}|\mathtt{M}_\mathtt{i}, \Lambda_\mathtt{i})\mathtt{p}(\mathtt{q}_\mathtt{l}^\mathtt{n}, \mathtt{x}_\mathtt{n}|\mathtt{q}_\mathtt{k}^{\mathtt{n}-1}\mathtt{X}_1^{\mathtt{n}-1}, \mathtt{M}_\mathtt{i}, \Lambda_\mathtt{i}), \tag{11}$$

где $P(q_l^n, X_1^n|M_i, \Lambda_i)$ представляет собой вероятность того, что частичная подпоследовательность наблюдений $X_1^n = \{x_1, x_2, ..., x_n\}$ была порождена моделью M_i, а в момент времени n наблюдалось состояние $q_l^n = S_l$ и был сгенерирован вектор наблюдений x_n.

Второй сомножитель в правой части равенства (10) можно представить в виде произведения вероятностей

$$p(q_l^n, x_n|q_k^{n-1}X_1^{n-1}, M_i, \Lambda_i) = p(x_n|q_l^n, q_k^{n-1}, M_i, \Lambda_i)p(q_l^n|q_k^{n-1}, M_i, \Lambda_i), \tag{12}$$

где первый сомножитель $p(x_n|q_l^n, q_k^{n-1}, M_i, \Lambda_i)$ представляет *эмиссионную вероятность*, а второй $p(q_l^n|q_k^{n-1}, M_i, \Lambda_i)$ – транзитивную вероятность. Обычно эмиссионную вероятность упрощают, чтобы снизить число свободных парамеров, полагая, что наблюдаемый акустический вектор x_n зависит только от текущего состояния процесса q_l^n, т.е. используют эмиссионную вероятность в виде $p(x_n|q_l)$.

Описанная стандартная СММ, как уже отмечалось, является довольно мощным инструментом, позволившим разработчикам существенно повысить качество распознавания речевого сигнала. Это демонстрирует целый ряд лабораторных систем распознавания слитной речи с большими словарями (1000-40000 слов), которые занимают высокое место в сравнительных испытаниях, проведенных в рамках проекта SQALE [73]. В экспериментах участвовали три системы, построенных на СММ:

- система распознавания Cu-HTK, которая была разработана Стивом Янгом (Steve Young) в Кэмбриджском университете в 1987 г. [http://htk.eng.cam.ac.uk],
- система распознавания LIMSI, разработанная в Laboratoire d'Informatique pour la Mecanique et les Sciences le l'Ingenieur во Франции,
- система Philips, разработанная в лаборатории человеко-машинного интерфейса фирмы Philips в Германии.

Завершая краткое описание СММ, необходимо отметить, что наряду с неоспоримыми достоинствами, такими как

- мощный математический аппарат,
- эффективное моделирование как временных, так и спектральных вариации речевого сигнала,
- достаточно гибкая топология – СММ могут легко включать не только фонологические правила или, например, строить модели слов из моделей фонов, но и позволяют использовать синтаксические правила,
- глубокая практическая проработка – разработаны мощные обучающие и распознающие алгоритмы, которые обеспечивают эффективное обучение на больших речевых базах данных и распознавание изолированных слов и слитной речи без адаптации под диктора,

исследования выявили целый ряд недостатков, преодолеть которые оказалось крайне трудно:

- слабые дискриминантные способности, так как во время обучения акустические модели формируются на основе критерия максимума правдоподобия, а не более точного максимума апостериорной вероятности;
- последовательности векторов наблюдений считаются статистически независимыми, т.е. некоррелированными, что неверно для речевого сигнала;
- кусочно-постоянный характер модели, т.е. каждое марковское состояние имеет стационарную статистику (независимо

от времени нахождения в данном состоянии распределения эмиссионных вероятностей одинаковы);

- априорный выбор топологии модели и статистических распределений;
- отсутствие эффективных и адекватных природе речевого сигнала моделей длительности состояний и их реализации в рамках марковских моделей;
- марковская модель полагается моделью первого порядка, т.е. состояние в момент времени n зависит только от предыдущего состояния в момент времени $n-1$;
- обучение и оптимизация лингвистической модели происходит отдельно от акустических моделей.

Перечисленные недостатки существенно ограничивающие возможности этого класса моделей [46], побудили исследователей к поиску альтернативных или дополняющих подходов к решению проблемы акустико-фонетического моделирования речевого сигнала.

Нейронные сети

Другим классом моделей, которые были использованы для акустико-фонетического моделирования речевого сигнала являются модели искусственных нейронных сетей (ИНС), структуры и принципы работы которых основываются на биологических моделях нервных систем, особенно на моделях головного мозга. Нейронные сети могут рассматриваться как разновидность самоорганизующихся алгоритмов и представляют собой множество однотипных и параллельно функционирующих элементов или нейронов, связанных между собой и "внешним миром" с помощью специально организованных связей. Нейрону в дискретные моменты времени по входным связям передается информация, на основе которой в соответствии с некоторыми принципами формируется выходной сигнал, который в свою очередь передается на входы других нейронов или во "внешний мир". Таким образом, основным элементом нейронной сети является нейрон.

Наиболее распространенной является модель нейрона МакКаллока-Питса (рис. 2), предложенная в 1943 г. [43,62], в соответствии с которой нейрон имеет набор входных связей и один выход, который может распараллеливаться.

Эта модель функционирует следующим образом: на вход нейрона подается входной вектор $\mathbf{x}(t) = \{x_1(t), x_2(t), ..., x_N(t)\}$, который скалярно умножается на весовой вектор $\mathbf{w}_k = \{w_{1k}, w_{2k}, ..., x_{Nk}\}$ или, другими словами, компоненты вектора $x_i(t)$ взвешиваются весовыми коэффициентами w_{ik} в соответ-

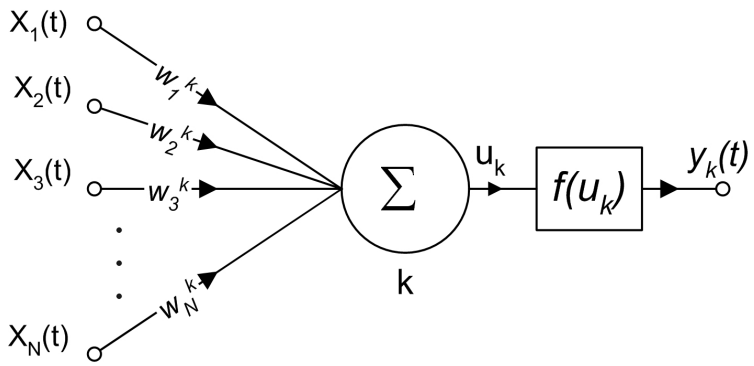

Рис. 2. Модель нейрона

ствии с формулой

$$u_k(t) = \sum_{i=0}^{N} w_{ik} x_i(t). \tag{13}$$

Выходной сигнал нейрона определяется как

$$y_k(t) = f(u_k(t)), \tag{14}$$

где $f(u_k(t))$ называется функцией активации нейрона. Чаще всего в качестве функции активации выбирается нелинейная непрерывная функция, например сигмоидальная функция

$$f(x) = \frac{1}{1 + e^{-\alpha x}} \tag{15}$$

где α – некоторый параметр, который влияет на форму функции активации и подбирается пользователем.

С конца 1980-х гг. многие исследователи начали активно использовать модели нейронных сетей в системах распознавания. Это отразилось на числе работ, посвященных распознаванию

речи с помощью нейронных сетей – оно возросло в несколько раз. Lippmann в 1989 г. написал обзор о состоянии моделей нейронных сетей в распознавании речи на конец 80-х гг. [40].

Самой известной и наиболее распространенной моделью нейронной сети является многослойный персептрон (МП), структурная схема, которого представлена на рис. 3.

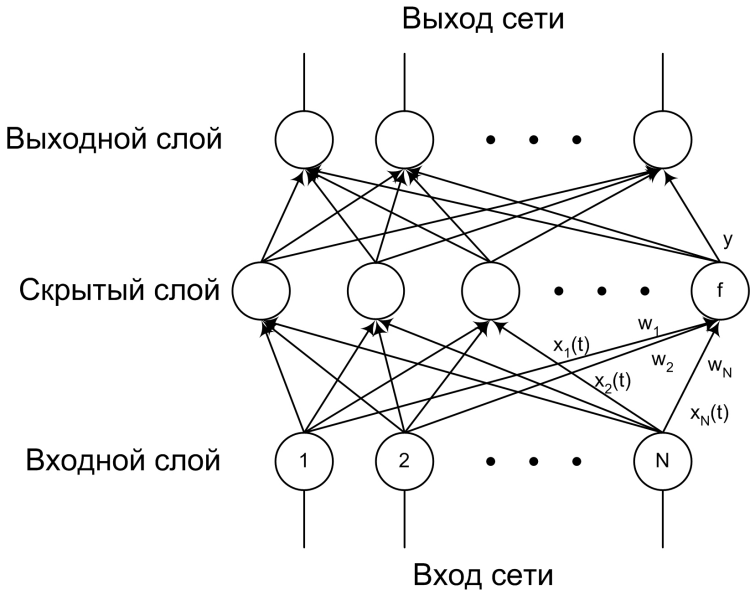

Рис. 3. Трехслойный персептрон

Элементы многослойного персептрона разделены на несколько слоев, внутри слоя элементы можно считать линейно упорядоченными и не взаимодействующими между собой. Каждый нейрон сети (кроме нейронов входного слоя – рецепторов) получает входной сигнал от каждого нейрона предыдущего слоя

и выходной сигнал нейрона (кроме последнего слоя) поступает на вход нейронов последующего слоя. Таким образом, МП является моделью со связями обеспечивающими распространение сигнала только вперед (без обратных связей) – от входа к выходу сети. Элементы промежуточных слоев называются скрытыми элементами, а слои – скрытыми слоями. Сами нейроны чаще всего функционируют в соответствии с моделью МакКаллока-Питса, в качестве функции активации выбирается сигмоидальная функция (14).

Наиболее известным алгоритмом обучения для МП является процедура, описанная Rosenblatt в 1959 г. [62], и как ее модификация предложенный Rumelhart в [63] алгоритм обратного распространения ошибки (Back Propagation Error), который позволяет осуществить управляемое обучение (обучение "с учителем").

BP-алгоритм является градиентным алгоритмом оптимизации, который минимизирует функцию стоимости (целевую функцию) между желаемым и сгенерированным выходом сети. Целью обучения является установление желаемого функционального соотношения входа и выхода путем коррекции весов связей между нейронами. После выбора некоторых начальных значений весов, в процессе обучения итерационно на сеть одновременно подаются входной и желаемый выходной (целевой) вектор. Сеть выполняет отображение входного вектора в выходной. Разность полученного и целевого вектора является ошибкой ε_k, т.е.

$$\varepsilon_k(t) = y_k^{trg}(t) - f(\mathbf{w}_k(t), \mathbf{x}(t)), \tag{16}$$

где $y_k^{trg}(t)$ – целевой выход k-го нейрона на t-м шаге алгоритма,

$\mathbf{w}_k = \{w_{1k}, w_{2k}, ..., x_{Nk}\}$ – весовой вектор k-го нейрона, $\mathbf{x}(t)$ – входной вектор и $f()$ – нелинейная функция активации нейрона. ε_k используется для подстройки $\{w_{kj}\}$ при ее обратном распространении от выхода сети ко входу. В качестве целевых используют различные функции, так, например, среднеквадратичную ошибку

$$E = \sum_{t=1}^{T} \|\mathbf{y}(t) - \mathbf{y}^{trg}(t)\|^2, \tag{17}$$

или функцию относительной энтропии

$$E_e = \sum_{t=1}^{T} \sum_{k=1}^{K} \left[y_k^{trg}(t) \ln \frac{y_k^{trg}(t)}{y_k} + (1 + y_k^{trg}(t)) \ln \left(\frac{1 - y_k^{trg}(t)}{1 - y_k(t)} \right) \right], \tag{18}$$

где $y_k^{trg}(t)$ – целевой, а $y_k(t)$ – наблюдаемый выход k-го нейрона выходного слоя на t-м шаге алгоритма, K – число нейронов в выходном слое и T – общее число обучающих образов.

Основным моментом в обучении сети является способ коррекции весов связей. Поскольку обучение проводится методом наискорейшего спуска, то уточнение весов связей проводится в направлении отрицательного градиента целевой функции, т.е.

$$w_{ij}(t+1) = w_{ij}(t) - \eta \frac{\partial E}{\partial w_{ij}(t)} x_i(t), \tag{19}$$

где η – коэффициент обучения, значение которого, как правило, выбирается из интервала [0,1]. Следует особо отметить, что использование градиентных методов для обучения МП гарантирует достижения только локального минимума на поверхности целевой функции, который может оказаться достаточно далеко

от глобального минимума. Выход из окрестности локального минимума при использовании простого алгоритма наискорейшего спуска невозможен. Для решения этой проблемы обычно используют обучение *с моментом* [57]. При таком методе процесс модификации весов определяется не только информацией о градиенте функции, но и фактическим трендом изменений весов Δw_{ij}, который вычисляется следующим образом:

$$\Delta w_{ij}(t+1) = -\eta \frac{\partial E}{\partial w_{ij}(t)} x_i(t) + \alpha \Delta w_{ij}(t), \tag{20}$$

где первое слагаемое соответствует обычному методу модификации весов, а второе является *моментом*, который отражает последнее изменение весов и не зависит от фактического значения градиента; α – коэффициент момента, значение которого выбирается из интервала $0 < \alpha < 1$. Как видно из (19), влияние момента особенно сильно проявляется в непосредственной близости к локальному минимуму, где значение градиента стремится к нулю, что приводит к возрастанию значений целевой функции и ее выходу из области локального минимума. Однако сильное влияние момента (при больших значениях α) может привести к нестабильности, т.е. расходимости алгоритма обучения.

В первых экспериментах [62] однослойный персептрон показал очень хорошие результаты при обучении в простых нелинейных задачах. Можно показать [5], что однослойный персептрон, как классификатор образов формирует в пространстве признаков дискриминантные гиперплоскости, которые при пересекающихся классах образов и слабо нелинейной пороговой функции минимизируют среднеквадратическую ошибку между y_k и y_k^{trg},

т.е. однослойные персептроны эквивалентны параметрическим гауссовым классификаторам (использование, которых приводит к оценке максимального правдоподобия). Другими словами, для двух классов, образы которых распределены по нормальному закону и в предположении, что признаки, описывающие образы, некоррелированны можно построить однослойный персептрон с такой же решающей функцией, как у параметрического гауссова классификатора.

Однако однослойный персептрон не может разделить образы, требующие для разделения более сложные поверхности в пространстве признаков. Так, например, однослойный персептрон не может решить проблему исключающего ИЛИ путем построения простой гиперплоскости.

