Методика тестирования систем автоматического синтеза и распознавания речи в целях определения коммерческой целесообразности их использования

Корсакова Н.С., Засыпкина К.А.
ЗАО «ЭсТуЭс Некст»

В данной статье рассматривается опыт создания тестового материала для определения качества систем синтеза и распознавания речи, а также демонстрируются сравнительные результаты тестирования нескольких промышленных систем с помощью этого материала.

• синтез речи • распознавание речи • методика тестирования • тестовый материал • база данных.

A new testing method of automatic speech synthesis and recognition systems is proposed. The authors focus mainly on the problem of choice of testing material which would be efficient for testing and comparison of commercial systems. Several particular speech synthesis and recognition systems are tested, results are presented and compared, so some conclusions concerning commercial viability of their use can be made.

• speech synthesis • speech recognition • testing method • testing material • database.

Введение

Рынок речевых технологий и средств компьютерной обработки речи — один из самых быстрорастущих на сегодняшний день. По данным компании J'son & Partners, он оценивается примерно в 3 млрд. долларов. Рост рынка, согласно аналитикам Voice Information Associates, составляет около 25% в год [1]. В настоящее время речевые технологии — это, прежде всего, технологии автоматического распознавания речи (ASR) и автоматического синтеза речи (TTS).

Выбор системы ASR или TTS для применения в какой-либо определенной сфере деятельности должен быть тщательным и взвешенным. Такая необходимость обусловлена следующими факторами:

1) денежные затраты. Ошибка в выборе техники (метода или алгоритма), которая будет использована в будущем приложении, может привести к лишним денежным затратам, которые могут быть вызваны необходимостью ликви-

дировать (сгладить) ошибки и неточности работы системы; исключением из результата работы системы лишней информации, переплатой за счёт реализации ненужных функций и т.д.;

2) временные затраты. При внедрении сложной технологии для реализации простых функций придётся затратить гораздо больше времени, чем требовалось бы, если бы разработчик ограничился только необходимыми методами и алгоритмами;

3) объём занимаемой памяти. При неверном выборе метода для реализации системы обработки речи возможны излишние затраты памяти для хранения информации, баз знаний, которых можно было бы избежать. Это, в свою очередь, приводит к избыточным денежным затратам;

4) качество. Существующие системы автоматического синтеза и распознавания речи различаются по качеству синтеза/распознавания. При выборе той или иной системы необходимо определить тип приложения, в которое будет внедрена технология, так как требования к качеству системы должны зависеть от сферы применения. Качество, в свою очередь, влияет на вышеперечисленные факторы, поэтому целесообразно перед выбором системы проверить ее работу на практике.

Несмотря на большое число разработок, проблемы синтеза и распознавания речи до сих пор считаются нерешенными, так как качество синтеза и распознавания только в отдельных случаях можно считать удовлетворительным и хорошим. Улучшению качества автоматического синтеза речи препятствует сложность разрешения языковой неоднозначности при автоматическом анализе текста, который используется в синтезе устной речи для расстановки пауз; определения главноударного слова в предложении; задания интонации вопроса, восклицания; для правильной расстановки ударения в словах [2]. На качество автоматического распознавания речи оказывают влияние условия прикладной области, в частности, состав и размер словаря. Причина невысокого качества распознавания кроется в вариативности речевого сигнала, которая обуславливается, например, индивидуальными особенностями дикторов, характеристиками каналов связи, а также влиянием окружающей обстановки [3].

Для оценки качества рассматриваемых систем обработки речи была разработана методика тестирования, которая может быть применена для оценки любых продуктов ASR и TTS. Данная методика направлена на оценку широкого класса систем автоматического синтеза и распознавания речи и не ограничена определённой предметной областью или сферой применения.

Тестирование систем ASR

Тестовый материал для систем автоматического распознавания речи состоит из двух частей: лексической и синтаксической. Лексическая часть предназначена для проверки точности распознавания изолированных слов, а синтаксическая – для проверки точности распознавания слитной речи. Тестовый материал предполагает оценку качества распознавания на русском и английском языках.

Лексическая часть для каждого языка состоит из квази-омонимов или так называемых «минимальных пар», а также других пар слов, способных вызвать трудности при распознавании. Материал фонетически сбалансирован, содержит основные реализации всех фонем обоих языков. Ср.:

выгореть-выгорать, вырвать-вырыть, довольно-довольный, заказ-закат, мол-мор, в охапку-в охотку, корпусной-корпусный и т.д.

