УДК 004.522

И.С. Кипяткова, А.А. Карпов

АНАЛИТИЧЕСКИЙ ОБЗОР СИСТЕМ РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ С БОЛЬШИМ СЛОВАРЕМ

Кипяткова И.С., Карпов А.А. **Аналитический обзор систем распознавания русской речи с большим словарем**.

Аннотация. Использование большого словаря необходимо для задачи стенографирования флективных языков, поскольку эти языки характеризуются наличием множества словоформ, образующих парадигму слова. В статье представлен обзор существующих систем распознавания речи, использующих большой и сверхбольшой словари, описаны методы и модели, применяемые в этих системах, приведены данные об их точности распознавания.

Ключевые слова: системы распознавания речи, сверхбольшой словарь.

Kipyatkova I.S., Karpov A.A. **An Analytical Survey of Large Vocabulary Russian Speech Recognition Systems.**

Abstract. The usage of large vocabulary is necessary for the inflective language dictation task, because in these languages there are lots of word-forms that comprise a word paradigm. In the paper, a survey of existing speech recognition systems that use large and extra-large vocabulary is presented, methods and models applying in these systems are described, data about recognition accuracy are given.

Keywords: speech recognition systems, extra-large vocabulary.

1. Введение. Одной из основных нерешенных проблем в области речевых исследований и технологий является автоматическое стенографирование или распознавание слитной разговорной речи. Процесс автоматического распознавания речи представляет собой преобразование акустического речевого сигнала, полученного от микрофона, в последовательность слов, которая затем может использоваться для интерпретации смысла речевого высказывания. В отличие от печатного текста или искусственных сигналов, естественная речь не допускает простого и однозначного членения на элементы (фонемы, слова, фразы), поскольку они не имеют явных физических границ. Границы слов в потоке речи автоматически могут быть определены лишь в ходе распознавания посредством подбора оптимальной последовательности слов, наилучшим образом согласующейся с входным потоком речи по акустическим, лингвистическим, семантическим и иным критериям.

Согласно принятой сейчас в мире классификации, малым словарем распознавания считается словарь, содержащий единицы и десятки слов. Задач и приложений, где используется малый словарь, достаточно много: распознавание последовательностей цифр, номеров телефо-

нов; системы речевого командного управления и т.д. Средний распознаваемый словарь содержит сотни слов. Такого словаря достаточно для работы большинства диалоговых или запросно-ответных систем. Большой словарь содержит тысячи и десятки тысяч слов, такие системы распознавания могут использоваться в автоматизированных справочных системах или системах диктовки текста в ограниченной предметной области (для аналитических языков). Словарь размером в сотни тысяч и миллионы слов считается сверхбольшим, он позволяет реализовывать системы стенографирования текста (включая синтетические языки).

Мировых исследований, посвященных разработке систем распознавания речи со сверхбольшим словарем, довольно мало. Это связано с тем, что для многих языков такой словарь был бы избыточным. При объеме словаря в 65 тыс. слов английского языка число внесловарных слов (*out-of-vocabulary words*) составляет 1,1 % [17]. Для флективных же языков, к числу которых относится и русский, из-за наличия большого числа словоформ для каждой парадигмы слова объем словаря распознавания и число существующих внесловарных слов возрастают на порядок по сравнению с аналитическими языками.

2. Системы распознавания слитной русской речи с большим и сверхбольшим словарем. Одна из первых систем распознавания русской речи с большим словарем была разработана в IBM [9]: обучение системы и распознавание слов производились на новостных текстах ТАСС. Объем словаря системы распознавания составил 36 тыс. слов. Для обучения использовались 30 тыс. фраз, произнесенных 40 дикторами (каждый диктор произнес в среднем по 750 фраз). В качестве акустических единиц использовалось 47 фонем. Триграммная модель языка была обучена на текстах в 40 млн слов. Коэффициент неопределенности модели языка [17] на тестовых данных равен 180. Для распознавания использовались 149 предложений, произнесенных 8 дикторами, которые не участвовали в записи обучающих данных. Также в работе [9] исследовалась модель языка, в которой происходило деление слова на основу и окончание, т.е. модель языка представлялась в виде цепочки из 6 частей. Тестирование проводилось на 149 предложениях из текстов ТАСС. Число неправильно распознанных слов составило менее 5 %.

