Актуальные задачи речевой акустики

Е.М. Максимов,
доктор технических наук

Ю.Н. Ромашкин,
кандидат технических наук

С.А. Лопатина,
кандидат физико-математических наук

В статье даётся оценка достигнутого уровня развития отечественных речевых технологий по основным прикладным задачам. Указываются существующие ограничения, обусловленные сложившимися представлениями о процессах речеобразования и слухового восприятия речи. Формулируются приоритетные научные направления по дальнейшему совершенствованию речевых технологий.

Современный уровень развития методов и средств цифровой обработки сигналов предоставляет широкие возможности для внедрения речевых технологий в различные сферы жизнедеятельности, связанные с речевой коммуникацией. Практический интерес к речевым технологиям обусловлен как коммерческими потребностями, так и необходимостью решения специальных задач, возникающих, например, в речевой криминалистике [1].

Данная статья посвящена анализу существующих методов и алгоритмов цифровой обработки речевых сигналов, а также постановке задач, которые, по нашему мнению, являются актуальными с точки зрения повышения эффективности обработки в соответствующих этим методам практических приложениях.

Анализ современного состояния и направлений развития речевых технологий рассмотрим в рамках следующей классификации прикладных задач, решаемых с помощью этих технологий:

- идентификация (верификация) диктора по устной речи;
- распознавание естественной речи (в отличие от распознавания команд или изолированных слов), в частности, распознавание «ключевых» слов в слитной речи;
- повышение разборчивости речи на фоне акустических помех и искажений;
- синтез естественной речи.

К настоящему времени методы решения перечисленных задач основаны на сложившихся представлениях и моделях речеобразования и слухового восприятия. Эти модели с определёнными ограничениями и упрощениями использованы

при статистическом синтезе существующих алгоритмов обработки и позволили достичь известных положительных результатов. Вместе с тем, достигнутый уровень решения задач с помощью этих методов является недостаточным, а в ряде случаев неудовлетворительным для их практического применения. Для достижения качественно нового уровня необходимы, с одной стороны, разработка методов и алгоритмов, адекватных механизмам речеобразования и восприятия, с другой стороны — исследования, направленные на уточнение и более глубокое изучение этих механизмов.

Рассмотрим в данном контексте уровень развитых речевых технологий, применяемых в сформулированных выше прикладных задачах.

Автоматическая идентификация (верификация) диктора

Доминирующие позиции при решении данной задачи в настоящее время занимает GMM-метод [2], основанный на многомерной гауссовской аппроксимации выборочных плотностей распределения ряда акустических параметров речи. В приложениях, реализующих этот метод, используется либо неадаптивная GMM-модель небольшой размерности (24–64), либо адаптивная с высокой размерностью 1024–2048. В качестве информативных признаков, характеризующих индивидуальные особенности речи диктора, наиболее часто используются мел-кепстральные коэффициенты, их первые и/или вторые производные по времени. Такой подход, по нашим оценкам, позволил в среднем обеспечить следующие значения основных показателей эффективности работы алгоритмов:

- область работоспособности по отношению сигнал/шум — не ниже 10–15 дБ;
- минимальная длительность речи для составления модели голоса диктора — порядка 1 мин.;
- минимальная длительность речи для идентификации –10–15 с;
- вероятность правильной идентификации — 0,85–0,95;
- вероятность ложной тревоги — не более 0,10–0,15.

Дальнейшее повышение эффективности алгоритмов автоматической идентификации (верификации) возможно, по нашему мнению, по двум направлениям. С одной стороны, требуется совершенствование существующего GMM-метода в части обеспечения его инвариантности к преобразованиям и искажениям речевого сигнала в канале передачи и приёма, свойствам акустических помех и шумов канала, расширения вектора используемых акустических параметров речи и оптимизации размерности пространства информативных признаков. Более перспективным и актуальным, однако, представляется добавление в алгоритм идентификации ряда лингвистических признаков, характеризующих индивидуальные особенности порождения и структуру речи каждого диктора. Необходим поиск таких лингвистических признаков, способов их математической формализации и оценки с помощью цифровых методов анализа речевого сигнала.

