УДК 004.934

ИССЛЕДОВАНИЕ СВОЙСТВ ПОКАЗАТЕЛЕЙ КАЧЕСТВА СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

А.А. Саакян

Рассмотрена проблема выбора показателей качества, характеризующих эффективность системы распознавания звуковой речи. Сформулированы требования к показателям качества и предложена методика их экспериментального исследования. Дан анализ свойств известных показателей качества, приведены результаты эксперимента, позволяющие выбрать наилучший из них.

Ключевые слова: распознавание речи, алгоритм распознавания, показатель качества, измерение качества.

ВВЕДЕНИЕ

Исследования в области машинного распознавания речи ведутся в настоящее время исключительно эвристическими методами по причине отсутствия полных моделей как речевого сигнала, так и процесса речеобразования. Подобные модели, существуй они, позволили бы аналитически решать задачу конструирования алгоритма распознавания с заданными характеристиками точности, надежности, дикторозависимости.

Альтернативой использованию полной модели речевого сигнала служит выдвижение гипотез о свойствах алгоритма распознавания, конструирование алгоритма согласно этим гипотезам и практическая проверка его эффективности. Существующие хорошо изученные модели отдельных аспектов и этапов процессов образования и восприятия речи человеком помогают сделать поиск гипотез направленным. В частности, знание анатомии голосового аппарата позволяет объяснить эффект коартикуляции — взаимного влияния соседствующих в речевом сигнале звуков. Хорошо изученная связь объективных физических характеристик звуковой волны с субъективными ощущениями, возникающими в ухе при ее восприятии, позволяет при машинной обработке речи описывать речевой сигнал с помощью этих моделей. При синтезе алгоритмов распознавания применяется также модель частотного разложения звуковой волны во внутреннем ухе.

Упомянутые модели имеют статус эвристик, поскольку, будучи отрывочными, они не объясняют всех свойств речевого сигнала, и при конструировании алгоритма извлечения из речевого потока определенной информации приходится привлекать дополнительные модели, обычно стохастические, требующие обучения. Характеристики синтезированного таким комбинированным способом алгоритма распознавания могут быть вычислены исключительно эмпирически с использованием тестовой выборки речевых записей. В связи с этим важен оптимальный выбор показателей качества системы распознавания (СР) речи, характеризующих заданные ее свойства и позволяющих сравнивать эффективность разных алгоритмов и оптимизировать их параметры. Синтез показателя качества (ПК) с требуемыми свойствами опять-таки затруднителен. Однако можно сформулировать требования к ПК, позволяющие из имеющегося их набора экспериментально выбрать наилучший. На решение этой задачи и нацелено настоящее исследование.

1. ТРЕБОВАНИЯ К ПОКАЗАТЕЛЮ КАЧЕСТВА

Показатель качества системы распознавания является вещественной функцией вида $Q(\vec{l}, T)$, где $\vec{l}$ — вектор параметров системы, T — тестовое множество, случайная выборка из совокупности S всевозможных речевых записей, включающей в себя всевозможные варианты произнесения всевозможных последовательностей слов. Вид и состав параметров $\vec{l}$, разумеется, различен для разных СР. Впрочем, большинство современных систем, ориентированных на преобразование звуковой речи в цепочку слов, имеет единообразную структуру и характеризуется сходным набором параметров: речевой сигнал моделируется как последовательность отдельных звуков (фонем), образующих конечный алфавит. Отдельные фонемы моделируются при помощи скрытых марковских моделей (СММ) [1, 2]. Параметры моделей вычисляются с использованием обучающей выборки. Алгоритм

обучения итеративный и обеспечивает достижение локального оптимума (на обучающей выборке) за конечное число итераций.

Однако на тестовой выборке, вследствие явления переобучения (сверхобучения), эффективность системы должна поначалу возрастать, достичь максимума, а затем начать уменьшаться. Таким образом, число итераций обучения является оптимизируемым параметром системы. Далее, перед началом практического применения СР может, при наличии такой возможности, быть адаптирована к особенностям голоса конкретного пользователя, что повышает точность распознавания. Прирост эффективности системы при адаптации зависит от количества адаптационного речевого материала, которое заранее (на этапе проектирования системы) неизвестно. Таким образом, количество адаптационного речевого материала также является параметром системы, определяющим ее эффективность.

Для измерения значения ПК $Q(\vec{l}, T)$ необходимо, задав исследуемой СР значения параметров $\vec{l}$, распознать с ее помощью тестовую выборку T и в результате получить, в зависимости от назначения системы, последовательность слов или отдельных звуков. Для тестовой выборки задана корректная транскрипция. Значение ПК вычисляется как степень близости распознанной последовательности к корректной. При исследовании свойств СР стремятся максимально разграничить действие разных факторов на результат работы, поэтому модель языка обычно выбирается максимально простой: цепочка элементов языка (слов или фонем) произвольной длины. Порядок элементов произвольный, априорная вероятность появления одинакова для всех элементов. Алфавит элементов в случае фонемной грамматики языка задан и меняться, разумеется, не может. При распознавании целых слов словарь языка формируется из всех слов, встречающихся в тестовой выборке, и дополняется до заданного размера случайно выбранными словами. Эффективность системы падает с увеличением числа допускаемых слов, поэтому объем словаря языка также является параметром системы.

