Компьютерные системы распознавания речи*

И.Л. Мазуренко

Описаны ограничения возможностей ЭВМ по моделированию процесса распознавания речи. Приведена классификация систем распознавания речи. Перечислены сравнительные характеристики известных компьютерных распознающих систем. Описан наиболее распространенный подход к моделированию распознавания речи на ЭВМ. Поставлены математические и технические проблемы, стоящие в этой области, и проанализирована перспектива их решения.

1. Введение

Задача распознавания речи состоит в восстановлении по звуковому сигналу слова естественного языка (из ограниченного словаря), произнесением которого является этот звуковой сигнал. Она обычно решается путем задания эталонов слов словаря и последующего сравнения звуковых сигналов с этими эталонами. Звуковой сигнал представляет из себя целочисленный вектор значений звукового давления, измеренного в равноотстоящие друг от друга моменты времени. Мощность пространства звуковых сигналов огромна (типичное значение мощности множества сигналов длительностью в 1 сек., используемых в компьютерных системах, равно 256^{11025} [4]). Для решения задачи распознавания обычно сначала равномерно разбивают сигнал на окна одинаковой длины. Окна преобразуют из временной области в частотную (например, с помощью преобразования Фурье [3]), чтобы близость окон относительно простых метрик (типа Евклидовой) соответствовала близости участков сигнала

*Работа выполнена при частичной поддержке гранта Миннауки (020105028)

"на слух". Затем решается задача нахождения соответствия между окнами звукового сигнала и окнами эталонов слов словаря. Сложность последней задачи заключается в том, что различные участки звукового сигнала в различных произнесениях одного и того же слова отличаются разной степенью сжатия или растяжения (вовсе не пропорционального). Для решения задачи нахождения соответствия между окнами сигналов традиционно используются методы динамического программирования ([2]). Создание компьютерных систем распознавания речи связано со множеством объективных трудностей, накладывающих на подобные системы искусственного интеллекта ряд ограничений.

Предельные возможности компьютера по распознаванию речи связаны прежде всего с тем, что человек, которого можно взять за эталон распознающей системы, распознает осмысленную речь, а компьютеру в полной мере это не дано. Компьютер принципиально не может с требуемой надежностью исправлять ошибки и неоднозначности распознавания, используя синтаксическую и семантическую связь слов предложения. Вместо этого в современных системах используется статистическая модель, задающая связь последовательных троек слов предложения. Кроме того, человек использует зачастую дополнительную, незвуковую информацию. Самым ярким примером здесь может служить так называемое "чтение по губам", которому могут обучиться глухие люди. Известно, что в шумной обстановке человеку легче распознавать речь, если он следит за губами говорящего. Человек воспринимает речь объемно, что позволяет ему производить шумоочистку и пространственное выделение сигнала более качественно, чем ЭВМ. Слуховой аппарат человека позволяет ему с точностью до полупространства определить направление на источник полезного сигнала и отделить его от остальных звуковых источников.

Фонетические модели, используемые в программировании алгоритмов на ЭВМ, не точны, так как не используют всего многообразия факторов. Для задания фонетических эталонов обычно используют статистические методы, предполагающие, что акустические параметры фонем распределены по нормальному закону. В реальности картина намного сложнее, что приводит к тому, что точная модель эталонов звуков и слов должна включать в себя множество эталонных элементов (по одному на каждый вариант произнесения).

Дополнительно, картина осложняется тем, что все известные алго-

ритмы распознавания речи являются дикторозависимыми. После настройки на голос одного диктора распознающие системы дают удовлетворительные результаты распознавания для этого типа голоса, но хуже работают на других голосах. Надежность распознавания речи человеком, напротив, не зависит от типа голоса диктора.

Все вышесказанное приводит к тому, что распознавание речи компьютером обладает ограниченной надежностью, существенно повысить которую вероятно не удастся в будущем ни путем совершенствования алгоритмов распознавания, ни путем увеличения вычислительных мощностей компьютера. Постоянно имея в виду это утверждение, можно приступать к анализу достижений в области распознавания речи, классификации стоящих в этой области задач и оценке перспектив их решения.

