Современные проблемы в области распознавания речи.

*** Галунов В.И., ** Соловьев А.Н.**

* Санкт-Петербургский Государственный Университет, ** AudiTech Ltd. (Санкт-Петербург)

В настоящее время научное сообщество вкладывает гигантское количество денег в развитие know-how и научно-исследовательские разработки для решения проблем автоматического распознавания и понимания речи. Это стимулируется практическими требованиями, связанными с созданием системы военного и коммерческого назначения. Не касаясь первого из них, можно указать, что только в европейском сообществе объем продаж систем гражданского назначения составляет несколько миллиардов долларов. При этом следует обратить внимание на то, что в практическом использовании отсутствуют системы, считающиеся по непонятным причинам вершиной развития систем автоматического распознавания речи. Это системы, которые можно назвать демонстрационными [1] и которые 50 лет назад назывались «фонетическими печатающими машинками». Их целью является перевод речи в соответствующий письменный текст.

Если рассматривать классическую схему «наука-технологии-практические системы» [2], то, прежде всего, надо определить те условия, в которых будет работать практическая система автоматического распознавания или понимания речи. Наиболее серьезные проблемы возникают при условиях:

- произвольный, наивный пользователь;
- спонтанная речь, сопровождаемая аграмматизмами и речевым «мусором»;
- наличие акустических помех и искажений, в том числе меняющихся;
- наличие речевых помех.

С другой стороны необходимо определить важность задачи, ее научную и прикладную фундаментальность, связь с другими областями знаний. При этом необходимо учитывать состояние научного-промышленного потенциала, его возможности. Ни для кого не секрет, что правильно поставленная задача – это уже половина решения.

В настоящее время в среде «речевиков» (а тем более «неречевиков») сложилось представление, что конечной и высшей целью является создание именно «фонетической печатающей машинки», а универсальным методом решения всех речевых проблем являются скрытые Марковские модели (HMM).

Постараемся проанализировать эти положения.

Первое – преобразование речи в текст.

Оставим за кадром прикладную необходимость «диктографа» для пользователей. Остановимся на возможностях и недостатках соответствующих систем автоматического распознавания речи (анонсируемые сегодня возможностью распознавания сотен и даже тысяч слов с надежностью до 98%; как это проверялось, и как это вообще возможно сделать – не ясно).

1. От пользователя требуется предварительная настройка системы на его голос от нескольких десятков минут до нескольких часов предварительного наговаривания текстов.
2. Некоторые проверки (см. например, [3]) не дают результатов лучших, чем 5% ошибок.
3. Так как слова, включенные даже в хорошо и аккуратно произносимый текст, оказываются как бы плавающими в океане омонимии, то количество ошибок (словесных) возрастает приблизительно в 5 раз. Беглое отслеживание таких ошибок, кроме случаев возникновения нелепых текстов, уже затруднительно. Аппарат коррекции ошибок в большинстве демонстрационных систем слабо отлажен.
4. Были упоминания, что даже для хорошо организованных спонтанно произнесенных текстов вероятность правильного распознавания слов не превышает одной трети.
5. Наконец, время обработки введенного отрезка речи в таких системах может занимать минуты.

Все сказанное говорит о том, что в качестве конечной цели предлагаемые демонстрационные системы «речь-текст» вряд ли представляют интерес. Это не исключает возможности использования их в качестве полигона для оценки научных идей, но в этом случае должны отчетливо излагаться те модели, которые закладываются в данные системы автоматического распознавания и каким образом должна проверяться их практическая перспективность. Таким образом, мы переходим на противоположный конец триады «практические системы – речевые технологии – речевая наука».

Второе – скрытые Марковские модели.

Рассмотрим с этой точки зрения тот основной метод, который используется в большинстве современных систем автоматического распознавания – метод скрытых

Марковских моделей. Сама вероятностная модель этого вида была предложена А.А.Марковым в 1913 г. для анализа письменных текстов [4] и прекрасно себя зарекомендовала в этой области [5]. С 70-х годов начались работы по адаптации этой модели к автоматическому распознаванию речи (см. классические работы [6, 7]). При этом мы никаким образом не может уйти от двух вопросов: на базе каких параметров вести анализ и сколько этих скрытых «сегментов» в речевом сообщении.