Рассмотрим требования к показателю качества.

- Пусть параметры $\vec{l}$ фиксированы, тогда $Q(\vec{l}, T)$ является функцией $Q_l(T)$ от случайного множества T: измеренное значение ПК зависит от случайной тестовой выборки. Для оценивания математического ожидания значения ПК и сравнения двух значений нужно применять статистические методы вычисления доверительных интервалов и проверки гипотез. При заданном уровне значимости доверительный интервал имеет тем меньшую (относительную) длину, чем меньшим рассеянием характеризуется случайная величина $Q_l(T)$. Очевидно, из двух ПК лучший из них обеспечивает большую точность оценки, т. е. меньшую длину доверительного интервала. Кроме того, предпочтительно выбрать ПК, для которого случайные величины $\{Q_l(T)\}_l$ имеют какое-нибудь стабильное, одинаковое и хорошо изученное распределение.
- Эмпирическое изучение системы, характеризующейся многими параметрами, затруднительно из-за огромного числа необходимых измерений. Если каждый параметр независимо от остальных принимает хотя бы сто значений, то уже при использовании всего трех параметров необходимо произвести 1 млн измерений. Для элиминирования действия явления «комбинаторного взрыва» стремятся отделить друг от друга влияние (на значение ПК) разных параметров и представить функцию $Q(\vec{l}, T)$ в виде комбинации (суммы, произведения и суперпозиции) функций отдельных параметров. Если ошибка подобной аппроксимации достаточно мала, можно исследовать зависимость $Q(\vec{l}, T)$ от каждого параметра отдельно. Соответственно, из двух ПК предпочтительней тот, который обеспечивает меньшую ошибку аппроксимации.
- Теоретически параметры СР таковы, что при фиксировании значений всех параметров $\vec{l}$, кроме одного, функция одного аргумента Q должна либо иметь локальный максимум, являющийся также и глобальным, либо монотонно зависеть от выбранного параметра. На практике зависимость показателей качества от значений параметров гораздо более сложная, что затрудняет оптимизацию параметров СР. Очевидно, из двух ПК предпочтительней тот, который лучше согласуется с теорией. Зафиксируем значения всех параметров $\vec{l}$, кроме l_k, равными $\vec{l}^*$. Функция $Q^*(l_k, T) = Q(\vec{l}, T)|l_p = l_p^*$, $p \neq k$ при фиксированной тестовой выборке T является дискретизированной (поскольку измерение значений ПК производится в конечном числе точек) функцией $q[i]$. Теоретический характер зависимости $Q^*(l_k, T)$ позволяет установить отношения, которые должны выполняться для пар элементов $q[i]$ и $q[j]$. Пусть, для определенности, функция $Q^*(l_k, T)$ должна монотонно возрастать по аргументу l_k. Тогда должно выполняться $q[i] < q[i + 1]$, $\forall i$. На практике для некоторых i условие монотонности нарушается: $q[i] \geqslant q[i + 1]$. Среднее число таких нарушений (при суммировании по всем аргументам) может служить мерой согласия ПК с теорией. При сравнении числа нарушений двух разных ПК, разумеется, необходимо осуществлять проверку значимости.

2. АППРОКСИМАЦИЯ МНОГОМЕРНОЙ ФУНКЦИИ ОДНОМЕРНЫМИ

Аппроксимация с приемлемой точностью многомерной функции одномерными представляет собой весьма нетривиальную задачу. Поскольку нас интересует не минимизация ошибки

аппроксимации, а сравнение значений ошибки для разных ПК, в настоящем исследовании мы ограничимся простейшей моделью аппроксимации. Рассмотрим функцию $F(\bar{x})$ n аргументов $\bar{x}$, заданную ее выборочно измеренными значениями $f(x_1, x_2, ..., x_n)$, где k-й аргумент x_k, независимо от остальных, принимает N_k возможных значений $\{x_k^i\}_{i=1,...,N_k}$. Представим ее в виде суммы функций, каждая из которых зависит только от одного аргумента, а также постоянной и ошибки аппроксимации: $f(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} f_i(x_i) + C + E(x_1, x_2, ..., x_n)$.

Вычислим дискретизованные функции $f_i(x_i)$, $i = 1, ..., n$, и постоянную C, минимизирующие сумму квадратов ошибки $E(x_1, x_2, ..., x_n)$.

Введем обозначения: $N = \prod_{j=1}^{n} N_j$, $\hat{N}_k = N_1 \cdot ... \cdot N_{k-1} N_{k+1} \cdot ... \cdot N_n = N/N_k$.