2. Современное состояние направления распознаванияе речи

Классификацию систем распознавания речи будем производить согласно новому стандарту в области программирования таких систем, принятому сейчас практически всеми известными разработчиками систем распознавания речи - Microsoft Speech API ([1]).

Согласно этому стандарту, системы распознавания речи различают по следующим признакам:

Интервал между отдельными словами. Если система распознает непрерывную речь, пользователь может произносить речевые фразы естественно, не делая паузы между словами. Непрерывное распознавание более предпочтительно, однако оно требует большей вычислительной мощности компьютеров, что приводит пока к малому числу таких систем. В системах, работающих с дискретной речью, пользователь при диктовке должен делать паузу между отдельными словами, обычно составляющую не менее 1/4 частиь секунды. Третьей разновидностью являются системы, выделяющие одно слово из интервала речи, даже если он состоит из нескольких непрерывно произнесенных слов (word-spotting, [1]).

Зависимость от диктора. Системы, обладающие относительной

независимостью от диктора, позволяют пользователю работать с системой без предварительной настройки, однако улучшают надежность распознавания после обучения. Независимость от диктора таких систем обычно достигается за счет хранения звуковых эталонов для всех наиболее типичных голосов носителей данного языка. Это, безусловно, требует в несколько раз большей производительности и объема памяти. Настройка на голос диктора дикторозависимых систем занимает обычно от 30 минут до нескольких часов. Это составляет главное неудобство для пользователя. Обычно дикторозависимые системы позволяют работать с относительной степенью надежности без предварительной настройки на голос конкретного пользователя. Третьей разновидностью систем по этому признаку являются системы, автоматически настраивающиеся на голос диктора по мере их использования. Системы последнего типа обладают двумя особенностями - им нужно знать, сделал ли пользователь ошибку, произнеся конкретное слово (иначе обучение будет неверным); после настройки на одного диктора такие системы перестают надежно работать с другими голосами.

Степень детализации при задании эталонов. Различают алгоритмы, в которых в качестве эталонов используются целые слова, и алгоритмы, использующие эталоны элементов слов. Сравнение целых слов дает большую точность, скорость, однако требует значительно большего объема памяти (пропорционально количеству слов в словаре) и обучения каждого слова. Алгоритмы сравнения элементов слов (фонем, слогов и т.п.) приходится применять в случае больших словарей, т.к. объем требуемой памяти пропорционален количеству этих эталонных элементов слов (например, звуков) и не зависит от объема словаря.

Размер словаря. Системы распознавания речи могут использовать большие или маленькие словари. Размер словаря системы распознавания почти не связан с реальным количеством слов, которые данная система может распознать. Он определяется количеством слов, требуемых для распознавания в данном конкретном состоянии системы. Системы, работающие с маленькими словарями (около 50 слов) позволяет пользователю давать простые команды компьютеру. Для диктовки текстов необходимы большие словари (несколько десятков тысяч слов). Если системы диктовки учитывают контекст для определения активного подсловаря в конкретном состоянии, то фактически они работают со словарями среднего размера (около 1000 слов [1]).

Несмотря на то, что в принципе возможна любая комбинация этих характеристик, в настоящее время наиболее популярными являются системы голосового управления компьютером и системы дискретной диктовки текстов.

Системы голосового управления компьютером ("Command and Control" в терминологии Microsoft Speech API [1]) - это системы дикторонезависимого распознавания непрерывно произносимых команд, составленных из слов ограниченного (до нескольких сотен слов) словаря. Для подобных систем, если пользователь произносит команду, не входящую в список, система либо выдает отказ от распознавания, либо выдает в качестве ответа похожую "на слух" команду. Список команд обычно интуитивно ясен в каждой конкретной ситуации. Согласно стандарту Microsoft Speech API , системы голосового управления должны работать успешно на компьютерах 486/66 МГц с 1 МБ свободной оперативной памяти.

Системы дискретной диктовки текстов ("Discrete Dictation" [1]), т.е. системы дикторозависимого распознавания дискретно произносимых слов из больших по объему (десятки тысяч слов) словарей. Подобные системы обычно требуют процессора Pentium/60 МГц и 8 МБ свободной оперативной памяти.