С увеличением количества слоев классификационные свойства персептрона качественно улучшаются. Двухслойный персептрон уже может решить проблему исключающего ИЛИ посредством формирования в качестве разделяющей выпуклой поверхности (как результата пересечения гиперплоскостей, формируемых элементами первого слоя), но возможности его также ограничены. Так, Минский и Пэйперт в своей работе [44] доказали, что двухслойный персептрон не может успешно представить или аппроксимировать функции вне очень узкого и специфического класса.

Использование трехслойного персептрона открывает еще большие возможности в аппроксимации отображения из одного конечно размерного пространства в другое, т.е. трехслойный персептрон может формировать разделяющие поверхности лю-

бой формы и получать любые, наперед заданные, непрерывные функции входных сигналов. В частности, с помощью выбора соответствующей решающей функции он может эмулировать любой традиционный детерминированный классификатор [39].

Теоретические основания для выводов о потенциальных свойствах трехслойного персептрона предоставляет результат А. Н. Колмогорова о возможности представления всякой действительной непрерывной функции переменных в виде суперпозиции конечного числа непрерывных действительных функций с глубиной вложения не более трех, в которой используется только линейное суммирование аргументов и непрерывно возрастающие функции одной переменной [2] или более поздние работы [28,41].

Основными мотивационными факторами к использованию многослойного персептрона послужили следующие преимущества нейронных сетей:

- Персептрон может осуществить дискриминантное обучение между речевыми единицами, которые представляют выходные классы персептрона. Персептрон не только обучается и оптимизирует параметры для каждого класса на данных, принадлежащих ему, но и пытается отклонять данные, принадлежащие другим классам.

- Персептрон может найти оптимальную комбинацию ограничений для классификации. При этом нет необходимости в строгих предположениях о распределении входных признаков (что обычно требуется в стандартных СММ).

- Персептрон – это структура с высокой степенью паралле-

лизма, что позволяет использовать параллельное оборудование.

Первые работы по использованию МП в системах распознавания речи [6,71,52] выявили один важный недостаток ИНС. Эти модели были разработаны для распознавания статических сигналов, а не для их последовательностей или сигналов, подверженных временной вариативности. Поэтому МП достаточно удачно использовался в качестве классификатора речевых классов, например таких, как изолированные слова [40], а попытки использовать его в системах распознавания слитной речи не увенчались успехом.

Позднее, чтобы учесть в моделях временную динамику, были предложены различные модификации МП. Так, в 1987 г. была предложена нейронная сеть с задержкой времени (Time-Delay Neural Network (TDNN)) [67,66,68] и рекуррентная нейронная сеть (Recurrent Neural Network (RNN)) [7,53,64].

Нейронная сеть с задерживанием времени (TDNN).

TDNN сеть (рис. 4) реализует одну из попыток использовать статический МП для распознавания динамической временной последовательности речевых данных путем преобразования временной последовательности в пространственную последовательность соответствующих нейронов. TDNN представляет собой МП, в котором в каждый момент времени на нейроны, образующие входной слой, подается не только текущий вектор параметров $\mathbf{x}(t)$ в момент времени t, но и часть последовательности векторов,

взятых с запаздыванием $X_{t-k}^{t} = \{\mathbf{x}(t-k), \mathbf{x}(t-(k-1)), ..., \mathbf{x}(t)\}$ и с опережением $X_{t}^{t+k} = \{\mathbf{x}(t), \mathbf{x}(t+1)), ..., \mathbf{x}(t+k)\}$. При этом получается, что активность каждого нейрона из скрытого слоя зависит от активности нейронов входного слоя на некотором конечном временном интервале X_{t-k}^{t+k}. Аналогично выходной слой связан со скрытым слоем. Как видно из рис. 4, активность выходного нейрона определяется активностью нейронов из скрытого слоя, взятых в моменты времени $t-1, t, t+1$. Число шагов, на которое МП "заглядывает" вперед и назад во времени, выбирается разработчиком модели. Для обучения сети с такой топологией также может использоваться BP алгоритм.

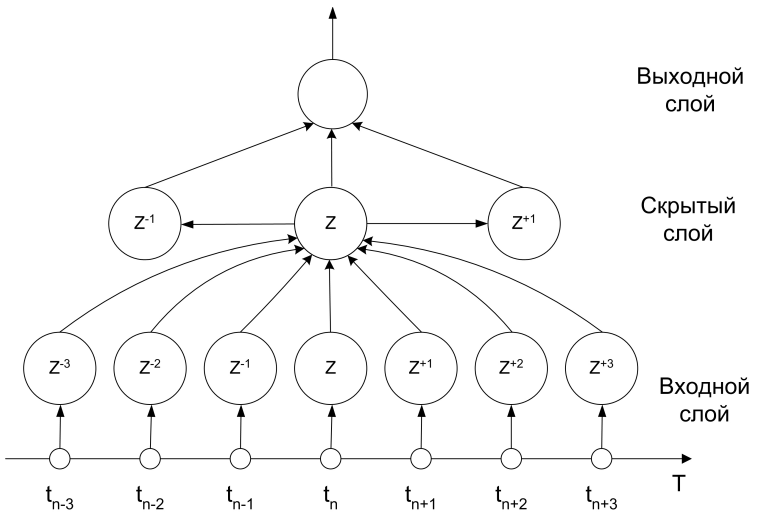

Рис. 4. Нейронная сеть с задерживанием времени

Одним из первых эту модель исследовал. Вайбел [69]. Лэнг и Хинтон[36] использовали TDNN в эксперименте по распо-

знаванию изолированных звуков “B, D, E, V” без подстройки под диктора. Для обучения сети использовался акустический материал, собранный от 100 дикторов мужчин. В результате была достигнута точность 7.8% ошибок. Последующие эксперименты с синтезом модульной сети [68,70], в которой каждый отдельный модуль представлял собой TDNN-сеть, специфицированную для распознавания звуков, показали возможность надежной идентификации всех согласных японского языка, изолированно произносимых дикторами – японцами. Точность распознавания в этих экспериментах достигла 95.9%. При этом точность распознавания гласных звуков в тех же экспериментах достигла 98.6%.

Рекуррентная сеть (RNN).

Другой способ использовать контекстную информацию состоит во введении связей между произвольными нейронами независимо от их топологии в сети. Однако, для того чтобы сеть оставалась МП, эти связи должны быть задержанными на один временной шаг. Это так называемые рекуррентные связи. Пример структуры такой сети приведен на рис. 5. Таким образом, активность нейронов зависит от активности нейронов на предыдущем уровне и от активности нейронов на предущем временном шаге. Такая сеть называется рекуррентной [63] или динамической [51] нейронной сетью (RNN).

Поначалу RNN мало использовались для систем распознавания речи из-за больших сложностей с обучением, анализом и

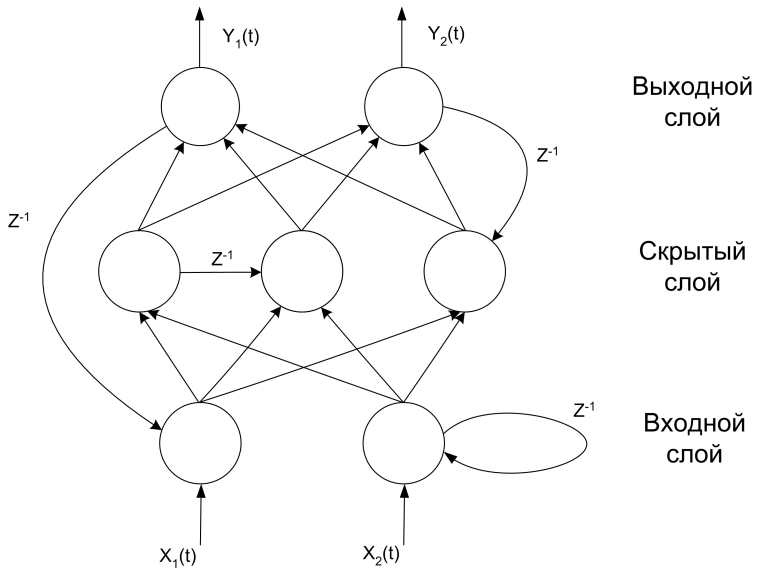

Рис. 5. Рекуррентная нейронная сеть

разработкой. Однако в результате ряда исследований было предложено несколько модификаций алгоритма ВР: рекуррентный ВР [54], ВР для последовательностей [26], рекуррентное обучение в реальном времени [72], время-зависимый рекуррентный ВР алгоритм [50,65] и, наиболее популярный, ВР во времени [63], которые значительно облегчили использование рекуррентных структур в системах распознавания речи [59].

Применение таких модификаций алгоритмов обучения многослойного персептрона привело к повышению качества распознавания кратковременных акустико-фонетических единиц, таких как фонемы, и лишь незначительно улучшили распознавание длительных последовательностей акустических наблюдений, которые необходимы для представления таких лингвистических единиц, как, например, слова. Теоретическое обоснование этого результата приводится в [16]. Кроме того, эти исследования выявили ряд существенных недостатков, которые не позволили сделать ИНС основной структурой для систем распознавания речи, это прежде всего:

- ИНС не имеют механизмов, которые бы адекватно представляли временную вариативность и последовательную природу речевого сигнала;
- для целого ряда параметров, определяющих динамику и топологию ИНС, пока не существует теоретических основ, позволяющих вычислить или выбрать эти параметры (они выбираются по усмотрению разработчика);
- несмотря на то, что разработан целый ряд алгоритмов,

которые ускоряют процедуру обучения, она остается очень ресурсоемким процессом.

Существование двух подходов СММ и ИНС, взаимно дополняющих и компенсирущих присущие им недостатки, в начале 90-х гг. привел исследователей к идее комбинировать эти структуры в рамках одной новой модели, которую определили как гибридную СММ/ИНС модель [20,24,27,37,45,48]. Гибридная модель позволяет эффективно объединить преимушества марковских моделей и нейронной сети, т.е. СММ обеспечивает возможность моделирования долговременных зависимостей, а ИНС обеспечивает непараметрическую универсальную аппроксимацию, оценку вероятности, алгоритмы дискриминантного обучения, уменьшение числа параметров для оценки, которые обычно требуются для стандартных СММ. Результатом использования таких гибридных структур явилось значительное повышение качества распознавания по сравнению со стандартными методами.

Гибридные модели ИНС и СММ

Описание гибридной модели МП и СММ

Как уже отмечалось выше, при использовании СММ в формуле (11) необходимо иметь оценку эмиссионной вероятности $p(x_n|q_l)$, которая представляет собой вероятность наблюдения

вектора x_n, при заданном гипотетическом СММ состоянии q_l. В начале 90-х гг. Боурлард и др. [17,18,20,45] предложили использовать многослойный персептрон для оценки вероятности $p(q_l|x_n)$, которая является апостериорной вероятностью СММ состоянии q_l при заданном наблюдаемом акустическом векторе x_n. Эту вероятность в соответствии с правилом Байеса, можно пересчитать в эмиссионную вероятность.

Формально это выглядит следующим образом. Пусть g_k при $k = 1, ..., K$ – выходное значение k-го нейрона выходного слоя персептрона, тогда g_k можно связать с дискретным СММ состоянием S_k. Теперь, если объединить множество параметров Θ_{HMM}, определенное для СММ с множеством параметров МП Θ_{MLP}, и использовать для обучения последовательность акустических векторов $X = \{x_1, x_2, ..., x_N\}$, размеченную в терминах состояний S_k, т.е. в момент времени n входным вектором для МП является акустический вектор x_n с меткой $q_n = S_k$. Тогда можно показать [17,25,58], что если:

- МП содержит достаточное количество скрытых нейронов, чтобы аппроксимировать функцию отображения входного вектора в выходной,
- МП не "переобучен" ("переобучение" выражается в слишком детальной адаптации весов к несущественным флуктуациям или нерегулярностям обучающих данных, что приводит к значительным погрешностям при распознавании),
- МП не находится в локальном минимуме, после процедуры обучения,

то оптимальное значение выхода МП является распределением вероятностей по дискретным СММ состояниям, которое обусловлено входным вектором

$$g_k(x_n, \Theta_{MLP}^{opt}) = p(S_k|x_n, \Theta_{HMM}), \tag{21}$$

где Θ_{MLP}^{opt} – множество параметров, полученное при обучении МП. Кроме того, в [17] было описано простое расширение предложенной модели с целью использования контекстной информации, т.е. в качестве входа для персептрона использовать последовательность из $2c + 1$ акустических векторов $X_{n-c}^{n+c} = \{x_{n-c}, ..., x_c, ..., x_{n+c}\}$. Тогда (21) можно переписать

$$g_k(x_n, \Theta_{MLP}^{opt}) = p(q_n = S_k|X_{n-c}^{n+c}, \Theta_{HMM}) \qquad \forall k = 1, ..., K. \tag{22}$$

Такое усовершенствование дает возможность учитывать корреляцию акустических векторов, что позволяет преодолеть ограничения, связанные со статистической независимостью векторов наблюдений.

Кроме того, в [17] предложено использовать в качестве входного параметра СММ состояние, вычисленное на предыдущем временном шаге:

$$g_k(x_n, \Theta_{MLP}^{opt}) = p(q_k^n|q_k^{n-1}, \Theta_{HMM}) \qquad \forall k = 1, ..., K. \tag{23}$$

Таким образом, в этой модели используется TDNN сеть, и структура такой системы представлена на рис. 6.

Предложенная вычислительная структура работает следующим образом. В каждый момент времени n на входной слой МП подается последовательность акустических векторов X_{n-c}^{n+c}

и СММ состояние на предыдущем временном шаге q_k^{n-1}, при этом на выходном слое будет формироваться распределение вероятностей по текущему состоянию СММ, обусловленное X_{n-c}^{n+c} и q_k^{n-1}.

Поскольку выходной вектор МП представляет собой аппроксимацию апостериорной вероятности, то $g_k(x_n, \Theta_{MLP}^{opt})$, является оценкой

$$p(q_k|x_n) = \frac{p(x_n|q_k)p(q_k)}{p(x_n)}, \tag{24}$$

которая неявно включает в себя эмиссионную вероятность $p(x_n|q_k)$ и априорную вероятность СММ состояния $p(q_k)$. Поскольку вероятность в (24) участвует как мультипликативный член, то это дает возможность изменять априорную вероятность состояния во время классификации без переобучения персептрона, нормировать выходные вероятности персептрона в зависимости от используемого обучающего речевого корпуса данных. И тогда, чтобы правдоподобие $p(x_n|q_k)$ можно было использовать в качестве эмиссионной вероятности для СММ, необходимо выход персептрона $g_k(x_n)$ поделить на относительную частоту встречаемости состояния S_k в обучающей выборке, что в результате дает нам оценку выражения

$$\frac{p(x_n|q_k)}{p(x_n)}. \tag{25}$$

При распознавании масштабирующий член $p(x_n)$ остается постоянным для всех состояний и не влияет на классификацию.

