ТЕХНИЧЕСКИЕ НАУКИ

УДК 004.896
MSC: 68T10; 68T50; 68T42
DOI: 10.35330/1991-6639-2020-6-98-20-33

СОВРЕМЕННЫЕ ПРОБЛЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ*

И.А. ГУРТУЕВА

Институт информатики и проблем регионального управления –
филиал ФГБНУ «Федеральный научный центр
«Кабардино-Балкарский научный центр Российской академии наук»
360000, КБР, г. Нальчик, ул. И. Арманд, 37-а
E-mail: iipru@rambler.ru

В предлагаемой работе приведен краткий обзор наиболее применяемых методик в области моделирования распознавания речи. Обсуждаются различные принципы транскрипции, разработанные консорциумом лингвистических данных. Описаны проблемы оценки уровня человеческой эффективности при решении задачи распознавания речи, проанализированы типичные ошибки, допускаемые при этом человеком. Показано, что люди демонстрируют высокий уровень согласованности при точной транскрипции предварительно подготовленной англоязычной речи и быстрой транскрипции разговорной телефонной речи. Показано также, что с возрастанием сложности речи возрастает показатель разночтений между двумя и более независимыми стенографистами. Приведены результаты сравнительного анализа ошибок, генерируемых речевой системой и допускаемых человеком. Проанализированы их сходства и различия. Перечислены современные проблемы автоматического распознавания речи, оценены перспективы их решения и определены направления будущих исследований.

Ключевые слова: искусственный интеллект, искусственные нейронные сети, распознавание речи, глубокое обучение, эффективность человека.

1. Введение

Развитие систем искусственного интеллекта в последние годы позволило достичь, а в некоторых случаях преодолеть человеческий уровень в решении широкого диапазона задач от игры в шахматы и Го [1, 2] до простых задач распознавания речи, таких как речь, начитываемая по подготовленному тексту [3], и распознавание перекрывающейся речи [4, 5] на ограниченных словарях. Большинство ранних разработок в области распознавания речи финансировалось *DARPA*. Оно же ставило задачи, которые выполнялись лингвистическим консорциумом (*CLD*) и Национальным институтом стандартов и технологий (*NIST*) [6]. Сначала задачи распознавания были простыми как, например, *Resource Management* [7] с маленьким словарем и тщательно контролируемой грамматикой; каждая последующая значительно усложнялась от распознавания речи по начитываемому тексту в задаче *Wall Street Journal* [8] к транскрипции новостных трансляций [9]. Одной из последних инициатив в данной области была задача распознавания разговорной телефонной речи. Распознавание такой речи особенно сложно вследствие ее спонтанности, неформальности, большого количества самопоправок, пауз и других нарушений беглости, часто встречающихся в ней. Ключевые разработки в данной области были сделаны такими организациями, как *AT&T* [10], *Yandex* [11], *IBM* [12], *Microsoft* [13], *BBN* [14], *SRI* [15], *LIMSI* [16], *Cambridge University* [17] и многими другими.

* Работа выполнена при финансовой поддержке грантов РФФИ №№ 18-01-00658, 19-01-00648

2. Обзор ключевых методик автоматического распознавания речи

Первые попытки по созданию автоматических систем распознавания речи были предприняты в 1950-1960-е годы. Отправной точкой начальных исследований были фундаментальные представления акустической фонетики. Но уже в 1959 году исходные позиции исследований принципиально изменились. В *University College* в Великобритании Фрай и Дин создали распознаватель четырех гласных и девяти согласных на основе статистической информации для учета допустимых фонемных последовательностей в английском языке [18]. Использовав данный подход, исследователи не только повысили усредненную точность распознавания фонем для слов, состоящих из двух и более фонем, но также впервые применили статистический синтаксис на уровне фонем в автоматическом распознавании речи. Интересно, что уже на начальном этапе работ по созданию речевых систем была решена одна из самых сложных задач, существующих в распознавании звуковых сообщений, – временная неоднородность в речевых событиях. Винцюк предложил применить методы динамического программирования для выравнивания по времени пары речевых высказываний (динамическое искажение времени) [19]. С конца 70-х динамическое программирование в разнообразных вариантах, включая алгоритм Витерби [20], стало доминирующим методом в автоматическом распознавании речи. Также в ряду успешных разработок конца 60-х следует указать исследования Редди в *Carnegie Mellon University* в сфере распознавания непрерывной речи на основе динамического отслеживания фонем [21].

В 70-е годы продолжалось интенсивное развитие идей распознавания паттернов [22] и методов динамического программирования [23], а также началась работа над серией экспериментов, нацеленных на разработку дикторонезависимых систем распознавания речи [24]. В *AT&T Bell Labs* для данной цели был использован широкий диапазон сложных классифицирующих алгоритмов, устанавливающих число паттернов, необходимых для представления всех вариантов различных слов для широкого круга пользователей. Проекты по созданию системы понимания и как ее составной части – системы распознавания речи, финансируемые *DARPA,* привели к созданию большого числа оригинальных систем и технологий [25]. Так, система *Hearsay I* использовала сематическую информацию для значительного снижения числа альтернатив, предлагаемых системой распознавания. Система *CMU Harpy* [26] распознавала речь по словарю из 1011 слов с достаточной точностью на основе концепции поиска по графам, в которой распознаваемая речь представлялась как связная сеть, полученная из лексических представлений слов с синтаксическими ограничениями и правилами словарных границ. Кроме того, при поддержке *DARPA* были разработаны системы *CMU's Hearsay II* и *BBN's HWIM (Hear What I Mean)* [25]. *Hearsay II* была создана с использованием концепции параллельных асинхронных процессов.