В работе [3] описывается система распознавания русской речи на базе комплекса программ SDT, который был разработан на предприятии ВНИИЭФ-СТЛ в ходе работ по контракту с фирмой Intel и предназначался для построения систем распознавания слитной речи с боль-

шим словарем. Первоначально он создавался для распознавания английского и китайского языков, затем на базе SDT был построен прототип системы распознавания слитной речи для русского языка. В данном прототипе не применялись алгоритмы, специально предназначенные для русского языка. Для построения акустических моделей использовались корпуса ISABASE и RuSpeech, записанные фирмой Cognitive Technologies, и корпус MultiSpeech, записанный фирмой Auditech, общая продолжительность речи около 80 ч. Было построено два вида внутрисловных моделей с 6 тыс. состояний: для HTK-совместимых признаков и признаков, вычисляемых согласно стандарту ETSI ES 202 050 v1.1.1 (Aurora). Модели для узкой транскрипции (114 монофонов) строились с использованием корпусов ISABASE и RuSpeech (51 тыс. фраз), для широкой транскрипции (55 монофонов) — с использованием всех трех корпусов (153 тыс. фраз). Для построения словаря (23 000 слов) и модели языка использовались совокупность предложений, записанных в корпусах RuSpeech и MultiSpech, и две статьи о речевых технологиях из газеты ComputeWorld. В качестве первой тестовой задачи использовалась тестовая часть корпуса RuSpeech. Она содержит 1 тыс. тестовых предложений, которые произносят 10 разных дикторов (5 женщин и 5 мужчин). Тексты взяты из газетных статей и новостных сайтов. В качестве второй тестовой задачи использовалось 500 тестовых предложений из первой задачи и 500 предложений из корпуса MultiSpeech. Эти тестовые фразы были исключены из обучения акустических моделей. Было добавлено 4 диктора (2 женщины и 2 мужчины). Результаты тестирования представлены в табл. 1.

Таблица 1. **Доля неправильно распознанных слов при использовании системы распознавания SDT, %**

Тестовая задача	Транскрипция	Число монофонов	HTK	Aurora
Первая	Широкая	114	2,7	3,5
	Узкая	55	2,3	3,1
Вторая	Широкая	114	10,8	6,8
	Узкая	55	6,1	6,3

В работе [15] для распознавания русской речи было предложено использовать графемы (минимальные единицы письменной речи, то есть буквы) вместо фонем (минимальных единиц звукового строя языка). Обучение и тестирование системы распознавания осуществлялось

на русской части корпуса GlobalPhone, представляющей собой чтение новостных статей. Корпус был разделен на три части:

1) для обучения акустических единиц (train) — 8170 фраз (продолжительность речи 17,0 ч);

2) для настройки параметров модели языка (dev) — 898 фраз (продолжительность речи 1,3 ч);

3) для распознавания (eval) — 1029 фраз (продолжительность речи 1,6 ч).

Модель языка была обучена на 19 млн слов из текстов, собранных из 6 газет, доступных on-line, относящихся к периоду с 1997 по 2004 г. Для триграммной модели коэффициент неопределенности на корпусе dev равен 1833. Доля неправильно распознанных слов дана в табл. 2.

Таблица 2. **Доля неправильно распознанных слов при использовании различных акустических единиц, %**

Подход	Тип корпуса	
	dev	eval
Фонемы	54,3	48,8
Графемы	55,1	51,4
Трифоны	33,0	33,5
Триграфемы	36,4	37,3
Триграфемы с увеличенным деревом	32,8	35,7

Результаты экспериментов, представленные в табл.2, показывают, что в общем случае при использовании графем вместо фонем ухудшается точность распознавания. Однако создание увеличенного дерева решений при обучении триграфем позволило ее повысить.

В работе [13] представлена система распознавания спонтанной чешской, словацкой и русской речи для обработки интервью очевидцев холокоста. В данной работе базовые транскрипции создавались автоматически с использованием определенного набора правил, при этом для многих слов генерировались несколько вариантов транскрипций для учета фонетических явлений слитной речи (например, ассимиляции согласных на границе слов). Затем создавались транскрипции, описывающие разговорные варианты произношения, а для русского языка еще и акцент, поскольку интервью были взяты не только у жителей России, а также у русских, живущих в Украине, Израиле, США. Кроме того, моделировались неречевые явления. Размер корпуса (продолжительность речи), использованного при создании акустических моделей для русского языка, составлял 100 ч. Модель языка представ-

ляла собой биграммную модель с применением методики возврата (*Katz's backing-off scheme*). При объеме словаря в 79 тыс. транскрипций неправильно распознано 38,57 % слов.