В сводной таблице 1 приведены характеристики наиболее известных сейчас систем распознавания речи ([5]).

Если оценивать существующие сейчас на рынке системы диктовки и голосового управления компьютером с точки зрения рядового пользователя, подходящего к компьютерным программам с позиций эффективности и удобства, то можно сделать следующие выводы.

Система голосового управления компьютером менее удобна, привычна и проста в обучении, чем клавиатура и мышь. Исключение могут составить лишь пользователи-инвалиды. Применение голоса для управления компьютером станет частью интерфейса лишь тогда, когда появятся принципиально новые мультимедиа-ориентированные операционные системы, изменится архитектура вычислительных машин (заметный шаг в этом направлении - появление процессоров MMX) и будут разработаны надежные системы очистки от стационарных и нестационарных помех.

Системы диктовки текстов являются пока привлекательными для покупателей в силу новизны предоставляющихся для пользователя возмож-

Название фирмы, название системы, цена (в долларах США)	Стандартный словарь; словарь, определяемый пользователем	Точность распознавания без обучения / после обуч.	Оптимальное использование после; скорость набора (сл./ мин.)	Для кого предназначена Система	Операционные системы, рекомендуемый объем оперативной памяти и памяти на диске	Совместимость со звуковыми картами	Особые достоинства и недостатки
Dragon Dragon Power Edition $1695	60,000 слов Позволяет заменить любое слово станд. Словаря на слово пользователя, но не более 60,000.	70% 95%	40 страниц прочитанного текста: 2 - 3 часа. До 90 слов в минуту. Средняя скорость 65 сл. / мин.	.	Windows 3.1, Windows 95, DOS 5.2, DOS 6.1 для настольных компьютеров, Windows NT 16 МБ 40 МБ	Множество, включая Sound Blaster 16	Достоинства: надежная Система. Недостатки: нужно переключаться от диктования к функции выполнения команд во время форматирования документов.
IBM IBM Voice TYPE 1.32 $995	23,000 2000 слов	90% 97%	2 недели До 125 слов в минуту	Офис, медицина юриспруденция, журналистика	Win 3.1, Windows 95, OS/2. 24 МБ 33 МБ Минимум. Дополнительно 30 МБ во время регистрации	Звук. карта IBM Proprietary включ. в комплект поставки. Не совм. с любой другой.	Достоинства: высокая скорость, позволяет разделить во времени процессы диктовки и исправления ошибок, высокая точность Недостатки: не так проста в управлении, как многие другие системы.
Kolvox OfficeTALK 3.0 $495(отд.) $995 (+ KVWin)	Зависит от используемого ПО, зависит от используемого ПО	Зависит от используемого ПО	Зависит от используемого ПО по распознаванию речи	Офис, домашнее использование	Windows 3.1x или Windows 95 24 МБ 32 МБ	Звуковая карта Megaphone ($285 дополнительно)	Достоинства: прекрасное добавление к любому ПО по распознаванию речи. Недостатки: требует соответствующее ПО. Не является отдельным продуктом .

Название фирмы, название системы, цена (в долларах США)	Стандартный словарь; словарь, определяемый пользователем	Точность распознавания без обучения / после обуч.	Оптимальное использование после; скорость набора (сл./ мин.)	Для кого предназначена Система	Операционные системы, рекомендуемый объем оперативной памяти и памяти на диске	Совместимость со звуковыми картами	Особые достоинства и недостатки
Kurzweil KVWin 2.0 $695	30,000 или 60,000 слов за ту же цену 10,000 или 20,000	90% 97%	3-4 часа .	Пользователи компьютеров, деловые люди	Windows 3.1 и 95, 16 МБ для 30,000 слов, 24 МБ для 60,000 слов 35 МБ	Creative Labs Sound Blaster 16	Достоинства: очень проста в использовании. Распознает непрерывную речь. .
Voice Connection Micro-Introvoice II $1695	1000 слов .	90% 98%	5 часов .	Промышленное производство, склады	Windows, DOS . .	.	Достоинства: исключительно маленькие, умещающиеся в руке устройства. Радио-частотная трансмиссия. Недостатки: 1000 слов - предельный размер словаря.
Philips Speech Processing Speech Magic	64000 слов, позволяет добавлять новые слова	95% после обучения		Здравоохранение, юриспруденция	Windows 3.1 и 95, 80486 66MHz, 16 MB, 500 MB	Требуется ISA-плата Speech Board LFN5201c сигнальным процессором	Достоинства: распознаватель непрерывной речи, учет контекстной зависимости при корректировке

Таблица 1.