Если рассмотреть признаки, выбираемые для описания речевых сигналов, то можно обнаружить, что с этой целью практически в произвольной последовательности используются все те признаки, которые опробовались в речевых исследованиях: уровни в спектральных полосах, те или иные формантные признаки, коэффициенты линейного прогноза, кепстральные признаки и т.д. Ни один из возможных наборов не дает явного преимущества, и результат практически зависит от аккуратного набора статистики. С выбором сегментов, для которых строится Марковская модель, ситуация складывается аналогичным образом. В первоначальных вариантах модели предполагалось, что соответствующие сегменты могут быть привязаны к фонемам, аллофонами или каким-либо другим фонетически оправданным элементам. Далее оказалось, как это было обнаружено еще на модели «динамического программирования», что все сегменты речевого потока практически одинаково информационны для распознавания крупных речевых единиц (слов, фраз). То есть сегменты, входящие в Марковскую модель практически лишились какой-либо лингвистической привязки и превратились в некоторые единицы, имеющие только вероятностный смысл.

Таким образом, построение систем автоматического распознавания речи на основе скрытой Марковской модели предполагает вероятностную организацию речевого поведения человека. Но такое предположение не является очевидным. Попытаемся рассмотреть речь как систему [8] и выделить те факторы, которые являются основными в ее формировании. Первым фактором и наиболее впечатляющим в речевой системе является ее продуктивность, т.е. возможность продуцировать сколь угодно большое количество информационных сообщений, обладающих разным смыслом. Эта особенность речи сохраняется в том культурном ее варианте, который называется письменной речью. Причем это основное отличие речи от коммуникационных систем у животных. Однако следует отметить еще один системообразующий фактор, который не присутствует в письменной речи, но является необходимым для реализации речевой системы в ее акустическом варианте. В первом приближении этот фактор можно назвать помехозащищенностью, т.е. первая его задача – обеспечить сохранение смысловой информации при различных вариантах акустических помех и искажений. Кроме того,

должна обеспечиваться достаточная точность передачи смысловой информации при различных вариантах нарушений (не только патологических, но чаще всего ситуационных) процессов речеобразования и речевосприятия. В отличие от письменной речи в обычном устном варианте нет возможности вернуться к началу текста и попытаться его заново осмыслить. Помехозащищенность обеспечивается целым рядом механизмов. Для систем автоматического распознавания речи следует указать, прежде всего, на использование двух механизмов:

1. Использование нескольких параллельно работающих способов выделения одних и тех же элементов речевого сигнала на базе анализа акустического сигнала. Примером может служить параллельное использование формантных и полосных признаков для идентификации фонетических элементов речевой структуры [9].
2. Параллельное независимое использование сегментного (фонемного) и целостного восприятия слов в потоке речи.

Самым существенным в сказанных представлениях о механизмах обеспечения помехоустойчивости является не то, что системы признаков задублированы и находятся на разных уровнях описания структуры речевого сигнала, а то, что в каждой отдельной реализации данного слова или фразы не обязательно проявление всего возможного набора признаков. Реализуемый в конкретной ситуации набор признаков будет определяться прагматическим, семантическим, помеховым и другим контекстами. Это сразу определяет невозможность использования вероятностных моделей для распознавания речевых сообщений. Это, конечно, не исключает использования вероятностных методов на низшем признаковом уровне.

Таким образом, первая и наиболее важная проблема в идеологии автоматического распознавания речи: какова должна быть общая модель распознавания, если отказаться от модной, но явно непродуктивной вероятностной модели.

Сказанное, конечно, не исключает возможности использования любой, даже самой простой модели для решения частных локальных задач распознавания речи. Таких задач очень много. Но если мы хотим решать глобальные задачи человеко-машинного речевого взаимодействия мы должны четко представлять механизмы, задействованные в речевой системе в целом.

