ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ АВТОМАТИЧЕСКОГО АНАЛИЗА РЕЧИ

УДК 004.522

А. А. Карпов, И. С. Кипяткова

МЕТОДОЛОГИЯ ОЦЕНИВАНИЯ РАБОТЫ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Представлена современная методология количественного оценивания результатов работы автоматических систем распознавания и диаризации речи. Приведены различные показатели и методы оценивания по критериям точности распознавания речи и скорости обработки речевого сигнала.

Ключевые слова: *автоматическое распознавание речи, точность распознавания речи, скорость обработки сигнала, критерии и показатели оценивания.*

Введение. Одной из основных проблем в работе систем автоматического распознавания речи является объективное количественное оценивание результатов распознавания, что имеет важное значение как для разработчиков, так и для конечных пользователей систем. Методология количественного оценивания производительности предназначена для сравнения и сопоставления различных систем распознавания, в ней выделяют критерий, показатель и метод:

— критерий — это область оценивания, т.е. то, что необходимо оценить: например, точность распознавания речи, скорость ее обработки, робастность и т.п.;

— показатель (мера или метрика) определяет конкретное свойство, которое оценивается для выбранного критерия: например, процент правильно распознанных слов, время обработки сигнала, уровень максимально допустимого шума при сохранении работоспособности и т.п.

— метод — это способ определения соответствующего значения для данного показателя: например, сравнение распознанных слов с последовательностью сказанных слов, оценка времени обработки в секундах и т.п.

При разработке систем автоматического распознавания речи, как правило, используются три набора данных: обучающий (“train”), отладочный (“dev”), оценочный или тестовый (“eval”). Обучающий набор данных (обычно это наибольшая часть речевых данных) применяется только для создания и обучения/тренировки системы; отладочный набор используется для настройки и адаптации параметров автоматической системы перед финальной стадией оценивания, этот набор данных должен иметь тот же формат, что и тестовые данные; оценочный набор содержит речевые данные, которые не использовались для обучения и настройки системы и доступны только при ее окончательной оценке.

Предметом рассмотрения в данной статье являются показатели точности и скорости распознавания речи.

Показатели точности распознавания речи. В системах автоматического распознавания речи основным показателем качества является точность распознавания, которая определяется как процент правильно распознанных слов (WRR — Word Recognition Rate) или, на-

оборот, неправильно распознанных слов (WER — Word Error Rate). Иногда также используется показатель ошибок распознавания фраз/предложений (SER — Sentence Error Rate), который является важным в диалоговых системах, где корректировка гипотезы распознавания невозможна в отличие от задачи диктовки текста. В последнее время в качестве основного показателя точности работы систем распознавания речи используется показатель WER, а именно, его абсолютное значение или относительное, если сравниваются различные модели/системы. Поскольку с развитием речевых технологий показатель WER все более приближается к нулю, то улучшение его значения более наглядно, чем повышение точности распознавания слов. Метод определения показателя WER состоит в выравнивании двух текстовых строк (первая — это результат распознавания, а вторая — запись того, что было сказано в действительности) с помощью алгоритма динамического программирования с вычислением расстояния Левенштейна [1]. Расстояние Левенштейна представляет собой „стоимость“ редактирования данных (минимальное количество или взвешенная сумма операций редактирования [2]) для преобразования первой строки во вторую с наименьшим числом операций ручной замены (S), удаления (D) и вставки (I) слов:

$$\mathrm{WER} = \frac{S + D + I}{T}, \quad \mathrm{WRR} = 1 - \mathrm{WER},$$

где T — количество слов в распознаваемой фразе.

Для оценивания результатов автоматического распознавания речи используется и такой показатель, как процент корректно распознанных слов (WCR — Word Correctly Recognized), который не учитывает ошибочные вставки слов, сделанные системой:

$$\mathrm{WCR} = \frac{H}{T} \cdot 100\ \%, \quad H = N - D - S,$$

где H — количество правильно распознанных слов.

WER — интуитивно понятный показатель качества распознавания для аналитических языков с достаточно простой морфологией, в которых грамматические значения однозначно выражаются самим словом (например, английский или французский). Однако синтетические языки (например, агглютинативные финский, турецкий или флективные русский, украинский) имеют богатую морфологию словообразования; в некоторых азиатских языках (китайском, корейском и т.п.) используются слоги взамен слов; в тайском языке отсутствуют явные разделители границ слов. Поэтому эти языки могут синтезировать достаточно длинные осмысленные словоформы из нескольких частей (морфем), определяющих грамматические признаки. Обычно конец словоформы произносится в беглой речи не так четко, как начальная часть слова, что приводит к акустической неопределенности и в среднем к более высоким по сравнению с аналитическими языками значениям показателя WER.