ностей, эффективной рекламной кампании и здорового исследовательского интереса к задаче. Однако реальные системы диктовки должны, очевидно, обладать следующими тремя свойствами: время набора текста с голоса, включая время на исправление ошибок, меньше времени набора того же текста с клавиатуры; пользователь не должен уставать от набора текста голосом больше, чем от набора того же текста с клавиатуры; стоимость системы диктовки окупается выигрышем во времени диктовки за сравнительно короткий период.

Тестирование существующих систем автором показывает, что они по большому счету не удовлетворяют ни одному из этих требований. Поэтому эти системы пока являются не более чем дорогими мультимедиа-игрушками, как, впрочем, и подавляющее большинство других продающихся сейчас мультимедиа-систем (например, систем обучения иностранным языкам).

Стоит, однако, упомянуть, что системы диктовки текстов на Западе нашли свое практическое применение в медицине. Это связано в первую очередь с тем, что область научных разработок для использования в медицине на Западе хорошо финансируется. Кроме того, задача здесь упрощается тем, что словари медицинских терминов в узкой предметной области имеют меньший объем, чем словари повседневного общения, а синтаксис и семантика диктуемых предложений чрезвычайно строгие, что повышает надежность распознавания. Системы диктовки текстов применяются в медицине зачастую тогда, когда руки и глаза диктующего заняты, например, во время операций. В этом случае до использования речевых технологий либо вообще не практиковалось документирование происходящего в реальном времени, либо приходилось задействовать дополнительные людские ресурсы. Применение распознавания речи значительно повысило эффективность работы врачей.

3. Наиболее распространенный подход к распознаванию речи

Для лучшего понимания особенностей задачи распознавания речи необходимо отметить, что все вышеперечисленные системы в принципе работают одинаково, используя одни и те же методы и алгоритмы. Раз-

ница в типе диктовки речи, размере словаря и т.п. обусловлена лишь спецификой задачи и имеющимися ограничениями по скорости вычислений и объему требуемой памяти.

Упрощенно процесс распознавания речи может быть описан в виде последовательности следующих основных шагов.

Шумоочистка и отделение полезного сигнала.

Методы, применяемые для решения данной задачи, можно условно разделить на четыре группы.

Методы первой группы обычно сводятся либо к выделению некоторых инвариантных относительно шума признаков, либо к обучению в условиях шумов или модификации эталонов распознавания с использованием оценки уровня шумов. Узким местом подобных методов является поразительный эффект ненадежной работы систем распознавания, настроенных на распознавание в шуме, в условиях отсутствия шумов [21].

В качестве примера методов из второй группы можно назвать коэффициенты линейного предсказания [3], кепстральные коэффициенты [3]. В качестве элементов эталонов в данном случае вместо численных значений используют вероятностные распределения (среднее + дисперсия). Для получения инвариантных признаков часто используют кратковременную функцию когерентности [23] вместо спектра Фурье или методы, связанные с моделированием слуховой системы человека [22].

Третья группа методов связана с цифровой обработкой сигнала. К методам этой группы можно отнести методы маскирования шумов (численные значения, сравнимые с характеристиками шума, игнорируются или используются с меньшими весовыми коэффициентами [12]) и методы шумоподавления с использованием нескольких микрофонов (очистка от низкочастотных шумов с использованием двух микрофонов на расстоянии $\cong 50$ см и высокочастотных - на расстоянии $\cong 5$ см, [13]).