Естественным представляется использование для построения систем автоматического распознавания моделей восприятия речи. В научном обиходе в настоящее время их реализовано довольно много. Рассмотрение этих моделей обнаруживает еще две проблемы.

Проблема номер два – каков принцип выбора первичного описания речевого сигнала? Здесь можно выделить три варианта подхода. Первый из них практически совпадает с принятым в большинстве систем автоматического распознавания и основан на статистическом анализе различных речевых акустических параметров [10]. Второй подход предполагает, что для распознавания речи необходим переход от акустических параметров к артикуляциям, которые лежали в основе порождения данного акустического сигнала. Это моторная теория [11] и теория прямого реализма [12]. Очевидно, что такая задача при нынешних уровнях нашего понимания механизмов речеобразования вряд ли разрешима. Третий подход, так называемая квантовая теория [13], представляется весьма перспективной. В этой теории акустические признаки делятся на 2 категории. Первый тип акустических признаков соответствует резкому изменению акустического сигнала при небольшом изменении артикуляционного тракта, второй тип синхронно плавному изменению сигнала с изменением артикуляции.

Третья проблема связана с тем, что если мы отказывается от простой линейной модели речевого сигнала, то становится не ясно, как должны взаимодействовать первичные признаки с другими речевыми уровнями: вербальным, семантическим, прагматическим, вероятностным и др.

Здесь мы подходим к еще одной совершенно темной области в системах распознавания речи, т.е. к верхним уровням распознавания – семантике и прагматике.

В современных системах распознавания речи задача понимания смысла чаще всего решается методом «снизу-вверх», т.е. сначала происходит распознавание речевых сегментов, а затем все распознанное поступает на семантический модуль. Как правило, сигнал на входе семантического блока представляет собой матрицу, составленную из векторов вероятности распознавания каждого сегмента потока речи, который соответствует при удачной сегментации какому-либо слову или словоформе. Дальнейшая работа семантического блока предполагает построение из этих векторов вероятности списка осмысленных предложений, ограниченных заданным порогом минимальной вероятности [14, 15]. Естественная речь зачастую аграмматична и практически сложно применить грамматику для построения высказывания, учитывая еще и то, что падежные окончания во флексивных языках чаще всего «заглатываются», т.е. не проговариваются достаточно четко. Поэтому используют другие разнообразные «улучшители» понимания как, например, учет предыстории, выявление контекста и падежно-ролевых отношений или использование различных статистически-вероятностных методов (частотности, ассоциативности и пр.). Как правило, на данном этапе используется обратная связь

семантического модуля с модулем распознавания: список поиска вероятных слов при распознавании пополняется ассоциативной лексикой с последующим пересчетом векторов вероятности. Повторяя цикл можно достичь более высокий процент правильного понимания смысла.

Особыми проблемами при таком подходе является, как уже указывалось, омонимия и так называемый «мусор» – слова, которых нет в словаре распознавания, а так же различного рода помехи как речевого, так и неречевого типа. Если степень омонимии можно уменьшить, выявляя и запоминая контекст сообщения, то проблема «мусора» не имеет простого решения, поскольку здесь помимо внешних помех необходимо выявлять и учитывать индивидуальные характеристики говорящего (хезитации, употребление эмотивных лексических элементов).

Возможен и другой, обратный подход к пониманию речи: подстройка модуля распознавания до обработки входного сигнала. Данный подход будет оправдан при использовании неинформационных функций коммуникации, где главной задачей является не передача информации как таковой, а выявление общей «семиотической ситуации». В этом случае важно не то, каким способом или словами выражается мысль, а то, что у адресата появится именно то состояние, которое будет индуцировать понимание (смысл). Этот подход видится более продуктивным, но для решения этой задачи необходимо учитывать не только весь семиозис, окружающий диалог, но и наличие у распознающей системы интеллекта – суммы знаний об окружающем мире и способность к анализу новой информации. Поэтому главной проблемой при таком подходе является правильное построение базы знаний, предварительное обучение системы, способность к адекватному анализу окружающей действительности.