Исследования в области распознавания речи 1980-х годов характеризуются сдвигом в методологии от парадигмы прямого распознавания образов к формальной концепции статистического моделирования. Нельзя не признать одну из ключевых технологий, разработанных в указанное десятилетие, – скрытое Марковское моделирование [27], ставшее самым применяемым методом буквально в каждой лаборатории мира. В этот период также была переосмыслена идея использования нейронных сетей в речевых системах, поскольку было разработано представление о связи технологии нейронных сетей и классических методов классификации образов [28]. Исследователи *IBM* создали языковую модель, определяющую вероятность возникновения упорядоченной последовательности из n языковых символов (фонем или слов), на основе статистических правил синтаксиса. Важность данной техники для разработки автоматических систем транскрибирования речи на больших словарях трудно переоценить. Кроме того, было разработано и реализовано широкое многообразие алгоритмов, основанных на сопоставлении конкатенированных паттернов отдельных слов, включая алгоритм построения многоуровневого сигнала Майерса и Рабинера [29], алгоритм с синхронизацией фреймов Ли и Рабинера [30], а также подход двух-

уровневого динамического программирования Сакое [23], однопроходный метод [31], каждый из которых обладает собственными преимуществами внедрения.

В 1990-е годы вновь происходит серьезное парадигматическое изменение в представлениях об автоматизации распознавания речевых сообщений. Задача распознавания образов, развивавшаяся в рамках байесовских представлений и, следовательно, требовавшая оценки распределения данных, трансформировалась в задачу оптимизации, включающую в себя минимизацию ошибки эмпирического распознавания [32]. Смена фундаментальных концептуальных представлений была обусловлена пониманием того факта, что функции распределения для речевого сигнала не могут быть выбраны или определены с высокой точностью и что теория байесовских решений становится неприменимой в данных условиях. Система распознавания должна, скорее, обладать наименьшим числом ошибок распознавания, чем демонстрировать наилучшую подгонку функции распределения известного набора данных. Описанная концепция минимизации ошибки породила большое число методов, в том числе критерий минимума ошибки классификации (*MCE*) и обучающий алгоритм обобщенного вероятностного падения (*GPD*) для минимизации целевой функции, аппроксимирующей процент ошибок. Кроме того, был разработан критерий максимума взаимной информации (*MMI*). При *MMI*-обучении взаимная информация между акустическим наблюдением и его корректным лексическим символом, усредненная по набору обучающих данных, максимизируется. Оба подхода приводят к более высокой эффективности распознавания по сравнению с подходом, основанным на максимуме сходства [33]. *DARPA* в 90-е годы сместило фокус исследований в области распознавания на задачу распознавания речи в естественном режиме. Среди наиболее сложных задач, предложенных *DARPA*, была задача создания и исследования речевого корпуса *Switchboard*, в котором собрана разговорная телефонная речь с большим количеством нарушений беглости (функциональные слова, паузы-хезитации, самопоправки). Большое количество разнообразных приложений, в том числе автоматическое голосовое индексирование документов и информационно-поисковые системы, было разработано на основе интеграции технологии транскрибирования широкого вещания и технологий извлечения информации. Особенно острую актуальность приобрела разработка методов повышения устойчивости функционирования систем при несоответствии обучающих и тестовых данных, в условиях высокой зашумленности, индивидуальных особенностях диктора, реверберациях и т.д. Большинство методов основано на максимуме сходства линейной регрессии (*MLLR*) [34], модели декомпозиции [35], композиции параллельной модели (*PMC*) [Gales, 36] и апостериорном структурном максимуме (*SMAP*) [37].

В 2000-х центральным оценочным критерием в разработках по созданию речевых систем стала телефонная разговорная речь. Эта задача крайне сложна вследствие высокой вариативности спонтанной разговорной речи, разнообразия акцентов, ограниченности частотного диапазона канала передачи, отсутствия контекста и т.д. Эффективность систем, все еще использовавших в значительной мере *GMM-HMM*-моделирование при решении данной задачи, стагнировала на отметке около 20% (в разных работах авторы указывают значение *WER* в пределах от 19.8% до 22% [38]). Процент ошибок не удавалось значительно снизить вплоть до 2010 года, когда были разработаны методы глубокого обучения, ставшие в распознавании речевых образов прорывной технологией [38, 39]. Разработчики *Microsoft* [38] продемонстрировали, что методы глубокого обучения, применяемые для акустического моделирования, способствуют значительному улучшению распознавания данных телефонных разговоров по сравнению с *GMM-HMM*-моделированием, а недавние разработки *IBM* показали дальнейшее продвижение [39] по сравнению с работами десятилетней давности. Как показывает диаграмма на рисунке 1, оптимизация одного из этапов обработки речи методами глубокого обучения позволила снизить процент ошибок до 16% (на 30% от предыдущего значения!). На основе данной методологии с использованием со-

временных архитектур сверточных и рекуррентных сетей для акустического [38] и языкового моделирования [39], а также их комбинаций разработчики *Microsoft* и *IBM* добились снижения процента ошибок сначала до уровня человека, а затем преодолели его (5.8% против 5.9% на данных *Switchboard* и 11.0% против 11.3% на данных *CallHome English*). *Saon* и др. получили лучшие результаты и провели отдельный эксперимент по оценке процента ошибок, допускаемых человеком на тех же тестовых данных (5.1% для *Switchboard*, 6.8% для *CallHome*) [39].

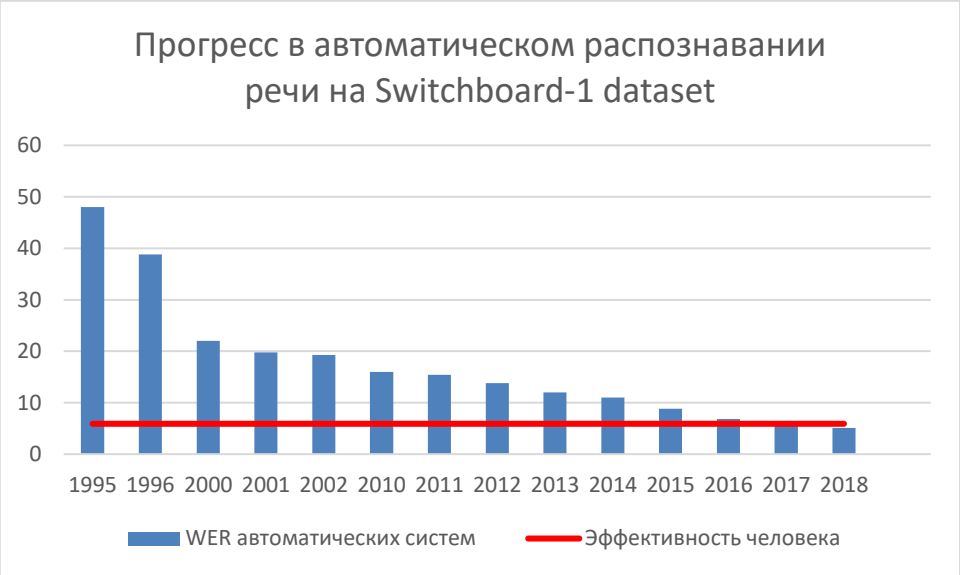

Рис. 1. *Повышение эффективности автоматического транскрибирования разговорной телефонной речи на материалах Switchboard-1 и эффективность человека при решении той же задачи*