Четвертая группа методов, применяемых для отделения полезного сигнала от посторонних шумов обычно связана с использованием массивов микрофонов, моделирующих направленный микрофон с переменным лучом направления (простейший метод "задержки и суммирования" или более сложный с модификацией весов микрофонов методом наименьших квадратов - метод Фроста [14], который позволяет подавлять и такой тип шумов, как искажения). Для изменения угла луча массива микрофонов используют методы настройки, наиболее известным из которых

является метод MUSIC (Шмидт, 1986 [18]). Используются также модификации этого алгоритма (ML-алгоритм, Зискинд и Вакс, 1988 [20] и более быстрый квази-ньютоновский алгоритм, Ватанабе, 1991 [19]).

Для повышения надежности распознавания речи в условиях малого отношения сигнал/шум, при большом объеме словаря или в ситуациях, когда необходимо работать без предварительной настройки на диктора, предпринимаются попытки использовать при распознавании также дополнительную (неречевую) информацию о дикторе.

Например, используются следующие виды информации: визуальная информация о движении губ и челюсти диктора, вводимая в компьютер при помощи видеокамеры [6]; информация о местоположении диктора, получаемая с помощью видеокамеры и служащая для ориентации направленного микрофона; информация о движении губ и челюсти диктора, вводимая более дешевыми способами (например с помощью отражательного фотодатчика [7]); информация об эмоциональном состоянии диктора (кожно-гальваническая реакция), используемая для уточнения результатов распознавания речи [8]; использование звукового канала костной проводимости [9]; регистрация выдыхаемого воздуха [10].

Преобразование входного речевого сигнала в набор акустических параметров.

Как отмечалось выше, обычно звуковой сигнал разбивают на окна одинаковой длины и преобразуют в частотную область с помощью дискретного преобразования Фурье или более сложного преобразования, после чего частотные параметры факторизуют с целью сокращения размерности. По физическому смыслу частотные параметры наиболее близки к тем, которые использует человек в процессе восприятия речи.

Приведение акустической формы сигнала к внутреннему алфавиту эталонных элементов.

Область значений акустических параметров речи разбивают на области сгущения, которые соответствуют элементам фонем, одинаковым для различных слов данного языка. Обычно таких областей для фиксированного языка насчитывают около 1000 [1], и если словарь системы распознавания содержит большее количество слов, с целью экономии памяти целесообразно в качестве эталонов системы распознавания рассматривать не слова, а соответствующие фонемные элементы. Совокупность таких эталонных элементов образует фонетическую кодовую книгу [15].

Примерные значения параметров эталонных элементов для всех дикторов данного языка известны заранее, и задача начального обучения состоит в уточнении значений этих параметров. В этом случае в процессе распознавания речи по акустическим параметрам каждого окна сигнала определяют ближайший к этому окну эталонный элемент.

Распознавание последовательности фонем и преобразование ее к тексту слов.

После определения вероятной последовательности эталонных элементов во входном сигнале необходимо восстановить по ней неизвестную последовательность фонем, являющуюся транскрипцией одного из слов словаря. Эта задача решается с помощью метода динамического программирования [2], когда в каждый момент времени определяется наиболее вероятная предполагаемая последовательность фонем в сигнале от начала слова до этого момента времени. Если акустические параметры преобразованы в вероятностные с использованием кодовой книги, а транскрипционные эталоны слов заданы в виде вероятностных автоматов, для распознавания обычно используются скрытые цепи Маркова [15]. В случае задания эталонов слов в виде последовательности значений акустических параметров без применения кодовой книги (обычно для маленьких словарей) применяют другую модель распознавания речи с использованием динамического программирования, называемую динамической деформацией времени [2].

Вероятностный автомат Маркова представляет из себя автомат с состояниями $S_1, S_2, \ldots S_N$, для которых определена матрица вероятностей перехода из состояния i в состояние j (обычно только элементы $P_{ii}, P_{ii+1}, P_{ii+2}$ этой матрицы отличны от 0). Все вероятности P_{ij} зависят только от пары номеров состояний (i, j). Кроме того, для каждой пары из $\{(i,j) : P_{ij}$ отлично от $0\}$ задана функция $Q_{ij}(x) : X \to \Re \cap [0,1]$, определяющая вероятность наблюдения эталонного элемента x из кодовой книги языка X в состоянии с номером j, при условии, что предыдущим состоянием было состояние с номером i. Если эффектом коартикуляции (влияния соседних звуков друг на друга) пренебречь, можно считать, что функции Q зависят только от номера последнего состояния.