В синтетических языках для оценивания точности автоматического распознавания речи могут применяться другие показатели: ошибки распознавания букв/символов [3], фонем (звуков речи) [4], слогов [5] или морфем [6]. Кроме того, для некоторых синтетических языков (например, русского) адекватным их структуре показателем является флективная ошибка распознавания слов (IWER — Inflectional Word Error Rate) [7], которая определяется следующим образом:

$$\mathrm{IWER} = \frac{S_{\mathrm{hard}} \cdot C_{\mathrm{hard}} + S_{\mathrm{soft}} \cdot C_{\mathrm{soft}} + D + I}{T}, \quad C_{\mathrm{soft}} < C_{\mathrm{hard}}, C_{\mathrm{hard}} \ge 1, 0 \le C_{\mathrm{soft}} < 1.$$

Показатель IWER приписывает вес C_{hard} всем неверным заменам слов, которые приводят к замене лексемы слова, т.е. к грубым ошибам распознавания (S_{hard} — количество ошибок), и меньший вес C_{soft} — всем негрубым ошибкам в словах, где было неверно распознано окончание словоформы, но основа слова распознана правильно (S_{soft} — количество негрубых ошибок).

При оценивании точности автоматического распознавания речи по показателю WER предполагается, что все слова во входной (поступающей на вход системы) фразе одинаково информативны и важны. Однако очевидно, что в системах, отличных от диктовки текста, например в диалоговых или в системах понимания (смысла) речи, некоторые значащие (ключевые) слова более важны, чем остальные (функциональные слова, предлоги, слова-заполнители и т.п.). В работе [8] предложено оценивать точность распознавания, используя взвешенный показатель неправильно распознанных слов (WWER — Weighted Word Error Rate), который определяется как

$$\text{WWER} = \frac{V_S + V_D + V_I}{V_T},$$

$$V_T = \sum_{W_i \in T} v_{W_i}, V_I = \sum_{\hat{W}_i \in I} v_{\hat{W}_i}, V_D = \sum_{W_i \in D} v_{W_i}, V_S = \sum_{s_j \in S} v_{s_j}, v_{s_j} = \max\left(\sum_{\hat{W}_i \in s_j} v_{\hat{W}_i}, \sum_{W_i \in s_j} v_{W_i} \right),$$

где v_{W_i} — вес слова W_i, которое является i-м во входной фразе, и $v_{\hat{W}_i}$ — вес слова $\hat{W}_i$, которое является i-м в гипотезе распознавания; s_j — j-й замененный фрагмент фразы (или одно слово) и v_{s_j} — вес данного фрагмента s_j.

Таким образом, согласно показателю WWER каждое слово может иметь различный вес (установленный экспертом или полуавтоматически) в соответствии с его влиянием на последующее понимание смысла сказанной фразы.

Национальным институтом стандартов и технологий (NIST, США) недавно был предложен такой показатель, как количество неправильно распознанных слов в речи каждого из дикторов (SAWER — Speaker Attributed Word Error Rate) — для задачи стенографирования совещаний [9], в которых предполагается участие нескольких дикторов. Данная задача объединяет технологии автоматического распознавания речи и диаризации голоса диктора (разметки звукового сигнала на фрагменты „кто и когда говорил“ — “Who Spoke When”) [10]. Результатом этой объединенной системы является текстовая транскрипция входного одноканального звукового сигнала для каждого распознанного слова с явным указанием на говорящего. Показатель SAWER определяется следующим выражением:

$$\text{SAWER} = \frac{S + D + I + V}{T},$$

где V — количество слов (или других языковых единиц), правильно распознанных системой автоматического распознавания речи, но с неправильной идентификацией диктора по результатам диаризации.