Достижение «человеческого паритета» в автоматическом транскрибировании телефонных разговоров поднимает вопрос о том, что следует считать человеческой точностью в данной задаче.

3. Оценка человеческого уровня эффективности при решении задачи транскрипции разговорной речи

Оценка уровня человеческой точности при решении задачи транскрипции речи важна не только для определения эффективности технологий распознавания речи, но также необходима при решении проблем стенографирования, создания оценочных протоколов и установления эталонов, соответствующих отраслевому стандарту профессионального производства стенограмм. Кроме того, она представляет самостоятельный научный интерес в области психолингвистических исследований, корпусной фонетики и т.п.

3.1. Методы стенографирования

Ранее по материалам самой цитируемой работы [40] принято было считать, что эффективность человека при транскрибировании речи составляет 4%. Но поскольку методы оценки автор не уточнил, указав лишь, что данный показатель достигается в «личном общении», были инициированы детальные исследования этого вопроса. Наибольший вклад – при поддержке большого числа спонсоров – в исследование данной проблемы внесли специалисты лингвистического консорциума при университете Пенсильвании, спроектировав с позиции универсального подхода для разных языков целый набор стенографических методик для создания стенограмм разного назначения от приближенных для обработки

больших объемов информации до детально аннотированных для малых. Каждая методика нацелена на соблюдение баланса между точностью, эффективностью, ресурсоемкостью и требованиями оценочной программы. Наиболее применяемыми стали методы быстрой и точной транскрипции, а также расширенный метод быстрой транскрипции.

Метод быстрой транскрипции включает в себя транскрипцию автоматически воспроизводимых сегментов речи без соблюдения правил капитализации и пунктуации, но с разметкой определенного набора нелексем. Поскольку данный метод был разработан для создания обучающих корпусов и оценки, разработчики исходили из предположения о том, что большой объем данных позволяет взвесить возможные недочеты недостаточно точной транскрипции и ускорить стенографирование. Метод быстрой транскрипции позволяет завершить за час транскрибирование пяти часов английской речи [41].

Расширенный метод быстрой транскрипции был разработан по программе *DARPA GALE* для ручного стенографирования вещания на арабском и китайском языках для обучения и разработки речевых систем. Метод быстрой транскрипции с расширением включает в себя определение тематических пределов, ручное аннотирование групп слов, функционирующих в спонтанной речи как единое смысловое целое, а также идентификацию диалектов для арабской и мандаринской речи, где это необходимо. То есть стенограмма, выполненная расширенным методом, значительно обогащена структурной информацией.

Наибольших временных затрат и многочисленных проверок качества требует метод точной транскрипции, разработанный для определения эталонов. Элементами точной транскрипции являются соблюдение стандартной орфографии и пунктуации, идентификация дикторов, аннотирование выдыхательных групп, мены коммуникативных ролей, паузхезитаций, имен собственных и, если это возможно, диалектов.

3.2. Методы оценки человеческого уровня эффективности и анализ ошибок, допускаемых человеком при решении задачи транскрипции

В 2004 году *LDC* проанализировал межсубъектную устойчивость, исследовав точные транскрипции новостного вещания на английском языке (*BN*) и телефонной разговорной речи (*CTS*) на материалах *RT-03* тестовых данных. Записи точно транскрибировались и сравнивались с помощью программного обеспечения, разработанного *LDC*, который загружает две стенограммы и помечает области согласованности. При решении задачи стенографии новостного вещания процент расхождений между двумя стенографистами (*WDR*) составил 1.3%. Причем 81% разночтений обусловлен незначительными различиями в пунктуации. Оставшиеся несоответствия обусловлены неверной записью слов, сокращениями, нарушениями беглости речи или разночтениями по морфологическому статусу слова. *WDR* для разговорной телефонной речи достигает 4.1-4.5%; детальный анализ различий показал, что 95% «вызовов арбитра» соответствуют сокращениям, быстрой или осложненной речи или нарушениям беглости [41].

В 2008 году *LDC* провел исследование межсубъектной устойчивости на речевых данных, полученных из конференц-залов с высоким процентом диалогов. Сравнения показали, что 64% дуально транскрибированных сегментов имели некоторое количество несогласованностей, варьирующих от экстремального несовпадения, когда один стенограф понял диктора принципиально иначе, чем другой, до незначительного – в пунктуации [41].

Таким образом, исследования показали хорошее межсубъектное согласование на материалах англоязычного новостного вещания и телефонной разговорной речи с ошибками, не критичными для разработки системы.

Однако эти области не репрезентативны для всего спектра подходов *LDC*, жанров и языков. Дальнейшие расширенные исследования были нацелены на определение базовой линии человеческой устойчивости. Речевой материал дополнен записями английской,

арабской и мандаринской речи в жанрах новостного и разговорного вещания, интервью, разговорной телефонной речи, совещания. В большинстве случаев анализировались быстрая транскрипция и точная транскрипция для каждой комбинации языка и жанра, а также для оценки влияния метода транскрипции на межсубъектную согласованность.