Вероятность дойти от начального состояния S_1 до конечного S_N по заданной цепочке переходов $S_1 S_{i_2} S_{i_3} S_{i_4} \ldots S_{i_{M-1}} S_N$ при заданной последовательности наблюдаемых эталонных элементов $x_1 x_2 \ldots x_M$ есть произведение вероятностей $P = Q_1 P_{1i_2} Q_{i_2} P_{i_2 i_3} Q_{i_3} P_{i_3 i_4} Q_{i_4} \ldots Q_{i_{M-1}} P_{i_{M-1} N} Q_N$.

Задача последнего этапа алгоритма распознавания речи состоит в нахождении наиболее вероятной последовательности состояний, то есть какого-нибудь локального минимума P по всем допустимым путям в автомате, ведущим из S_1 в S_N. Значение вероятности в точке минимума считается вероятностью того, что данная на вход автомата последовательность эталонных элементов задает произнесение именного того слова, эталоном которого является этот вероятностный автомат. Для нахождения этого минимума применяется алгоритм Витерби (разновидность динамического программирования [15]) - для каждого момента времени $t = 1, 2, \ldots, M$ и каждого состояния S_i считается минимум вероятности по всем путям, ведущим из S_1 в S_i по итерационной формуле

$$P_i(t) = \min_{j:P_{ij}\neq 0} \{P_j(t-1)P_{ji}Q_{ji}(x_t)\} \qquad (1)$$

Для вычисления вероятности $P_N(M)$ с помощью алгоритма Витерби [15] требуется применить формулу (1) αNM ($\alpha = const$) раз, что намного меньше, чем число вычислений в полноперебором алгоритме, если учесть, что все далекие от главной диагонали элементы матрицы $\|P_{ij}\|$ равны 0.

Обычно вместо вероятности P в формуле (1) используют $-\log P$, при этом все умножения заменяются на сложения, а min - на max. Функции $Q_{ij}(.)$ задаются несколькими способами. Если в качестве входных элементов используют элементы кодовой книги, функции $Q_{ij}(.)$ (или без учета коартикуляции $Q_i(.)$) можно задать в виде прямоугольной матрицы значений этой функции на всех элементах кодовой книги. Такие марковские модели называют дискретными [15]. Если на вход автомата подаются значения акустических параметров $(p_1, p_2, \ldots, p_n)$ на окнах звука, то каждому состоянию автомата Маркова сопоставляют либо вектор пар $((\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_n, \sigma_n))$ среднего значения и дисперсии каждого параметра в данном состоянии, либо пару (вектор средних значений, матрица ковариации), также определяющие область значений акустических параметров, а функцию $Q_i(x)$ определяют как функцию расстояния Махаланобиса

$$\rho(x, p) = (x - \mu_p)^t \Sigma_p^{-1} (x - \mu_p)$$

(здесь μ - вектор средних, Σ - матрица ковариации). Такие модели Маркова называют непрерывными [15].

В настоящий момент самыми сложными элементами при построении системы распознавания речи являются, как это не покажется странным, не распознающие алгоритмы - их подробные описания можно прочитать в монографиях и патентах, предшествующих появлению той или иной коммерческой системы распознавания, а построение акустической модели языка и начальное обучение эталонов слов словаря, чаще всего являющихся вероятностными автоматами Маркова. Как правило, для построения достоверной с вероятностной точки зрения модели того или иного языка необходимо проведение многолетней работы больших высокооплачиваемых коллективов по сбору и анализу акустических данных огромного числа носителей данного языка. Необходимо тщательно учесть все типы голосов и акцентов, имеющихся у носителей языка, и для каждой разновидности голоса и акцента получить достоверную оценку элементов кодовой книги данного языка. Не менее сложная задача - это построение эталонов слов. Для этого необходимо, чтобы каждое слово словаря (а их может быть около $100,000$) было произнесено каждым представителем данного типа диктора несколько десятков раз, иначе полученный вероятностный автомат будет статистически недостоверен. Наконец, для успешного применения синтаксических и семантических зависимостей между словами предложений необходимо построить некоторую грамматику, в той или иной мере отражающую строение предложений языка.