Автоматическое распознавание естественной речи

Задача автоматического распознавания естественной речи (включая распознавание «ключевых» слов в слитной речи) является наиболее сложной, поскольку в наибольшей степени требует адекватных моделей порождения, восприятия и понимания речи. Исследованиями различных аспектов проблемы распознавания речи занимаются многие зарубежные и отечественные организации на протяжении нескольких десятилетий. Но, несмотря

на значительные успехи лабораторных разработок в этой области, практическое применение такие системы нашли в очень узкой области.

Современные подходы к решению данной задачи широко используют аппарат скрытых марковских моделей, параметрическое представление акустических параметров речи для различных элементов слитной речи (аллофонов, дифонов, трифонов и т.д. [3]). Далее осуществляется поиск наиболее вероятной последовательности распознанных звуков или слов с учётом принятой модели конкретного языка.

Принятый подход позволил, по нашим оценкам, в среднем обеспечить следующие значения показателей эффективности работы алгоритмов:

- область работоспособности по отношению сигнал/шум — не ниже 15–20 дБ;
- вероятность правильного распознавания «ключевых» слов в слитной речи — примерно 0,8–0,9 (при объёме рабочего словаря порядка 100 слов и вероятности ложной тревоги — не более 0,2–0,3);
- вероятность правильного распознавания слитной речи — около 0,6–0,7 (при объёме рабочего словаря порядка 20–30 тысяч слов).

Надёжность существующих систем автоматического распознавания речи зависит от многих факторов.

Речь, состоящая из изолированных слов или произносимая в замедленном темпе, более проста для распознавания. В быстрой естественной речи некоторые фонемы «смазываются» или просто «проглатываются», возрастают также коартикуляционные эффекты.

Наличие акустически похожих слов также затрудняет распознавание. Системы, рассчитанные на большой словарь, требуют больше времени на принятие решения. Уменьшение этого времени обычно производится за счёт упрощения алгоритма, что приводит к увеличению ошибок.

Сложность звукового строя конкретного языка в значительной степени определяется его фонетическим составом и правилами порождения слов. Например, звуковой строй японского языка гораздо проще для распознавания, чем французского, а русского и английского — сложнее французского.

Существенно влияют на надёжность автоматического распознавания речи воздействие внешних акустических помех, наличие амплитудно-частотных и временных искажений речевого сигнала в канале приёма и передачи, изменения психофизического состояния говорящего, артикуляционные дефекты в речи. Явление интерференции (смешение языков, взаимовлияние языков) давно интересует исследователей, но только сейчас делаются попытки их формализации. Необходима дальнейшая систематизация фонетических, лексических, синтаксических, просодических отклонений в речи иностранцев.

Без решения всего комплекса этих проблем получение качественно новых результатов при автоматическом распознавании речи представляется маловероятным.

Повышение разборчивости на фоне акустических помех и искажений

Эффективность алгоритмов выделения речи на фоне помех в значительной мере определяется акустическими условиями приёма речи и статистическими свойствами помех. Задачу подавления квазистационарных коррелированных помех и повышения разборчивости речи на фоне таких помех можно считать решённой как в теоретическом, так и прикладном плане на основе применения алгоритмов адаптивной фильтрации [4]. При наличии некоррелированных помех разработаны методы, основанные на теории марковской нелинейной фильтрации [5]. Однако предположения и допущения, используемые при реализации алгоритмов марковской фильтрации, не в полной мере адекватны механизмам речеобразования и восприятия, что обусловливает искажения речевого сигнала, хотя отношение сигнал/шум может быть существенно повышено. Эти же недостатки свойственны методам спектрального вычитания. Они позволяют увеличивать отношение сигнал/шум, однако, разборчивость речи при этом либо совсем не повышается, либо даже снижается вследствие появляющихся в результате обработки заметных искажений речи. Для условий приёма речи на фоне нестационарных помех (музыкальных, речевых и т.п.) пригодных для практического применения алгоритмов фильтрации пока не разработано.