Наиболее детально были проанализированы англоязычные стенограммы, выполненные методом быстрой транскрипции. Разночтения классифицировались по трем категориям: «ошибка стенографиста», «незначительные различия» и «вызов арбитра». Разночтения размечались как ошибки стенографиста в случае, когда стенографист опускал слово или высказывание, изменял порядок слов, вставлял слово или не понимал его. Около 15% разночтений, найденных в быстро транскрибированных англоязычных стенограммах, были признаны ошибками. 65% рассогласований принадлежат категории «незначительные» (различия в капитализации, пунктуации, аннотировании шумов, функциональных слов). 20% были размечены как «вызовы арбитра», когда аннотатор не может определить, кто из стенографистов прав, и имели место в областях с нарушением беглости, затрудненной или ускоренной речи.

Предварительные результаты для всех языков и жанров, как показывает таблица 1, согласуются с результатами исследований RT-03: стенограммы подготовленной речи в целом более устойчивы, чем стенограммы спонтанной. Показано также, что для большинства языков, за исключением новостного вещания на китайском, методы точной транскрипции приводят к более высоким показателям согласованности по сравнению с методами быстрой транскрипции. Подготовленная речь также транскрибируется с более высоким показателем межсубъектной согласованности, вне зависимости от метода транскрипции, чем речь спонтанная.

Таблица 1

ОЦЕНКИ ЧЕЛОВЕЧЕСКОЙ ЭФФЕКТИВНОСТИ ПРИ РЕШЕНИИ ЗАДАЧИ ТРАНСКРИПЦИИ РЕЧЕВЫХ СООБЩЕНИЙ РАЗНЫХ ЖАНРОВ ДЛЯ АНГЛИЙСКОГО, АРАБСКОГО И МАНДАРИНСКОГО ЯЗЫКОВ [41]

язык	***жанр***	***WDR (TT), %***	***WDR (PT), %***
английский	разговорная телефонная речь	4.1-4.5	9.63 (5 пар)
	совещание	-	6.23 (4 пары)
	интервью	n/a	3.84 (22 пары)
	новостное вещание	1.3	3.5 (6 пар)
	ТВ-вещание	n/a	6.3 (6 пар)
мандаринский	новостное вещание	7.40 (23 пары)	6.14 (18 пар)
	ТВ-вещание	9.06 (24 пары)	9.45 (4 пары)
арабский	новостное вещание	3.13 (14 пар)	3.42 (16 пар)
	ТВ-вещание	3.93 (12 пар)	8.27 (18 пар)

Собственные эксперименты по исследованию эффективности человека проводятся также специалистами *Microsoft* и *Appen (IBM)*. *Microsoft* использует двухпроходную транскрипцию [38]. На первом проходе стенографист транскрибирует сегменты диалога, соответствующие высказыванию, что, с одной стороны, облегчает задачу, поскольку говорящие более четко разделены, а с другой – усложняет, так как участники коммуникации не чередуются и контекст может отсутствовать. На втором проходе второй слушатель контролирует данные для исправления ошибок. Затем, применив инструменты оценки *NIST*, вычислили процент ошибок и получили 5,9% на части тестового набора данных *SWB* и 11,3% на части *CallHome (CH)* из тестового набора *NIST 2000*.

Сотрудники *Appen* [39] измерили человеческие ошибки на том же наборе данных, но с использованием более сложного процесса. Согласно выбранному протоколу над стенограммами независимо работали три специалиста. Их работа проверялась четвертым старшим стенографистом. Одним из принципиальных отличий при проведении эксперимента в *IBM* была осведомленность стенографистов о проведении эксперимента. Всего транскрибирующие выполняли от 12 до 18 прослушивающих проходов. Окончательный результат был получен на основе выбора стенографиста (с контролем качества) с самым низким *WER* по данным испытаний. Как отмечалось ранее, полученные оценки *WER* были равны 5.1% и 6.8% соответственно. Значительно более низкие оценки для *CH* могут быть следствием того, что расшифровщики имели доступ ко всему разговору, что дает возможность стенографисту «адаптироваться» к данным более эффективно.

Таблица 2

ЭФФЕКТИВНОСТЬ ЧЕЛОВЕКА ПРИ ТРАНСКРИБИРОВАНИИ РАЗГОВОРНОЙ ТЕЛЕФОННОЙ РЕЧИ НА МАТЕРИАЛАХ *SWITCHBOARD* И *CALLHOME* ПО РАЗНЫМ ОЦЕНКАМ

	WER человека на SWB	*WER человека на CH*
IBM	5.1	6.8
Microsoft	5.9	11.3
LDC	4.1-4.5 (ТТ) 9.63 (БТ)	-

Важно понимать, что разговорная речь имеет высокую степень неопределенности. Например, разговорное произношение высоко вариативно и часто сокращается. Еще одним источником двусмысленности являются отсутствие контекста и разные уровни осведомленности общающихся о теме разговора (особенно в случае *CH*). При наличии неопределенности уровень ошибок, характеризующий разночтения, может быть снижен путем согласованного устранения неоднозначности (обсуждения стенографистами в процессе работы), хотя это не обязательно отражает истинное согласие (отсутствие разночтений), основанное на понимании речи.

4. СИСТЕМА VS ЧЕЛОВЕК: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ОШИБОК

Согласно исследованиям, проведенными *NIST* с использованием тестирования *Wilcoxon* и *Matched Pairs Sentence Segment Word Error Rate*, проценты ошибок при решении задачи транскрипции системой и человеком на уровне высказывания разнятся незначительно. Но при этом для определения направления дальнейших исследований необходимо выяснить, являются ли сложности для автоматического транскрибера также и сложностями для человека. Исследования, приведенные в работе [38], показали, что на материалах разных корпусных подмножеств, каждый из которых насчитывает 40 дикторов, количество ошибок, допускаемых системой и человеком на уровне дикторов, коррелирует с коэффициентами 0.65 на материалах *SWB* и 0.73 на материалах *CH*. Таким образом, свойства речи как функции содержания, диктора или технических характеристик канала, вызывающие затруднения для автоматической системы, являются также затруднительными и для человека.

В типах ошибок, допускаемых человеком и машиной при решении задачи распознавания разговорной телефонной речи, также наблюдается усредненное сходство. Таблицы 3-5 показывают десять наиболее часто встречающихся ошибок (замен, пропусков и вставок), допускаемых автоматической системой и стенографистами при транскрибировании разговорной речи.