Однако следует понимать, что процент неправильного распознавания — это в действительности только количественный показатель точности распознавания (количество ошибок распознавания на фразу или слово), но не вероятность распознавания слова во фразе, так как показатель WER не ограничивается интервалом вероятности [0; 1] и не имеет верхнего предела. Например, представим, что кто-то произнес фразу, состоящую из 10 слов, но система ее полностью распознала неправильно и предложила гипотезу из 15 других слов. В этом случае WER=150 % (S=10, I=5, H=D=0), и, следовательно, показатель точности WRR отрицательный (т.е. –50 %), что не имеет смысла. Для того чтобы решить эту проблему, недавно были предложены другие показатели, в частности: ошибка распознавания соответствий (MER — Match Error Rate) и показатель потери информации, содержащейся в словах (WIL — Word Information Lost) [11], основанные на величине относительной потери информации и определяемые следующим образом:

$$\text{MER} = \frac{S + D + I}{T_P = H + S + D + I} = 1 - \frac{H}{T_P}; \quad \text{WIL} = 1 - \frac{H^2}{T \cdot T_O}, \text{ если } H >> S + D + I,$$

где T_O — количество слов в гипотезе распознавания; однако оба этих показателя редко применяются, так как обеспечивают обычно несколько меньшую точность распознавания по сравнению со стандартными показателями.

Все названные выше показатели учитывают только одну наилучшую гипотезу распознавания каждой произнесенной фразы, и совсем не обязательно, что этот единственный результат распознавания окажется действительно правильным. Однако некоторые системы автоматического распознавания речи (например, фонетический декодер) способны выдавать сразу несколько гипотез распознавания с наибольшими вероятностями — так называемый список N лучших гипотез (N-best List). Дополнительным показателем для оценки таких результатов является показатель ошибок распознавания слов в списке лучших гипотез [12], который оценивается путем выбора из N гипотез, ранжированных по уменьшению оценки правдоподобия, единственной гипотезы, имеющей наименьший уровень ошибок. Показатель WER гипотезы с минимальным уровнем ошибок по каждой входной фразе выбирается как основной результат распознавания, характеризующий процент ошибок распознавания в списке N лучших гипотез.

При моделировании и распознавании речи на основе теории вероятностей также используются доверительные интервалы для того, чтобы показать значимость результатов. При оценивании результатов автоматического распознавания речи доверительный интервал (confidence interval) иногда указывается вместе со средним значением показателя WER (например, WER=18,5 ± 2,3 %). В общем случае доверительные интервалы показывают, во-первых, какое значение показателя WER можно ожидать при изменении набора тестовых данных, во-вторых, насколько значимым является предложенное улучшение модели распознавания. Однако на практике доверительные интервалы показателя WER оказываются весьма широкими, что объясняется высокой вариативностью речи и голоса дикторов, а также речевыми сбоями (некоторые фразы распознаются с нулевым показателем WER, но другие приводят к очень высокому уровню ошибок). Большинство производимых улучшений в моделях автоматического распознавания речи не вызывают изменения значений, выходящих за пределы доверительного интервала, из-за ограниченности наборов тестовых данных, что несколько снижает значимость результатов. Однако как новые, так и базовые методы распознавания речи обычно оцениваются разработчиками исходя из одинаковых тестовых данных (эти речевые данные не являются в разных сравниваемых моделях распознавания независимыми); в этом случае при количественной оценке точности распознавания доверительные интервалы могут не рассматриваться. Но в случае когда модели распознавания тестируются с использованием различных и независимых тестовых наборов, требуется вычисление доверительного интервала дополнительно к среднему значению показателя WER [13].

Показатели скорости распознавания речи. Второй важный критерий работы системы автоматического распознавания речи — скорость обработки речи. Скорость обработки вычисляется, как правило, с использованием меры, называемой показателем скорости (SF — Speed Factor) и также известной как показатель реального времени (RT — Real Time) [9], который определяется отношением общего времени обработки, требуемого для анализа всей записанной речи на одном ядре процессора, к длительности исходного анализируемого аудиосигнала. Например, если 10-минутный аудиофайл обрабатывается системой распознавания речи в течение 5 минут, то SF=0,5 RT, если файл обрабатывается в течение 20 минут, то SF=2,0 RT, что значительно хуже. Скорость обработки может быть также указана в абсолютных значениях времени (например, количество минут/секунд для обработки входного сигнала), однако это не является наглядным. Другим показателем скорости автоматического распознавания речи может быть период ожидания обработки отсчета (SPL — Sample Processing Latency) [9]. Этот показатель означает максимальное количество аудиоданных, которое алгоритм распознавания должен обработать до выдачи результата о первом отсчете сигнала.

При создании обладающей (сверх)большим словарем системы автоматического распознавания речи, которая работает в реальном масштабе времени с использованием микрофона (онлайн режим), часто требуется найти компромисс между точностью распознавания и скоростью обработки. Настройка некоторых параметров системы может улучшить точность распознавания, но уменьшить скорость обработки. В этом случае может быть полезным график зависимости показателя WER от скорости распознавания в некоторых контрольных точках [14]; результаты анализа этого графика позволяют выбрать оптимальные параметры системы.