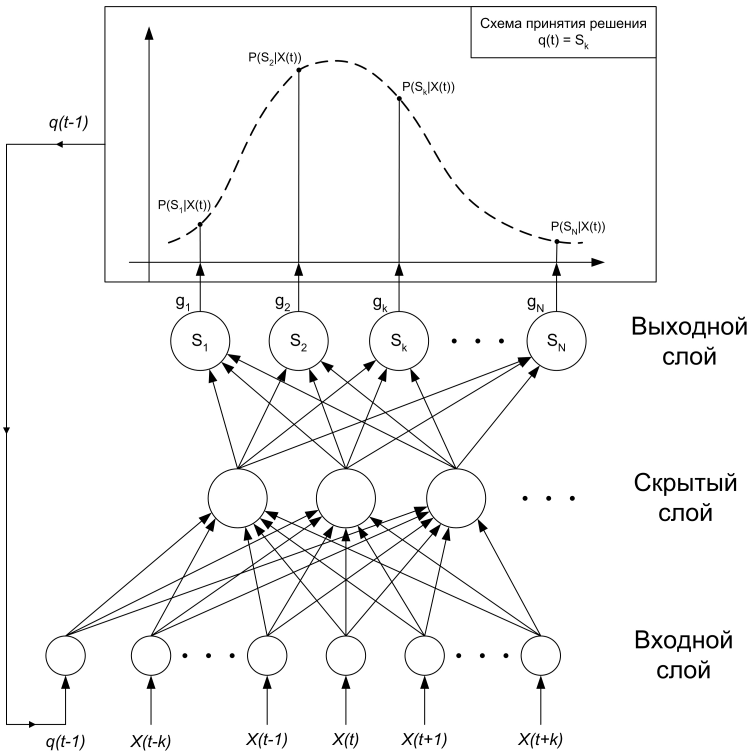

Рис. 6. Оценка вероятности с помощью TDNN сети

Описание гибридной модели рекуррентной сети и СММ.

Аналогичная модель была предложена Робинсоном и др. [30,31,59,60], которые использовали рекуррентную сеть вместо TDNN сети, также для оценки эмиссионных вероятностей СММ.

Авторы предложили в основных уравнениях для линейных динамических систем заменить линейные матричные операторы на нелинейную сеть с обратными связями, и в результате была получена вычислительная структура, приведенная на рис. 7. Текущий акустический вектор x_n поступает на вход сети совместно с текущим вектором состояния u_n. Эти векторы проходят через стандартную сеть без обратных связей, чтобы получить выходной вектор g_n и следующий вектор состояния u_{n+1}. Если определить комбинированный входной вектор как z_n, а матрицу весов связей сети как $\mathbf{W}$ и $\mathbf{V}$, тогда

$$z_n = \begin{bmatrix} 1 \\ x_n \\ u_n \end{bmatrix}, \tag{26}$$

$$g_n^k = \frac{\exp\left(W_k z_n\right)}{\sum_j \exp W_j z_n}, \tag{27}$$

$$u_{n+1}^k = \frac{1}{1 + \exp -V_k z_n}, \tag{28}$$

Включение 1 в (26) дает возможность создать смещение для обеспечения нелинейности. Аналогично модели Боурларда с использованием TDNN сети выход рекуррентной сети представляет собой оценку апостериорной вероятности СММ состояния q_k^n в

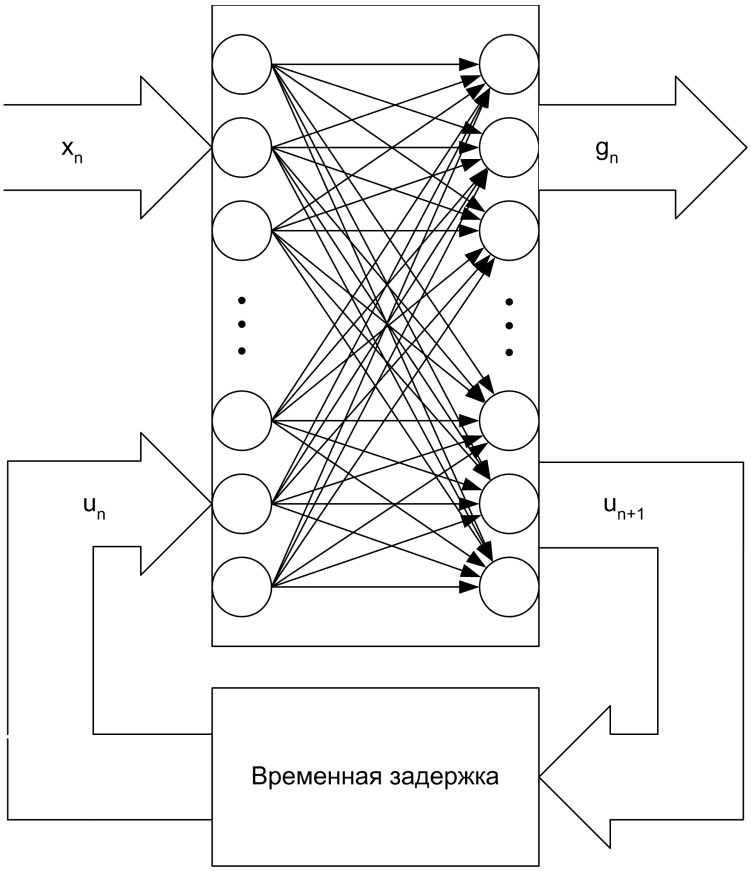

Рис. 7. Рекуррентная нейронная сеть

момент времени n:

$$g_k^n = P(q_k^n | X_1^n, u_0). \tag{29}$$

Теоретические основания для такой интерпретации приведены в работе [42].

При использовании рекуррентной сети для оценки эмиссионной вероятности в гибридноймодели можно получить довольно большой акустический контекст, за счет использования вектора внутреннего состояния u_n.

Как уже отмечалось выше, при использовании СММ делаются предположения о том, что наблюдения статистически независимы и марковский процесс первого порядка, т.е.

$$p(x_n | Q_1^n, X_1^{n-1}) = p(x_n | q_k^n), \tag{30}$$

где $Q_1^n = \{q_1, q_2, ..., q_n\}$ – последовательность СММ состояний в моменты времени $t = 1, 2, ...n$. Использование рекуррентной сети позволяет сократить число предположений, т.е.

$$p(x_n | Q_1^n, X_1^{n-1}) = p(x_n | q_n, X_1^{n-1}), \tag{31}$$

что позволяет учитывать акустический контекст для локальной модели наблюдений. Тогда, переформулировав (10) для модели M_i с учетом (31), получим

$$p(X_1^L | Q_1^L, M_i, \Lambda_i) = \prod_{l=1}^{L} p(x_l | X_1^{l-1}) \frac{P(q_l | x_l)}{P(q_l | X_1^{l-1})}. \tag{32}$$

Так как сомножитель $p(x_l | X_1^{l-1})$ не зависит от последовательности фонов, то на этапе распознавания его можно игнорировать.

Поскольку рекуррентная сеть используется для оценки $P(q_l|x_l)$, то необходимо вычислить оставшийся член $P(q_l|X_1^{l-1})$. Один из простейших способов вычисления – это предположить, что текущее состояние не зависит от наблюдаемого контекста [61], т.е.

$$P(q_l|X_1^{l-1}) = P(q_l), \tag{33}$$

где $P(q_l)$ можно определить как относительную частоту встречаемости состояния q_l в обучающей выборке, т.е. получаем результат аналогичный модели Боурларда.

Обучение гибридной модели

Обучение гибридной модели заключается в оценке параметров как марковской цепи, так и весов нейронной сети. Пока не существует алгоритма, который бы позволил одновременно оценить оба множества параметров и для СММ, и для нейронной сети. Кроме того, поскольку для нейронной сети используется обучение “с учителем”, то требуется значительный объем акустических данных, размеченных вручную, который в настоящее время отсутствует.

Боурлард предложил итерационную процедуру обучения, которая стартует с начальной разметки обучающих акустических данных. На этих данных происходит обучение сети. Далее совместно, используя обученную сеть для оценки эмиссионных вероятностей и алгоритм Витерби, происходит переразметка обучающих данных. На полученной разметке снова происходит

обучение сети и итерация повторяется. Начальная сегментация может быть получена с помощью стандартной СММ или просто делением последовательности акустических наблюдений на равные сегменты, причем каждый сегмент должен быть помечен соответствующим СММ состоянием. Аналогичный метод был предложен в [24].
При использовании гибридных моделей с рекуррентными сетями Т. Робинсон [61] предложил вариант обучения с использованием алгоритма Витерби для оценки параметров системы, который изложен ниже.

Параметры системы модифицируются, используя алгоритм Витерби для максимизации логарифма правдоподобия наиболее вероятной последовательности состояний для обучающих данных. Первый проход алгоритма Витерби делается, чтобы разметить последовательность векторов параметров в терминах СММ состояний. Затем параметры системы подстраиваются, чтобы увеличить правдоподобие последовательности векторов параметров. Эта максимизация происходит в два этапа:

1) максимизация эмиссионных вероятностей,

2) максимизация транзитивных вероятностей.

Эмиссионные вероятности максимизируются, используя метод градиентного спуска, а транзитивные вероятности – переоценкой моделей длительностей. Таким образом, обучающий цикл состоит из следующих шагов:

Шаг 1. Расстановка меток фонов на каждый фрейм обучающих

данных. Эта начальная разметка обычно выполняется экспертом вручную.

Шаг 2. На основе ручной разметки строится модель длительности фонов и вычисляется априорная вероятность фона, которая используется для преобразования выхода рекуррентной сети в оценку правдоподобия.

Шаг 3. Аналогично на основе ручной разметки производится обучение рекуррентной сети.

Шаг 4. Используя параметры, вычисленные на шаге 2, и рекуррентную сеть, обученную на шаге 3, выполняется разметка дополнительных обучающих данных и переход к шагу 2.

В экспериментах [61] было установлено, что для обучения достаточно четырех итераций.

Тестирование гибридных моделей

Гибридные модели использовались в довольно большом числе систем. И показали хорошие результаты. Боурлард и коллеги в период 1988-1994 гг. провели целый ряд успешных экспериментов по применению гибридной модели в системах распознавания речи [17]. Так, например, в систему распознавания слитной речи DECIPHER [21], которая использовалась для задачи управления ресурсами проекта DARPA. Система DECIPHER представляла собой дикторо-независимую систему распознавания слитной речи

построенную на скрытых марковских моделях. Размер словаря составлял 998 слов, с использованием модели языка для пар слов, перплексия равнялась 60, а без модели языка – равнялась 998. Кроме того, использовалось множество вероятностных произносительных транскрипций для слов, фонологическое и акустическое моделирование кросс-слов, контекстно зависимые модели фонов с множеством плотностей.

В системе DECIPHER были использованы как контекстно-независимые, так и контексто-зависимые модели. В первом случае многослойный персептрон был интегрирован в контекстно независимую модель. Базовая система имела 69 моделей фонов с одним распределением эмиссионных вероятностей, каждое слово имело одну произносительную транскриприцию. Модели фонов представляли собой CMM, состояшую из двух или трех состояний с параметрическим связыванием плотностей вероятностей. Этот гибрид сравнивался с CMM системой DECIPHER, в которой эмиссионные вероятности моделировались Гауссовыми смесями. При этом DECIPHER использовался в качестве стартовой системы для получения начальной фонетической разметки на первой итерации обучения многослойного персептрона.

В результате экспериментов было получено значительное улучшение качества распознавания по сравнению с контексто-независимой системой, основанной на CMM. Так на одном из тестовых множествах (Februar91) гибридная контекстно-независимая модель продемонстрировала уровень ошибок 5.8%, что значительно лучше контекстно-независимой CMM модели, уровень ошибок которой составил 11% [47]. Кроме того, в

одном из экспериментов была использована совместная оценка эмиссионных вероятностей как персептроном, так и Гауссовыми смесями, причем для комбинирования этих вероятностей были использованы несколько эвристик, например, вида

$$\log\left(P(x|q_j)\right) = \lambda_1 \log\left(\frac{P_{mlp}(q_j|x)}{P(q_j)}\right) + \lambda_1 \log\left(P_{gm}(x|q_j)\right), \qquad (34)$$

где P_{mlp} обозначает вероятность, оцененную с помощью персептрона, а P_{gm} – с помощью Гауссовых смесей. Набор коэффициентов λ_i был выбран одним для всех состояний. Такой способ оценки продемонстрировал наилучшее качество при уровне ошибок порядка 5.5%.

Аналогичные эксперименты были проведены Робинсоном, использовавшим гибрид СММ и рекуррентной сети в системе распознавания слитной речи ABBOT (Cu-Con), которая была успешно протестирована в рамках проекта November 1993 ARPA Wall Street Journal Test, а также в европейском проекте SQALE (Speech Quality Assessment for Linguistic Engineering)[73]. SQALE был посвящен сравнению нескольких ведущих мировых систем распознавания, таких как Cu-Con и Cu-HTK, созданы в Cambridge University Engineering Department (Великобритания), LIMSI из Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingenieur (Франция) и PHILIPS the Man-Machine-Interface group with Philips Research Laboratories (Германия). Системы Cu-HTK, LIMSI и PHILIPS построены на базе СММ и используют для акустического моделирования непрерывные плотности.

Система Cu-Con использует для акустического моделирова-

ния четыре рекуррентные нейронные сети [32]. Каждая сеть состояла из одного слоя, и ее выход на каждом временном фрейме был вектор оценок вероятностей фонов, при этом в качестве обратной связи использовался 256-размерный вектор состояния, который заводился на входной слой. Полученные таким образом четыре вероятности с выхода каждой сети далее сливались в одну вероятность для каждого фона на каждом входном фрейме. Кроме того, используемые рекуррентные сети обучались для оценки контекст-классов для каждого фона, основанного на векторе состояния каждой сети. Затем выходы такого оценивателя сливались и умножались на контексто-независимую вероятность фона, чтобы получить апостериорную контекстно-зависимую вероятность фона. Контексты выбирались с использованием решающей процедуры на основе кластеризующего дерева [35]. В качестве модели языка системы использовали триграммы и биграммы. Результаты сравнительных экспериментов для американского английского языка при использовании триграмм и биграмм приведены в табл. 1.

Хеннеберг и коллеги [29] предложили усложнение теоретических основ, сформулированных Боурлардом и Морганом путем обобщения локальной апостериорной вероятности на глобальную апостериорную вероятность модели, сформулированную как новый обучающий алгоритм для гибридной модели. Это расширение базируется на работе Франко и коллег [23], в которой контекстно-независимая СММ была заменена на модель, позволяющую интегрировать в себя акустический контекст. Введение зависимости от контекста можно факторизовать по теореме

Система	триграммы	биграммы
Cu-Con	12.9%	17.0%
Cu-HTK	13.2%	16.7%
LIMSI	13.5%	17.2%
PHILIPS	14.7%	20.3%

Таблица 1.