Для каждого значения x_k^i каждого аргумента x_k вычислим среднее арифметическое $f_k'(x_k^i)$ значений функции $f(x_1, x_2, ..., x_n)$, таких, что x_k равен x_k^i, а остальные аргументы пробегают все свои значения. Получим n дискретных функций:

$$f_k'(x_k^i) = \frac{1}{\hat{N}_k} \sum_{x_1, ..., x_{k-1}, x_{k+1}, ..., x_n} f(x_1, ..., x_{k-1}, x_k^i, x_{k+1}, ..., x_n) = f_k(x_k^i) + C + \frac{1}{\hat{N}_k} \sum_{x_1, ..., x_{k-1}, x_{k+1}, ..., x_n} \sum_{\substack{j=1 \\ j \neq k}}^{n} f_j(x_j) + \frac{1}{\hat{N}_k} \sum_{x_1, ..., x_{k-1}, x_{k+1}, ..., x_n} E(x_1, ..., x_{k-1}, x_k^i, x_{k+1}, ..., x_n) =$$

$$= f_k(x_k^i) + C + \sum_{\substack{j=1 \\ j \neq k}}^{n} \frac{1}{N_j} \sum_{x_j} f_j(x_j) + \frac{1}{\hat{N}_k} \sum_{x_1, ..., x_{k-1}, x_{k+1}, ..., x_n} E(x_1, ..., x_{k-1}, x_k^i, x_{k+1}, ..., x_n).$$

Запишем также выражение для среднего значения $\hat{f}$ функции $f(x_1, x_2, ..., x_n)$ по всем значениям всех аргументов $\bar{x}$:

$$\hat{f} = \frac{1}{N} \sum_{x_1, x_2, ..., x_n} f(x_1, x_2, ..., x_n) = \sum_{j=1}^{n} \frac{1}{N_j} \sum_{x_j} f_j(x_j) + C + \frac{1}{N} \sum_{x_1, x_2, ..., x_n} E(x_1, x_2, ..., x_n).$$

Очевидно, что для каждого набора значений аргументов $x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}$ справедливо:

$$\sum_{k=1}^{n} f_k'(x_k^{i_k}) - (n-1)\hat{f} = \sum_{k=1}^{n} f_k(x_k^{i_k}) + nC + (n-1) \sum_{j=1}^{n} \frac{1}{N_j} \sum_{x_j} f_j(x_j) - \frac{n-1}{N} \sum_{x_1, x_2, ..., x_n} E(x_1, x_2, ..., x_n) + \sum_{k=1}^{n} \frac{1}{\hat{N}_k} \sum_{x_1, ..., x_{k-1}, x_{k+1}, ..., x_n} E(x_1, ..., x_{k-1}, x_k^{i_k}, x_{k+1}, ..., x_n) - (n-1) \sum_{j=1}^{n} \frac{1}{N_j} \sum_{x_j} f_j(x_j) - (n-1)C =$$

$$= \sum_{k=1}^{n} f_k(x_k^{i_k}) + C + \sum_{k=1}^{n} \frac{1}{\hat{N}_k} \sum_{x_1, ..., x_{k-1}, x_{k+1}, ..., x_n} E(x_1, ..., x_{k-1}, x_k^{i_k}, x_{k+1}, ..., x_n) - \frac{n-1}{N} \sum_{x_1, x_2, ..., x_n} E(x_1, x_2, ..., x_n).$$

Отсюда

$$f(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}) = \sum_{i=1}^{n} f_i(x_k^{i_k}) + C + E(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}) = \sum_{k=1}^{n} f_k'(x_k^{i_k}) - (n-1)\hat{f} - \sum_{k=1}^{n} \frac{1}{\hat{N}_k} \sum_{x_1, ..., x_{k-1}, x_{k+1}, ..., x_n} E(x_1, ..., x_{k-1}, x_k^{i_k}, x_{k+1}, ..., x_n) + \frac{n-1}{N} \sum_{x_1, x_2, ..., x_n} E(x_1, x_2, ..., x_n) + E(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}).$$

Или, обозначив три последних слагаемых как $E'(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n})$,

$$f(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}) = \sum_{i=1}^{n} f_i(x_k^{i_k}) + C + E(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}) = \sum_{k=1}^{n} f_k'(x_k^{i_k}) - (n-1)\hat{f} + E'(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}).$$

Каждая из функций f_k' зависит только от одного из аргументов, $(n-1)\hat{f}$ — константа для заданной выборки, $E'(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n})$ — ошибка аппроксимации. Получено искомое разложение. Убедимся, что константа $(n-1)\hat{f}$ выбрана оптимальным образом. Вычислим минимум суммы (по выборочным значениям) квадратов значений ошибки от-

носительно неизвестного приращения C' постоянного члена разложения:

$$f(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}) = \sum_{k=1}^{n} f_k'(x_k^{i_k}) - (n-1)\hat{f} + C' + E'(x_1^{i_1}, x_2^{i_2}, ..., x_n^{i_n}),$$

$$\sum_{x_1, x_2, ..., x_n} (E'(x_1, x_2, ..., x_n))^2 = \sum_{x_1, x_2, ..., x_n} \left(f(x_1, x_2, ..., x_n) - \sum_{k=1}^{n} f_k'(x_k) + (n-1)\hat{f} - C' \right)^2 \to \min.$$

$$NC'^2 - 2C' \sum_{x_1, x_2, ..., x_n} \left(f(x_1, x_2, ..., x_n) - \sum_{k=1}^{n} f_k'(x_k) + (n-1)\hat{f} \right) + \sum_{x_1, x_2, ..., x_n} \left(f(x_1, x_2, ..., x_n) - \sum_{k=1}^{n} f_k'(x_k) + (n-1)\hat{f} \right)^2 \to \min.$$

