Направление 210200

ПРОЕКТИРОВАНИЕ И ТЕХНОЛОГИЯ ЭЛЕКТРОННЫХ СРЕДСТВ

Магистерская программа 210200.05

Информационные технологии проектирования электронных средств

Руководитель программы д.т.н., проф. Муромцев Ю. Л.

Караульных Д. В.

ТЕХНОЛОГИИ РАСПОЗНАВАНИЯ РЕЧИ

Работа выполнена под руководством к.т.н.Орлова В. В.

ТГТУ, Кафедра «Конструирование радиоэлектронных и микропроцессорных систем»

Переход от графического пользовательского интерфейса (GUI) к свободному общению с компьютером представляется вполне естественным по причине того, что большинству из нас проще высказать свою мысль вслух, чем записать ее, кроме того, современные компьютерные интерфейсы (GUI и командная строка) задают пользователю жесткие ограничения, которых можно было бы избежать, научив компьютер понимать нашу речь.

Для распознавания необходимо записать человеческую речь, автоматически разобрать ее на минимальные составляющие, свериться с

базой сэмплов, подставить нужные фонемы, а потом собрать из фонем слова, расставив знаки препинания на основе анализа просодических эффектов.

Но на практике человеческая речь — понятие не точное и наш мозг постоянно выполняет сложнейшую работу по распознаванию образов. Разработчикам систем распознавания речи пришлось столкнуться с тем, что люди в массе своей говорят неразборчиво и не всегда в приемлемых шумовых условиях. Кроме того:

- Язык (а разговорный язык — тем более) не является постоянным. В большинстве языков имеется несколько диалектов, и даже в рамках одного диалекта существует несколько равноправных вариантов произношения одного и того же слова.

- У каждого из нас есть свои речевые особенности, которые могут затруднить распознавание речи.

- В естественной речи содержатся звуки и слова-паразиты («эээ», «ммм», «как-бы» и т.д.), которые необходимо отфильтровывать.

- В реальной жизни редко встречаются идеальные условия для записи звука: как правило, запись речи сопровождается шумами разной природы, которые мешают выделить голосовой сигнал для дальнейшей обработки.

Кроме этого сюда добавляются эффекты коартикуляции (а их правильная обработка в задаче распознавания речи куда критичнее, чем в задаче синтеза; если там мы рискуем лишь тем, что слово будет звучать ненатурально, то здесь в результате неправильной обработки коартикуляции система не найдет нужное слово в словаре.

Существующие технологии не позволяют решить эти проблемы в комплексе. Поэтому в зависимости от поставленной задачи техники распознавания речи меняются. Однако у них есть много общего (например, почти все современные системы распознавания речи используют для поиска нужных фонем скрытые модели Маркова [1]).

Долгое время системы распознавания требовали, чтобы пользователь выговаривал каждое слово отдельно, однако появились пакеты, умеющие обрабатывать так называемую слитную речь. Но системе попрежнему требуется время на обработку услышанного, и гораздо эффективнее выдавать ей законченные предложения (если они короткие) или более-менее самостоятельные фрагменты предложений. Во многих современных пакетах распознавания есть синтаксические и семантические модули, и подобная разбивка облегчит распознавание, одновременно улучшив качество. Иными словами, «слитная речь» в данном случае является синонимом диктовки.

Другой важный критерий — привязка к пользователю. На самом деле практически все современные системы распознавания речи являют-

ся обучаемыми. Разница только в том, что дикторо-независимую систему обучил производитель, заложив в неё сотни, или тысячи примеров. Поскольку у таких систем — при прочих равных условиях — требования к компьютерным ресурсам намного выше, а производительность хуже, то на потребительском рынке большей популярностью пользуются системы, которые пользователь после покупки подгоняет «под себя». Тем не менее, приложений, для которых важна именно независимость от пользователя, более чем достаточно — автоматические корпоративные колл-центры, например, должны быть универсальны.

Третий критерий – размер словаря. Чем меньше словарь, тем проще обучить систему и сделать ее дикторонезависимой. Единственное исключение из этого правила – голосовой набор в мобильных телефонах. Эта система снабжена очень маленьким словарем — но, с другой стороны, и системные требования у нее крайне скромны, раз она работает на мобильном телефоне.

