УДК 004.352.243

Т.В.Колтакова

ОСОБЕННОСТИ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Воронежский институт высоких технологий

В данной статье рассмотрены основные проблемы, связанные с речевой акустикой, а также актуальные вопросы в области распознавания речи. Отмечается, что трудности распознавания речевых сигналов обусловлены их вероятностными свойствами, а также многообразием воздействующих помех. Речь рассматривается с системной точки зрения, как объект обладающий иерархической структурой и имеющий некоторые свойства, определяемые не только свойствами входящих в него частей. Пользователям требуется проводить предварительную настройку систем на их голос. Системообразующие факторы продемонстрированы на примере акустических свойств речи и системы фонетических признаков.

Ключевые слова: распознавание речи, система распознавания, обработка сигналов, помеха.

В настоящее время область разработки САРР (систем автоматического распознавания речи) сместилась в основном к созданию систем автоматического понимания речи. При этом открывается совершенно новые перспективы к развитию речевой науки уже не в эмпирическом или измерительном, а в рациональном, т.е. познавательном плане [1].

Идея первых разработок САРР сводилась к предположению, что устная речь полностью подобна письменной, и она состоит из дискретных временных сегментов. Проблема в этом случае сводится к правильному выбору алфавита признакового описания этих сегментов.

Основной проблемой в области исследований в области речевой коммуникации, таким образом, становится реальная организация структуры речевого сообщения, которая отчетливо носит не вероятностный характер [2, 3].

Только для европейских сообществ объемы продаж систем, нацеленных на гражданское назначение дают несколько миллиардов долларов. Может сложиться представление, что в качестве конечной и высшей цели стремятся к созданию именно «фонетических печатающих машинок», а в качестве универсального метода для решения различных проблем в области передачи речевых сигналов можно считать скрытые Марковские модели (HMM), который на основе статических методов позволяет определить указанный алфавит и в нем можно работать с речью аналогично тому, как работают с письменными текстами.

Тяжело решаемые проблемы появляются для условий:

- произвольных, наивных пользователей;

- когда используется спонтанная речь, которая сопровождается аграмматизмами и в которой есть речевой «мусор»;
- возникновения акустических помех и искажений, в которых может происходить изменение параметров;
- возникновения речевых помех.

От пользователей необходимо осуществление предварительных настроек системы на их голос, которая по длительности составляет от десяти минут до нескольких часов, для того, чтобы проводить процессы предварительного наговаривания текстов. В некоторых проверках не получаются результаты лучшие, чем 5% ошибок. Поскольку слова, которые включены в тексты, произносимые аккуратным способом, получаются как бы размазанными по омонимической области, то число ошибок в словах возрастает приблизительным образом в 5 раз. Проводить процессы по беглому отслеживанию подобных ошибок уже сложно. Процедуры коррекции ошибок в большинстве демонстрационных системах работают довольно плохо. Некоторые исследователи установили, что даже когда есть хорошо созданные тексты, которые спонтанным образом произносят, то величина вероятности, что слова правильно распознают не будет больше, чем одна треть. Время, соответствующее работе с определенным отрезком речи в подобных системах – в среднем, не более 10 минут [4, 5].

Это не исключает возможностей применения их в для того, чтобы исследовать различные научные теории.

Если сделать рассмотрение признаков, которые выбираются для того, чтобы сделать описание речевых сигналов, то мы можем увидеть, что с такими целями практически в произвольных последовательностях применяют все те признаки, которые были испытаны в исследованиях речевых сигналов: уровни по спектральным полосам, определенные формантные признаки, коэффициенты, связанные с линейным прогнозом, кепстральные характеристики и др. Ни для одного из тех наборов, которые возможны, мы не можем говорить о явном преимуществе, и результаты практически зависят от того, насколько аккуратен набор статистики. Все сегменты в речевом потоке являются одинаковым образом информационными для того, чтобы распознавать крупные речевые единицы (слова, фразы). То есть для сегментов, входящих в Марковские модели мы можем отметить, что отсутствует какая-либо лингвистическая привязка и мы можем говорить о единицах, которые имеют лишь вероятностный смысл, то есть предполагается вероятностная организация в речевом поведении людей.

