УДК 81'322.6

А. И. Соломенник, А. О. Таланов, М. В. Соломенник,
О. Г. Хомицевич, П. Г. Чистиков

ОЦЕНКА КАЧЕСТВА СИНТЕЗИРОВАННОЙ РЕЧИ: ПРОБЛЕМЫ И РЕШЕНИЯ

Рассмотрены различные аспекты проблемы оценки результатов работы систем синтеза речи. Приведен краткий обзор существующих методик оценки качества.

Ключевые слова: *синтез речи, качество синтезированной речи, сравнение систем синтеза речи.*

Введение. Синтезированная речь в последние годы все больше используется в различных сферах, например, в банковских системах голосового самообслуживания, транспортных компаний, при проведении телефонных опросов. Синтезированными голосами „говорят“ мобильные устройства, озвучиваются аудиокниги. Поэтому задача оценки качества синтезированной речи и сравнения систем синтеза между собой становится как никогда актуальной. Однако в этой области существует немало проблем. Основной можно назвать субъективность оценок: кто-то обращает внимание на тембр синтезированного голоса или на ошибки в произношении, для кого-то голос слишком „роботизирован“ или, наоборот, „излишне живой“ и непредсказуемый. В настоящей статье будут рассмотрены существующие подходы к объективной оценке качества синтеза в целом и его отдельных компонентов.

Оценка качества лингвистической обработки. Системы синтеза могут сравниваться и оцениваться по следующим объективным параметрам, отражающим решение задач лингвистической обработки в системах синтеза:

1) выделение предложений в тексте и разбиение их на отдельные слова; разметка текста на буквы, специальные символы, цифры и знаки пунктуации;

2) нормализация текста — расшифровка сокращений, аббревиатур, цифровых обозначений, номеров телефонов, дат, времени и т.п.;

3) определение места ударения и морфо-грамматических характеристик слов в предложении, для этого обычно используется словарь и/или набор правил или статистические модели;

4) снятие омонимии (омографии), т.е. выбор одной из нескольких словоформ, соответствующих тому или иному слову текста. Эти словоформы могут различаться ударением, наличием буквы „ё“ или грамматическими характеристиками;

5) построение сегментной транскрипции по правилам или по словарю (в зависимости от языка).

Для объективной оценки качества лингвистической обработки может быть создан тестовый текст либо использован фрагмент готового текстового корпуса, на основе которого вычисляется процент ошибок по каждому из параметров [1]. Эти оценки могут быть получены автоматически, если для сравнения доступен нормализованный и размеченный системой синтеза текст.

Оценка просодической обработки. К такой обработке относится использование тех компонентов, которые придают тексту интонационное оформление, т.е. происходит деление текста на просодические единицы — синтагмы, определение длины пауз между синтагмами и выбор интонационного оформления для каждой из синтагм. Затем происходит вычисление физических параметров — длительности, частоты основного тона (ЧОТ), энергии — на основе полученных данных. Деление на синтагмы и выбор интонации могут осуществляться как по правилам, так и на основе статистических моделей, причем этап выбора интонационного типа в последнем случае может быть пропущен, система сразу на основе имеющихся данных может переходить к предсказанию требуемых физических параметров звуков.

На этом этапе объективная количественная оценка представляет собой уже менее тривиальную задачу, поскольку вышеуказанные параметры в естественной речи могут варьироваться. При оценке расстановки пауз могут отдельно учитываться места, где пауза необходима, возможна и недопустима [2]. Выбор интонационной модели внутри одной системы обозначений может быть оценен таким же образом (по набору допустимых вариантов), но при сравнении разных систем интонация обычно оценивается уже в выходных звуковых файлах по параметру схожести с естественной речью.

Оценка акустического модуля. Возможные проблемы в работе акустического модуля существенно зависят от технологии синтеза: например, в формантном, аллофонном или дифонном компилятивном синтезе это может быть общая заметная неестественность (роботизированность) звучания одновременно с неудачными отдельными звуками; в компилятивном синтезе методом Unit Selection (US) — различимые стыки звуков, призвуки, несоответствие интонационного оформления логически обусловленному контекстом, причем ошибки обычно неравномерно распределены по тексту; в синтезе, основанном на статистическом моделировании (HMM), — роботизированность всей речи или звуков определенного типа, в то время как резких „скачков“ тона или энергии, как в синтезе US, обычно не наблюдается. Если используется значительная модификация звука, в синтезированной речи появляются заметные призвуки и эффект роботизированности.