Байеса

$$P(X|q_j, c_k) = \frac{P(q_j|X, c_k)P(X|c_k)}{P(q_j|c_k)}, \tag{35}$$

где X – вектор акустических наблюдений, q_j – CMM состояние и $c_k(k = 1, 2, ..., K)$ – рассматриваемый контекст. В уравнении (35) величина $P(q_j|X, c_k)$ оценивается с помощью многослойного персептрона. В модели Боурларда использовался один персептрон для оценки всех вероятностей состояний $P(q_j|X), q_j \in S$, а в [23] использовались K многослойных персептронов, где K – это длина рассматриваемого контекста. Таким образом, k-й персептрон обучается для оценки величины $P(q_j|X, c_k)$ для всех $q_j \in S$. Для обучения используется стандартный алгоритм *back propagation*. Величину $P(X|c_k)$ аналогично можно факторизовать в виде

$$P(X|c_k) = \frac{P(c_k|X)P(X)}{P(c_k)}, \tag{36}$$

где $P(X|c_k)$ оценивается многослойным персептроном аналогично модели, используемой Боурлардом. Величина $P(c_k)$ вычис-

ляется как частота встречаемости k-го контекста в обучающем корпусе акустических данных.

Заключение

Описанная гибридная модель нашла применение во многих системах распознавания слитной речи с большими словарями и продемонстрировала очень неплохие результаты по сравнению с системами, построенными на основе каждой из моделей, составляющих гибрид, в отдельности. Исследования, проведенные с описанными системами показали, что несмотря на относительную простоту структуры они обладают целым рядом потенциальных преимуществ (по сравнению со стандартными СММ), которые были подтверждены на практике:

- Точность модели – оценка вероятностей с помощью нейронной сети не требует детальных предположений о форме статистических распределений, которые должны быть промоделированы. В результате можно получить более точные акустические модели.

- Дискриминантная способность: с помощью нейронной сети значительно проще реализовать дискриминантное обучение.

- Учитывание контекста – поскольку описанные модели нейронных сетей могут использовать акустический контекст, то

локальная корреляция акустических векторов может быть учтена при вычислении распределений вероятностей. По различным причинам нечто подобное трудно реализовать в стандартных СММ.

- Экономное использование параметров (снижение размерности системы) – так как все распределения вероятностей представлены тем же множеством разделяемых параметров. Хорошо известно, что более "экономично" моделировать границы между акустическими классами, чем поверхности функций плотностей (т.е. правдоподобий).

- Гибкость – использование нейронных сетей для оценивания акустической вероятности позволяет легко сочетать разнообразные параметры, например такие, как смесь непрерывных и дискретных измерений.

- Комплементарность – в некоторых системах нейронная сеть снабжает дополнительной информацией базовую СММ систему. Так, например, в одном из экспериментов комбинация СММ с нейронной сетью (названной, "сегментной нейронной сетью") позволила значительно повысить качество распознавания [8,9].

Однако несмотря на достигнутые успехи, необходимо продолжать исследовательские работы, направленные на разработку гибридных сруктур, позволяющих проводить глобальное дискриминатное обучение, т.е. на разработку моделей, основанных на одновременном оценивании обоих множеств параметров как

СММ, так и нейронной сети при использовании одного критерия оптимизации. Кроме того, пока остаются открытыми вопросы, связанные с адаптацией таких систем, т.е. адаптации к диктору или адаптации к каналу связи, также необходимо повышать устойчивость систем при работе в шумной обстановке.

Литература

1. *Елинек Ф.* Распознавание непрерывной речи статистическими методами // ТИИЭР. 1976. Т. 64. № 4. С. 131-160.

2. *Колмогоров А.Н.* О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного и сложения // ДАН АН СССР. 1957. Т. 114. № 5. С. 953-956.

3. *Левинсон С.Е.* Структурные методы автоматического распознавания речи // ТИИЭР. 1985. Т. 73. № 11. С. 100-128.

4. *Маккоул Дж., Рукос С., Гиш Г.* Векторное квантование при кодировании речи // ТИИЭР. 1985. Т. 73. № 11. С. 19-61.

5. *Ту Д., Гонсалес Р.* Принципы распознавания образов // Пер. с англ. под ред. Ю.И. Журавлева. М.: Мир, 1987. 411 с.

6. *Цыпкин Я.З.* Обучение и адаптация в автоматических системах // М.: Наука, 1968. 400 с.

7. *Almeida L.B.* A Learning Rule for Asynchronous Perceptrons with Feedback in a Combinatorial Environment // In: 1st International Conference on Neural Networks. IEEE. 1987. II-609.

8. *Austin S., Zavaliagkos G., Makhoul J., Schwartz R.* Speech recognition using segmental neural nets // IEEE ICASSP, San Francisco, March 1992, pp. I-625-628.

9. *Austin S., Zavaliagkos G., Makhoul J., Schwartz R.* Improving state-of the-art continuous speech recognition system using the N-best paradigm with neural networks // Proceedings DARPA Speech and Natural Language Workshop, Harriman, NY (Morgan Kaufmann, Los Altos, CA). 1992. pp. 180-184.

10. *Bahl L.R. and Jelinek F.* Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition // IEEE Trans. Informat. Theory. 1975. Vol. IT-21, pp. 404-411.

11. *Baker J.K.* The DRAGON system - An overview // IEEE Trans. on Acoust. Speech Signal Process. 1975. Vol. ASSP-23. No. 1. pp. 24-29

12. *Baum L.E., Petrie T.* Statistical inference for probabilistic functions of finite state Markov chains // Ann. Math. Stat. 1966. Vol.37. pp. 1554-1563.

13. *Baum L.E., Egon J.A.* An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology // Bull. Amer. Meteorol. Soc. 1967. Vol. 73. pp. 360-363.

14. *Baum L.E., Petrie T., Soules G., and Weiss N.* A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains // Ann. Math. Stat. 1970. Vol 41. No. 1. pp. 164-171.

15. *Baum L.E.* An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes // Inequalities. 1972. Vol. 3. pp. 1-8.

16. *Bengio Y., Simard P., Frasconi P.* Learning long-term dependencies with gradient descent is difficult // IEEE Transaction on Neural Networks 5 (2) (1994) pp. 157-166. (Special Issue on Recurrent Neural Networks, March 94).

17. *Bourlard H., Morgan N.* Connectionist Speech Recognition. A Hybrid Approach // The Kluwer International Series in Engineering and Computer Science, Vol. 247, Kluwer Academic Publishers, Boston, 1994.

18. *Bourlard H., Morgan N.* Continuous speech recognition by connectionist statistical methods // IEEE Transaction on Neural Networks. 1993. Vol. 4. No. 6. pp. 893-909.

19. *Bourlard H., Morgan N.* Hybrid connectionist models for continuous speech recognition // In: C.H. Lee, F.K.

Soong, K.K. Paliwal (Eds), Automatic Speech and Speaker Recognition: Advanced Topics, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, USA 1996.

20. *Bourlard H., Wellekens C.* Links Between Markov Models and Multilayer Perceptrons // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1990. Vol. 12. No. 12. pp. 1167-1178.

21. *Cohen M., Murveit H., Bernstein H., Price P., Weintraub M.* The DECIPHER speech recognition system // IEEE ICASSP, Albuquerque, 1990. pp. 77-80.

22. *Dempster A.P., Laird N.M., and Rubin D.B.* Maximum likelihood from incomplete data via the EM algorithm // J. Roy. Stat. Soc. 1977. Vol. 39, No. 1. pp. 1-38.

23. *Franco H., Cohen M., Morgan N., Rumelhart D., Abrash V.* Context-dependent connectionist probability estimation in a hybrid hidden Markov model-neural net speech recognition system // Computer Speech and Language. 1994. 8. pp. 211-222.

24. *Franzini M.A., Lee K.F., Waibel A.* Connectionist Viterbi training: a new hybrid method for continuous speech recognition // IEEE ICASSP 1990, pp. 425-428.

25. *Gish H.* A probabilistic approach to the understanding and training of neural network classifiers // IEEE ICASSP 1990.

pp. 1361-1364.

26. *Gori M., Bengio Y., R. De Mori* BPS: a learning algorithm for capturing the dynamical nature of speech // Proceedings of the International Joint Conference on Neural Networks, Washington, DC, IEEE, New York, 1989, pp. 643-644.

27. *Haffner P., Franzini M.A., Waibel A.* Integrating time alignment and neural networks for high performance continuous speech recognition // IEEE ICASSP 1991. pp. 105-108

28. *Hecht-Nielsen R.* Kolmogorov's mapping neural network existence theorem // in IEEE First International Conference on Neural Networks, pp. III:11-14, San Diego: SOS Printing.

29. *Henneberg J., Ris C., Bourlard H., Renals S., Morgan N.* Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems // Proceedings of EUROSPEECH, 1997. Vol. 4, Rhodi, pp. 1951-1954.

30. *Hochberg M. M., Renals S. J., Robinson A. J., Kershaw D. J.* Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system // Proceedings of CSLP, Yokohama, 1994. pp. 1499-1502.

31. *Hochberg M. M., Renals S. J., Robinson A. J., Cook G. D.* Recent improvements to the ABBOT large vocabulary csr system // IEEE ICASSP, Detroit, 1995. pp. 62-72.

32. *Hochberg M., Renals S. & Robinson A.* ABBOT: the CUED hybrid connectionist-HMM large vocabulary recognition

system // Proceedings of the Spoken Language Technology Workshop, 1995. pp.170-178, Austin, TX, U.S.A.

33. *Jelinek F.* A fast sequential decoding algorithm using a stack // IBM J. Res. Develop., 1969. Vol. 13. pp. 675-685.

34. *Jelinek F., Bahl L.R., and Mercer R.L.* Design of a linguistic statistical decoder for the recognition of continuous speech // IEEE Trans. Informat. Theory, 1975. Vol. IT-21. pp. 250-256.

35. *Kershaw D. J., Hochberg M. M., Robinson A. J.* Context dependent classes in a hybrid recurrent network-HMM speech recognition system // Cambridge University Engineering Department, Technical Report, CUED/F-INFENG/TR.217.1995.

36. *Lang K.J., Hinton G.E.* The development of the time-delay neural network architecture for speech recognition // Technical Report CMU-CS-88-152, Carnegie-Mellon University, 1988

37. *Levin E.* Word recognition using hidden control neural architecture // IEEE ICASSP 1990.

38. *Levinson S.E., Rabiner L.R., and Sondhi M.M.* An introduction to the application of the theory of probabilistic function of a Markov process to automatic speech recognition // Bell Syst. Tech. Journal, Apr. 1983. Vol. 62, no.4, pp. 1035-1074.

39. *Lippmann R.P.* Neural nets for computing // IEEE ICASSP 1988. Vol. 1, pp. 1-6.

40. *Lippmann R.P.* Review of neural networks for speech recognition // Neural Computing, 1989. 1. pp. 1-38.

41. *Lorentz G.G.* The thirteenth problem of Hilbert // In F.E. Browder (Ed), Proceedings of Symposia in Pure Mathematics, Vol. 28, pp. 419-430. Providence, RI: American Mathematical Society.

42. *McCullagh P., Nelder J. A.* Generalized Linear Models // London: Chapman and Hall, 1983.

43. *McCulloch W. S., Pitts W. H.* A logical calculus of ideas immanent in nervous activity // Bull. Math. Biophysics, 1943. Vol. 5. pp. 115-119.

44. *Minsky M., Papert S.* Perceptrons // Cambridge: MIT Press. 1969.

45. *Morgan N., Bourlard H.* Continuous speech recognition using multilayer perceptrons with hidden Markov models // ICCASP 1990, pp. 413-416

46. *Morgan N., Bourlard H.* Neural Network for Statistical Recognition of Continuous Speech // Proceedings of IEEE, 1995. Vol.83, No. 5. pp.742-770.

47. *Morgan N., Bourlard H.* Hybrid neural network/ hidden Markov model system for continuous speech recognition // Intl. Journal of Pattern Recognition and Arificial Intelligence, Special Issue on Advances in Pattern Recognition Systems

using Neural Networks (I. Guyon and P. Wang, Eds.). 1993. Vol. 7, No. 4.

48. *Niles L.T., Silverman H.F.* Combining hidden Markov models and neural networks classifiers // IEEE ICASSP 1990. pp. 417-420.

49. *Paul D.B., Baker J.K., Baker J.M.* On the interaction between true source, training and testing language models // IEEE ICASSP 1991. pp. 569-572.

50. *Pearlmutter B. A.* Learning state space trajectories in recurrent neural networks // Neural Comput. 1989. No. 1. pp. 263-269.

51. *Pearlmutter B. A.* Dynamic Recurrent Neural Networks // Technical Report CMU-CS-88-191, Carnegie-Mellon University, Computer Science Dept. Pittsburg, PA. 1990.

52. *Peeling S.M. and Moore R.K.* Experiments in Isolated Digit Recognition Using Multi-Layer Perceptron // Technical Report 4073, Royal Speech and Radar Establishment, Malvern, Worcesber, Great Britain, 1987.

53. *Pineda F.J.* Generalization of Back-Propagation to Recurrent Neural Networks // Physical Review Letters 59, 1987. pp. 2229-2232.

54. *Pineda F.J.* Recurrent back-propagation and the dynamical approach to adaptive neural computation // Neural Computing, 1989. No. 1. pp. 161-172.

55. *Rabiner L. R., Juang B.-H. and Lee C.-H.* An Overview of Automatic Speech Recognition // In: C.H. Lee, F.K. Soong, K.K. Paliwal (Eds), Automatic Speech and Speaker Recognition: Advanced Topics, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, USA 1996.

56. *Rabiner L.R.* A tutorial on hidden Markov models and selected application in speech recognition // Proceedings of the IEEE, 1989, Vol. 77, 2. pp. 257-286. Русский перевод: Л.Р. Рабинер, Скрытые Марковские модели и их применение в избранных приложениях при распознавании речи: Обзор. ТИИЭР. 1989. Т. 77. № 2 февраль. С. 86-120.

57. *Rahim M. R.* Artificial Neural Networks for Speech Analysis/Synthesis // Chapman&Hall, 1994

58. *Richard M.D., Lippmann R.P.* Neural network classifiers estimate Bayesian a posteriori probabilities // Neural Computation, 1991. No. 3. pp. 461-483.

59. *Robinson A.J., Fallside F.* Static and dynamic error propagation network with application to speech coding // In: D.Z. Anderson (Ed.), Neural Information Processing System, American Institute of Physics, New York, Denver, CO, 1988, pp. 635-641.

60. *Robinson T.* An application of recurrent nets to phone probability estimation // IEEE Transaction on Neural Networks, 1994. Vol. 5. No. 2. pp. 298-305.