Для работы со сверхбольшими словарями (объемом более 1 млн слов) в работе [4] предлагается двухпроходный алгоритм распознавания дискретной и слитной речи. На первом этапе применяется фонетический стенограф, на выходе которого получается набор распознанных фонем. Заранее в процессе обучения из словаря транскрипций создается индекс от троек фонем к транскрипциям. Ключом индекса является тройка фонем. Каждое вхождение в таблицу содержит список транскрипций, в которые входит тройка фонем ключа вхождения. Выход фонетического стенографа делится на пересекающиеся тройки фонем со сдвигом в одну фонему. Получающаяся тройка фонем становится запросом к базе данных. Ответ на один запрос состоит из списка транскрипций, в которые входит данная тройка фонем. Этот список копируется в подсловарь для второго прохода алгоритма. Следующий запрос из потока добавляет новую порцию транскрипций, при этом подсчитывается число повторений, чтобы можно было вычислить ранг слова в подсловаре. Все транскрипции в полученном подсловаре упорядочиваются согласно рангу слова (счетчику повторений). Первые N транскрипций заносятся в окончательный подсловарь для второго прохода алгоритма. Таким образом, подсловарь для распознавания содержит не более N транскрипций с наивысшими рангами, где N — фиксированное число. На втором проходе речевой сигнал распознается в условиях ограниченного подсловаря. Для распознавания слитной речи дополнительно используется граф формирования слов. Каждая тройка фонем из ответа фонетического стенографа порождает промежуточную вершину с номером, синхронным времени появления тройки фонем. Транскрипции вставляются между промежуточными вершинами так, чтобы порождающие тройки фонем оказались в одной колонке по вертикали. В случае, когда пересекаются транскрипции одного слова, порожденные разными тройками фонем, ранги этих транскрипций увеличиваются на единицу. Для уменьшения сложности графа слов используется ограничение N для числа слов в каждый момент времени. При этом удаляются слова с малыми рангами. Для экспериментов по распознаванию речи использовались словари объемом 15 тыс., 95 тыс. и 1987 тыс. слов. Акустические модели обучались на выборке из 12 тыс. звуковых записей из словаря объемом в 2037 слов и фраз, произнесенных одним диктором. Для проверки надежности распознавания слитной речи было записана 1 тыс. фраз с числами от 0 до

999. Для словаря объемом в 1987 тыс. слов точность распознавания составила 83 %. Для словарей объемом 15 тыс. и 95 тыс. слов проведено сравнение точности распознавания с использованием базовой системы и системы с двухпроходным декодером. Эксперименты показали, что при использовании двухпроходного декодера точность распознавания ухудшилась, однако время распознавания сократилось. Результаты экспериментов приведены в табл. 3.

Таблица 3. **Точность распознавания цифр с применением двухпроходного декодера**

Система	Параметр распознавания	Объем словаря, тыс. слов		
		15	95	1987
Базовая	Точность, %	96,5	92,6	—
	Время, с	36,0	205	—
С двухпроходным декодером	Точность, %	85,3	84,9	83,0
	Время, с	2,3	9,4	160,0

В работе [1] описан процесс обучения системы автоматического распознавания речи в новостных передачах. Текстовый корпус для создания модели языка собирался из Интернет-источников. Для русского языка использованы 25 сайтов, объем корпуса составил 129,5 млн словоформ. По текстовому корпусу созданы три частотных словаря:

1) общий,

2) имен собственных,

3) имен нарицательных.

Для русского языка выбран словарь объемом 213 тыс. слов, покрывший 98 % текста. В качестве модели использовались *n*-граммы со значением $n = 1, 2, 3$. Полученные частотные словари для каждого языка транскрибированы специально разработанной программой — автоматическим транскриптором на основе грамматики GTT (*Grammar for Text Transcription*). Для русского языка использовался расширенный вариант SAMPA (Speech Assessment Methods Phonetic Alphabet), в который были добавлены степени редукции гласных для заударного слога, второго и последующих предударных слогов, за исключением позиции абсолютного конца и абсолютного начала. Разработка транскриптора (реализация транскрибирования) проводилась в четыре этапа в соответствии с языковыми формами:

1) для изолированных слов по правилам литературной нормы,

2) для слитной речи по правилам литературной нормы,

3) для изолированных слов разговорной речи,

4) для слитной разговорной речи.

