6Ш

УДК 681.142.66

В.Ю. Шелепов
Институт проблем искусственного интеллекта

Новые методы в пофонемном распознавании речи

Статья состоит из двух частей. Первая носит концептуальный характер и посвящена распознавателю изолированных слов, который не требует от пользователя предварительного обучения распознаванию каждого слова и основан на автоматическом построении транскрипций и эталонов слов распознаваемого словаря. Вторая часть посвящена описанию новых алгоритмов сегментации речевого сигнала, используемых в пофонемном распознавании.

В настоящее время в отделе фундаментальных проблем распознавания речевых образов Института проблем искусственного интеллекта развиваются два подхода к пофонемному распознаванию речи. Один из них использует автоматическую транскрипцию слов распознаваемого словаря и на ее основе автоматическое же построение эталонов из векторов пофонемной кодовой книги. (Будем в дальнейшем называть такие эталоны «транскрипционными»). Далее осуществляется распознавание по этим эталонам с помощью алгоритма DTW [1-3]. В словаре из 100 слов:

один, два, три, четыре, пять, шесть, семь, восемь, девять, ноль, разделить, сложить, вычесть, пробел, точка, открыть, закрыть, умножить, запятая, равно, неравно, кавычка, двоеточие, данные, перейти, если, цикл, стоп, читать, печатать, пробить, писать, конец, пространство, перемотать, назад, истинно, ложно, описание, индекс, оператор, нет, или, также, пометить, больше, меньше, программа, подпрограмма, функция, реальный, целый, точный, комплексный, логический, вхождение, внешний, вызвать, возвратить, размер, общий, эквивалентно, кодировать, экспонента, логарифм, синус, косинус, корень, абсолют, арктангенс, тангенс, остаток, минимум, цифра, латинский, русский, ввести, степень, градус, минута, секунда, влево, вправо, вверх, вниз, слушай, интеграл, радиан, квадрат, куб, натуральный, арксинус, массив, начало, примечание, иначе, собственный, шаг, значение, пока

такой распознаватель в 50% случаев правильно определяет произнесенное слово. В остальных случаях он ставит его в списке кандидатов на распознавание, выводимом в порядке увеличения DTW-расстояний, не далее, чем на десятое место. В нашей программе заложена возможность с помощью двойного щелчка мыши на нужном слове этого списка заменить им результат распознавания в окне, где набирается текст, и одновременно заменить ранее существовавший транскрипционный эталон эталоном, построенным по результату произнесения. Такой эталон будем называть «голосовым». Использование голосового эталона, как показывает опыт, обеспечивает в дальнейшем распознавание рассматриваемого слова, близкое к 100-процентному. Подобные же закономерности

наблюдались нами при работе со словарем из 2632 наиболее употребительных русских слов и словарем из 5500 слов для набора математических текстов. Изложенное позволяет сформировать систему голосового набора текста типа «Voice Type» (голосовая пишущая машинка), которая не требует от пользователя предварительного обучения путем наговаривания каждого слова распознаваемого словаря. Он сразу может приступать к набору нужного ему текста, заменяя по ходу транскрипционные эталоны голосовыми и тем самым постоянно улучшая уровень распознавания. Для большей наглядности слова, для которых уже созданы голосовые эталоны, выделяются в вышеупомянутом списке кандидатов красным цветом. После того как все слова словаря «покраснеют», пользователь будет иметь систему с весьма высоким качеством распознавания.

Первоначальный распознаватель с транскрипционными эталонами можно рассматривать как систему, которая осуществляет предварительное распознавание, заменяя в каждом случае полный словарь во много раз меньшим списком ближайших кандидатов на распознавание. Далее можно пойти путем создания каскада таких распознавателей, которые используют различные системы признаков. Хорошо также показало себя применение к списку кандидатов распознавателя, основанного на сегментации записанного слова с последующим распознаванием выделенных фонем. Некоторые аспекты, связанные с этим распознавателем затронуты в статье [4]. Важнейшую роль в его работе играет успешная сегментация записанного речевого сигнала, то есть разбиение его на участки, отвечающие отдельным аллофонам.

В статье [5] описаны некоторые подходы к проблеме сегментации речевого сигнала. Полагаем, что за время, прошедшее после публикации этой работы мы существенно продвинулись к цели. Теперь мы имеем систему, которая с весьма высокой надежностью выделяет все границы между отрезками, отвечающими последовательным аллофонам, возможно, проставляя при этом небольшое количество лишних меток. Переходя к описанию соответствующих алгоритмов, обратим внимание на то, что для правильной работы тех из них, которые предназначены для выделения пауз, важно пользоваться 8-битной оцифровкой сигнала.

Пусть

$x_k = x(k),\ k = 1, 2, ..., 10000$ –

последовательность значений 8-битной оцифровки речевого сигнала с частотой дескритизации 20050 Гц.

Мы начинаем с разбиения сигнала на квазипериоды. Эта процедура описана в работе [6]. Она основана на минимизации функции:

$$L(k) = \sum_{i=0}^{k-1} \left| x(\boldsymbol{i}_0 + i + k) - x(\boldsymbol{i}_0 + i) \right|, \qquad (1)$$

по всем k таким, что

$Min \le k \le Max$.