61. *Robinson T., Hochberg M., Renals S.* The use of recurrent neural networks in continuous speech recognition // In: C.H. Lee, F.K. Soong, K.K. Paliwal (Eds), Automatic Speech and Speaker Recognition: Advanced Topics, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, USA 1996.

62. *Rosenblatt F.* Principles of Neurodynamics // Spartan Books, New York, 1959. Русский перевод: Розетблатт Ф. Принципы нейродинамики (перцептрон и теория механизмов мозга). М.: Мир, 1965. 480 с.

63. *Rumelhart D. E., Hinton G. E., Williams R. J.* Learning internal representations by error propagation // In: Rumelhart, D. E., G. E. Hinton, (eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1 Foundations., chapter 8. Bradford Books/MIT Press, Cambridge, MA, 1986 ISBN 0-262-18120-7.

64. *Rumelhart D. E., Hinton G. E. and Williams R. J.* Interactive Processes in Speech Perception: The TRACE Model // In: Parallel Distributed Processing: Vol. 2, Psychological and Biological Models, eds. D. E. Rumelhart and J.L. McClelland. Cambridge, MA: MIT Press. 1986.

65. *Sato M.* A real time learning algorithm for recurrent analog neural networks // Biol. Cybernet. 62, 1990. pp. 237-241

66. *Sawai H., Waibel A., Miyatake M., Shicano K.* Spotting Japanese SV-syllables and phonemes using time-delay neural

networks // IEEE Internation Confernce on Acoustic, Speech and Signal Processing, ICASSP 1989, Vol. 1, pp. 25-28.

67. *Tank D.W., Hopfield J.J.* Concentrating information in time: analog neural network with application to speech recognition problems // Int.conf.on neural networks, ICNN, 1987. pp. 455-468.

68. *Waibel A.* Modular construction of time-delay neural networks for speech recognition // Neural Comput. 1989. No. 1. pp. 39-46.

69. *Waibel A., Hanazawa T., Hinton G., Shikano K. & Lang K.* Phoneme Recognition Using Time-Delay Neural Networks // IEEE Transaction on Acoustic Speech Signal Processing Vol. 37, 1989, pp. 328-339.

70. *Waibel A., Sawai H., Shikano K.* Modularity and scaling in large phonemic neural networks // IEEE Transaction Acoustic Speech Signal Processing Vol. 37, 1989, pp. 1888-1898.

71. *Widrow B., Hoff M.* Adaptive switching circuits // Proc. IRE WESCON Convention Record, 1960. pp. 107-115.

72. *Williams R. J., Zipser D.* A learning algorithm for continually running fully recurrent neural networks // Neural Comput. 1989. No. 1. pp. 87-111.

73. *Young S.J., Adda-Dekker M., Aubert X.* Multilingual large vocabulary speech recognition: the European SQALE project // Computer Speech and Language, 1997, 11, pp. 73-89.

Выделение незнакомых слов и акустических событий при распознавании речи[1]

В.Я.Чучупал

Аннотация

В работе содержится анализ существующих методов обнаружения незнакомых слов на основе мер сходства, приведен новый алгоритм выявления незнакомых слов на основе оценок правдоподобия для наблюдаемого речевого сигнала и результаты численных измерений оценок эффективности нового алгоритма. Ключевые слова: распознавание речи, обнаружение незнакомых слов, оценки достоверности результатов распзнавания.

Введение

Автоматическое распознавание речи постепенно становится привычным атрибутом жизни современного человека. Однако широкого использования технологий распознавания речи в повседневной деятельности, несмотря на многочисленные прогнозы известных специалистов [1], пока не наблюдается. Одной из основных причин этого является недостаточная робастность существующих методов распознавания речи. В данном случае под робастностью понимается свойство сохранения характеристик

[1]При финансовой поддержке РФФИ, грант 04-01-00588

системы распознавания речи в реальных условиях эксплуатации, например при наличии шумов и изменений характеристик канала связи.

В современных методах распознавания речи моделирование речевого потока осуществляется при помощи инвентарей акустических моделей звуков речи и произносительных транскрипций слов. Оценка параметров моделей звуков производится на обучающих выборках – речевых корпусах данных. В реальной ситуации речевой сигнал, подлежащий распознаванию, наблюдается в каналах связи с иными, по сравнению с обучающим корпусом данных, свойствами. Сигнал, который поступает с постоянно включенного микрофона, содержит шумы, речь посторонних лиц, различные нарушения речевого потока (смех, кашель, оговорки) или слова, которые не входят в словарь системы распознавания речи. Влияние этих факторов приводит к ошибкам распознавания.

Система распознавания речи обычно служит для реализации интерфейса человека с некоторой прикладной системой. В этом случае требуется обеспечить достоверность не столько текста сообщения, сколько его смысла. Поэтому логичным представляется оформление результата распознавания речи в виде нескольких, наиболее вероятных альтернатив о тексте высказывания, ранжированных по степени достоверности. Например, каждое предполагаемое слово может сопровождаться численной оценкой достоверности распознавания.

Вопросы повышения робастности систем распознавания и формирования оценок достоверности результатов в силу важно-

сти исследуются достаточно давно [2] – [4]. Если задача состоит в том, чтобы найти незнакомые слова или акустических события, то существует два основных пути решения.

Во–первых, это явное моделирование всего звукового потока, например за счет введения специальных слов – 'заполнителей', которые не входят в словарь системы. При этом словарь системы может быть расширен за счет моделирования некоторых распространенных невербальных речевых событий (кашель, аспирация, заполненные паузы). Это – наиболее естественный путь, который соответствует технологии моделирования и обработки речевого потока на основе марковских цепей. У него, однако, есть существенные недостатки. Возможность достаточно адекватно представлять все акустическое многообразие вокруг нас с помощью ограниченного числа заполнителей представляется маловероятной. Кроме того, для распознавания небольших словарей использование заполнителей неэкономично.

Другой способ выявления незнакомых слов основан на использовании величин локальных мер сходства, например оценок правдоподобия. Величина меры сходства для слова сравнивается с ее ожидаемым значением и на основании этого сравнения делается заключение о достоверности распознавания слова. Такой подход обычно просто реализуется и требует мало дополнительных ресурсов, поскольку вычисление оценок правдоподобия является частью стандартной процедуры обработки и распознавания речевого сигнала.

К недостаткам этого подхода относится то, что наблюдаемые значения величин правдоподобия для конкретного слова весьма

вариативны. Они существенно зависят не только от фонемного состава слова, но и акустических характеристик голоса диктора, свойств канала связи, акустико–фоновой обстановки и параметров моделей звуков речи. Более того, эти оценки могут существенно зависеть также от случайных факторов, связанных, например, с малым (3–5 мс) смещением окна анализа.

Поэтому на практике используются относительные оценки мер сходства, выраженные в виде отношения правдоподобия, где наблюдаемая величина нормируется средним значением мер сходства для данного слова. Более хорошие результаты получаются при построении сложных оценок достоверности, которые комбинируют несколько видов мер сходства [2].

Отметим, что оба способа выявления незнакомых слов, в принципе, эквивалентны, так как в обоих случаях принимаемое решение основано на величине правдоподобия наблюдаемых данных. В первом случае величина правдоподобия используется для нахождения мер сходства и принятия решения на основании их величин, во втором случае величина правдоподобия используется неявно – маловероятное слово при распознавании будет заменено другим, более правдоподобным.

Целью настоящего исследования является анализ существующих методов обнаружения незнакомых слов на основе мер сходства, разработка и численное исследование эффективности нового алгоритма выявления незнакомых слов на основе оценок правдоподобия для наблюдаемого речевого сигнала при заданном множестве акустико-фонетических моделей.

С практической точки зрения, методы, основанные на ис-

пользовании оценок мер сходства можно условно разделить на три группы в соответствии с видом используемой априорной информации:

- дискриминантные методы, которые основаны на модификации параметров моделей таким образом, чтобы максимально разнести меры сходства для ‘своих’ и ‘чужих’ слов;
- методы, основанные на апостериорных оценках мер сходства словарных слов, получаемых на настроечной выборке;
- методы с использованием априорной информации о счетах, например из параметров акустико-фонетических моделей и произносительных транскрипций слов.

Наиболее точные, по опубликованным результатам, методы основаны на использовании дискриминантных [9] моделей. Такой подход предполагает наличие достаточно представительных настроечных выборок для ‘своих’ и ‘чужих’ слов, что на практике может быть затруднительным. Кроме того, полученные дискриминантными методами оценки параметров существенно зависят от настроечных данных и являются ‘оптимальными’ только для них.

Более простым и естественным представляется путь к построению методов обнаружения незнакомых слов на основе существующих в системе распознавания оценок параметров моделей звуков.

Стандартные акустические счета

При распознавании речевой сигнал разбивается на участки (кадры) длительностью 10–30 мс. На кадре производится оценка вектора параметров, который рассматривается как наблюдение. Для каждого наблюдения (вектора параметров) вычисляется его правдоподобие в условиях данной акустической модели. Будем называть акустическим счетом величину правдоподобия наблюдаемых параметров, а также производные от нее функции, которые не зависят от других, неизвестных, параметров.

Существует несколько вариантов оценок счета при известной длительности состояний слова. Наиболее часто используются средний счет:

$$\tilde{S}(w) = \frac{1}{N_w} \sum_{j=1}^{N_w} s_j, \tag{1}$$

где N_w – длительность (число наблюдений) слова w, s_j – счет наблюдения j и центрированный средний счет:

$$\tilde{S}(w) = \frac{1}{N_w} \sum_{j=1}^{N_w} (s_j - s) \tag{2}$$

где s_j – средний счет, который вычислен для длительного промежутка времени.

Обе оценки широко используются при решении проблемы выявления незнакомых слов, однако несколько лучшие результаты были получены при использовании дважды нормированного счета [7]:

$$S_2(w) = \frac{1}{N_w} \sum_{j=1}^{N_w} \frac{1}{N_s} \sum_{j=1}^{N_s} s_j \tag{3}$$

где N_w – число состояний в модели слова w, N_s – длительность состояния s, s_j – счет наблюдения j.

По аналогии с (2) можно определить дважды нормированный центрированный счет:

$$S_2(w) = \frac{1}{N_w} \sum_{j=1}^{N_w} \frac{1}{N_s} (s_j - \tilde{s}), \tag{4}$$

где $\tilde{s}_j$ – средний апостериорный счет на наблюдаемом высказывании.

Оценка эффективности аналитических счетов и заполнителей

Эффективность акустических счетов как меры достоверности результата распознавания может быть описана частотой совершаемых ошибок первого и второго рода, в данном случае пропусков незнакомых слов (когда оно распознается как корректное слово) и ложных тревог, когда корректное, словарное слово интерпретируется как незнакомое. В том случае частота ошибок распознавания равна

$$ERR = \frac{I + D}{N} * 100\% \tag{5}$$

где I – число вставленных слов, D – число пропусков, N – общее число слов (корректное).

Потенциальная эффективность счета наглядно отображается операционной характеристикой приемника, либо характеристикой обнаружения – DET кривой [6], где ось абсцисс соответствует проценту ложных тревог, а ось ординат – пропусков, взятых в логарифмическом масштабе.

Другой характеристикой эффективности счета является пословная ошибка распознавания, которая широко используется при оценке систем распознавания речи (WER – word error rate) [5]. Пословная ошибка вычисляется как:

$$ERR = \frac{I + S + D}{N} * 100\% \tag{6}$$

где I – число вставленных слов, S – число замен, D – число пропущенных слов и N – общее (корректное) число слов.

Сравнительное исследование эффективности счетов (1)—(4) выполнялось на материале корпуса данных FaVoR, который содержит речевой материал от 1673 дикторов. Словарь корпуса включает цифры от 0 до 9, записанные как в раздельном произношении, так и слитно (последовательности из 5 цифр) и служебные команды: ‘да’, ‘нет’, ‘старт’ и ‘стоп’. Корпус записан в естественной, достаточно шумной акустико–фоновой обстановке (среднее отношение сигнал/шум равно 15 дБ) с присутствием значительного количества различных незнакомых слов.

Для оценки эффективности счетов (1)—(4) корпус данных был разделен на три части: обучающая выборка (654 человека), настроечная выборка (197 человек) и тестовая выборка (822

человека). Всего тестовая выборка состояла из 1019 записей, которые содержали произнесения 29552 слов.

Поскольку характеристики существующей системы достаточно хорошие: с вероятностью правильного распознавания выше 95%, при оценке операционных характеристик проблемой являлась оценка случаев ‘ложных тревог’, когда для данного слова нужно найти примеры его распознавания при произнесении другого слова. Поскольку таких данных не хватало, они создавались искусственным путем. Из словаря удалялось слово, близкое по акустическим характеристикам к проверяемому. Затем выполнялся тест на распознавание, который и содержал нужное количество ложных тревог.

На рис.1 и 2 приведены DET кривые для счетов (1)—(4), которые были расчитаны для относительно длинного слова ‘восемь’ и короткого слова ‘два’.

Очевидно, что дважды нормированный (3) и дважды нормированный нормализованный (4) счета в обоих случаях существенно лучше разделяют ‘свои’ и ‘чужие’, а дважды нормированный нормализованный счет – самый эффективный.

Так как величины счетов (1)—(4) зависят от произносительной транскрипции, то оптимальные пороги для принятия решения являются индивидуальными для каждого варианта произнесения слова.

Сравним эффективность обнаружения незнакомых слов на основе счетов с методами обнаружения на основе моделей заполнителей. Для корпуса данных FaVoR можно предложить набор заполнителей, который достаточно адекватно отражает

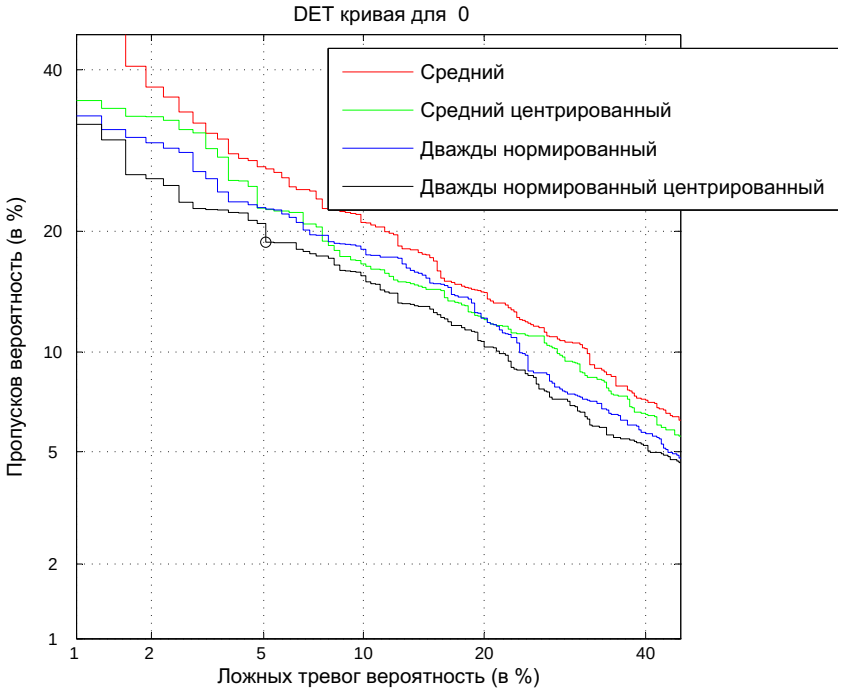

Рис. 1. DET характеристики акустических счетов для цифры 0

реально наблюдаемые незнакомые слова. Это различные варианты нарушений речевого потока (сильные вдохи, выдохи, кашель, смех, шлепки губами).