Точность транскрипции доходила до 99 %. Объем речевого корпуса для обучения акустических моделей превысил 200 ч, в записи принимали участие 3280 дикторов. Была достигнута точность распознавания 60–70 % слов в зависимости от качества звуковых файлов.

В СПИИРАН разработана система распознавания речи SIRIUS [2], в которой введен дополнительный уровень представления языка и речи — морфемный уровень. За счет разделения словоформ на морфемы словарь распознаваемых лексических единиц значительно сократился, так как в процессе словообразования часто используются одни и те же морфемы. Разработана также модель компактного представления словарей сверхбольшого объема для систем распознавания речи флективных языков (в частности, русского) на базе двухуровневого морфофонемного префиксного графа (ДМПГ) [6]. По сравнению с базовыми подходами в модели ДМПГ число структурных элементов (узлы фонем, основ, концовок и дуги) сокращается более чем на порядок. В процессе декодирования русской слитной речи модель ДМПГ обеспечивает формирование на выходе распознавателя только грамматически правильных слов и позволяет увеличить скорость распознавания речи. В системе SIRIUS при распознавании слитной русской речи точность превышает 90 % на словаре объемом около 2000 слов из конкретной предметной области.

Распознаванием речи с большим и сверхбольшим словарем также занимаются в компании «Центр речевых технологий» (ЦРТ). Для этого используется модель языка, основанная на делении слов на основание и окончание [12]. В итоге словарь состоит из 85 тыс. оснований и 4 тыс. окончаний, покрывая таким образом 1300 тыс. словоформ.

Также разработкой систем распознавания речи занимается компания Vocative. В статье [8] представлены результаты распознавания русской речи для задач распознавания ключевых слов, названий фильмов, станций метро, дат рождения, стран. Объем словаря такой системы небольшой — 20–100 слов, а точность распознавания в зависимости от задачи — 80–100 %.

Проведенный обзор систем распознавания русской речи показывает, что на данный момент фактически не существует подобных систем со сверхбольшим словарем. Однако для многих других флективных языков проводятся исследования по созданию систем распознавания речи со сверхбольшими словарями, что показывает необходимость

создания таких систем и для русского языка. Поэтому в следующем разделе представлено несколько систем распознавания речи со сверхбольшим словарем для других флективных языков.

3. Системы распознавания слитной речи с большим и сверхбольшим словарем для других языков. В системе распознавания украинской речи, разработанной для стенографирования заседаний Верховной Рады Украины [5], в качестве акустических единиц применялись 56 украинских контекстно-независимых фонем (включая фонему-паузу). Для обучения использовались записи продолжительностью 28 ч, содержащие 211 тыс. слов. Записана речь 208 дикторов. Распознавание производилось на образцах выступлений депутатов Верховной Рады Украины, записанных в отличные от обучающей выборки дни. Для распознавания использовались записи длиной в 8 ч, содержащие 69 тыс. слов. Всего использовались записи 118 дикторов. Записи 36 дикторов не встретились в обучающей выборке, т.е. эти дикторы оказались неизвестными для системы распознавания. Словарь был составлен из текстов стенограмм заседаний Верховной Рады Украины. С официального сайта Верховной Рады были загружены все стенограммы заседаний начиная с 1991 г., что превысило 100 Мб информации. Результирующий текст разделен на две части:

1) стенограммы за весь указанный период, кроме 2002–2003 гг. (14,6 млн слов),

2) стенограммы 2002–2003 гг. (409 тыс. слов).

По первой части корпуса составлен частотный словарь из 156 тыс. слов. Доля текста, покрываемого словами с частотами упоминания выше 50, превышает 94 %, что соответствует словарю в 15 тыс. слов. Была создана биграммная модель языка с использованием обратных (*back-off*) коэффициентов. Точность распознавания для различных объемов словаря представлена в табл. 4. Авторы статьи [5] делают вывод, что наиболее оптимален словарь в 15 тыс. слов, поскольку снижение точности распознавания незначительно по сравнению с распознаванием при использовании максимального словаря.