И для стенографистов, и для речевой системы ошибки типа «вставка» и «пропуск» очень сходны. В частности, одной из наиболее частотных ошибок, допускаемых стенографистами, является пропуск слова «я». Далее, стенографисты реже допускают ошибки-

замены, чаще ошибки-пропуски. Относительно высокий процент ошибок-пропусков, возможно, отражает склонность человека избегать вывода неопределенной информации или соответствовать требованиям производительности труда в профессии стенографиста. Во всех случаях число вставок относительно невелико.

Таблица 3

НАИБОЛЕЕ ЧАСТОТНЫЕ ОШИБКИ-ПРОПУСКИ, ДОПУСКАЕМЫЕ ЧЕЛОВЕКОМ И СИСТЕМОЙ (*КОЛИЧЕСТВО СЛУЧАЕВ ВОЗНИКНОВЕНИЯ ОШИБКИ: ПРОПУЩЕННОЕ СЛОВО) [38]

материалы CallHome		***материалы Switchboard***	
система	***человек***	***система***	***человек***
44: i	73: i	31: it	34: i
23: it	59: and	26: i	30: and
29: a	48: it	19: a	29: it
29: and	47: is	17: that	22: a
25: is	45: the	15: you	22: that
19: he	41: %bcack	13: and	22: you
18: are	37: a	12: have	17: the
17: oh	33: you	12: oh	17: to
17: that	31: oh	11: are	15: oh
17: the	30: that	11: is	15: yeah

Таблица 4

НАИБОЛЕЕ ЧАСТОТНЫЕ ОШИБКИ-ВСТАВКИ, ДОПУСКАЕМЫЕ ЧЕЛОВЕКОМ И СИСТЕМОЙ (*КОЛИЧЕСТВО СЛУЧАЕВ ВОЗНИКНОВЕНИЯ ОШИБКИ: ОШИБОЧНО ВСТАВЛЕННОЕ СЛОВО) [38]

материалы CallHome		***материалы Switchboard***	
система	***человек***	***система***	***человек***
15: a	10: i	19: i	12: i
15: is	9: and	9: and	11: and
11: i	8: a	7: of	9: you
11: the	8: that	6: do	8: is
11: you	8: the	6: is	6: they
9: it	7: have	5: but	5: do
7: oh	5: you	5: yeah	5: have
6: and	4: are	4: air	5: it
6: in	4: is	4: in	5: yeah
6: know	4: they	4: you	4: a

Анализ ошибок-замен показывает, что лидируют в данном классе ошибок короткие функциональные слова, дискурсивные маркеры и заполненные паузы. В частности, наиболее частотная ошибка замещения, допускаемая системой, заключается в неверном замещении междометий, заполняющих паузы («%hesitation», метка междометий «эээ» и «эм» в различных написаниях) междометиями, выражающими согласие с партнером по коммуникации («% bcack», метка «ээээээ», «м-м-м» и т. д.). Данная ошибка замещения гораздо реже встречается у человека. Указанные слова составляют наиболее частотный тип ошибки по многим причинам: они часто возникают в речи, часто реализуются в сокращенной форме. Кроме того, более затруднительны для системы вследствие фонетического сходства. Дополнительный вклад в частотность ошибки данного типа вносит неверная разметка обучающих данных. И, главное, данные междометия имеют противоположные прагматические функции в коммуникативных намерениях. Заполнение пауз служит для сигнала о желании либо начать, либо продолжить говорить. Междометия согласия, с дру-

гой стороны, показывают, что говорящий слушает и что другой оратор может продолжать. Разумеется, используя все имеющиеся фонетические, просодические и контекстные ключи при распознавании, люди хорошо осознают их различия. Система распознавания в отличие от человека использует только стандартные акустические и фонетические модели. Это свидетельствует о том, что для коррекции этого недостатка необходимо моделирование диалогового контекста.

Таблица 5

НАИБОЛЕЕ ЧАСТОТНЫЕ ОШИБКИ-ЗАМЕНЫ, ДОПУСКАЕМЫЕ СИСТЕМОЙ И ЧЕЛОВЕКОМ (*КОЛИЧЕСТВО СЛУЧАЕВ ВОЗНИКНОВЕНИЯ ОШИБКИ: СЛОВО/ОШИБОЧНАЯ ГИПОТЕЗА) [38]

материалы CallHome		***материалы Switchboard***	
система	***человек***	***система***	***человек***
45: (%hesitation)/(%bcack)*	12: a/the	29: (%hesitation)/(%bcack)	12: (%hesitation)/hmm
12: was/is	10: (%hesitation)/a	9: (%hesitation)/oh	10: (%hesitation)/oh
9: (%hesitation)/a	10: was/is	9: was/is	9: was/is
8: (%hesitation)/oh	7: (%hesitation)/hmm	8: and/in	8: (%hesitation)/a
8: a/the	7: bentsy/bensi	6: (%hesitation)/i	5: in/and
7: and/in	7: is/was	6: in/and	4: (%hesitation)/(%bcack)
7: it/that	6: could/can	5: (%hesitation)/a	4: and/in
6: in/and	6: well/oh	5: (%hesitation)/yeah	4: is/was
5: a/to	5: (%hesitation)/bcack	5: a/the	4: that/it
5: aw/oh	5: (%hesitation)/oh	5: jeeze/jeeze	4: the/a

5. СОВРЕМЕННЫЕ ПРОБЛЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Эффективность речевых систем была значительно улучшена с применением методов глубокого обучения. Человеческий уровень достигнут при решении задачи транскрипции телефонной разговорной речи на материалах *SWB,* но говорить о решении задачи распознавания спонтанной речи еще рано. Важно перейти от ограниченного распознавания речи к универсальным системам, устойчивым в осложненных условиях эксплуатации.