Заключение. Представлен аналитический обзор различных критериев, количественных показателей и методов, применяемых для оценки результатов работы систем автоматического распознавания и диаризации речи. Рассмотрены основные и альтернативные показатели качества, такие как точность и корректность распознавания речи, ошибка распознавания фраз, слов и символов, скорость обработки речевого сигнала и ряд других.

Статья подготовлена по результатам исследований, проводимых при поддержке Минобрнауки РФ (федеральная целевая программа „Исследования и разработки“, госконтракт № 07.514.11.4139); совета по грантам Президента РФ (проект № МК-1880.2012.8) и Российского фонда фундаментальных исследований (проект № 12-08-01265-а).

СПИСОК ЛИТЕРАТУРЫ

1. *Levenshtein V. I.* Binary codes capable of correcting deletions, insertions and reversals // Sov. Phys. Dokl. 1966. Vol. 6. P. 707—710.
2. *Khokhlov Y., Tomashenko N.* Speech recognition performance evaluation for LVCSR system // Proc. of the 14th Intern. Conf. “Speech and Computer” SPECOM—2011, Kazan, Russia. 2011. P. 129—135.
3. *Kurimo M., Creutz M., Varjokallio M., Arsoy E., Saraclar M.* Unsupervised segmentation of words into morphemes — Morpho challenge 2005 Application to automatic speech recognition // Proc. Interspeech-2006, Pittsburgh, PA. 2006. P. 1021—1024.
4. *Schlippe T., Ochs S., Schultz T.* Grapheme-to-phoneme model generation for indo-european languages // Proc. ICASSP-2012, Kyoto, Japan. 2012.
5. *Huang C., Chang E., Zhou J., Lee K.* Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition // Proc. Interspeech-2000, Beijing, China. 2000. P. 818—821.
6. *Ablimit M., Neubig G., Mimura M., Mori S., Kawahara T., Hamdulla A.* Uyghur morpheme-based language models and ASR // Proc. of the 10th IEEE Intern. Conf. on Signal Processing ICSP-2010, Beijing, China. 2010. P. 581—584.
7. *Karpov A., Kipyatkova I., Ronzhin A.* Very large vocabulary ASR for spoken russian with syntactic and morphemic analysis // Proc. Interspeech-2011, Florence, Italy. 2011. P. 3161—3164.
8. *Nanjo H., Kawahara T.* A new ASR evaluation measure and minimum bayes-risk decoding for open-domain speech understanding // Proc. ICASSP-2005, Philadelphia, PA. 2005. P. 1053—1056.
9. The US NIST 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan [Электронный ресурс]: <http://www.itl.nist.gov/iad/mig/tests/rt/2009/>.
10. *Ронжин А. Л., Будков В. Ю.* Система протоколирования дикторов на базе алгоритма определения речевой активности в многоканальном аудиопотоке // Речевые технологии. 2010. № 3. С. 98—102.
11. *Morris A. C., Maier V., Green P.* From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition // Proc. Interspeech- 2004, Jeju Island, Korea. 2004. P. 2765—2768.
12. *Tran B.-H., Seide F., Steinbiss T.* A word graph based N-best search in continuous speech recognition // Proc. ICSLP-96, Philadelphia, PA. 1996. P. 2127—2130.
13. *Vilar J. M.* Efficient computation of confidence intervals for word error rates // Proc. ICASSP-2008, Las Vegas, NV. 2008. P. 5101—5104.
14. *Hruz M., Campr P., Dikici E., Kindirouglu A., Krnoul Z., Ronzhin Al., Sak H., Schorno D., Akarun L., Aran O., Karpov A., Saraclar M., Zelezny M.* Automatic fingersign to speech translation system // J. on Multimodal User Interfaces. 2011. Vol. 4, N 2. P. 61—79.

Сведения об авторах

Алексей Анатольевич Карпов — канд. техн. наук; СПИИРАН, лаборатория речевых и многомодальных интерфейсов; E-mail: karpov@iias.spb.su

Ирина Сергеевна Кипяткова — канд. техн. наук; СПИИРАН, лаборатория речевых и многомодальных интерфейсов; E-mail: kipyatkova@iias.spb.su

Рекомендована СПИИРАН

Поступила в редакцию
10.06.12 г.