Таблица 4. **Точность распознавания слов украинской речи, %**

Объем словаря, тыс. слов	64	50	30	20	15	10	5
Точность распознавания	68,59	68,54	68,38	67,79	67,15	65,49	62,18

Отметим также несколько систем распознавания с большим и сверхбольшим словарями для других флективных языков. В табл. 5 представлены результаты распознавания чешской речи для различных областей [11].

Таблица 5. **Точность распознавания чешской речи**

Задача	Стиль речи	Словарь (тип и объем)	Точность, %
Изложение медицинских заключений	Диктовка	Предметно-ориентированный, 130 тыс. слов	97
Изложение судебных заключений	То же	Предметно-ориентированный, 300 тыс. слов	95
Радионовости (записанные в студии)	Чтение профессиональными дикторами	Общий, 350 тыс. слов	94
Радионовости (записанные на улице)	То же	Общий, 350 тыс. слов	89
Ток-шоу (записанные на студии)	Разговор	Общий, 350 тыс. слов	78

В работе [11] представлена система распознавания спонтанной чешской речи, предназначенная для распознавания телефонных разговоров. В ней для учета вариативности произношения в спонтанной речи создаются дополнительные варианты транскрипций. Модель языка основана на Интернет-материалах и транскрибированных устных разговорах. Для тестирования системы использовано 500 фраз, точность распознавания составила 48,5 %.

Зачастую в системах распознавания речи с большим и сверхбольшим словарем объем последнего сокращается путем деления слов на морфемы. В работе [10] представлены результаты распознавания при делении слов на морфемы: для финского языка точность составила 73,2 %, для эстонского — 34 %. Когда в системе распознавания не использовалось такое деление, точность распознавания эстонского языка составила 62,5 % [14]. Также морфемная модель языка использовалась для распознавания турецкой речи, его точность составила 42,6 % [7]. В работе [16] основанная на морфемах система распознавания венгерского языка тестировалась для трех задач по следующим видам речи:

1) спонтанной разговорной (объем словаря 20 тыс. слов), достигнута точность распознавания 50 %;

2) выступлениям на пресс-конференциях (объем словаря 92 тыс. слов), точность распознавания 70 %;

3) радионовостям (объем словаря 285 тыс. слов), точность распознавания 79 %.

4. Заключение. Использование сверхбольших словарей особенно актуально для задачи распознавания русской речи, поскольку флективный русский язык характеризуется наличием множества словоформ, образующих парадигму слова. Анализ публикаций показал, что на данный момент фактически не существует систем распознавания русской речи со сверхбольшим словарем. Кроме того, во всех проанализированных исследованиях не учитывались особенности русского языка при построении и применении языковых моделей, а использовались только базовые статистические методы обработки текста. Поэтому необходимо разрабатывать новые методы и подходы для создания оригинальных статистических моделей русского языка для распознавания слитной речи со сверхбольшим словарем.

Литература

1. *Викторов А.Б., Грамнщкий С.Г., Гордеев С.С., Ескевич М.В. и др.* Универсальная методика подготовки компонентов обучения систем распознавания речи // Речевые технологии. Народное образование. 2009. № 2. С. 39–55.
2. *Карпов А.А., Ронжин А.Л., Ли И.В.* SIRIUS — система дикторонезависимого распознавания слитной русской речи // Известия ТРТУ. 2005.№ 10. С. 44–53.
3. *Кибкало А.А., Лотков М.М., Рогожкин И.Г., Туровец А.А.* Разработка системы распознавания русской речи // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 2003. Вып. 3. С. 8–20.
4. *Пилипенко В.В.* Распознавание дискретной и слитной речи из сверхбольших словарей на основе выборки информации из баз данных // Искусственный интеллект. 2006. № 3. С. 548–557.
5. *Пилипенко В.В., Робейко В.В.* Автоматизированный стенограф украинской речи // Искусственный интеллект. 2008. № 4. С. 768–775.
6. *Ронжин А.Л.* Топологические особенности морфофонемного способа представления словаря для распознавания русской речи // Вестник компьютерных и информационных технологий. 2008.№ 9. С. 12–19.
7. *Arisoy E., Dutagaci H., Arslan L.M.* A unified language model for large vocabulary continuous speech recognition of Turkish // Signal Processing. 2006. Vol. 86, № 10. P. 2844–2862.
8. *Bolotova O., Gusev M., Smirnov V.* Speech Recognition System for the Russian Speech // Proc. of 12th Intern. Conf. on Speech and Computer SPECOM. Moscow. Russia. 2007. P. 475–480.
9. *Kanevsky D., Monkowski M., Sedivy J.* Large Vocabulary Speaker-Independent Continuous Speech recognition in Russian Language // Proc. Intern. Workshop SPECOM'96. St.Petersburg. Russia. 1996. P. 117–121.
10. *Kurimo M., Hirsimäki T., Turunen V.T., Virpioja S. et al.* Unsupervised decomposition of words for speech recognition and retrieval // Proc. of 13th Intern. Conf. «Speech and Computer», SPECOM'2009. St. Petersburg. 2009. P. 23–28.