Здесь Min и Max – числа, которые заранее определяются в соответствии с высотой голоса диктора. Для автора этой статьи при оцифровке сигнала с частотой 22050 Гц высота основного тона соответствует длине квазипериода приблизительно равной 150 отсчетам. Поэтому в нашем случае мы полагали

Min=60, Max=200. Для дикторов с низким басом или дикторов женщин фигурирующие здесь и ниже конкретные числа должны быть изменены.

Условимся для простоты все полученные отрезки называть «квазипериодами» (хотя для шипящих и пауз они таковыми не являются).

Нормируем каждый квазипериод по амплитуде. Если

$$x_1, x_2, \ldots x_n -$$

последовательность значений сигнала на квазипериоде, то полагаем:

$$M = \max(x_1, x_2, \ldots x_n), \; m = \min(x_1, x_2, \ldots x_n),$$

$$y_i = \frac{x_i - m}{M - m}, \qquad i = 1,2,\ldots,n$$

Если мы имеем 2 соседних квазипериода одинаковой длины, то вычислим величину:

$$\sum_{i=1}^{n} |y_{i+n} - y_i|,$$

которая характеризует насколько один квазипериод отличается от другого. В случае отличия квазипериодов по длине, описанную величину строим аналогично, обозначая через n длину более короткого квазипериода. Вычислив такие величины для всех последовательных пар рассматриваемых квазипериодов, места смены аллофонов следует искать среди отсчетов, которым соответствуют локальные максимумы полученной последовательности. Пусть это отсчеты:

$$n_1, n_2, \ldots n_L \qquad (2)$$

Сигнал разбивается на блоки, границами которых служат отсчеты с номерами (2).

Далее, на каждом из блоков подсчитывается число нестрогих минимумов исходного сигнала l и строится величина:

$$a = \frac{n}{l}, \qquad (3)$$

где n-полное число отсчетов в блоке. Поскольку произнесение глухих взрывных звуков «п, к, т» связано с кратковременным перекрыванием ротовой полости, соответствующие участки речевого сигнала содержат отрезки «пауз» – кратковременное отсутствие речи. При 8-битной записи характерным свойством пауз является большое количество участков постоянства сигнала, так что в формуле (3) число l не на много меньше числа n. В результате приходим к следующему критерию:

Если для блока выполняется неравенство:

$$a < 2,$$

то относим его к паузе. Отметим, что этот критерий может быть использован в модуле записи для определения начала и конца речевого сигнала.

Для выделения шипящих сигнал обрабатывается высокочастотным фильтром с частотой среза 1500 Гц. Пусть

$$x_k^{'}, \quad k = 1, 2, \ldots, 10000 -$$

последовательность значений профильтрованного сигнала. Нумеруя отсчеты в пределах блока от 1 до n, строим величину:

$$b = \frac{\sum_{i=1}^{n} |x_i - 127|}{\sum_{i=1}^{n} |x'_i - 127|}. \qquad (4)$$

Она характеризует изменение энергии блока при указанной фильтрации. Характерным свойством шипящих является то, что для них изменение энергии при указанной фильтрации не велико и мы получаем следующий критерий.

Если для блока выполняется неравенство:

$b < 2$,

то относим этот блок к шипящей. В начале и в конце цепочки блоков, относящихся к шипящим, ставим метки. То же делаем для цепочки блоков, относящихся к паузе.

Выделив шипящие и паузы, переходим к сегментации голосовых звуков. Для этого рассматриваем значения сигнала, соответствующие последовательности (2):

$x_{n_1}, x_{n_2}, \ldots, x_{n_L}$.

Находим в ней все локальные максимумы, не относящиеся к выделенным отрезкам шипящих и пауз, и в соответствующих местах сигнала проставляем метки, считая их границами между участками голосовых аллофонов.

Далее распознаватель должен перейти к определению фонем, отвечающих отрезкам сигнала, выделенным при сегментации. Методам, которые мы при этом используем, будет посвящена отдельная статья. Распознаванию слова по полученной цепочке фонем посвящена статья [4].

Программную реализацию описанных в данной работе подходов и алгоритмов осуществили В.А. Грабовая, Д.Г. Старушко, Е.Е. Федоров.

Литература

1. Грабовая В.А., Федоров Е.Е., Шелепов В.Ю. О системе распознавания русской речи с автоматическим построением эталонов // Искусственный интеллект. – 2000. – № 1.
2. Дорохин О.А., Федоров Е.Е., Шелепов В.Ю. Некоторые подходы к пофонемному распознаванию русской речи и распознаванию больших словарей // Искусственный интеллект. – №2. – 2000. – С. 329-333.
3. Дорохин О.А., Засыпкин А.В., Червин Н.А., Шелепов В.Ю. О некоторых подходах к проблеме компьютерного распознавания устной русской речи // Труды Международной конференции «Знания, диалог, решение». – Том 1. – Ялта. – 1997. – С.234-240.
4. Божко Д.В., Грабовая В.А., Шелепов В.Ю. Интерпретатор распознанной цепочки фонем, которая может содержать ошибки // Искусственный интеллект. – 2001. – № 3.
5. Дорохин О.А., Старушко Д.Г., Федоров Е.Е., Шелепов В.Ю. Сегментация речевого сигнала // Искусственный интеллект. – № 3. – 2000. – С. 450-458.
6. Федоров Е.Е., Шелепов В.Ю. Защита речевых распознавателей от шума и посторонней речи // Искусственный интеллект. – 2001. – № 3.

Материал поступил в редакцию 11.07.01.