Минимизируемая функция является выпуклой вниз параболой, достигающей своего минимума в точке

$$C' = \frac{1}{N} \sum_{x_1, x_2, ..., x_n} \left(f(x_1, x_2, ..., x_n) - \sum_{k=1}^{n} f_k'(x_k) + (n-1)\hat{f} \right) = \hat{f} - \frac{1}{N} \sum_{x_1, x_2, ..., x_n} \sum_{k=1}^{n} f_k'(x_k) + \frac{N}{N}(n-1)\hat{f} = n\hat{f} - \frac{1}{N} \sum_{x_1, x_2, ..., x_n} \sum_{k=1}^{n} f_k'(x_k) =$$

$$= n\hat{f} - \frac{1}{N} \sum_{k=1}^{n} \hat{N}_k \sum_{x_k} f_k'(x_k) = n\hat{f} - \sum_{k=1}^{n} \hat{f} = 0.$$

Если необходимо изучить поведение зависимости $f(x_1, x_2, ..., x_n)$ только от некоторых аргументов (пусть для определенности это будут первые m аргументов) аппроксимацию следует произвести в виде:

$$f(x_1, x_2, ..., x_n) = \sum_{i=1}^{m} f_i(x_i) + g(x_{m+1}, x_{m+2}, ..., x_n) + C + E(x_1, x_2, ..., x_n).$$

Для этого подмножество аргументов x_{m+1}, x_{m+2}, ..., x_n представляется в виде одного аргумента x'_{m+1}, после чего повторяются все приведенные рассуждения для разложения функции $f(x_1, x_2, ..., x_m, x'_{m+1})$ на сумму функций отдельных аргументов:

$$f(x_1, x_2, ..., x_m, x'_{m+1}) = \sum_{i=1}^{m} f_i(x_i) + g(x'_{m+1}) + C + E(x_1, x_2, ..., x_m, x'_{m+1}).$$

3. ОПИСАНИЕ ИССЛЕДУЕМЫХ ПОКАЗАТЕЛЕЙ КАЧЕСТВА

Как уже говорилось, качество системы распознавания связывается с расхождением между распознанной и априорной последовательностями слов (фонем). Для этого нужно любой паре последовательностей поставить в соответствие вещественное число, равное, например, расстоянию Левенштейна между ними. Расстояние Левенштейна [3] между двумя строками (последовательностями символов) равно минимальному числу операций вставки, удаления и замены символа, необходимому для преобразования первой строки во вторую. При сравнении транскрипций символом служит элемент языка — слово или фонема. Расстояние Левенштейна вычисляется с помощью алгоритма динамического программирования, минимизирующего сумму $S + D + I$, где I, D и S — число операций вставки, удаления и замены соответственно. Обозначим через H число совпадений, а через N_1 и N_2 — число символов в первой и второй строках. Очевидно, выполняются тождества: $N_1 = H + S + D$, $N_2 = H + S + I$. Выбор в качестве метрики расстояния Левенштейна не является строго обоснованным. Однако он интуитивно понятен: для систем диктовки текста расстояние Левенштейна интерпретируется как стоимость редактирования неверно распознанной строки. Соответственно, средняя точность систем диктовки определяется относительным количеством ошибочно распознанных слов (Word Error Rate) [2]:

$$WER = (S + D + I)/N_1,$$

где N_1 — число символов в корректной транскрипции. Этот ПК используется в подавляющем большинстве исследований.

Распространение ПК *WER* на системы, предназначенные не для транскрибирования звуковой речи, а для ее интерпретации, т. е. проблемно-ориентированные системы, оснащенные речевым интерфейсом (к которым относятся, например, автоматизированная система для бронирования билетов по телефону, робот, воспринимающий речевые команды, и т. п.), наталкивается на ряд противоречий. Речевой поток в таких системах состоит из отдельных, достаточно независимых друг от друга фрагментов [2]. Ошибка в распознавании фрагмента исправляется повторным вводом фрагмента целиком. Редактирование невозможно, соответственно, не имеет смысла выражать качество распознавания через среднюю стоимость редактирования.

Качество распознавания системы интерпретации речи должно быть связано с количеством полученной из речевого сообщения информации. Следовательно, средняя точность должна принимать значения в интервале [0; 1], тогда как значе-

ние *WER* не ограничено сверху. Пусть в результате распознавания некоторого сообщения $H = 0$. Количество полученной информации равно нулю, однако значение *WER* не фиксировано и может неограниченно возрастать при увеличении *I*. Более того, даже в случае транскрибирования речи, значение *WER* не является истинным размером среднего штрафа: вводимый в компьютер текст, так же, как и в системах прочих типов, делится на фрагменты — предложения. Если стоимость редактирования очередного ошибочно распознанного предложения чересчур велика, лучшим способом коррекции будет повторный ввод предложения целиком. Таким образом, в случае диктовки текста значение ПК также должно быть ограничено снизу. При исследовании свойств алгоритмов распознавания обычно используется не словарная, а фонемная грамматика в целях уменьшения влияния структурных свойств языка на результат. Для фонемной грамматики, независимо от назначения разрабатываемой СР, показатель *WER* не имеет интерпретации в виде среднего размера штрафа за редактирование ошибок.