11. *Nouza J., Silovsky J.* Adapting Lexical and Language models for Transcription of Highly Spontaneous Spoken Czech / Eds. P. Sojka et al. // TSD 2010. LNAI 6231. Berlin-Heidelberg, 2010. P. 377–385.
12. *Oparin I, Talanov A.* Stem-Based Approach to Pronunciation Vocabulary Construction and Language Modeling for Russian // Proc. of 10th Intern. Conf. SPECOM, Patras, Greece, 2005. P. 575–578.
13. *Psutka J., Ircing P., Psutka J.V., Hajič J. et al.*. Automatic Transcription of Czech, Russian, and Slovak Spontaneous Speech in the MALACH Project // Proc. of Eurospeech. Lisboa. Portugal. Sept. 4–8. 2005. P. 1349–1352.
14. *Ragni A.* Initial Experiments with Estonian Speech Recognition // Proc. of the 16th Nordic Conf. of Computational Linguistics NODALIDA-2007. Nivre J. et al. (Eds). Tartu. 2007. P. 249–252.
15. *Stuker S., Schultz T.* A grapheme Based Speech Recognition System for Russian // Proc. Intern. Conf. SPECOM'2004. St.Petersburg. Russia. 2004 P. 297–303.
16. *Tarjan B., Mihajlik P.* On Morph-Based LVCSR Improvements // Proc. of 2nd Intern. Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU-10). 2010. P. 10–16.
17. *Whittaker E.W.D.* Statistical Language Modelling for Automatic Speech Recognition of Russian and English. PhD thesis. Cambridge University. 2000, 140 p.

Кипяткова Ирина Сергеевна — младший научный сотрудник лаборатории речевых и многомодальных интерфейсов Учреждения Российской академии наук Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: автоматическое распознавание речи, статистические модели языка. Число научных публикаций — 15. kipyatkova@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081. Научный руководитель — канд. техн. наук А.А. Карпов.

Kipyatkova Irina Sergeevna — junior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition statistical language models. The number of publications — 15. kipyatkova@iias.spb.su; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081. Scientific adviser — PhD A.A. Karpov.

Карпов Алексей Анатольевич — канд. техн. наук**,** старший научный сотрудник лаборатории речевых и многомодальных интерфейсов Учреждения Российской академии наук Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: автоматическое распознавание речи, многомодальные интерфейсы, аудиовизуальное распознавание речи. Число научных публикаций — 100. karpov@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Karpov Alexey Anatolyevich — PhD, senior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, multimodal interfaces, audio-visual speech recognition. The number of publications — 100. karpov@iias.spb.su; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Поддержка исследований. Данное исследование поддержано Советом по Грантам Президента РФ (проект МК-64898.2010.8), фондом РФФИ (проекты № 08-08-00128, 09-07-91220-СТ, 10-08-00199) и Комитетом по науке и высшей школе Правительства Санкт-Петербурга.

Рекомендовано лабораторией речевых и многомодальных интерфейсов, заведующий лабораторией д-р техн. наук, доц. А.Л. Ронжин.
Статья поступила в редакцию 08.11.2010.