В целом существующие методы выделения речи на фоне аддитивных помех умеренной интенсивности (отношениях сигнал/шум около 10 дБ) обеспечивают, по результатам наших оценок, следующие выигрыши в слоговой разборчивости речи:

- при наличии квазистационарных коррелированных помех — 20–30%;
- некоррелированных шумоподобных помех — 8–10%
- нестационарных помех — 5–7%
- реверберации речи — до 10%.

Синтез речи

В настоящее время наибольшее развитие получили методы синтеза речи на основе артикуляционного [6], лингво-акустического подходов и синтеза по правилам [7]. Основу этих методов составляют достаточно подробная математическая модель артикуляторного тракта, модели речеобразования по разным речевым элементам (фонемам, аллофонам, слогам, дифонам, трифонам и т.д.), а также различные рекуррентные модели с линейным предсказанием. При адекватном наборе исходных элементов артикуляционный и лингво-акустический подходы обеспечивает качественное воспроизведение спектрального состава речи, а набор правил — возможность формирования её естественного просодического оформления.

Проблема состоит в достижении натуральности звучания, приближающейся к естественной речи, устранении недостатков существующих методов при стыковках элементов речи между собой, управлении просодическими характеристиками формируемого сообщения, модификации индивидуальных особенностей синтезируемого голоса. Требуется также углубление лингвистических и акустических знаний о процессах речеобразования для уточнения и расширения набора правил, используемых при управлении процессом синтеза.

Для повышения качества синтеза речи первоочередной задачей является детальное теоретическое и инструментальное исследование формирования динамических процессов речеобразующего тракта, связанных с эффектами коартикуляции, переходом от звука к звуку. Необходимо также определение физических, лингвистических и психологических параметров, создающих натуральность синтезируемой речи. В настоящее время почти не изучены

механизмы восприятия синтетической речи. Поэтому более или менее удачные коммуникативные, модальные, стилевые и эмоциональные интонации в программном синтезе получаются пока не на основе познанных закономерностей, а скорее интуитивно или методом подбора параметров.

Пока предлагаемые решения в большинстве своём служат только стартовыми позициями в решении проблемы, что отражает наше сегодняшнее представление о речеобразовании как системе в целом. Без исследования процессов формирования естественной спонтанной речи практическое применение синтетической речи ограничено.

Литература

1. *Галяшина Е.И.* Слуховая перцепция как базовый метод фоноскопии. Речевые информационные технологии, 2003.
2. *Reynolds D.A., Rose R.C.* Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. // IEEE Trans. on Speech and Audio Proc., 1995. Vol.3, №.1. pp. 72–83.
3. *Потапова Р.Г.* Речь: коммуникация, информация, кибернетика. М.: Радио и связь, 1997. 528с.
4. *Уидроу Б., Стирнз С.* Адаптивная обработка сигналов. М.: Радио и связь, 1989. 440с.
5. *Маркел Дж.Д., Грей А.Х.* Линейное предсказание речи. М.: Связь, 1980. 308с.
6. *Сорокин В.Н.* Синтез речи. М.: Наука, 1992.
7. *Darrow B.* Research Spurs Development of Talking Machines // Design News, 1984. V. 40, №.12.

Максимов Е.М. —

1952 г.р., доктор технических наук. Государственное учреждение «Войсковая часть 35533».

Ромашкин Ю.Н. —

1953 г.р., кандидат технических наук. Московский государственный институт радиотехники, электроники и автоматики (технический университет).

Лопатина С.А. —

1961 г.р., кандидат физико-математических наук. Государственное учреждение «Войсковая часть 35533».