Несмотря на указанные недостатки, показатель *WER* в настоящее время де-факто стал стандартом в области распознавания речи. Предложенные в работе [4] два новых альтернативных показателя *MER* (Match Error Rate) и *WIL* (Word Information Lost) остались практически незамеченными исследователями, по-видимому, вследствие отсутствия массового интереса к проблеме выбора ПК в силу привычки. Рассмотрим определение и свойства этих показателей. Показатель *MER* определяется как среднее число ошибочных пар:

$$MER = \frac{S + D + I}{H + S + D + I} = 1 - \frac{H}{N},$$

где $N = H + D + S + I$ — число пар символов.

В результате применения алгоритма динамического программирования к двум строкам, в общем случае, разной длины они как бы выравниваются друг относительно друга, образуя $H + D + S + I$ пар символов, из которых H корректных (образованных одинаковыми символами из первой и второй строк) и $D + S + I$ некорректных, образованных различающимися или вовсе не имеющими соответствия символами.

Показатель *WER* рассматривает априорную и распознанную транскрипции как набор независимых пар (аналогично показатель *MER*), являющихся реализациями двумерной случайной величины, и трактует точность распознавания как меру зависимости между компонентами этой случайной величины. В результате преобразований выводится формула, в которую входят только величины H, I, D и S (вследствие сделанных в ходе рассуждений в работе [4] допущений формула применима только в случаях, когда I, D и S существенно меньше H):

$$WIL = 1 - \frac{H^2}{N_1 N_2}.$$

Легко видеть, что показатели и *MER*, и *WIL* ограничены и принадлежат интервалу [0; 1]. Вычисление новых ПК не сложнее вычисления показателя *WER*. Исследуем теперь экспериментально свойства всех трех ПК согласно сформулированным в § 1 соображениям, для чего прежде всего составим методику эксперимента.

4. МЕТОДИКА ЭКСПЕРИМЕНТА

Для удобства представления мы будем оперировать ПК *Acc* (accuracy), *Mac* (match accuracy) и *WIP* (word information preserved): $Acc = 1 - WER$, $Mac = 1 - MER$, $WIP = 1 - WIL$.

Цель эксперимента состоит в исследовании свойств трех ПК при распознавании тестовой выборки как фонемной, так и словарной грамматиками. Для проведения измерений сконструируем систему распознавания русской речи, использующую для моделирования речевого сигнала СММ. Для исследования свойств ПК абсолютные значения точности несущественны. Поэтому ограничимся простейшими, монофонными СММ, не учитывающими контекст (каждая фонема моделируется одной и той же СММ, независимо от соседствующих фонем; исходя из этих же соображений, от моделирования контекстной зависимости часто отказываются при сравнении эффективности разных акустических признаков, т. е. способов параметрического описания речевого сигнала). Алфавит фонем основывается на фонетической транскрипционной системе русской речи Р.И. Аванесова [5], также применявшейся при разработке речевых баз ISABASE [6] и RuSpeech [7]. Для создания и обучения СММ используется свободно распространяемый инструментарий HTK [8].

Набор параметров СР состоит из числа t итераций обучения, объема a адаптационной речевой выборки (в секундах), объема n активного словаря распознавания (для словарных грамматик) и размера штрафа p за каждый распознанный символ. При распознавании величина p умножается на число символов в распознанной транскрипции и прибавляется к значению ее логарифма правдоподобия. Оптимальный выбор значения p позволяет найти баланс между числом ошибок вставки и удалений и минимизировать суммарное число ошибок. Таким образом, показатель качества имеет вид $Q(t, a, p, T)$ для фонемной грамматики и $Q(t, a, p, n, T)$ для словарной. Для заданных (t, a) или (t, a, n) существует оптимальное значение p, максимизирующее данный ПК. Однако оперировать этим параметром в виде функции $p(t, a)$ или

$p(t, a, n)$ довольно неудобно. Поэтому для удобства этот параметр обычно полагают постоянным. Вычислим для каждого из ПК погрешность, с которой можно принять $p(t, a) = p^*$, где $p^* = \text{const}$. Для этого можно аппроксимировать ПК суммой двух функций, одна из которых зависит от (t, a, n), а вторая от p:

$$Q(t, a, p, n) = Q_1(t, a, n) + Q_p(p) + E(t, a, p, n),$$

соответственно, $p^* = \arg\max_p Q_p(p)$.

Поскольку требуется найти только оптимальное значение p^*, можно воспользоваться формулой погрешности аппроксимации:

$$E(p) = \sqrt{\sum_{t, a, n} (Q^*(t, a, n) - Q(t, a, p, n))^2},$$

$p^* = \arg\min_p E(p)$, где $Q^*(t, a, n) = \max_p Q(t, a, p, n)$.