При использовании заполнителей потока решение выносилось без использования величин порогов, т.е. принималось по наилучшей гипотезе. Слово считалось пропущенным, если оно отсутствовало в предложении на выходе декодера, но фактически было произнесено. Слово считалось вставкой, если оно присутствовало

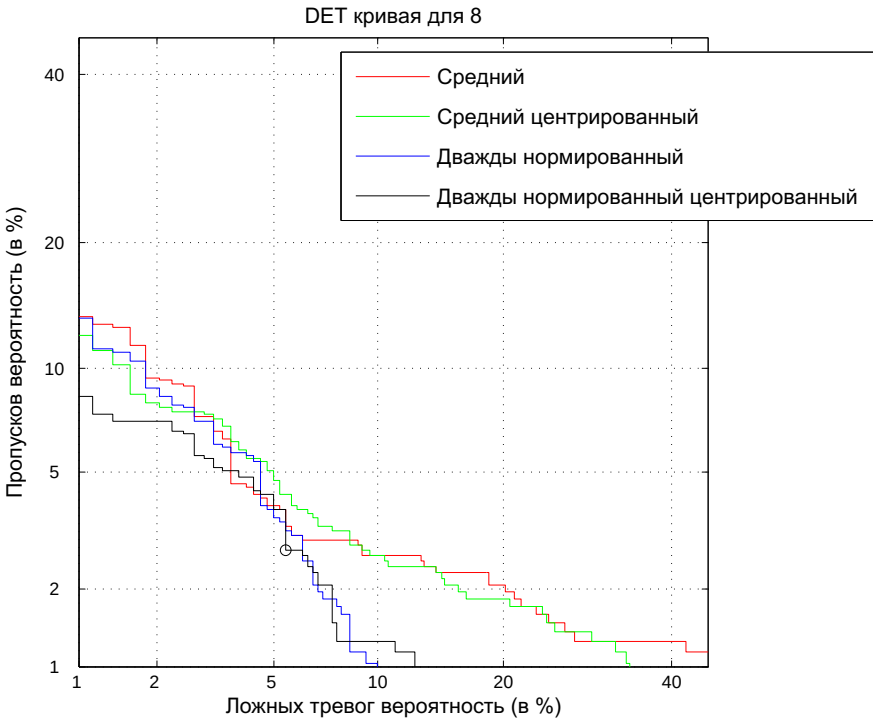

Рис. 2. DET характеристики акустических счетов для цифры 8

в предложении на выходе декодера, но реально не произносилось. В качестве меры эффективности использовалась пословная ошибка WER (5). Использовался дважды нормированный счет (3). Предварительная оценка порогов этого счета проводилась на настроечной части корпуса данных, пороги при этом выбирались так, чтобы ошибка второго рода (распознавание корректного слова как незнакомого) не превышала 0.5%.

Результат экспериментов приведен в табл.1.

Таблица 1. Пословная ошибка распознавания при использовании порога обнаружения и моделей заполнителей

Пороги	Заполнители	Ошибка	Замены	Вставки	Пропуски
Нет	Нет	4,89	465	899	84
Нет	Есть	3,75	467	546	95
Есть	Нет	3,98	385	310	483
Есть	Есть	3,77	387	223	506

Как видно из приведенных результатов, использование заполнителей приводит в случае 'точечного' решения к более высокой точности обнаружения незнакомых слов. Почему заполнители лучше, чем аналитические счета? Потому что акустические условия (множество дикторов, например) тестовых записей не идентичны условиям обучающей и настроечной выборок. Поэтому требуется адаптация значений счетов, даже если используются относительные счета. Использование заполнителей неявно такую адаптацию значений осуществляет.

Адаптация акустических счетов по наблюдениям

Величина акустического счета характеризует степень соответствия наблюдаемых параметров выбранным моделям. Поскольку обучение (оценка параметров) моделей обычно проводится на

данных, которые включают голоса самых разных людей, в разнообразных акустико–фоновых условиях, можно ожидать более точного соответствия моделей наблюдениям, если эти модели либо акустические счета будут адаптированы к конкретной ситуации. Примером такой адаптиции счетов являются центрированные счета (2) и (4), значения которых получены вычитанием ‘среднего’ значения.

Использование ‘среднего’ может быть заменено на текущие оценки смещения счетов относительно их ожидаемых значений, которые могут быть рассчитаны аналитически, исходя из произносительной транскрипции слов и параметров акустических моделей.

На рис.3 показана последовательность 860 наблюдаемых акустических счетов для слова ‘восемь’. Помимо разброса значений относительно среднего можно также отметить наличие слабого тренда, что приводит к заключению о возможности использования метода адаптивной коррекции среднего.

Попробуем ввести линейную коррекцию значения счета. Будем оценивать текущее значение среднего счета как

$$\tilde{m}(w, t+1) = \tilde{m}(w, t) + \alpha * (\tilde{m}(w, t) - S(w, t)), \qquad (7)$$

где $\tilde{m}$ – оценка среднего, $S(w, t)$ – наблюдаемое значение счета, α – коэффициент адаптации.

Начальное приближение для среднего значения счета $S(w, 0)$ и величина отклонения решающего порога от среднего σ оцениваеются по настроечной выборке. Таким образом, решающий порог в момент t :

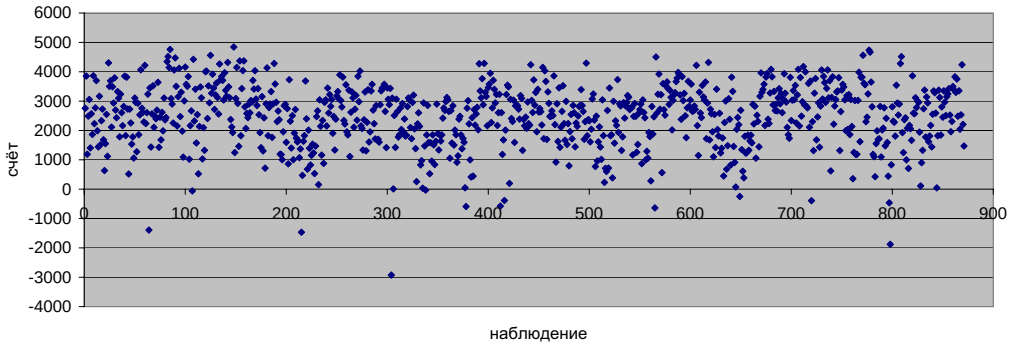

Рис. 3. График счетов 860 акустических реализаций слова ‘восемь’

$$th(w,t) = \tilde{m}(w,t) - \sigma(w) \tag{8}$$

Результат численного эксперимента с использованием адаптивного порога (7) приведен в табл.2. Начальное значение порога определялось на настроечных данных из расчета ошибки второго рода не более 0.5%.

Таким образом, пословная ошибка распознавания при использовании адаптивного порога уменьшилась на (3,75-3,16)/3,75 = 15,7% по сравнению с ошибкой при использовании заполнителей и на (4,89-3,16)/4,89 = 35,3% по сравнению с ошибкой при использовании фиксированного порогового значения.

Таблица 2. Ошибка при использовании адаптивного порога

α	Заполнители	Ошибка(%)	Замены	Вставки	Пропуски
0	Нет	3,45	387	235	400
0.5	Нет	3,16	423	286	227
1.0	Нет	3,22	432	328	192

Оценка начальных значений счетов по параметрам моделей звуков

Существенным недостатком рассмотренных акустических счетов является то, что оценка порогового значения требует наличия настроечной выборки, т.е. какого-то количества произнесений слова в аналогичной акустико-фоновой обстановке. На практике не всегда можно обеспечить наличие настроечных данных.

Рассмотрим метод формирования оценки счета только на основе параметров акустических моделей звуков.

В предположении, что наблюдаемые векторы параметров x_i независимы, а распределения счетов моделей звуков s_i нормальные: $p(s) = N(m, \sigma)$, распределение счетов $s(w)$ слов $w = (s_1, s_2, ..., s_{N_w})$ также будет нормальным, причем

$$m_x = \sum_{i=1}^{N_w} m_i \tag{9}$$

$$\sigma_w^2 = \sum_{i=1}^{N} \sigma_i^2 \qquad (10)$$

На рис.4 приведены гистограммы распределения логарифмов счета (3) для реализаций цифр 8 и 0.

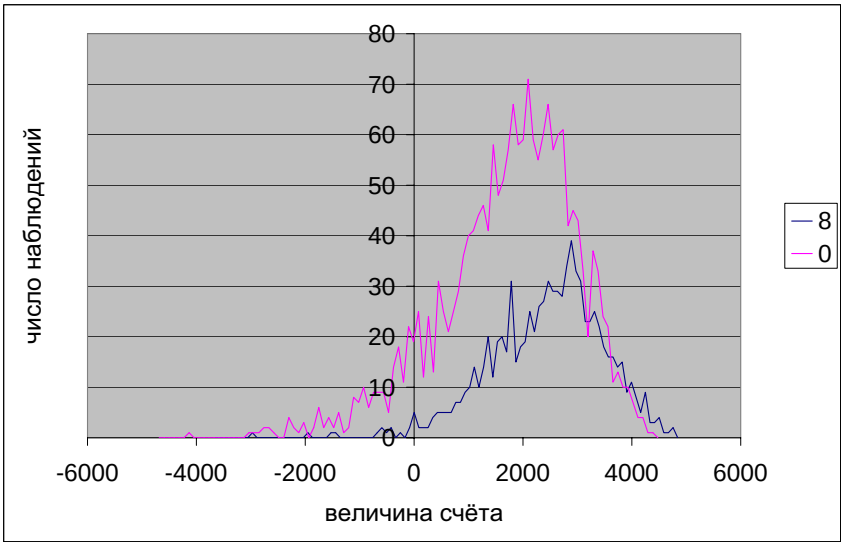

Рис. 4. Гистограммы распределения счетов для цифр 0 и 8

Очевидно, что распределение наблюдаемых величин счетов не является нормальным (в [8] предполагается, что это – гамма-распределение). Тем не менее, в качестве начального приближения для среднего счета слова в (7) можно взять сумму средних счетов составляющих слово моделей, заданных произносительной транскрипцией, а для отклонения – корень из суммы квадратов

отклонений. Оценка среднего дважды нормированного счета, его дисперсии и величины порога для принятия решения будут иметь вид

$$m(w) = \frac{1}{N_w} \sum_{s=1}^{N_w} \frac{1}{N_s} \sum_{j=1}^{N_s} m_j, \tag{11}$$

$$\sigma(w) = \sqrt{\frac{1}{N_w} \sum_{s=1}^{N_w} \frac{1}{N_s} \sum_{j=1}^{N_s} \sigma_j^2}, \tag{12}$$

$$th(w) = s(w) - m(w) - \sigma(w). \tag{13}$$

На следующих рис. 5 и рис. 6 приведены результаты, полученные при использовании счетов (11)-(13).

Очевидно, что использование аналитических оценок счетов практически не ухудшило показатели эффективности.

Отметим, что переход к использованию акустических счетов (11)-(13) предполагает принятие решения на основе отклонения наблюдаемого счета от оценки счета, которая основана на произносительной транскрипции слова и параметрах моделей звуков. Если произносительная транскрипция слова не соответствует реальному произношению, то решение о наличии незнакомого слова будет использовать неправильный порог.

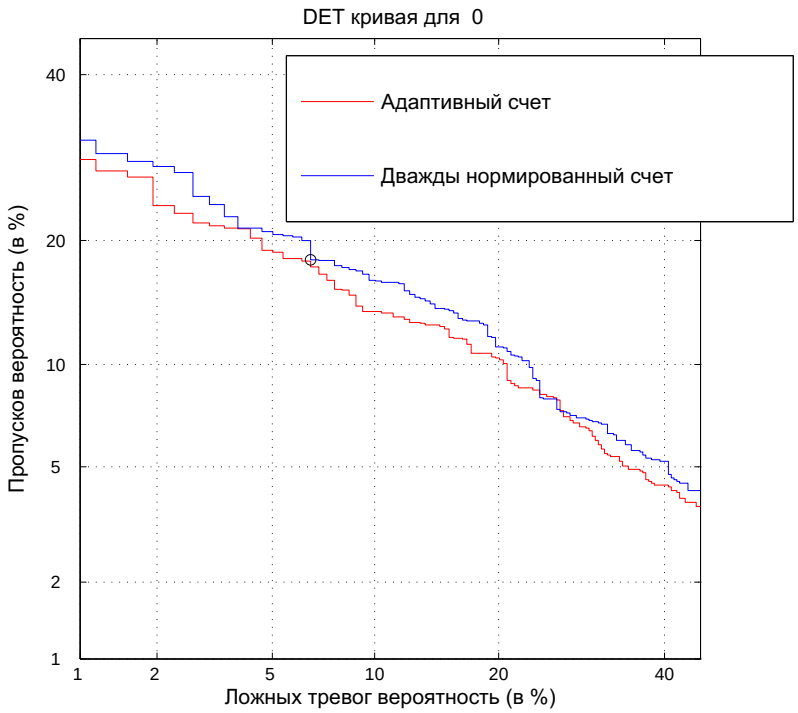

Рис. 5. DET кривая счета (11)-(13) для цифры 0

Заключение

В работе проведено исследование методов обнаружения незнакомых слов или акустических событий, которые основаны на использовании численных характеристик - вероятностных счетов. Эти счета представляют собой модифицированные оценки логарифма правдоподобия для наблюдаемого речевого сигнала и являются основой для принятия решений при распознавании речи.