5.1. ПРОБЛЕМЫ РЕАЛИЗАЦИИ

Прорывные успехи в автоматическом транскрибировании разговорной телефонной речи связаны главным образом с применением двунаправленных рекуррентных сетей. Но данный алгоритм характеризуется довольно высоким временем задержки, поскольку не позволяет осуществить необходимые вычисления, прежде чем пользователь завершит высказывание. То есть время задержки определяется длиной высказывания. Требование в несколько десятков миллисекунд при решении данной задачи нельзя назвать чрезмерным, так как чаще всего распознавание речи является лишь первым этапом в целой серии вычислений. При этом поскольку алгоритмы, понижающие показатели времени задержки, повышают количество вычислительной мощности, необходимо учитывать целесообразность улучшения точности распознавания речи.

Вышесказанное не означает, что не следует проводить исследования по улучшению точности и оптимизации вычислительной мощности. Это означает, что способ эффективного использования глубокого обучения в распознавании речи остается нерешенной проблемой.

5.2. ПРОБЛЕМА УЧЕТА ОШИБОК

Как показывает сравнительный анализ ошибок, допускаемых автоматическими системами и человеком, ошибки людей при расшифровке речи менее критичны. Потому эффективность системы, оцениваемая как соотношение суммы чисел ошибок подстановок (S), удалений (D), вставок (I) к числу слов в высказывании

$$WER = \frac{S + D + I}{|W|} 100\%$$

(/*W*/ – число слов в последовательности *W*), не стоит использовать в качестве основного, так как нельзя назвать его объективным. Важнее оценивать количество высказываний, смысл которых был искажен, то есть оценить показатель семантических ошибок.

5.3. Проблема шумоподавления

Проблема распознавания речи в условиях сильного зашумления возникла в связи с необходимостью речевого управления и обеспечения надежной связи с диспетчером при использовании тяжелой техники. Сравнение эффективности человека и модели *Deep Speech 2* от *Baidu* при разных показателях *SNR* (*signal-to-noise ratio*) показало, что процент ошибок автоматического распознавания возрастает не менее чем в восемь раз [41]. Более того, при *SNR*, достигающем значения в пределах 0-6 дБ, вероятность ошибки при выявлении только факта речевой активности составляет 1%. Классические методы с использованием нескольких микрофонов не демонстрируют высокой надежности. Методы, использующие дополнительные неакустические источники информации, не находят широкого применения, так как возникает необходимость надевать гарнитуру с микрофонами и датчиками. Данная задача в распознавании речевых ответов также пока остается открытой.

5.4. Проблема учета акцентов

Сравнительный анализ стенограмм, выполненных человеком и сгенерированных системой *Deep Speech 2*, показал, что человек допускает больше ошибок при распознавании акцентной речи [41]. Одним из возможных объяснений этого может быть то, что в эксперименте участвовали носители американского английского. Жители соответствующих регионов, вероятно, справились бы с распознаванием акцентов родных стран успешнее. Но вне зависимости от признания данных результатов достаточно или недостаточно объективными создание речевой системы для английского языка с учетом вариатива американских акцентов требует создания речевого корпуса объемом в пять тысяч часов. И простое увеличение обучающих баз данных нельзя признать эффективным решением данной проблемы.

5.5. Автоматическая сегментация смешанного аудиосигнала

Разработки алгоритмов независимого от диктора автоматического сегментирования многоканальной речи ведутся специалистами в области речевых технологий в сотрудничестве с нейрофизиологами не только для создания интеллектуальных приложений, но и для создания слуховых аппаратов с функцией декодирования слухового внимания.

Классические методы формирования луча для усиления целевого звукового сообщения требуют использования нескольких микрофонов и применимы, когда существует достаточное пространственное разделение между источниками. Методы глубокого обучения более эффективны для решения задачи разделения источников, но ограничены «известными» дикторами, чьи голоса используются на этапе предварительного обучения, и неустойчивы в условиях эксплуатации, когда число говорящих увеличивается. Сеть глубоких аттракторов успешно выполняет разделение источников на основе проецирования частотно-временного представления смешанного аудиосигнала в многомерное пространство, где представления дикторов разделяются четче. Но данный алгоритм характеризуется высоким временем задержки. Перспективными представляются разработки по решению проблемы независимого сегментирования с использованием новой онлайн-реализации сети глубоких аттракторов, ведущиеся в Колумбийском университете [42]. Но пока задача универсального разделения речи остается одной из самых сложных проблем обработки речи.

5.6. Проблема учета контекста

Человек при решении задачи распознавания речи использует фонетические, просодические, визуальные и контекстные ключи. Автоматическая система распознавания, в отличие от человека, использует только структурный контекст в акустическом и языковом моделировании. Разработчики современных речевых систем используют дополнительную неречевую информацию. Например, сведения о геолокации пользователей при голосовом поиске в картах, доступ к списку контактов для распознавания имен собственных. Эффективность распознавания при этом, безусловно, возрастает. Но ни одна из существующих сегодня речевых систем не использует моделирование диалогового контекста. Исследования по изучению возможностей анализа и включения диалогового контекста только начинаются.

5.7. Проблема переноса на другие языки

Наибольшие успехи в автоматизации транскрипции достигнуты для языков германской группы, в первую очередь английского. С учетом различий между тоновыми и нетоновыми языками или агглютинативными и флективными и т.д., по сути, разработчики распознающих систем для языков разных групп сталкиваются с необходимостью решения новой задачи, поскольку построение надежной системы сводится не только к созданию достоверного акустического классификатора, но также требует создания эталонов, обучения наиболее часто употребляемым словам и разработки грамматической модели. Существующие методы не достигли достаточного уровня обобщения, и их применение для резко разнящихся языков сомнительно.

6. Заключение

Резюмируя вышеизложенное, можно сказать, что последние разработки в области автоматического распознавания речи очень успешны. Но, сообщая о достижении человеческого уровня эффективности, важно уточнить оценочные метрики, эталон человеческой эффективности и содержание тестовых материалов, на базе которых проводился сравнительный анализ, поскольку результат оценивания работы технологий распознавания речи во многом зависит от системы оценки. Проблемой также оказалась задача оценки уровня распознавания речи человеком и, разумеется, связанные проблемы стандартизации лингвистических описаний, согласования форматов представления информации лингвистических ресурсов разного назначения.