Степень влияния результата работы акустического модуля на общее качество синтеза сложно переоценить. Поэтому в технологии синтеза основное внимание уделяется именно развитию технологии получения результирующего речевого потока. Опосредованно оценить качество работы этого модуля можно на основе оценки качества синтеза (или оценки общего впечатления), поскольку он формирует выходной сигнал на основании работы предыдущих модулей.

Оценка общего качества синтеза. Методы оценки качества синтеза можно в первую очередь разделить на две большие группы: субъективные (MOS-оценка) и инструментальные.

К первой относятся разного рода тесты, опросники, заполняемые экспертами — специалистами либо наивными слушателями. При создании опросников обычно используются рекомендации P.85 ITU-T „Метод субъективной оценки качества речи устройств речевого вывода“ [3]. В них используется MOS-оценка по пятибалльной шкале по нескольким категориям: общее впечатление, слуховое усилие, естественность, понимание смысла сообщения, темп, разборчивость, приятность голоса. На основе этих критериев принимается решение о приемлемости голоса (для определенных задач) по двубалльной шкале.

Однако проведение такого тестирования является довольно трудоемкой задачей. Для того чтобы ускорить процесс оценки и сделать его более доступным на каждом шаге разработки систем синтеза, создают различные инструментальные (или объективные) методы оценки качества синтеза. Такие методы основываются как на автоматическом сравнении синтезированной речи (с использованием различных мер близости) с „живой“ речью того же диктора [4, 5], так и на построении дикторонезависимых моделей естественной речи и различных методах оценки того, насколько синтезированная речь к ним приближена [6]. При этом исходным является предположение о том, что в естественной речи невозможны резкие скачки в частоте основного тона, энергии или спектральных составляющих, характерные для систем конкатенативного синтеза. В работе [7] предлагается инструмент, позволяющий оценивать качество просодической обработки на основании данных о значениях ЧОТ и длительности звуков речи.

Кроме упомянутых ранее характеристик речи в системе оценки качества синтезированной речи могут быть использованы следующие признаки, оценка которых может выполняться автоматически [8]:

— интонированность/монотонность определяется по изменению производной ЧОТ;

— ритмичность — параметр, который может характеризовать разные аспекты речи, прежде всего он определяется паузами, разбивающими речь на относительно равномерные отрезки;

— мелодичность — параметр, отражающий долю голосовых (вокализованных) фрагментов речи.

Подходы и системы оценки. В России для объективной оценки качества синтезированной речи чаще всего используется ГОСТ Р 50840-95 „Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости“. Наряду с ним используются различные тесты отдельных компонентов [9], но единого стандарта оценки пока нет. В работе [10] был предложен подход к комплексной оценке систем синтеза русской речи, однако он не имеет широкой известности и практически не применяется.

Одним из главных событий в сфере синтеза речи является Blizzard Challenge — „соревнования“ синтезаторов. Голоса для сравнения систем синтеза создаются на основе одних и тех же звуковых баз данных, предоставляемых перед началом соревнований. По прошествии времени, отведенного на создание голосов, участникам выдается набор текстов, синтезированные звуковые файлы для которых необходимо предоставить организаторам для оценки. В 2010 г. соревнования проводились для корпусов речи на английском и китайском языках [11]. В качестве дополнительного задания в 2012 г. предлагалось разработать собственный метод оценки качества синтеза и провести оценку [12].

По образцу Blizzard Challenge для испаноязычных синтезаторов был организован конкурс Albayzin [13]. Существует стандартизованный набор тестов для синтеза речи на французском языке, разработанный в ходе национального проекта EvaSy (Evaluation of speech synthesis systems — оценка систем синтеза речи) [14].

Заключение. Оценка качества синтезаторов в последние годы является предметом широких исследований; за рубежом активно ведутся работы по стандартизации оценок. Для русскоязычных синтезаторов существуют отдельные перспективные разработки, на основании

которых должен быть выработан единый стандарт качества синтезированной речи. Представленный обзор наработок в этой области за последние несколько лет является первым шагом к выработке такого стандарта.

Система оценки качества синтезированной речи может применяться для решения следующих задач.

Тестирование системы синтеза в процессе разработки. К системе оценки предъявляются следующие требования: она должна быть автоматической; иметь достаточно высокое быстродействие; может оцениваться как на соответствие голосу конкретного диктора, так и на соответствие общим параметрам речевого сигнала. Для анализа должны быть доступны результаты всех этапов синтеза, и проверка должна осуществляться с использованием промежуточной информации, генерируемой системой в явном виде.