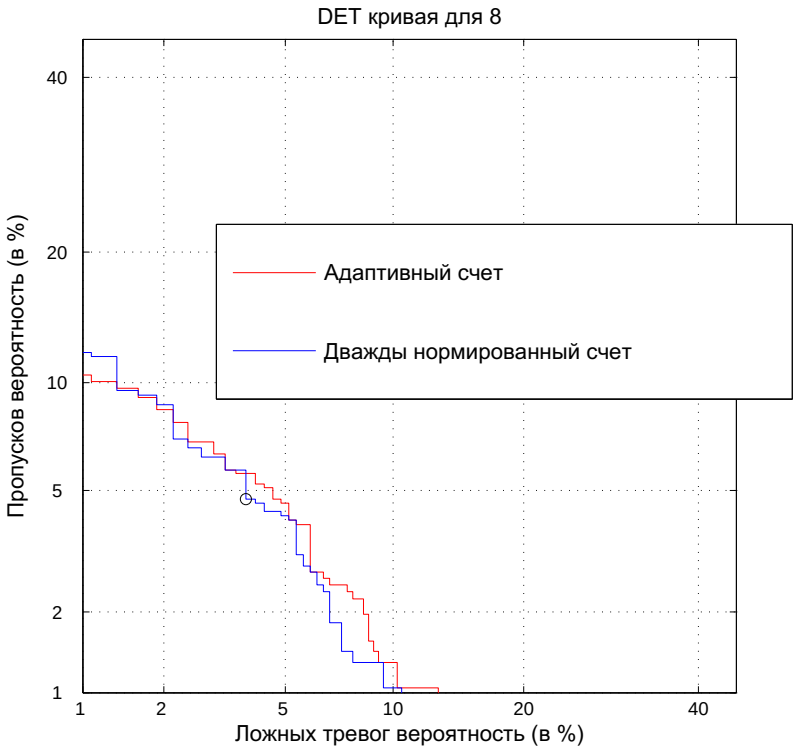

Рис. 6. DET кривая счета (11)-(13) для цифры 8

В экспериментах на речевом корпусе данных FaVoR показано, что наиболее эффективной, в терминах количества ошибок первого и второго рода, интегральной (вычисляемой для всего слова) оценкой, является дважды нормированный центрированный счет.

Это обстоятельство объясняется известным недостатком моделирования речевого сигнала с помощью скрытых марковских моделей, когда соседние наблюдения предполагаются независимыми. На практике соседние наблюдения – это значения

параметров речевого сигнала, вычисляемые с короткими интервалами в 5-20 мс, на протяжении которых артикуляторный тракт существенных изменений, как правило, не претерпевает. Таким образом, значения соседних векторов параметров не являются независимыми.

Использование интегрального счета вида (1), на основе которого принимается решение о распознавании в процедуре Витерби, приводит к увеличению 'вклада' стационарных длительных звуков, таких как гласные, и преуменьшению влияния коротких нестационарных звуков, например взрывных.

С другой стороны, использование дважды нормированного счета (3) 'уравнивает' вклад отдельных звуков в интегральную оценку, что приводит к улучшению эффективности принятия решения о распознавании.

Основные препятствием для использования счетов при обнаружении незнакомых слов является высокая вариативность их абсолютных величин при изменении акустико-фоновой обстановки, канала связи или голоса.

Центрирование счета, которое заключается в устранении среднего значения, улучшает эффективность при распознавании.

Применение описанных выше акустических счетов в системе распознавания речи предполагает выбор порогов для принятия решения о том, является ли анализируемый акустический образ словарным словом или нет. Оценка порогов может быть выполнена разными способами. Наиболее простой, точный, но практически неудобный способ – провести оценку интегральных счетов и выбор порогов по настроечным данным. Полученные

счета и пороги можно рассматривать как апостериорные. Для проведения оценки нужно иметь настроечные данные, причем желательно в аналогичных акустико-фоновых условиях. Если такие данные есть, существует возможность существенно улучшить характеристики системы распознавания, что подтверждено экспериментами (табл.1).

В работе предложена адаптивная процедура коррекции порогового значения счета в соответствии с текущими значениями счетов (11)-(13) и показано, что ее использование позволяет повысить точность распознавания.

Проведение специальных измерений счетов слов на настроечной выборке является отдельной, не всегда доступной процедурой. Кроме того, этот подход игнорирует тот факт, что в системе распознавания уже есть оценки параметров моделей звуков, на основе которых можно построить оценки счетов слов по их произносительным транскрипциям.

Исходя из параметров моделей звуков и произносительных транскрипций слов, можно оценить априорные значения средних счетов словарных слов. В работе предложен метод оценки априорных дисперсий счетов, выбора априорного порога, а также процедура адаптации средних значений и порогов в соответствии с наблюдаемым речевым сигналом. Показано, что эффективность предложенных априорных оценок счетов и метода адаптации порога практически соответствует использованию апостериорных оценок пороговых счетов на настроечной выборке.

Литература

1. “Bill Gates predictions about speech recognition a historical review.htm” in Matthew Paul Thomas blog at http://mpt.net.nz/archive/2005/12/30/gates.

2. Timothy J.Hazen, Stephanie Seneff and Joseph Polifroni. Recognition confidence scoring and its use in speech understanding systems,Computer Speech and Language, 2002, 16, 49-67.

3. Sui. M, Gish, H. Evaluation of word confidence for speech recognition systems. Computer Speech and Language, 13, 299-319.

4. Bazzi, I, Glass, J. Modeling out of vocabulary words for robust speech recognition. Proc. ICASSP 2000, Beijing, China, Vol. 1, pp.401-404.

5. Young, S. The HTK BOOK. Ver. 2.1. Cambridge University, 1997.

6. Программа DETware. Национальный институт стандартов США, NIST, http://www.nist.gov.

7. Rachna Vijay Vargiya. Keyword spotting using normalization of posterior probability of confidence measures. Ms. Thesis in Computer Science, 2005, USA.

8. Joel Pinto, H.N.V.Sitaram. Confidence measures in speech recognition based on probability distribution of Likelihoods. http://www.hpl.hp.com/techreports/2005/HPL-2005-144.pdf

9. Lou Boves, Johan Koolwaaij. Weighting phone confidence measures for automatic speech recognition. Proc. COST 249, Gent, Belgique, May 2000

Обнаружение новых слов и невербальных событий при распознавании речи

Нгуен Минь Туан

Аннотация

В работе приведен обзор современных методов проверки корректности результатов распознавания. На основе этого обзора и выводов реализован алгоритм GdAlg для формирования оценок достоверности результатов распознавания, основанных на отношении правдоподобия. Приведены результаты численных испытаний алгоритма на речевом корпусе данных Favor. Ключевые слова: распознавание речи, обнаружение незнакомых слов, оценки достоверности результатов распознавания.

Постановка задачи

Целью автоматического распознавания речи является аккуратное и эффективное преобразование речевого сигнала в текстовое сообщение независимо от индивидуальности голоса диктора и окружающей среды. К сожалению, при применении системы распознавания в естественных условиях коммуникации большое количество ошибок возникает вследствие появления незнакомых для системы слов и неречевых акустических событий. Ошибки в виде пропусков, вставок и замен могут привести к нежелательному результату для дальнейшей обработки. Полностью

устранить эти ошибки невозможно, поэтому разумной целью является создание методов проверки корректности распознанных слов.

В начале статьи приведен обзор современных методов проверки корректности распознавания. На основе этого обзора и выводов реализован алгоритм GdAlg формирования оценок достоверности результатов распознавания, основанных на отношении правдоподобия. Статья содержит подробное описание алгоритма и результаты его численной проверки на корпусе данных Favor.

Обзор методов проверки корректности результатов распознавания

Двумя наиболее широко применяемыми подходами к решению проблемы построения оценок корректности распознавания являются: вычисление оценок правдоподобия и использование моделей - заполнителей.

Метод вычисления оценок правдоподобия состоит в том, что для каждого слова на выходе из распознавателя вычисляется числовая характеристика $Cm(w)$. Эта характеристика сравнивается с некоторым порогом τ_w. Если значения характеристики больше, чем значение порога, то слово считается правильно распознанным. В противном случае соответствующая часть сигнала считается шумом или незнакомым словом. Оценки

правдоподобия можно условно разделить на три группы: простые характеристики, апостериорная вероятность и отношения правдоподобия. Подробно группы оценок описаны в следующих разделах.

Модели заполнителей часто применяются при необходимости иметь дело с незнакомыми словами. Подход заключается в создании одной или нескольких акустических моделей для 'обобщенных' слов, представляющих все незнакомые слова, или неречевые акустические события.

Простые характеристики

К простым характеристикам для оценки правдоподобия распознаваемого слова относится любая числовая характеристика, получаемая из распознавателя в процессе декодирования [1], [2]. Эти характеристики могут иметь акустическую или грамматическую природу. В качестве признаков для проверки корректности распознанного слова берутся такие характеристики, у которых функция распределения вероятности для правильно распознанных слов существенно отличается от функции распределения вероятности для неправильно распознанных слов. Примеры таких характеристик:

– нормированная акустическая оценка

$$Cm(w) = \frac{1}{T_w} \log P(Y_w \mid \lambda_w), \tag{2.1}$$

где w - распознанное слово, Y_w - сегмент сигнала соответствующего слова w, T_w - длина сегмента Y_w, λ_w - акустическая модель слова .

– плотность гипотез. Для каждого слова и временного фрейма определяется формула плотности

$$D(w,t) =\mid \{w : [w,s,e] \in WG \wedge s \leqslant t \leqslant e\} \mid, \tag{2.2}$$

где WG - словный граф (Word Graph), получаемый после процесса декодирования, s, e - начало и конец сегмента сигнала слова w соответственно. Тогда для каждого слова w плотность гипотез вычисляется следующим образом:

$$Cm([w,s,e]) = \frac{1}{e-s+1}\sum_{t=s}^{e} D(w,t). \tag{2.3}$$

Для достижения более хорошего результата применяется комбинация нескольких, взаимно независимых характеристик. Для комбинирования характеристик чаще всего используются линейный дискриминантный анализ [3], метод опорных векторов [4], нейронные сети [5]. Существуют и более простые методы комбинирования, например использования нелинейной функции

$$Cm(w) = exp(\alpha_1 \log Cm_1(w) + ... + \alpha_n \log Cm_n(w)), \tag{2.4}$$

где $Cm_1(w), ..., Cm_n(w)$ - числовые характеристики слова w, $\alpha_1, ..., \alpha_n$ - коэффициенты, удовлетворяющие условию $\alpha_1 + ... + \alpha_n = 1$.

Апостериорная вероятность

Статический подход к решению проблемы распознавания речи основывается на правиле решений по Байесу:

$$W^* = \arg\max_W P(W \mid Y) = \arg\max_W \frac{P(Y \mid W)P(W)}{P(Y)}, \qquad (2.5)$$

где $P(W)$ - вероятность модели языки, $P(Y \mid W)$ - вероятность акустической модели, $P(Y)$ - вероятность наблюдения сигнала. Если все три вероятности известны, то апостериорная вероятность $P([w, s, e] \mid Y)$ для конкретного слова w, с началом и концом в моментах времени s и e соответственно, легко вычисляется по формуле

$$P([w, s, e] \mid Y) = \sum_{W:[w,s,e]\in W} \frac{P(Y \mid W)P(W)}{P(Y)}. \qquad (2.6)$$

Эта апостериорная вероятность могла бы непосредственно использоваться как характеристика для определения корректности распознавания слова w. Теоретически вероятность $P(Y)$ имеет вид

$$P(Y) = \sum_W P(Y \mid W). \qquad (2.7)$$

На практике невозможно оценить точное значение вероятности $P(Y)$ и ее рассматривают как величину, которая не зависит от выбора конкретной цепочки слов. Таким образом, решения, приминаемые в процессе декодирования, базируются на ненормированных оценках. Эти оценки пригодны для сравнения

конкурирующих цепочек слов, но не для проверки корректности распознавания каждого слова в цепочке. Имеются несколько алгоритмов, которые аппроксимируют значение $P(Y)$ с помощью списка N лучших гипотез (N-best List) или словного графа [1]. Пример при использовании словного графа:

$$Cm(w) \approx \frac{\sum_{W \in WG:[w,s,e] \in W} P(Y \mid W) P(W)}{\sum_{W \in WG} P(Y \mid W)}. \tag{2.8}$$

Отношение правдоподобия

Метод основывается на критерии Неймана-Пирсона о проверке гипотез [7]. Пусть имеются распознанное слово w и соответствующий ему сегмент сигнала Y_w, тогда существуют две гипотезы:

H_0 (нулевая гипотеза): сегмент сигнала Y_w корректно распознан как слово w.

H_1 (альтернативная гипотеза): сегмент сигнала Y_w некорректно распознан как слово w.

и соотношение правдоподобия:

$$LR(Y_w \mid w) = \frac{P(Y \mid H_0)}{P(Y \mid H_1)}. \tag{2.9}$$

Если значение $LR(Y_w \mid w)$ больше значение порога τ, то принимается гипотеза H_0, в противном случае принимается гипотеза H_1. Таким образом, при известных вероятностях $P(Y \mid$

H_0) и $P(Y \mid H_1)$ мы можем определить, является ли слово w на выходе из распознавателя корректно распознанным.

Чтобы использовать решение на основе критерия Неймана-Пирсона для каждого слова w из словаря системы строятся две акустические модели λ_w^c (целевая модель) и λ_w^a (альтернативная модель) такие, что $P(Y \mid H_0) = P(Y \mid \lambda_w^c)$ и $P(Y \mid H_1) = P(Y \mid \lambda_w^a)$ для любого сегмента сигнала Y. Если такие модели удалось построить, то в качестве оценки правдоподобия для проверки корректности распознавания слова w можно взять функцию

$$Cm(w) = \frac{1}{T_w} \log \frac{P(Y \mid \lambda_w^c)}{P(Y \mid \lambda_w^a)}, \tag{2.10}$$

где Y - сегмент сигнала, распознанный как слово w ; T_w - длина сегмента Y.

В большинство случаев на практике создаются целевые и альтернативные модели не для отдельных слов, а для частей слов (монофонов, трифонов), из которых составляются все слова словаря системы распознавания. В этом случае, для слова $w = u_1...u_N$, состоящего из N частей, оценка правдоподобия вычисляется по формуле

$$Cm(w) = \frac{1}{N} \sum_{i=1}^{N} LR(u_i), \tag{2.11}$$

где

$$LR(u) = \frac{1}{T_u} \log \frac{P(Y_u \mid \lambda_u^c)}{P(Y_u \mid \lambda_u^a)}. \tag{2.12}$$

Модели слов-заполнителей

В отличие от трех предыдущих методов, где проверка корректности распознанных слов выполняется после процесса декодирования, при использовании моделей заполнителей распознавание сегментов сигнала, представляющих незнакомые слова или неречевые акустические события, осуществляется на этапе декодирования.

Метод основан на том, что в словарь системы распознавания добавляются так называемые 'обобщенные' слова [8]. Роль этих слов состоит в том, чтобы любой сегмент сигнала незнакомого слова или неречевого акустического события был распознан системой как одно 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 и обобщенных слов. Обобщенные слова затем отбрасываются, и оставшаяся часть цепочки считается результатом распознавания.