Если разработчики, занимающиеся синтезом речи, начинали с копирования человеческого голосового аппарата и только потом разработали систему компилятивного синтеза, «собирающую» нужные слова из обрывков фонем, то системы распознавания речи имеют мало общего с тем, как распознает речь человеческий мозг. Скрытые модели Маркова, которые стали применять для распознавания в 1970-е гг., оказались эффективным средством для поиска нужных фонем, но они не являются панацеей и не способны решить все проблемы распознавания речи. У современной науки весьма неясные представления о глубинных процессах, отвечающих за распознавание речи в нашем мозге, так что делать какие-то выводы о качестве систем распознавания мы можем лишь потому, что есть задачи, которые им совсем не под силу:

- Они не умеют автоматически распознавать язык диктора. Любой человек, хоть раз слышавший итальянскую речь, скорее всего, узнает ее, услышав снова (при этом он может не иметь ни малейшего представления о самом языке). Машина так не умеет, она применяет заложенную в нее языковую модель, независимо от того, на каком языке с ней говорит человек.

- Они не умеют выделять речь по-настоящему. Качество распознавания в шумном окружении падает чуть ли не вдвое. Главным средством борьбы с шумами являются механизмы подавления, которые эффективны далеко не всегда. Сосредоточиться на речи собеседника, отсеяв все остальные звуки, как необязательные для распознавания, и уж тем более выделить речь одного человека из диалога система не может;

- Распознают они не очень хорошо. Человек легко поймет общий смысл сказанного и большинство слов, даже если у собеседника очень

сильный акцент. В то же время система распознавания, выполняющая сравнение элементов фонем в этом случае будет давать сбои.

И, наконец, самое главное. Хотя при распознавании используются элементы синтаксического и семантического анализа, нужно признать, что машины из того, что мы им говорим, ничего не понимают [2].

Основным подходом к проблеме распознавания речи в настоящее время является ИМЗ-подход. Он базируется на иерархическом (И) принципе обработки информации и на использовании многозначных решений (МЗ) на всех уровнях этой обработки. Опыт исследований показывает, что для достижения приемлемой для практики надежности распознавания речи требуется решение проблемных задач на всех уровнях. А это требует больших затрат и времени. Поэтому выдвигается ряд промежуточных, но важных для практики задач:

1. распознавание отдельно произносимых слов;
2. выделение ключевых слов в потоке речи;
3. распознавание слитной речи, составленной из слов заданного словаря.

Оказалось, однако, что и решение перечисленных задач для произвольного диктора или неограниченного словаря требует серьезных усилий и остается еще целый ряд принципиальных вопросов, требующих глубокой проработки [3].

Основной техникой для многих систем распознавания речи является статистический метод, называемый скрытым марковским моделированием. Такие системы разрабатываются во многих центрах и способны на "хорошее распознавание слов речи, не используя тренировку распознавания акустической речи". Данный результат был получен тестированием системы на данных, полученных из министерства обороны США, содержащих записи тысяч телефонных переговоров. В масштабах ограниченного тестирования вероятность правильно обнаруженных 22 ключевых слов варьировалась от 45 до 60% при условии допущения 10 ложных положительных результатов на ключевой слово в час. Таким образом, если 1000 ключевых слов было использовано во время часового переговора, будет, по крайней мере, 300 пропущенных ключевых слов и 220 ложных обнаружений. Другими словами качество распознавания не позволяет практически использовать систему.

Одно из применений системы распознавания речи – автоматический анализ и обнаружение в телефонных переговорах ключевых слов, которые могут отнести произносящего их человека к преступникам или террористам, что становится в наше время особенно актуально. Из-за отсутствия привязки к говорящему и нечеткости выделенного сигнала из перехваченных телефонных переговоров, скорее всего даже лучшие алгоритмы и быстрые процессоры, чем используемые сейчас,

будут давать худшие результаты, чем получаемые в современных хорошо обученных системах [4].

Из вышесказанного можно сделать выводы:

- современные системы распознавания не являются полностью дикторонезависимыми, и их приходится настраивать под конкретного человека;
- для качественной работы требуются большие вычислительные мощности, поэтому эти системы нельзя встраивать в различные устройства;
- качество распознавания не является высоким.

Для устранения перечисленных недостатков необходима разработка новых алгоритмов, основанных на анализе процессов распознавания речи человеком, а также проектирование новых аппаратных средств, предназначенных для этой цели.

Список литературы

1. http://www.cs.berkeley.edu/~murphyk/Bayes/rabiner.pdf
2. Таран О., Мирошниченко С., Гуриев В. Ничего никому не скажу//Компьютерра-2005.- №36.-C&C Computer Publishing Limited.-78 с.
3. http://www.agentura.ru/equipment/radio/nepr/
4. http://impb.psn.ru/~sychyov/html/sound00.shtml