Оценка собственной системы синтеза речи в сравнении с конкурентами. Для этого может применяться как автоматическая дикторонезависимая оценка [6], так и оценка экспертов. В данном случае может быть затруднен доступ к результатам синтеза: для коммерческих приложений обычно доступны только интерактивные демоверсии, при помощи которых можно получить образцы звука низкого качества с фоновой музыкой и др. в целях защиты от коммерческого использования, или же доступны только заранее подготовленные примеры. Для корректного сравнения результатов работы синтезаторов необходимо использовать их полнофункциональные версии.

Участие в конкурсах, проводимых независимыми компаниями. Система оценки может быть не автоматической, но автоматизированной. Для оценки системы могут привлекаться большие человеческие ресурсы (например, заинтересованные пользователи Интернета). Хоть внутренняя структура систем синтеза и останется закрытой, будет возможно получение промежуточных результатов работы системы в унифицированном виде. Системы синтеза могут тестироваться на одной и той же голосовой базе, на основе которой строится синтезированный голос.

СПИСОК ЛИТЕРАТУРЫ

1. *Sproat R., Black A. W., Chen S., Kumar S., Ostendorf M., Richards C.* Normalization of non-standard words // Computer Speech and Language. 2001. Vol. 15. P. 287—333.

2. *Хомицевич О. Г., Соломенник М. В.* Автоматическая расстановка пауз в системе синтеза русской речи по тексту // Матер. Междунар. конф. „Диалог“. 2010.

3. Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, ITU-T Rec. Int. Telecom. Union. 1994. 85 p.

4. *Vepa J., King S., Taylor P.* Objective distance measures for spectral discontinuities in concatenative speech synthesis // Proc. Intern. Conf. Spoken Language Processing. September, 2002. P. 2605—2608.

5. *Stylianou Y., Syrdal A.* Perceptual and objective detection of discontinuities in concatenative speech synthesis // Proc. Intern. Conf. Acoustics, Speech, and Signal Processing. June, 2001. P. 837—840.

6. *Falk T. H., Möller S.* Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems // IEEE Signal Proc. Letters. 2008. Vol. 15. P. 781—784.

7. *Norrenbrock C. R., Hinterleitner F., Heute U., Moller S.* Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals // IEEE Signal Proc. Letters. 2012. P. 255—258.

8. *Киселев В. В., Давыдов А. Г., Ткаченя А. В.* Система определения эмоционального состояния диктора по голосу // Междунар. науч.-техн. конф. „Открытые семантические технологии проектирования интеллектуальных систем“ (OSTIS-2012) / Под ред. *В. В. Голенкова*. Минск: БГУИР, 2012. С. 355—358.

9. *Гецэвіч Ю. С.* Алгарытмы лінгвістычнай апрацоўкі тэкстаў для сінтэзу маўлення на беларускай і рускай мовах: Дыс. … канд. тэхн. навук. Мінск, 2012. 191 с.

10. *Русанова О. А.* Исследование и разработка методов анализа и оценки качества синтезированной устной речи. Дис. … канд. техн. наук. Красноярск, 2004. 107 с.

11. [Электронный ресурс]: <http://festvox.org/blizzard/blizzard2010.html>.

12. [Электронный ресурс]: <http://www.synsig.org/index.php/Blizzard_Challenge_2012_Rules>.

13. *Méndez F. et al.* The Albayzín 2010 Text-to-Speech Evaluation // Fala2010. 2010. P. 317—340.

14. [Электронный ресурс]: <http://www.technolangue.net/article.php3?id_article=202>.

Сведения об авторах

Анна Ивановна Соломенник — ООО „Речевые технологии“, Минск; научный сотрудник; E-mail: solomennik-a@speechpro.com

Андрей Олегович Таланов — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела синтеза речи; E-mail: andre@speechpro.com

Михаил Васильевич Соломенник — канд. техн. наук; ООО „Речевые технологии“, Минск; ведущий инженер-программист; E-mail: solomennik-m@speechpro.com

Ольга Гурьевна Хомицевич — PhD; ООО „ЦРТ“, Санкт-Петербург; старший научный сотрудник; E-mail: khomitsevich@speechpro.com

Павел Геннадьевич Чистиков — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: chistikov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.