Assert-asset, await-awake, back-bag, bald-balk, mad-maid, file-fire, journal-journey, sport-spot, you-your, etc.

Синтаксическая часть представляет собой адаптированный вариант материала, часто используемого для определения качества передачи речи, взятого из ГОСТа 16600-72 [4], и состоит из трёх групп для каждого языка, которые различаются по типу предложений:

простые, сложноподчинённые и сложносочинённые предложения. Материал также фонетически сбалансирован — в каждой группе фраз встречаются все фонемы русской речи и их основные варианты произношения. Слова для предложений взяты из разряда нейтральной лексики. Ср.:

Фильм снимают целый год. Самолёт оказался в воздушной яме.
I cannot find candies. Phil, focus please.

Оператор стирал старые записи, как вдруг неожиданно раздался звонок.
Prove it before I give you that promise.

Диктора поразило это сообщение, но план уже утверждается в области.
I came to talk, but you took everything wrong.

Кроме того, тестовый материал может быть адаптирован под готовые приложения с учётом сферы применения, типа приложения, объёма и специфики словаря и прочих факторов.

Зачитывая тестовый материал, участники тестирования проверяют результат распознавания и фиксируют ошибки (замены, вставки и выпадения). Затем производится подсчет ошибок, и вычисляются WER (Word Error Rate) [5] и WES (Word Error Rate per Sentence) [6] для слов и предложений соответственно по формулам:

$$WER = \frac{sub + del + ins}{nwords} \quad 100\%; \quad WES = \frac{sub(s) + del(s) + ins(s)}{nwords(s)} \quad 100\%,$$

где *sub*, *del* и *ins* – это количество замен, выпадений и вставок соответственно, *nwords* – количество слов, (s) означает «в предложении». Вычисляется WER для каждой группы слов, WES – для каждого предложения, а затем среднее WES для каждой группы предложений. Также выявляются основные ошибки и трудности распознавания.

По разработанной методике нами были протестированы четыре системы автоматического распознавания речи. Для каждой ASR были вычислены WER и WES, проведено их сравнение, сделаны выводы о качестве распознавания каждой системы. Кроме того, были выявлены наиболее частые ошибки, возникающие при распознавании. Рис. 1 отражает долю ошибок каждой системы от общего числа тестовых единиц при распознавании самых «проблемных» звуков. На диаграмме четко видно, что худшее качество показала система ASR 2, тогда как ASR 3 допустила меньше всего ошибок при распознавании. Эти результаты подтверждаются также числовыми показателями WER и WES.