Итак, можно сказать, что уровень человеческой эффективности достигнут разработчиками *Microsoft* и *IBM* на тестовых материалах *Switchboard*, то есть при решении задачи транскрибирования разговорной телефонной речи, и составляет 5.8-5.1%.

Однако о решении задачи распознавания *спонтанной* разговорной телефонной речи все-таки пока еще рано говорить, хотя, по оценкам специалистов *Microsoft,* процент ошибок автоматической системы ниже человеческой эффективности при решении данной задачи и составляет 11.0% против 11.3%. Сотрудники *IBM* заявили о достижении автоматическими системами *WER* в 10.3%. Но по их собственным оценкам, человеческая эффективность при транскрибировании спонтанной телефонной речи между родственниками и близкими друзьями значительно ниже – 6.8%. Поскольку результаты экспериментов *IBM* по исследованию границ человеческой точности приблизились к результатам измерений разночтений *LDC* на других тестовых данных, возможно, стоит признать их более точными, пока не подведена черта в решении вопроса об уровне человеческой эффективности.

Сравнительный анализ наиболее частотных ошибок, допускаемых системой и человеком при решении задачи распознавания, показал в целом их сходство. Но ошибки, допускаемые человеком, менее критичны, реже искажают смысл высказывания. Таким образом, можно считать, что направление исследований верно, к сожалению, в автоматическом распознавании речи еще много открытых проблем, самыми сложными из которых являются проблемы шумоподавления и диаризации.

REFERENCES

1. Campbell M., Hoane A.J., Hsu F.-h. Deep Blue // Artificial intelligence. 2002. Vol. 134. Pp. 57-83.

2. Silver D., Huang A., Maddison C. J.,Guez A., Sifre L., Van Den Driessche G., Schrittwieser J., Antonoglou I., Panneershelvam V., Lanctot M., et al. Mastering the game of Go with deep neural networks and tree search // Nature. 2016. Vol. 529. Pp. 484-489.

3. Amodei D., Anubhai R., Battenberg E., Case C., Casper J., Catanzaro B., Hen J., Chrzanowski M., Coates A., Diamos G., et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin // arXiv preprint arXiv:1512.02595. 2015.

4. Kristjansson T.T., Hershey J.R., Olsen P.A., Rennie S.J., Gopinath R.A. Super-human multi-talker speech recognition: the IBM 2006 Speech Separation Challenge system // Proc. Interspeech. 2006. Vol. 12. P. 155.

5. Weng C., Yu D., Seltzer M. L., Droppo J. Single-channel mixed speech recognition using deep neural networks // Proc. IEEE ICASSP. 2014. Pp. 5632-5636.

6. Pallett D.S. A look at NIST's benchmark ASR tests: past, present, and future // IEEE Automatic Speech Recognition and Understanding Workshop. 2003. Pp. 483-488.

7. Price P., Fisher W.M., Bernstein J., Pallett D.S. The DARPA 1000-word resource management database for continuous speech recognition // Proc. IEEE ICASSP. 1988. pp. 651–654.

8. Paul D.B., Baker J.M. The design for the wall street journal-based csr corpus // Proceedings of the workshop on Speech and Natural Language. 1992. Pp. 357-362.

9. *Graff D., Wu Z., MacIntyre R., Liberman M.* The 1996 broadcast news speech and language-model corpus // Proceedings of the DARPA Workshop on Spoken Language technology. 1997. Pp. 11-14.

10. Ljolje A. The AT&T 2001 LVCSR system // NIST LVCSR Workshop. 2001.

11. Philppov D. *Interaktivnoye golosovoye redaktirovaniye teksta s pomoshch'yu novykh rechevykh tekhnologiy ot Yandeksa* [Interactive Voice Text Editing Using New Speech Technologies from Yandex]. https://habr.com/ru/company/yandex/blog/243813/. 2014.

12. Chen S.F., Kingsbury B., Mangu L., Povey D., Saon G., Soltau H., Zweig G. Advances in speech transcription at IBM under the DARPA EARS program // IEEE Trans. Audio, Speech, and Language Processing. 2006. Vol. 14. Pp. 1596-1608.

13. Seide F., Li G., Yu D. Conversational speech transcription using context-dependent deep neural networks // Proc. Interspeech. 2011. Pp. 437-440.

14. Matsoukas S., Gauvain J.-L., Adda G., Colthurst T., Kao C.-L., Kimball O., Lamel L., Lefevre F., Ma J.Z., Makhoul J., et al. Advances in transcription of broadcast news and conversational telephone speech within the combined ears bbn/limsi system // IEEE Transactions on Audio, Speech, and Language Processing. 2006. Vol. 14. Pp. 1541-1556.

15. Stolcke A., Chen B., Franco H., Gadde V. R. R., Graciarena M., Hwang M.-Y., Kirchhoff K., Mandal A., Morgan N., Lei X., et al. Recent innovations in speech-to-text transcription at SRI-ICSI-UW // IEEE Transactions on Audio, Speech, and Language Processing. 2006. Vol. 14. Pp. 1729-1744.

16. Gauvain J.-L., Lamel L., Schwenk H., Adda G., Chen L., Lefevre F. Conversational telephone speech recognition // Proc. IEEE ICASSP. 2003. Vol. 1. Pp. 1-212.

17. Evermann G., Chan H. Y., Gales M. J. F., Hain T., Liu X., Mrva D., Wang L., Woodland P.C. Development of the 2003 CU-HTK conversational telephone speech transcription system // Proc. IEEE ICASSP. 2004. Vol. 1. Pp. 1-249. 2004.

18. Fry D.B. Theoretical aspects of mechanical speech recognition // J. British Inst. Radio Engr. 1959. Pp. 211-229.

19. Vintsyuk T.K. *Raspoznavaniye slov ustnoy rechi metodami dinamicheskogo programmirovaniya* [Speech discrimination by dynamic programming] // Kibernetika. 1968. 4 (2). Pp. 81- 88.

20. Viterbi A.J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm // IEEE Trans. Information Theory, IT- 13. 1967. Pp. 260-269.

21. Reddy D.R. An approach to computer speech recognition by direct analysis of the speech wave // Tech. Report No. C549, Computer Science Dept., Stanford Univ. 1966.