Достаточно старым естественно-научным направлением, в котором пытаются сделать объяснение строения речевых сигналов на фонетических уровнях считают акустическую теорию речеобразования, ее научное

оформление сделал Гельмгольц. В качестве основных особенностей модели можно отметить такие:

- В процессах речеобразования выделяют две составляющие, которые являются независимыми: возбуждают звуковой сигнал и создается соответствующее фонетическое качество звукового сигнала, поскольку происходит возбуждение резонансных частот в артикуляционных трактах (по Гельмгольцу) или осуществляется фильтрация (для современного рассмотрения).
- В характеристиках фонетического качества звука выделяют форманты, позволяющие сделать вывод о том, какие из значений резонансных частот в артикуляционных трактах рассматриваются (или говорят о полюсах передаточной функции в артикуляционном фильтре) или какие могут быть максимумы в спектрах речевых сигналов.

Анализ классической литературы показывает, что в голосовых источниках существуют свои полюсы и нули, что в явном виде влияет на то, какое качество создаваемых речевых сигналов, если существуют форманты, влияющие существенным образом на фонетическое качество, то в таких случаях для них характерными будут максимумы в спектрах.

Более 50 лет назад сформировали на базе экспериментальных способов теорию, которая связана с методиками расчета характеристик разборчивости речи, где в качестве основы принят сигнал, имеющий полосное представление. Российскими исследователями Варшавским Л.А. и Литваком И.М. была определена гипотеза, говорящая о том, что характеристики фонетического качества звуков связаны с тем, как соотносятся уровни мощности по спектральных полосам, а форманты (максимальные значения в спектрах) можно сделать доступными в речеобразующих аппаратах на основе того, что методом получают требуемые полосные соотношения. Есть определенное число систем, которые параллельным образом функционируют, с применением звукоразличительных признаков. Поскольку есть несколько систем признаков, относящихся к разным типам, мы можем обеспечить устойчивость речевых сигналов как коммуникативных систем при воздействиях помех, шумов и искажений в широких диапазонах.

Можно увидеть, каким образом проявляются индивидуальные особенности в произношении не только для высокочастотных или других не фонемообразующих формант, но и для сдвигов относительно средних значений в низкочастотных формантах, которые фонетически значимы. Во-вторых, индивидуальные вариации в системах частот формант определяются не только какие индивидуальные особенности строения артикуляционных трактов, соответствующих данным типам звуков, но и

контекстуальной динамикой создания такого строения в процессах речепроизводства, которые являются характерными для данных дикторов.

Понятно, что вследствие просто физических ограничений люди в процессах речеобразования не имеют возможностей для управления управлять большим количеством спектральных составляющих. Это позволяет говорить о том, что в качестве первичной цели в создании речевых сигналов скорее будет общая форма спектров. Но при этом нельзя исключать фонетико-различительные характеристики формант, а еще может потребоваться параллельная система признаков.

Идеологию акустических теорий в речеобразовании принимают как каноническую, хотя можно отметить факты, которые не будут согласовываться с такими теориями:

- Гласные фонемы, акустические характеристики которых хорошо согласуются с известными правилами, практически не содержат полезную информацию о речевых сообщениях, хотя они характеризуются значительно большей интенсивностью, чем согласные, их как раз и рассматривают в качестве основных переносчиков информации.
- Целая совокупность экспериментов позволяет утверждать то, что наиболее важную роль в процессах передачи речи можно связать с переходами между фонемами.
- Независимость источников звуков и артикуляторных фильтров не подтверждается в результате экспериментальных исследований.
- Возможно рассмотрение акустических процессов при процессах речеобразования, которые принципиальным образом отличаются от модели Гельмгольца – Фонта. В качестве примера можно привести модуляционную модель.

Акустическую речь можно свести к фонетическим письменным видам (и наоборот), но нельзя ее считать ограниченной лишь линейными фонетическими структурами. Часто она их и не имеет, а только можно изобразить их как фонетическую цепочку. Целое не будет представлять простую сумму своих составляющих, а обладает определенными новыми свойствами, которые не выводятся из подобных составляющих.