Для возможности сравнения значений погрешности $E(p^*)$ разных ПК нужно перевести их в относительную шкалу:

$$E(p) = \frac{2\sqrt{\sum_{t, a, n} (Q^*(t, a, n) - Q(t, a, p, n))^2}}{\sqrt{\sum_{t, a, n} (Q^*(t, a, n) + Q(t, a, p, n))^2}} \cdot 100\ \%. \quad (1)$$

Ошибка $E(t, a, p, n)$ также переводится в относительную шкалу:

$$E_{\text{отн}}(t, a, p, n) = \frac{E(t, a, p, n)}{\max\limits_{t,a,p,n} Q(t, a, p, n) - \min\limits_{t,a,p,n} Q(t, a, p, n)} \cdot 100\ \%.$$

Значения ПК измеряются для следующих значений параметров СР: $t \in [3; 72]$, $a \in (0, 60, 120, 300, 600, 1200)$, $p \in [-120; 0]$ (с шагом 0,25), $n \in (20, 40, 60, 100, 200)$. Тестовая выборка T состоит из 58 речевых записей суммарным объемом ~360 с. После того, как для каждого из ПК получена выборка $Q(t, a, p, T)$ или $Q(t, a, p, n, T)$, производится ее статистическая обработка, в ходе которой вычисляются (для каждого ПК отдельно):

— согласие каждой выборки $Q_{t,a,p}(T)$ ($Q_{t,a,p,n}(T)$) с нормальным распределением (по критерию χ^2 [9]);

— доверительный интервал для математического ожидания $Q(t, a, p, n)$;

— оптимальное значение штрафа p^* и соответствующая ему ошибка аппроксимации.

После этого параметр p фиксируется и исследуется выборка $Q(t, a, p^*, T)$ ($Q(t, a, p^*, n, T)$): вычисляется ошибка аппроксимации при разложении на функции независимых аргументов и степень согласия с теоретическим характером функции.

5. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Длина (в относительных единицах) доверительного интервала для математического ожидания всех трех ПК для фонемной грамматики почти не зависит от (t, a); независимость также и от p можно принять только для показателей *Mac* и *WIP*. График зависимости (для фонемной грамматики) длины доверительного интервала от p, усредненной по (t, a), представлен на рис. 1, *а*. Для всех значений (t, a, p) длина доверительного интервала для показателя *Mac* меньше той же величины для ПК *Acc* и *WIP* (т. е. нет необходимости проверять значимость соответствующей гипотезы). Практически полная независимость длины доверительного интервала от значений всех параметров СР, что можно считать справедливым для ПК *Mac* и *WIP* (для фонемной грамматики), но не *Acc*, означает, что при использовании тестовой выборки фиксированного размера точность оценки математического ожидания ПК не зависит от значений параметров СР. Зависимость (усредненной по (t, a, n)) длины доверительного интервала от p для словарной грамматики приведена на рис. 1, *б*. Длина интервала по-прежнему минимальна у ПК *Mac*, однако уже не является независимой от параметров СР величиной ни для одного ПК.

Графики оценки зависимости $Q_p(p)$, полученной в результате аппроксимации многомерной функции ПК одномерными, приведен на рис. 2. Максимум по p у ПК *Acc*, *Mac* и *WIP* для фонемной грамматики достигается при значениях −20, −15,25 и −16,75, соответственно (−74, −74, −74 для словарной грамматики). Зависимости ошибки $E(p)$, вычисленные по формуле (1), изображены на

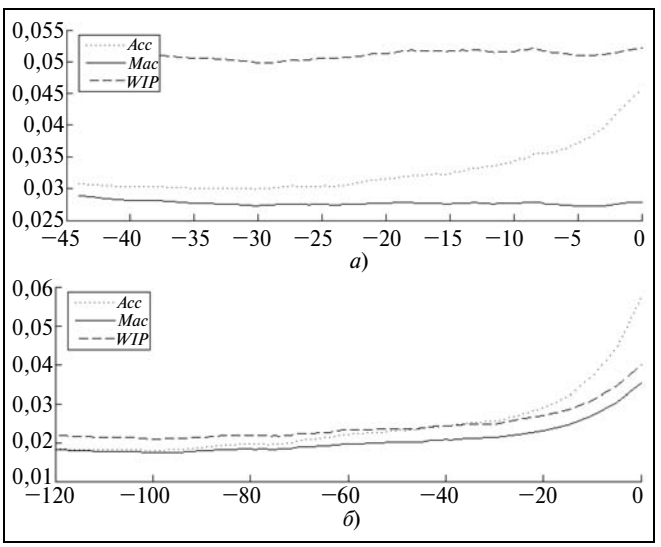

Рис. 1. Длина доверительного интервала в относительных единицах:
а — фонемная грамматика; *б* — словарная грамматика

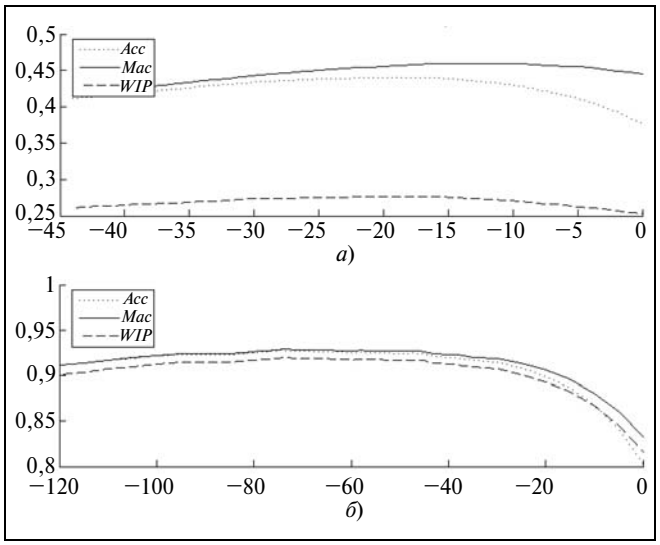

Рис. 2. График оценки зависимости $Q_p(p)$:
а — фонемная грамматика; *б* — словарная грамматика