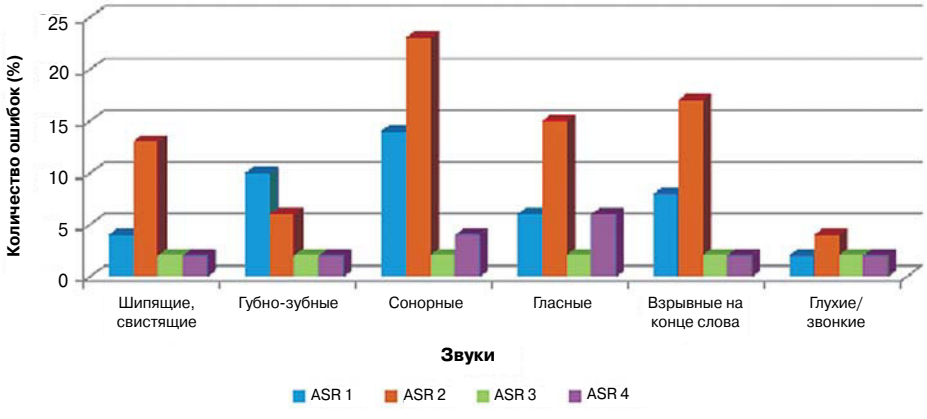

Рис. 1

Тестирование систем TTS

Тестовый материал для систем автоматического синтеза речи разделен на две части: лексическую и синтаксическую. В лексическую часть входят слова, обладающие определёнными характеристиками, способными повлиять на качество синтеза. С их помощью проверяется качество синтеза изолированных слов. Синтаксический блок включает в себя разного типа предложения для проверки качества синтеза слитной речи. Тестовый материал предполагает оценку качества распознавания на русском и английском языках. Оба блока содержат труднопроизносимые звукосочетания: сонорные звуки на стыке слов и слогов, гласные на стыке слов и слогов, шипящие и свистящие звуки и их сочетания, сочетания взрывных, взрывные на конце слова, и т.д. Выбор в пользу такого рода сочетаний звуков сделан на основе опыта работы с системами автоматического синтеза речи и обусловлен существованием проблем, с которыми в этой связи приходилось сталкиваться. Данная методика является комбинацией и адаптацией различных методов тестирования систем TTS, таких как DRT (Diagnostic Rhyme Test), MRT (Modified Rhyme Test), направленные на проверку качества синтеза начальных и/или конечных согласных; DMCT (Diagnostic Medial Consonant Test), проверяющий синтез интервокальных согласных; PB (Phonetically Balanced Word Lists), MOS (Mean Opinion Score), определяющие общее качество синтезированной речи, и т.д [7].

Первый блок тестового материала разделен на 6 групп:

- многосложные слова;
- сложные слова и аббревиатуры;
- слова с сочетанием нескольких гласных подряд;
- имена и фамилии;
- географические названия;
- числительные.

Поскольку данное тестирование проводится для оценки систем широкой сферы применения, конкретный языковой материал берется из числа общеупотребительной и деловой лексики. Материал фонетически сбалансирован и содержит звукосочетания, способные вызвать проблемы при синтезе. Ср.:

Группа №1.1: многосложные слова
Аббревиатура, деформировать, инфраструктура, сертификат, и т.д.
Группа №1.1 (eng): polysyllables
Abbreviation, character, government, handkerchief, significance, etc.
Группа №1.2: сложные слова и аббревиатуры
Работоспособность, ОАО, США, бухучёт, мосгорсуд, и т.д.
Группа №1.2 (eng): complex words and abbreviations
Ballot-paper, crisscross, SOS, PhD, etc, VIP, D.C, etc.
Группа №1.3: слова с сочетанием нескольких гласных подряд
Актуальный, закоулок, зоопарк, киоск, материал, хаотичный и т.д.
Группа №1.3 (eng): words with several vowels in a row
Alien, alleviate, cooperate, curiosity, dubious, loathe, moreover, etc.
Группа №1.4: имена и фамилии
Пётр, Илья, Анастасия, Ксения, Коваленко, Кривич, Залевская, и т.д.
Группа №1.4 (eng): first and last names
Duncan, Bernard, Anna, Alice, Katherine, Smith, Cameron, Darwin, etc.
Группа №1.5: географические названия
Санкт-Петербург, Швейцария, Средняя Азия, Ближний Восток, и т.д.
Группа № 1.5 (eng): geographical names
New York, Los Angeles, Manchester, Wales, Middle East, Asia, etc.
Группы №1.6 и №1.6 (eng): числительные / Numerals
1, 2, 3, 4, 5, 10, 11, 12, 13, 15, 20, 26, 100, 200, 1000.

Второй блок тестового материала состоит из предложений разного типа. Языковой материал данного блока является адаптированным материалом ГОСТа 16600-72: Передача речи по трактам радиотелефонной связи [4] и ГОСТа Р 50840-95: Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости [8]. Из предложенных в ГОСТах списков предложений были выбраны предложения, содержащие нейтральную лексику, и далее скомпонованы так, чтобы получились сложноподчинённые и сложносочинённые предложения. Следует отметить, что в получившихся предложениях встречаются все вышеперечисленные звукосочетания. Ср.:
Солнце ещё находится в зените. *Фильм снимают целый год*.
The room was remarkably groom and grouse.
Мальчик разбил санки о скамейку, а ведь они были заново выкрашены красной краской.
My telephone was too old, and I took a new one.
Этап был завершён, когда был собран киноаппарат.
Before you brought this book, my bookcase was not so heavy.

При тестировании систем синтеза речи с помощью первого блока участникам эксперимента предлагается оценить разборчивость и естественность синтезированной речи по пятибалльной шкале и, если была отмечена ошибка синтеза, указать, какая ошибка была допущена синтезатором, зафиксировав кириллицей или латиницей, используя специальные обозначения (таблица 1), вариант произношения, предложенный системой.