В связи с вышесказанным встает проблема переноса компьютерных систем распознавания речи, работающих сейчас главным образом на моделях языков германской группы (английском, немецком, французском, итальянском и т.п.) на другие группы языков, например, славянские или азиатские. По сути, перед разработчиками распознающих систем встает задача построения таких систем заново, практически с нуля, поскольку львиную долю времени и средств при разработке новой системы занимает процесс построения достоверной акустической модели, эталонов слов и грамматики языка. При построении систем, распознающих русский язык, например, придется не только строить новую акустическую модель и обучать словарь наиболее используемых русских слов, но и строить модели грамматики русского языка, которые, как несложно предположить, будут на порядок сложнее модели триграм, использующейся сейчас для задания грамматики языка английского. Именно в силу этого на рынке нет и в ближайшем времени вряд ли появится до-

стойная система диктовки русских текстов - работа над ее построением требует слишком больших финансовых вложений.

4. Проблемы и перспективы их разрешения

Суммируя вышесказанное, можно перечислить следующие проблемы, стоящие перед разработчиками систем распознавания речи.

Проблема подавления стационарных и нестационарных помех.

Существующие сейчас системы диктовки текстов и голосового управления компьютером практически не применяют в своей работе алгоритмы шумоподавления. Это связано с тем, что компьютерные речевые системы чаще всего используются в условиях дома или офиса, где уровень внешних помех не очень высок. Однако шумоподавление используют в системах речевого управления реальными техническими устройствами, например, в авиации. Учитывая то, что отсутствие шумоподавления в компьютерных речевых системах сказывается на проценте ошибок при распознавании (например, если пользователь делает громкий выдох, система всегда идентифицирует его как одно из слов словаря, да еще и пытается учесть контекст), можно в ближайшем будущем ожидать перенесения алгоритмов и устройств шумоподавления на персональные компьютеры.

Проблема перехода к распознаванию непрерывной речи.

Эта проблема связана скорее с недостатком вычислительных мощностей персональных компьютеров, делающим пока системы непрерывной диктовки слишком дорогими и потому непопулярными. Повсеместный переход на распознавание непрерывной речи - дело ближайших десяти лет, хотя при этом задача распознавания дискретной речи не теряет силы.

Проблема учета контекста.

В настоящий момент для учета контекста (синтаксиса, семантики) при восстановлении последовательности произнесенных слов пользуются простыми, не зависящими от языка грамматиками, позволяющими, однако, вследствие богатого обучающего материала (в распоряжении раз-

работчиков имеется огромное количество текстов на данном языке, хранящихся в электронном виде) в достаточной степени учитывать связь слов в предложениях на естественном языке. В будущем следует ожидать усложнения грамматик, разработанных с учетом специфики языков, и разработки соответствующих процедур обучения.

Проблема поиска новых звуковых параметров.

Сейчас для распознавания речи используют в основном спектральные параметры речи - быстрое преобразование Фурье, спектр линейного предсказания, кепстральные коэффициенты и т.п. Эти параметры обладают как рядом преимуществ (соответствие восприятию звука человеком, возможность применения Евклидовой метрики или ее вероятностного аналога - расстояния Махаланобиса для сравнения окон звука), так и недостатков (зависимость спектральных параметров от голоса диктора). Естественно, будут продолжены исследования по поиску инвариантных относительно типа голоса и влияния шумов речевых параметров.

Проблема переноса на другие языки.

Вышеизложенная проблема переноса на другие языки имеет скорее не научный, а коммерческий характер. Возможность появления систем распознавания русской речи в будущем необходимо связывать с удешевлением вычислительных средств и появлением потребности в таких системах. Единственной пока областью серьезного коммерческого применения систем диктовки на Западе является медицина, и поэтому появление подобных систем у нас произойдет не раньше, чем на Западе появится удобная и, главное, практически применимая система диктовки для повседневных нужд.

Проблема поиска новых алгоритмов восстановления последовательности произнесенных звуков.