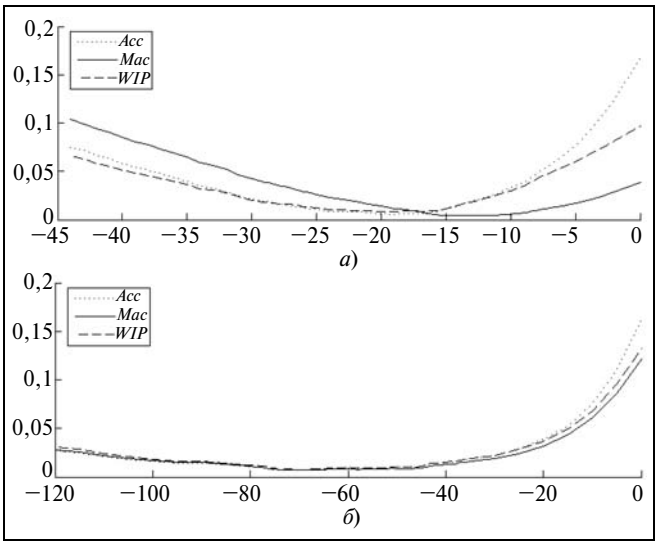

Рис. 3. Зависимости ошибки $E(p)$:
а — фонемная грамматика; *б* — словарная грамматика

рис. 3. Кроме точечной оценки p^*, вычислена также более надежная, интервальная: интервал $[p_1, p_2]$ такой, что значения $Q_p(p)$ или $E(p)$ (для первого или второго критерия оптимизации, соответственно) в этом интервале отличаются от оптимального не более чем на 1 %. Точечные и интервальные оценки p^* для ПК *Acc*, *Mac* и *WIP* и обеих грамматик, полученные двумя способами, приведены в табл. 1. Оценки, вычисленные двумя способами, мало различаются. В качестве итоговых значений p^* примем среднее значение границ интервальных оценок, являющихся пересечением интервалов, вычисленных двумя способами. Как видно из табл. 1, ошибка аппроксимации, вычисленная по формуле (1), минимальна у ПК *Mac*. Показатели *Acc* и *WIP* характеризуются большей ошибкой.

При вычислении согласия выборок $Q_{t,a,p}(T)$ ($Q_{t,a,p,n}(T)$) с нормальным распределением уровень значимости критерия ведет себя крайне нестабильно, даже будучи усредненным по параметрам *t*, *a* и *n* (рис. 4). Можно видеть, что в окрестности оптимальных значений *p* уровень значимости в среднем выше, чем вблизи граничных. Для примерного сравнения уровней значимости для различных ПК вычислим его среднее в окрестности оптимальных значений *p*. В качестве окрестностей возьмем вычисленные ранее интервальные оценки. Полученные результаты (см. табл. 1) позволяют заключить, что для фонемной грамматики степень согласия с нормальным распределением у ПК *Acc* и *Mac* примерно одинакова и несколько больше, чем у ПК *WIP*. Для словарной грамматики показатель *Mac* имеет преимущество перед показателями *Acc* и *WIP*. Ввиду нестабильности уровня значимости эта гипотеза для всех трех ПК должна приниматься с осторожностью.

Дальнейшие эксперименты производились с выборкой, где значение штрафа зафиксировано:

Таблица 1

Оценки оптимальных значений p^* и средний в окрестности p^* уровень значимости критерия χ^2

Показатель качества	Уровень значимости критерия χ^2	Оценки, вычисленные через аппроксимацию ПК одномерными функциями		Оценки, вычисленные по формуле (1)		
		Интервальная	Точечная	Интервальная	Точечная	Погрешность аппроксимации, %
Фонемная грамматика						
Acc	0,51	[−22; −15,5]	−20	[−20; −16,75]	−20	1,17
Mac	0,48	[−15,75;10,25]	−15,25	[−15,5; −10,75]	−15,5	0,81
WIP	0,39	[−20,5; −15,5]	−16,75	[−20; −16,25]	−20	2,52
Словарная грамматика						
Acc	0,12	[−76; −69]	−74	[−76; −64]	−74	0,65
Mac	0,30	[−76; −69]	−74	[−75; −64]	−69	0,53
WIP	0,13	[−76; −69]	−74	[−74; −65]	−74	0,64