РЕФЕРАТ

Кипяткова И.С., Карпов А.А. **Аналитический обзор систем распознавания русской речи с большим словарем.**

В статье представлен обзор систем распознавания русской речи с большим словарем. Одна из первых таких систем распознавания русской речи имела словарь объемом 36 тыс. слов; при тестировании этой системы на фразах, взятых их обучающего корпуса, точность распознавания составила более 95 %. Другая система, построенная на базе комплекса программ SDT, дает точность более 90 % при объеме словаря 23 тыс. слов. В системе распознавания речи, разработанной для обработки интервью очевидцев холокоста, для русского языка использовался словарь объемом 79 тыс. транскрипций, при этом процент неправильно распознанных слов составил 38,57 %. Также в статье описывается система, в которой для распознавания русской речи вместо фонем и трифонов используются соответственно графемы и триграфемы. Затем описывается двухпроходный алгоритм распознавания дискретной и слитной речи для работы со сверхбольшими словарями. Применение этого алгоритма уменьшает точность распознавания, однако время распознавания значительно снижается. Также представлена система распознавания речи в новостных передачах, объем словаря этой системы составляет 213 тыс. слов, точность распознавания — 60–70 %. В системе распознавания речи SIRIUS введен дополнительный уровень представления языка — морфемный, что позволяет сократить объем словаря и ускорить обработку речи. На словаре объемом около 2 тыс. слов точность распознавания составляет более 90 %. В компании «Центр речевых технологий» для распознавания речи с большим и сверхбольшим словарем используют модель языка, основанную на делении слова на основание и окончание. Также в статье представлена система распознавания речи, разработанная компанией Vocative.

Кроме того, в статье описаны несколько систем с большим и сверхбольшим словарем для распознавания других флективных языков: украинского, чешского, финского, эстонского, турецкого и венгерского. Представленная в статье система распознавания украинской речи разработана для стенографирования заседаний Верховной Рады Украины. Эксперименты для тестирования этой системы проводились при различном объеме словаря, при этом максимальный словарь содержал 156 тыс. слов. По результатам экспериментов был выбран словарь объемом в 15 тыс. слов, поскольку при его использовании точность распознавания снижалась незначительно по сравнению с вариантом использования максимального словаря.

Проведенный обзор показал, что на данный момент фактически не существует систем распознавания русской речи со сверхбольшим словарем, а в разрабатываемых системах не учитываются особенности русского языка при построении языковых моделей. Поэтому необходимо разрабатывать новые методы и модели для создания моделей русского языка для распознавания слитной речи со сверхбольшим словарем.

SUMMARY

Kipyatkova I.S., Karpov A.A. **An Analytical Survey of Large Vocabulary Russian Speech Recognition Systems.**

In the paper, a survey of large vocabulary Russian speech recognition systems is presented. One of the first such Russian speech recognition systems had 36 K vocabulary; recognition accuracy of this system was over 95 % when testing on phrases from a training corpus. Another system built on STD software complex gives the accuracy more than 90 % with a vocabulary of 23 K words. In the speech recognition system built for processing of interviews of Holokost witness, the vocabulary of 79 K transcriptions was used for Russian speech recognition; the word error rate of this system was 38.57 %. Also a system, where graphemes and trigraphemes were used instead of phonemes and triphones, is described. Then a two-pass algorithm of discrete and continuous speech for operating with an extra-large vocabulary is described. The usage of this algorithm decreases the recognition accuracy, however recognition time reduces significantly. Also a system of recognition of broadcast news is presented, vocabulary capacity of this system is 231 K words; recognition accuracy is 60–70 %. In the SIRIUS speech recognition system, an additional level of language representation (a morphemic level) is introduced, that allows to reduce vocabulary capacity and to fasten speech processing. The recognition accuracy is above 90 %. In Speech Technology Center, a language model based on division of words on stems and endings is used for large and extra-large vocabulary speech recognition. Also in the paper, a speech recognition system developed by Vocative is mentioned.

Furthermore, in the paper, some other systems with large and extra-large vocabularies are described in order to recognize speech in other inflective languages: Ukrainian, Czech, Finnish, Estonian, Turkish, and Hungarian. An Ukrainian speech recognition system presented in the paper is developed for stenographing of speeches in the Verkhovna Rada of Ukraine. Experiments on testing of this system were made with different vocabulary capacities, including a maximal vocabulary of 156 K words. According with experiments a 15 K vocabulary was chosen since it provides a minimal decrease of recognition accuracy in comparison with recognition with full vocabulary.

The survey has shown that at the moment in fact there is no Russian speech recognition system with an extra large vocabulary, and some developing systems do not take into account specific features of the Russian language, when creating language models. Therefore, it is necessary to develop new methods and models to create Russian language models for large vocabulary continuous speech recognition.