Сейчас для сравнения последовательности акустических параметров с эталонами слов словаря используется только три метода - самый распространенный метод скрытых марковских процессов, тесно связанный с ним метод динамической деформации времени (применяемый на словарях небольшого размера) и стоящий несколько поодаль метод с использованием нейронных сетей. Можно надеяться, что в недалеком будущем появятся принципиально новые математические методы в области распознавания речи.

Работа выполнена на кафедре Математической теории интеллектуальных систем механико-математического факультета Московского Государственного Университета им. М. В. Ломоносова.

Автор выражает благодарность своему научному руководителю кандидату физико-математических наук Дмитрию Николаевичу Бабину за помощь в написании настоящего обзора.

Список литературы

[1] *Microsoft Speech SDK 3.0.* Бета-версия. Документация.

[2] Винцюк Т.К. *Анализ, распознавание и интерпретация речевых сигналов.* - Киев, 1985 г.

[3] Рабинер Л.Р., Шафер Р.В. *Цифровая обработка речевых сигналов*: пер. с англ. - М.: Радио и связь, 1981 г.

[4] Мазуренко И.Л. *О сокращении перебора в словаре речевых команд в составе системы распознавания речи.* В сб.: Интеллектуальные системы, т.2, Москва, 1997 г.

[5] Internet: электронные страницы фирмы 21st Sentury Eloquence (http:// www.voicerecognition.com/).

[6] Патент США N 4769845.

[7] Патент США N 5473726 *Audio and amplitude modulated photo data collection for speech recognition.*

[8] Патент США N 5539861 *Speech recognition using bio-signals.*

[9] Патент США N 5151944 *Headrest and mobile body equipped with same.*

[10] Патент США N 5680505 *Recognition based on wind direction and magnitude.*

[11] *Цифровая обработка акустических сигналов*, Выч. центр АН СССР, Москва, 1989.

[12] Matsumoto-H. Imai-H. *Comparative study of various spectrum matching measures on noise robustness.* Proceedings of ICASSP-86. Tokyo, Japan. pp. 769-772. April 1986.

[13] Dal Degan-N. Prati-C. *Acoustic Noise Analysis and Speech Enhancement Techniques for Mobile Radio Applications.* Signal Processing. vol. 15, pp. 43-56. 1988.

[14] Frost-O-L. *An Algorithm for Lineary Constrained Adaptive Array Processing.* Proceedings of the IEEE. vol. 60, no. 8, pp. 926-935. August 1972.

[15] Rabiner, L. R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition* , Proceedings of the IEEE 1989.

[16] *Discrete-Time Processing of Speech Signals*, authored by J. R. Deller, J. G. Proakis, and J. H. L. Hansen, Macmillan Publishing Company.

[17] *Readings in Speech Recognition*, edited by A. Waibel and K. F. Lee, Morgan Kaufmann Publishers, Inc. San Mateo, CA, copyright 1990.

[18] Schmidt-R-O. *Multiple Emitter Location and Signal Parameter Estimation.* IEEE Transactions on Antennas and Propagation. vol. AP-34, no. 3, pp. 276-280. March 1986.

[19] Watanabe-H. Suzuki-M. Nagai-N. Miki-N. *Maximum likelihood bearing estimation by quasi-Newton method using a uniform linear array.* ICASSP 91: 1991 International Conference on Acoustics, Speech and Signal Processing. Toronto, Ont., Canada. pp. 3325-8 vol.5. IEEE. 14-17 April 1991.

[20] Ziskind-I. Wax-M. *Maximum likelihood localization of multiple sources by alternating projection.* IEEE Transactions on Acoustics, Speech and Signal Processing. vol.36, no.10. pp. 1553-60. Oct. 1988.

[21] Fried-N. Cuperman-V. *Evaluation of Speech Recognition Equipment in a Vehicular Environment.* IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. pp. 455-458. 1-2 June 1989.

[22] Ghitza-O. *Auditory nerve representation as a front-end for speech recognition in a noisy environment.* Computer Speech and Language. vol. 1 pp. 109-130.

[23] Juang-B-H. *Speech Recognition in Adverse Environments.* Computer Speech and Language. vol. 5, pp. 275-294.