22. Velichko V.M., Zagoruyko N.G. *Avtomaticheskoye raspoznavaniye ogranichennogo nabora ustnykh komand* [Automatic recognition of 200 words] // Int. J. Man- Machine Studies. 1970. 2. Pp. 223.

23. Sakoe H., Chiba S. Dynamic programming algorithm optimization for spoken word recognition // IEEE Trans. Acoustics, Speech, Signal Proc, ASSP-26 (1). 1978. Pp. 43-49.

24. Rabiner L. R., et. al. Speaker independent recognition of isolated words using clustering techniques // IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27. 1979. Pp. 336-349.

25. Klatt D. Review of the ARPA speech understanding project // J.A.S.A. 1977. 62(6). Pp. 1324-1366.

26. Lowerre B. The HARPY speech understanding system // Trends in Speech Recognition, W. Lea, Ed., Speech Science Pub. 1990. Pp. 576-586.

27. Rabiner L. R., Juang B. H. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliff. New Jersey. 1993.

28. Katagiri S. Speech pattern recognition using neural networks // W. Chou and B.-H. Juang (Eds.) Pattern Recognition in Speech and Language Processing, CRC Press. 2003. Pp. 115-147.

29. Myers C.S., Rabiner L.R. A level building dynamic time warping algorithm for connected word recognition // IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-29. 1981. Pp. 284-297.

30. Lee C.H., Rabiner L.R. A frame synchronous network search algorithm for connected word recognition // IEEE Trans. Acoustics, Speech, Signal Proc. 1989. 37 (11). Pp. 1649-1658.

31. Bridle J.S., Brown M.D. Connected word recognition using whole word templates // Proc. Inst. Acoust. Autumn Conf. 1979. Pp. 25-28.

32. Juang B.-H., Furui S. Automatic speech recognition and understanding: A first step toward natural human-machine communication // Proc. IEEE, 88, 8. 2000. Pp. 1142-1165.

33. Chou W. Mimimum classification error (MCE) approach in pattern recognition // Chou W., Juang B.-H. (Eds.) Pattern Recognition in Speech and Language Processing. CRC Press. 2003. Pp. 1-49.

34. Leggetter C.J., Woodland P.C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models // Computer Speech and Language. 1995. 9. Pp. 171-185.

35. Varga A.P., Moore R.K. Hidden Markov model decomposition of speech and noise // Proc. ICASSP. 1990. Pp. 845-848.

36. Gales M. J. F., Young S.J. Parallel model combination for speech recognition in noise // Technical Report, CUED/FINFENG/ TR135. 1993.

37. Shinoda K., Lee C.H. A structural Bayes approach to speaker adaptation // IEEE Trans. Speech and Audio Proc. 2001. 9, 3. Pp. 276-287.

38. Stolcke A., Droppo. J. Comparing Human and Machine Errors in Conversational Speech Transcription. /Interspeech. 2017. Pp. 137-141. DOI: 10.21437.

39. Saon G., Kurata G., Sercu T., Audhkhasi K.,Thomas S., Dimitriadis D., Cui X., Ramabhadran B., Picheny M., Lim L.-L., Roomi B., Hall P. English Conversational Telephone Speech Recognition by Humans and Machines // INTERSPEECH 2017 DOI: 10.21437.

40. Lippmann R.P. Speech recognition by machines and humans // Speech Communication. 1997. Vol. 22. Issue 1. Pages 1-15. https://doi.org/10.1016/S0167-6393(97)00021-6.

41. Glenn M. L., Strassel S. M.,H. Lee, Maeda K., Zakhary R., Li X. Transcription Methods for Consistency, Volume and Efficiency // Proceedings of the International Conference on Language Resources and Evaluation, LREC. 2010. Malta.

42. Hannun A. Speech Recognition Is Not Solved. https://awni.github.io/speech-recognition. 2017.

43. Han C., O'Sullivan J., Luo Y., Herrero J., Mehta A.D., Mesgarani N. Speaker-independent auditory attention decoding without access to clean speech sources // Sci Adv. 2019 May 15;5(5):eaav6134. doi: 10.1126/sciadv.aav6134. PMID: 31106271; PMCID: PMC6520028.

MODERN PROBLEMS OF AUTOMATIC SPEECH RECOGNITION*

I.A. GURTUEVA

Institute of Computer Science and Problems of Regional Management –
Branch of Federal public budgetary scientific establishment «Federal scientific center
«Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences»
360000, KBR, Nalchik, 37-a, I. Armand St.
E-mail: iipru@rambler.ru

This paper provides a concise review of the most applied methods in speech recognition. Various principles of transcription developed in the Linguistic Data Consortium are discussed. The problems in evaluating the human level of efficiency in solving the problem of speech recognition are described. The typical errors made by a human are analyzed. It has been shown that transcribers demonstrate a high level of consistency with accurate transcription of pre-prepared English speech and fast transcription of conversational telephone speech. It is also shown that with increasing complexity of speech, the word disagreement rate increases. The results of a comparative analysis of errors generated by the speech system and those made by humans are presented. Their similarities and differences are analyzed. The modern automatic speech recognition problems are listed, the prospects for their solution and the directions of future research are estimated.

Keywords: deep learning, artificial intelligence, artificial neuron networks, speech recognition, human parity.

Работа поступила 30.11.2020 г.

Сведения об авторе:

Гуртуева Ирина Асланбековна, н.с. отдела «Компьютерная лингвистика» Института информатики и проблем регионального управления – филиала Кабардино-Балкарского научного центра РАН.
360000, КБР, г. Нальчик, ул. И. Арманд, 37-а.
E-mail: gurtueva-i@yandex.ru

Information about the author:

Gurtueva Irina Aslanbekovna, researcher of the Department of Computer Linguistics of the Institute of Computer Science and Problems of Regional Management of KBSC of the Russian Academy of Sciences.
360000, KBR, Nalchik, I. Armand street, 37-a.
E-mail: gurtueva-i@yandex.ru

* The work was carried out with the financial support of the RFBR grants No. No 18-01-00658, 19-01-00648