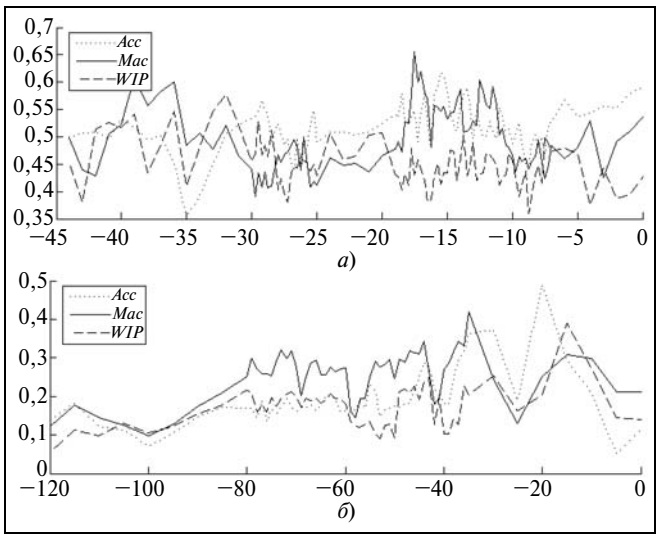

Рис. 4. Степень согласия выборки с нормальным распределением: *а* — фонемная грамматика; *б* — словарная грамматика

$p = p^*$. Разложение функций ПК на независимые функции отдельных аргументов t, a (t, a, n) дает примерно одинаковые ошибки для всех трех ПК (и обеих грамматик). Значения ошибок аппроксимации представлены в табл. 2. Число нарушений монотонности для трех ПК совпадает полностью (для преобразования в относительную шкалу эта величина разделена на общее число измерений, см. табл. 2). Таким образом, применение этого критерия позволяет лишь заключить, что новые ПК не хуже показателя *WER*. Поскольку результаты измерений одинаковы, проверять значимость не имеет смысла.

В целом по итогам эксперимента ПК *Mac* продемонстрировал наилучшие результаты. В сравнении с двумя другими ПК он характеризуется меньшей длиной доверительного интервала для математического ожидания и ошибки аппроксимации при выборе постоянного значения p^* штрафа. Помимо этого, для фонемной грамматики длина доверительного интервала не зависит от параметров СР.

Таблица 2

Ошибки аппроксимации одномерными функциями и относительное число нарушений монотонности в расчете на общее число измерений, %

Характеристика	Показатель		
	Acc	*Mac*	*WIP*
Фонемная грамматика			
Ошибка аппроксимации: $Q(t, a) = Q_t(t) + Q_a(a)$	20,1	16,9	14,0
Число нарушений монотонности	26,2		
Словарная грамматика			
Ошибка аппроксимации:			
$Q(t, a, n) = Q_t(t) + Q_a(a) + Q_n(n)$	67,6	67,1	69,3
$Q(t, a, n) = Q_1(t, a) + Q_n(n)$	34,0	33,3	
$Q(t, a, n) = Q_t(t) + Q_1(a, n)$	31,8	31,5	34,0
$Q(t, a, n) = Q_1(t, n) + Q_n(a)$	22,4	22,3	24,0
Число нарушений монотонности	23,8		

ЗАКЛЮЧЕНИЕ

Как уже отмечалось, цель этой работы состояла не в анализе, насколько правильно связывать качество алгоритмов распознавания речи с оптимальным выравниванием корректной и распознанной транскрипций относительно друг друга (этот вопрос следует считать открытым), а в выборе наилучшего из трех предложенных к настоящему времени показателей качества, базирующихся на метрике Левенштейна. Выполненное экспериментальное исследование свойств трех ПК позволяет сделать вывод о преимуществе метрики *MER* (Match Error Rate) перед метриками *WER* (Word Error Rate) и *WIL* (Word Information Lost) как при исследовании свойств алгоритмов распознавания, так и при вычислении итоговой оценки эффективности систем, ориентированных на интерпретацию речевых сообщений, и рекомендовать для использования именно метрику *MER*. Поскольку простейшая аппроксимация многомерных функций ПК одномерными не выявила существенных расхождений в ошибке (аппроксимации), в дальнейших исследованиях свойств ПК целесообразно применить более сложную модель комбинации одномерных зависимостей, включающую в себя суммирование и суперпозицию функций.

ЛИТЕРАТУРА

1. *Rabiner L.R.* A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognitions // Proc. of the IEEE. — 1989. — Vol. 77, N 2. — P. 257—286.
2. *Huang X.D., Acero A., Hon H.* Spoken Language Processing: a guide to theory, algorithm and system development. — New Jersey: Prentice-Hall, Inc., 2001.
3. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. АН СССР. — 1965. — Т. 163, № 4. — С. 845—848.
4. *Morris A.C., Maier V., Green P.D.* From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition // Proc. ICSLP. — 2004.
5. *Аванесов Р.И.* Русское литературное произношение. — М., 1972.
6. *База* речевых фрагментов русского языка «ISABASE» / Д.А. Богданов и др. // Интеллектуальные технологии ввода и обработки информации: Сб. тр. Ин-та системного анализа РАН. — М., 1998. — С. 74—85.
7. *http://www.cognitive.ru/innovation/voice-recog.htm* (дата обращения 26.05.2009).
8. *http://htk.eng.cam.ac.uk* (дата обращения 26.05.2009).
9. *Ван дер Варден Б.Л.* Математическая статистика / Под ред. И.В. Смирнова. — М.: Иностранная литература, 1960.

Статья представлена к публикации членом редколлегии В.Н. Новосельцевым.

Саакян Артем Александрович — гл. специалист,
ФГУП «ЦНИИ «Электроприбор», г. Санкт-Петербург,
г. Москва, ☎(812) 232-59-15, 499-78-01, ✉ ArtemSaak@mail.ru.