Рассмотрим речь в качестве системы и сделаем выделение тех факторов, которые можно считать как базовые в ее формировании.

Первым фактором в речевых системах можно считать их продуктивность, то есть возможности продуцирования сколь угодно большого количества информационных сообщений, которые несут разный смысл.

Еще одним системообразующим фактором, который не существует в письменной речи, но можно считать в качестве необходимого для того,

чтобы реализовать речевые системы в их акустических вариантах – является помехозащищенность. Под помехоустойчивостью понимают не только возможности обеспечения функционирования коммуникационных систем, когда существуют внешние мешающие факторы в каналах связи, но и субъективные факторы: характеристики психологического состояния людей, возможности отвлечения внимания.

Ряд экспериментальных данных позволяет указать на то, что речеслуховые системы обладают совокупностью специфических свойств, которые не будут связанными с общеслуховыми восприятиями.

Передача лингвистической информации таким образом обеспечивается рядом параллельно действующих систем различительных признаков. Необходимо 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. Это практически не дает возможности успешного применения в автоматическом распознавании речи методов, основанных на применении эталонов.

Сразу встает проблема в идеологии автоматического распознавания речи: какова должна быть общая модель распознавания, если отказаться от модной, но, очевидным образом, непродуктивной вероятностной модели. Естественным представляется применение для построения систем автоматического распознавания моделей восприятия речи.

Вторая проблема это принцип выбора первичного описания речевого сигнала. Либо это статистический анализ различных речевых акустических параметров, либо для распознавания речи необходимо перейти от акустических параметров к артикуляциям, либо выбор за новой перспективной квантовой теорией. Согласно которой первый тип акустических признаков соответствует резкому изменению акустического сигнала при небольшом изменении артикуляционного тракта, второй тип синхронно плавному изменению сигнала с изменением артикуляции.

Логично задать вопрос: каким образом необходимо, чтобы взаимодействовали первичные признаки и другие речевые уровни: вербальным, семантическим, прагматическим, вероятностным и др.

В современных САРР задача понимания смыла чаще всего решается распознаванием речевых сегментов, а затем поступлением всего распознанного на семантический модуль. Однако, естественная речь зачастую аграмматична и на практике сложно применить грамматику для построения высказывания. Приходится использовать разнообразные «улучшители» понимания как, например, учет предыстории, выявление контекста и падежно-ролевых отношений или использование различных

статистически-вероятностных методов (частотности, ассоциативности и пр.). Плюс ко всему список поиска вероятных слов при распознавании пополняется ассоциативной лексикой. Особыми проблемами является «мусор» – слова, которых нет в словаре распознавания, а так же различного рода помехи как речевого, так и неречевого типа.

Исходя из всего вышесказанного, необходимо рассматривать и другой подход к пониманию речи: проведение подстройки модуля распознавания до процессов обработки входных сигналов. Для решения подобной задачи требуется учитывать не только весь семиозис, окружающие диалоги, но и существование у распознающих систем интеллекта. Основной проблемой при таких подходах можно считать правильное формирование базы знаний, проведение предварительного обучения системы, способность к тому, чтобы проводить адекватный анализ окружающей действительности. Важно также отслеживать, чтобы распознавание не сводилось лишь к тому, чтобы был простой поиск ключевых слов, в соответствие которым из ограниченного числа «смыслов» подыскиваем наилучший.

Одной из нерешенных проблем также является называемый в теории восприятия речи cocktail-party эффект или более точно – анализ акустических сцен в условиях сильной зашумленности. Несмотря на достаточно пристальное внимание научного мира к проблеме помехозащищенности, заметных улучшений по решению данной проблемы нет [6].

Наиболее перспективные сферы применения речевых технологических устройств связаны с их взаимодействием с человеком и требуют по своей сути повторения в технологической системе методов работы с речью, с речевой информацией, используемых человеком. Компрессия речи практически в настоящее время сосредоточилась на методах лобовой аппроксимации речевого сигнала без учета его акустико-информационной структуры.