Таблица 1

Обозначение	Пример	Значение
Буква в верхнем регистре	парАметрический	Смещение ударения в слове
Замена буквы на 0	пар0метрический	Пропуск звука
Вставка звуков	параометрический	Посторонние звуки
Знак → перед словом (в предложении)	Мы поехали на → станцию	Ускорение темпа на слове
Знак ← перед словом	Мы поехали на ← станцию	Замедление темпа на слове
Знак ↑ перед/после/в середине слова	Мы поеха ↑ ли на станцию	Восходящая интонация
Знак ↓ перед/после/в середине слова	Мы поеха ↓ ли на станцию	Нисходящая интонация
Знак ~ перед/после слова	Мы поехали на ~ станцию	Прерывистая интонация
Знак ~ в середине слова	Мы поехали на ста~нцию.	Посторонние шумы/невнятность
Знак |	Мы поехали | на станцию	Лишняя пауза
Знак X	Мы X поехали на станцию	Отсутствие необходимой паузы

На втором этапе осуществляется проверка плавности воспроизведения предложений, расстановки фразовых ударений, а также просодические характеристики синтезируемой речи. Участники тестирования, помимо вышеописанных критериев, используют для оценки такие критерии как скорость, паузы и интонация. Интонация оценивается по двум шкалам: «верная-неверная», «ровная-прерывистая», расстановка пауз – по шкале «верно-неверно».

После этого для каждого блока производится подсчёт доли верно синтезированных слов/предложений; слов/предложений, воспроизведённых с минимальными искажениями; и слов/предложений, сильно искажённых при синтезе, от общего числа слов/предложений. Естественно, оценка является субъективной и опирается на восприятие синтезированных слов и предложений участниками тестирования. На завершающем этапе тестирования проводится сбор и анализ полученных результатов, после чего составляется сводная таблица, в которой указываются системы TTS, средний процент точного

воспроизведения, незначительно искажённого, сильно искажённого и основные ошибки синтеза как для слов, так и для предложений.

Все вышеперечисленное было проделано для четырех систем автоматического синтеза речи. Основными типами ошибок оказались:

- смещение ударения (например, *АнастАсия* вместо *АнастасИя*);
- выпадение звука (например, *Euro0* вместо *Europe*);
- замена звука (например, *иристпруденция* вместо *юристпруденция*);
- неровная интонация (например, *оператор стирает ↑ старые записи*).

Процент этих ошибок от общего числа тестовых единиц для каждой системы TTS представлен на рис. 2.

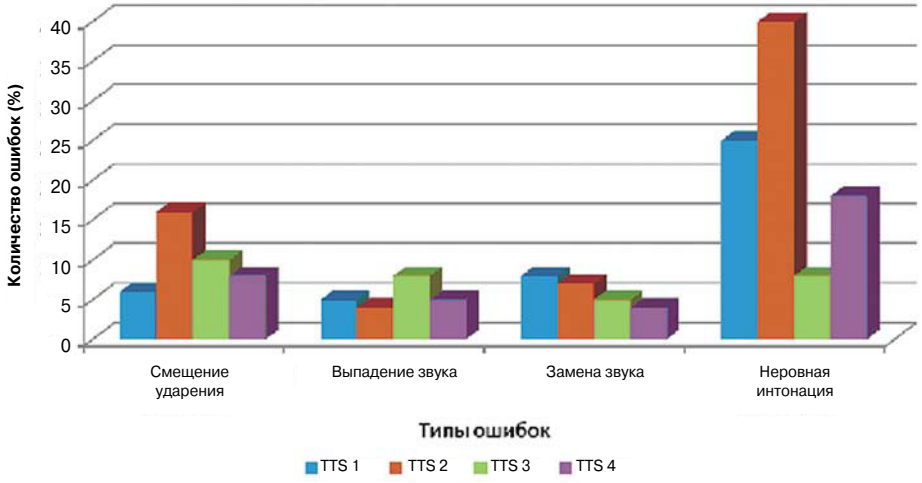

Рис. 2

Кроме того, было выявлено, что количество ошибок для разных тестовых групп различается. Например, самыми сложными для синтеза оказались сложные слова и аббревиатуры, тогда как числа синтезировались сравнительно хорошо: у двух из четырех систем ошибок не зафиксировано, о чем свидетельствует рис. 3.