Вывод

Методы, основанные на вычислении простых характеристик, просты и не требуют больших вычислительных и временных ресурсов. В то же время во многих экспериментах было показано, что простые характеристики обладают высокой корреляционной

зависимостью. Поэтому комбинироваиие простых характеристик часто не приводит к заметному улучшению результата по сравнению с использованием отдельных характеристик.

Методам апостериорных вероятностей для вычисления оценки правдоподобия необходим словный граф (Word Graph) или список N лучших гипотез (N-best List). При большом словаре распознавателя построение словного графа (Word Graph) или списка N лучших гипотез (N-best List) приводит к большому объему вычисления и низкой производительности системы.

Главная проблема методов, основанных на отношении правдоподобия, заключается в моделировании альтернативных моделей. Это объясняется тем, что пространство акустических событий, моделируемое альтернативными моделями, очень большое и сложное. Алгоритмы обучения целевых и альтернативных моделей также играют важную роль в эффективности метода.

Недостатком подхода с использованием моделей словзаполнителей является высокая вероятность ошибки, когда слова, входящие в словарь распознавателя, распознаются как обобщенные слова. Кроме этого встает и вопрос об оптимальном выборе алфавита обобщенных слов.

Метод отношения правдоподобия представляется автору самым перспективным и интересным, поскольку проблема моделирования альтернативных моделей имеет множество потенциальных решений. В следующем разделе описывается алгоритм GdAlg, который является первым шагом исследования автора методов, основанных на отношении правдоподобия.

Описание алгоритма GdAlg

Акустические модели и оценки корректности

В качестве базисной акустической единицы моделирования выбран трифон, то есть акустическая реализация фонемы при заданных предшествующих и последующих фонемах. Каждый трифон моделируется с помощью непрерывной скрытой Марковской моделей [9]

Пусть w - распознанное слово на выходе из декодера, которое надо проверить на корректность. Слово w состоит из N трифонов, $w = u_1...u_N$.

Y_w - сегмент сигнала, соответствующий распознанного слова w. Y_w состоит из N подсегментов $Y_w = Y_{u_1}...Y_{u_N}$, где Y_{u_i} - сегмент сигнала трифона u_i

$\lambda^c = \{\lambda^c_u\}$ - целевые скрытые марковские модели, определенные для каждого трифона u.

$\lambda^a = \{\lambda^a_u\}$ - альтернативные скрытые марковские модели, определенные для каждого трифона u. Все модели λ^c и λ^a являются непрерывными скрытыми марковскими моделями.

Для каждого трифона u слова w и сегмент сигнала Y_u определим функцию соотношения

$$R(u, Y_u) = \frac{1}{T_{Y_u}} \log \frac{P(Y_u \mid \lambda^c_u, Q^c_u)}{P(Y_u \mid \lambda^a_u, Q^a_u)}, \qquad 3.1$$

где T_{Y_u} - длина сегмента сигнала Y_u, $Q_u^c = \arg\max_Q P(Y_u, Q \mid \lambda_u^c)$, $Q_u^a = \arg\max_Q P(Y_u, Q \mid \lambda_u^a)$ - оптимальные последовательности состояний скрытых марковских моделей после декодирования.

Определим оценку достоверности трифона u как

$$U(u, Y_u) = \frac{1}{1 + \exp(-0.5R(u, Y_u))} \tag{3.2}$$

и оценку достоверности слова w как

a) арифметическое среднее оценок достоверности трифонов $u_1, ..., u_N$:

$$Cm(w) = \frac{1}{N} \sum_{i=1}^{N} U(u_i, Y_{u_i}) \tag{3.3}$$

) геометрическое среднее оценок достоверности трифонов $u_1, ..., u_N$:

$$Cm(w) = \exp(\frac{1}{N} \log \sum_{i=1}^{N} U(u_i, Y_{u_i})). \tag{3.4}$$

Обучение целевых и альтернативных моделей

Обучение состоит в изменение параметров $\lambda = \{\lambda^c, \lambda^a\}$ таким образом, чтобы увеличить среднее значение $Cm(w)$для всех корректно распознанных слов w и уменьшить среднее значение $Cm(w)$ в противном случае. К числу параметров, подлежащих изменению, относятся веса смесей, математические ожидания и дисперсии гауссовых распределения.

Введем функцию стоимости для сегмента сигнала Y_u , распознанного как трифон u:

$$F_u(Y_u, \lambda_u) = \frac{1}{1 + \exp(-0.5\delta(Y_u)R(u, Y_y))}, \qquad 3.5$$

где

$$\delta(Y_u) = \begin{cases} -1, Y_u \text{ - корректно распознан;} \\ 1, Y_u \text{ - некорректно распознан.} \end{cases}$$

Тогда задача обучения параметров целевых и альтернативных моделей сводится к задаче оптимизации функции $F_u(Y_y, \lambda_u)$ для каждого трифона u. Эта задача решается методом градиентного спуска:

$$\lambda_u^* = \lambda_u - e\nabla(\frac{1}{N_u}\sum_{i=1}^{N_u} F_u(Y_u^i, \lambda_u)), \qquad 3.6$$

где N_u - количество сегментов сигнала, распознанных как трифон u, e - коэффициент спуска.

Таким образом, получаем следующий алгоритм оценки достоверности результата распознавания

Алгоритм GdAlg

Шаг 1. На наборе обучающих данных $\{Y\}$ производится распознавание. Распознанные цепочки трифонов выравниваются с правильными цепочками и каждый трифон помечается как корректно или некорректно распознанный. Трифон u с началом и концом в моментах времени s и e считается корректно распознанным, если в правильной цепочке имеется трифон u с началом и концом в моментах времени s' и e', такой что $s \leq 0.5(s' + e') \leq e$.

Шаг 2. Начальные значения параметров целевой и альтернативной моделей каждого трифона u инициализируются алгоритмом Баума-Велша. Корректно и некорректно распознанные как трифон u сегменты сигнала используются для инициализации значений целевой и альтернативной моделей соответственно.

Шаг 3. Для каждой пары целевой и альтернативной моделей трифона u обновляем параметров по формуле (3.6).

Шаг 4. Повторять шаг 3, пока величины изменений параметров не станут меньше заданной константы ε.

Численные эксперименты

Численные эксперименты выполнялись на речевом корпусе FaVoR [10]. В эксперименте как базовая система использован инструментарий HTK на основе скрытых марковских моделей. Алгоритм обучение параметров целевых и альтернативных моделей реализован с помощью средства разработки Visual Studio 2005.

Речевой корпус данных

Речевой корпус данных FaVoR содержит записи слитной речи 1673 дикторов. Все записи оцифрованы с частотой дискретизации 22,050 кГц. Словарь корпуса состоит из 14 слов и содержит цифры от 0 до 9, и служебные слова "да "нет "старт"и "стоп".

В записях речевого корпуса имеется большое количество различных акустических шумов.

Акустические модели модулей распознавания и проверки корректности обучались на выборке из 987 записей (405 дикторов). Для тестирования используется выборка, состоящая из 428 записей (151 дикторов).

В качестве вектора признаков сигнала использовался 42–мерный вектор, состоящий из 12 мел–кепстральных коэффициентов (MFCC), логарифма энергии сигнала, их первых и вторых производных.

Акустические модели

Каждый трифон модуля распознавания моделируется непрерывной скрытой марковской моделей с тремя состояниями и верхней диагональной матрицей переходов. Вероятность эмиссии каждого состояния представляется как смесь 16 нормальных (Гауссовских) функций плотности вероятности. Акустическая модель паузы моделируется с помощью эргодической скрытой марковской модели с тремя состояниями и смесью из 32 нормальных распределений для каждого состояния. Целевые и альтернативные скрытые марковские модели для каждого трифона имеют по три состояния и верхние диагональные матрицы переходов. Вероятность эмиссии каждого состояния моделируется с помощью смеси из восьми гауссовских функций плотности вероятности. Сходимость обучения целевых и

альтернативных моделей алгоритмом GdAld достигается после нескольких итераций.

Результаты эксперимента

К ошибкам первого рода или пропускам цели относятся слова, которые корректно распознаны, но отклонены алгоритмом GdAlg. К ошибкам второго рода или ложным тревогам относятся слова, которые являются вставками или заменами, но приняты алгоритмом GdAlg как корректно распознанные.

На рис. 1 показана параметрическая кривая ошибки (DET-curve), построенная для двух способов формирования оценок достоверности распознавания слова из оценок достоверности составляющих их трифонов: геометрического и арифметического среднего. Очевидно, что при вычислении оценки достоверности слова геометрическое среднее оценок достоверности составляющих трифонов дает более хороший результат. На рис. 2 показана кривая зависимости суммы ошибок первого и второго рода от значения порога.

Наилучший результат получается при использовании оценки на основе геометрического среднего и выбора значения порога равного 0.45. В этом случае алгоритм GdAlg отклонил 85.3% неправильно распознанных слов и отбросил 6.5% правильно распознанных слов.

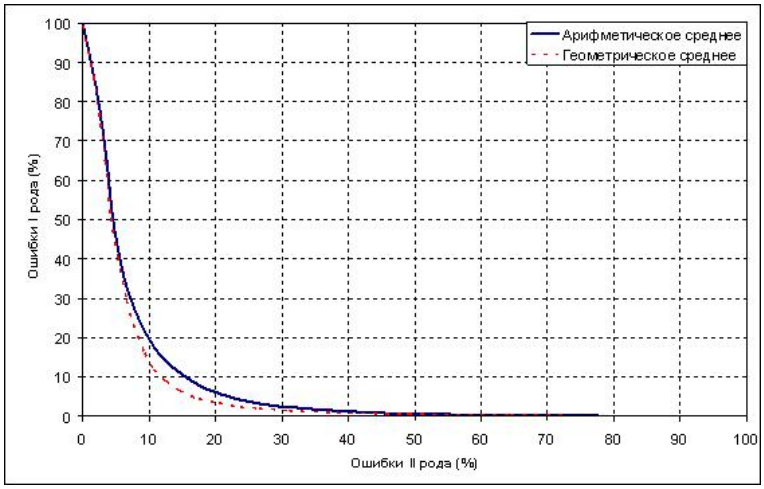
Арифметическое среднее
Геометрическое среднее
Ошибки I рода (%)
Ошибки II рода (%)
0
10
20
30
40
50
60
70
80
90
100

Рис. 1

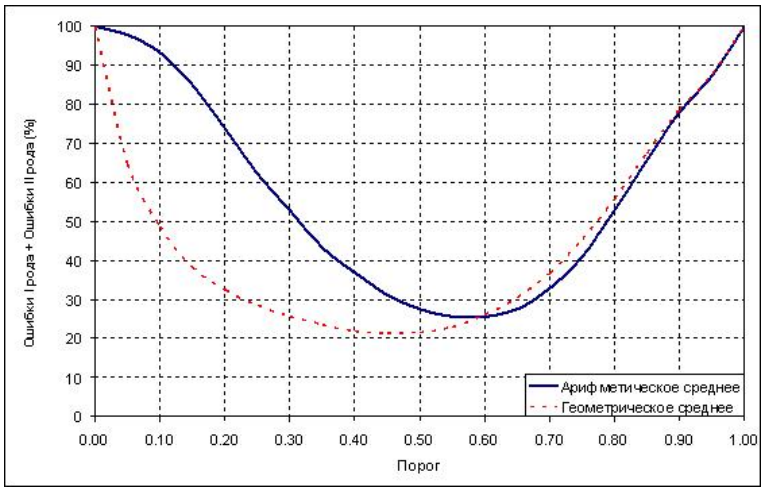
Арифметическое среднее
Геометрическое среднее
Ошибки I рода + Ошибки II рода (%)
Порог
0
10
20
30
40
50
60
70
80
90
100
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

Рис. 2

Заключение

Приведен обзор современного состояния исследований в области формирования оценок достоверности результатов автоматического распознавания речи Представлен алгоритм проверки корректности распознанных слов, основанный на соотношении правдоподобия. Его применение на корпусе речевых данных FaVoR показало значительное сокращение количества ошибок II рода при умеренном малом увеличении ошибок I рода. Показано, что дальнейшие работы следует вести в направление улучшения алгоритма обучения целевых и альтернативных моделей и комбинирования данного подхода проверки корректности с подходом с моделями заполнителей.

Литература

1. T. Kemp, T. Schaaf. Estimating confidence using word lattices, Eurospeech–97, 1997.

2. T. J. Hazen, S. Seneff, J. Polifroni, Recognition Confidence Scoring and It's Use in Speech Understanding Systems, Computer Speech and Language, 2002.

3. R. A. Sukar. fRejection for Connected Digit Recognition Based on GPD Segmental Discrimination, IEEE Proc. ICASSP, 1994.

4. R. Zhang, A. I. Rudnicky. Word Level Confidence Annotation Using Combinations of Features, Proc. of 7 European Conference on Speech Communication and Technology, 2001.

5. L. Mathan, L. Miclet. Rejection of Extraneous Input in Speech Recognition Applications, Using Multi-Layer Perceptrons and The Trace of HMMs, Proc. of International Conference on Acoustics, Speech and Signal Processing, 1991.

6. F. Wessel, R. Schluter, K. M., H. Ney, Confidence Measures for Large Vocabulary Continuous Speech Recognition, IEEE Transactions on Speech and Audio Processing 2001, 2001.

7. R. C. Rose, B. H. Juang, C.H. Lee. A Training Procedure for Verifying String Hypothesis in Continous Speech Recognition, Proc. of International Conference on Acoustic, Speech and Signal Processing, 1995.

8. I. Bazzi, Modelling Out-of-Vocabulary Words for Robust Speech Recognition, Ph.D. Thesis, 2002.

9. S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland. The HTK Book, Cambridge University Engineering Department, 2001.

10. Десятчиков А.А, Ковков Д.В., Лобанцов В.В. Комплекс алгоритмов для устойчивого распознавания человека // Известия РАН. Теория и системы управления, 2006. №6. С. 1-12.

11. X. D. Huang, Y. Ariki, M. A Jack. Hidden Markov Models for Speech Recognition, Edinburgh University Press, 1990.

12. L. Rabiner, B. H. Juang. Fundamentals of Speech Recognition, Prentice-Hall, 1993.

13. G. A. H. Abrego, Confidence Measures for Speech Recognition and Utterance Verification, Ph.D. Thesis, 2000.

14. N. Moreau, D. Jouvet, Use of A Confidence Measure Based on Frame Level Likelihood Ratios for the Rejection of Incorrect Data, Eurospeech'99, 1999.

15. J. Pinto, R. N. V. Sitaram, Confidence Measures in Speech Recognition Based on Probability Distribution of Likelihoods, HP Labs Technical Report, 2005.

16. L. Smidl, J. V. Psutka, J. Zahradil, Keyword Spotting with Triphone Based Filler Model, Specom 2005, 2005.

17. N. B. Yoma, C. Molina, F. Hueinupan, J. Inzunza, Combining Word Features with Bayes-Based Confidence Measure in Speech Recognition, Specom 2005, 2005.

Содержание