В речевом сообщении обязательно должна быть избыточность не только вследствие прямого запараллеливания тех признаков, которые являются различительными, но и она позволяет функционировать речевой коммуникации для различных типах помех, как поступающих извне в систему, так и существующих внутри системы. Однако, заранее предполагаемая избыточность в акустико-параметрическом обеспечении речевых сообщений определяет необязательность в полном параметрическом обеспечении речевого сообщения для каждого конкретного этапа в речевых коммуникациях. Можно предположить участие в процессе речевых коммуникаций трех компонентов: источник информации, канал связи, приемник информации. Соответственно в

каждой из конкретных ситуаций речевых общений довольно ясно, чем можно мы можем пожертвовать в структурах речевых сообщений.

При этом наиболее неясным вопросом в речевых коммуникациях на настоящий момент остается вопрос о способах временной организации речевых сигналов, распадающийся на два:

- реализуется ли реальным образом линейный формат в речевых сигналах (аналогично тому, как это происходит в письменной речи для фонетически ориентированных типов письменности).
- где в речевых акустических сигналах расположена основная смысловая, поведенчески полезная информация.

Если мы сумеем понять, как организован процесс речевой коммуникации, то только тогда мы сможем четко поставить конкретные акустические проблемы. С другой стороны, если удастся разобраться в акустических механизмах речеобразования, то станет понятна и структура речевого сигнала.

ЛИТЕРАТУРА

1. Галунов В.И. Помехоустойчивость как системообразующий фактор речи. Проблемы и методы экспериментально-фонетических исследований. СПб., 2002.
2. Галунов В.И., Гарбарук В.И. Акустическая теория речеобразования и системы фонетических признаков. "100 лет экспериментальной фонетике в России". Материалы международной конференции. СПб, 2001.
3. Н.Б.Покровский Расчет и измерение разборчивости речи.-Связьиздат, 1962.
4. В.Н.Сорокин. Теория речеобразования.-М., Радио и связь,1985.
5. Фролов В.Е. Перспективы развития автоматического распознавания слитной русской речи. XXXIV Международная филологическая конференция, 2005.
6. Д.В.Разумихин, 2000. Использование нейронных сетей на уровне семантики в системе распознавания речи. IV всероссийская конференция "Нейрокомпьютеры и их применение, с.208-210.

T.V.Koltakova

THE FEATURES OF AUTOMATIC SPEECH RECOGNITION

Voronezh Institute of High Technologies

In the paper the basic of speech acoustic and actual items in the field of speech recognition. It is noted that the difficulty of recognition of speech signals associated with their probabilistic properties and diversity impact of interference. Users need to pre-configure systems on their voice. The system factors are shown on the example of acoustics features of speeches and the system of fonetic signs.

Keywords: the speech recognition system of recognition, handling of signals, disturbance.

REFERENCES

1. Galunov V.I. Pomekhoustoychivost' kak sistemoobrazuyushchiy faktor rechi. Problemy i metody eksperimental'no-foneticheskikh issledovaniy. SPb., 2002.
2. Galunov V.I., Garbaruk V.I. Akusticheskaya teoriya recheobrazovaniya i sistemy foneticheskikh priznakov. "100 let eksperimental'noy fonetike v Rossii". Materialy mezhdunarodnoy konferentsii. SPb, 2001.
3. N.B.Pokrovskiy Raschet i izmerenie razborchivosti rechi.- Svyaz'izdat, 1962.
4. V.N.Sorokin. Teoriya recheobrazovaniya.-M., Radio i svyaz',1985.
5. Frolov V.E. Perspektivy razvitiya avtomaticheskogo raspoznavaniya slitnoy russkoy rechi. XXXIV Mezhdunarodnaya filologicheskaya konferentsiya, 2005.
6. D.V.Razumikhin, 2000. Ispol'zovanie neyronnykh setey na urovne semantiki v sisteme raspoznavaniya rechi. IV vserossiyskaya konferentsiya "Neyrokomp'yutery i ikh primenenie, pp.208-210.