Примитивным способом использования второго способа распознавания могут служить фреймовые семантические модели, когда адресант сам принудительно выбирает «тему разговора», и при этом подключается та или иная модель диалога с соответствующей лексикой и грамматикой [15, 16]. Распознавание ограничивается поиском ключевых слов, в соответствие которым из ограниченного числа «смыслов» подыскивается наилучший.

И наконец еще одна нерешенная проблема, которое в теории восприятия речи носит название cocktail-party эффект или более расширенно – анализ акустических сцен. Этот эффект основан на способности слушающего сосредоточить внимание на выделенном источнике звука в условиях сильной зашумленности. В настоящий момент в автоматическом распознавании речи разрабатываются в основном методы подавления

сравнительно гладких помех и искажений. Хотя помехоустойчивость систем распознавания считается одним из основных направлений в области создания систем практического направления [18], заметных прорывов в этой области не наблюдается. Возможно и здесь было бы интересно рассмотреть те модели, которые рассматриваются при анализе слухового восприятия [19, 20].

Список цитируемой литературы:

1. V.Galunov, V.Taubkin, 1999. Speech technologies and speech science. SpeeCom'99, p.10-13.
2. V.I.Galunov, G.V.Galunov, 2001. Science perspectives of speech technology. SpeeCom 2001, 143-145.
3. В.И.Галунов, В.И.Гарбарук, 2001. Акустическая теория речеобразования и система фонетических признаков. Материалы международного конгресса 100 лет экспериментальной фонетике в России, с.58-62.
4. А.А.Марков, 1916. Об одном применении статистического метода. Известия АН, сер.6, X, №4, 239.
5. F.Elinek, 1976. Распознавание непрерывной речи статистическими методами. ТИИЭР 64, №4, с.131-160.
6. F.Elinek, 1985. Разработка экспериментального устройства, распознающего раздельно произнесенные слова. ТИИЭР 73, №11, с.91-99.
7. В.И.Галунов, 2002. Помехоустойчивость как системообразующий фактор речи. Проблемы и методы экспериментально-фонетических исследований, с.205-300.
8. В.И.Галунов, 2003. Речь как система. Труды XIII сессии РАО, т.3, с.19-21.
9. D.Kraft, 1979. Speech perception. J.Phonetics, 7, p.279-312.
10. P.K.Kuhl, P.Inverson, 1995. Linguistic experience and the "perceptual magnet effect". In W.Strange (Ed). Speech perception and linguistic experience, p.121-154.
11. A.M.Liberman, I.G.Waftingly, 1985. The motor theory of speech perception revised. Cognition 21, p.1-36.
12. G.A.Fowler, 1986. An event approach in the study of speech perception from direct-realist perception. J.Phonetics, 14, p.3-28.
13. K.N.Stevens, 1989. On the quantal theory of speech. J.Phonetics, 17, 3-15.
14. Д.В.Разумихин, 2001. Разработка системы понимания устной речи в диалоге. Международная конференция по компьютерной лингвистике "Диалог 2001", с.323-329.
15. Д.В.Разумихин, 2000. Использование нейронных сетей на уровне семантики в системе распознавания речи. IV всероссийская конференция "Нейрокомпьютеры и их применение, с.208-210.
16. Д.Разумихин, А.Соловьев, 2003. Системы автоматического распознавания речи с различными моделями организации диалога. XIII сессия российского акустического общества, т.3, с141-144.
17. Soloviev A.N., Victorova K.O., Razumikhin D.V., 2002. About using non-informational functions in models of speech communication. SpeeCom 2002, p.27-31.
18. Robust Methods for Speech Recognition in Advise Conditions. Proc. Conf., Finland 1999.
19. David F. Rosenthal, Hiroshi G. Okuno. Computational Auditory Scene Analysis. Publisher: Lawrence Erlbaum Assoc. ISBN: 0805822836
20. Albert S. Bregman. Auditory Scene Analysis: The Perceptual Organization of Sound. Publisher: The MIT Press. ISBN:0262022974