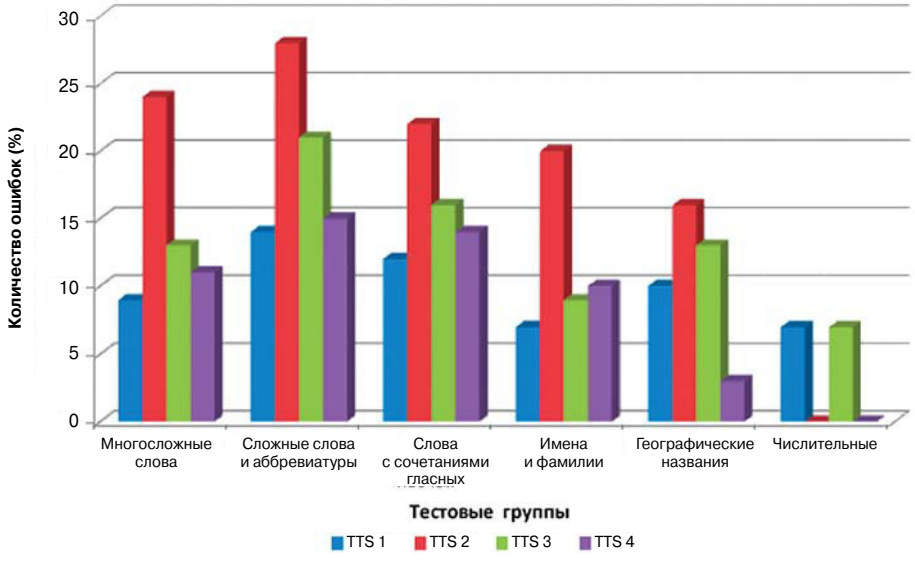

Рис. 3

Выводы

Разработанная методика тестирования систем автоматического синтеза и распознавания речи позволяет выяснить, какими недостатками и достоинствами обладают существующие системы ASR и TTS, сравнить их, а также определить коммерческую целесообразность их использования. Кроме того, после проведения такого тестирования можно сделать общие выводы об основных проблемах, связанных с автоматическим синтезом и распознаванием речи, и способах их преодоления. Важно отметить, что данная методика может иметь успешное коммерческое применение, так как она позволяет сделать выбор в пользу той или иной системы ASR и TTS для ее внедрения в конкретное решение. Результатом разработки данной методики стали зарегистрированные базы данных для тестирования систем ASR и TTS [9].

Список литературы

1. *Грамматчиков А*. Поговорить с компьютером. [Электронный ресурс]. Режим доступа: http://expert.ru/expert/2007/44/pogovorit_s_ komputerom/

2. *Русанова О.А*. Исследование и разработка методов анализа и оценки качества синтезированной устной речи: дисс... канд. техн. наук / Краснояр. гос. техн. ун-т. Красноярск. 2004. 107 с.

3. *Нгуен Минь Туан*. Разработка алгоритмов построения оценок достоверности для систем распознавания речи: дисс... канд. техн. наук / Вычисл. центр РАН. М.: 2008. 102 с.

4. ГОСТ. 16600-72. Передача речи по трактам радиотелефонной связи. Требования к разборчивости речи и методы артикуляционных измерений. Переизд. 1973. Взамен ГОСТ 16600-71; Введ. 27.09.72. М.: Госстандарт России. 1973. 75 с.

5. Word Error Rate. [Электронный ресурс]. Режим доступа: http://en.wikipedia.org/wiki/Word_error_rate

6. *Strik H., Cucchiarini C., Kessens J.M*. Comparing the recognition performance of CSRs: in search of an adequate metric and statistical significance test, Proc. ICSLP-2000, Beijing. 2000. P. 740–743

7. *Lemmetty S*. Review of Speech Synthesis Technology, Master's Thesis, Helsinki University of Technology. 1999. 104 p.

8. ГОСТ. Р 50840-95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. Введ. 01.01.97. М.: Госстандарт России. 1996. 230 с.

9. Свидетельство о государственной регистрации базы данных № 2012620510. «База данных материала для тестирования систем синте-за и распознавания речи для оценки качества». Правообладатель: ЗАО «ЭсТуЭс Нэкст». Зарегистрировано в Реестре баз данных 5 июня 2012 года.

Сведения об авторах

Корсакова Н.С. —
лингвист, ЗАО «ЭсТуЭс Некст». info@s2snext.com

Засыпкина К.А. —
директор по проектам, ЗАО «ЭсТуЭс Некст». info@s2snext.com