М.Х. КАРАБАЛАЕВА

СИСТЕМА РАСПОЗНАВАНИЯ ЦЕЛЫХ СЛОВ С ИСПОЛЬЗОВАНИЕМ ДИНАМИЧЕСКОГО ВЫРАВНИВАНИЯ ВРЕМЕНИ

В статье описана система распознавания целых слов, позволяющая с высокой точностью распознавать отдельно произнесенные слова из ограниченного словаря путём сравнения их с эталонными образцами.

Автоматическое распознавание устной речи – традиционная задача искусственного интеллекта. Ею начали с энтузиазмом заниматься еще на заре возникновения информатики как науки так же, как и задачей автоматического перевода с одного языка на другой. Однако по прошествии нескольких десятилетий результаты многочисленных исследовательских групп и в той, и в другой области остаются довольно скромными. Две ключевые задачи распознавания речи – достижение стопроцентной точности на ограниченном наборе команд хотя бы для одного дикторского голоса и независимое от диктора распознавание произвольной слитной речи с приемлемым качеством – не решены, несмотря на полувековую историю их разработки.

Главная особенность речевого сигнала в том, что он очень сильно варьируется по многим параметрам: длительность, темп, высота голоса, искажения, вносимые большой изменчивостью голосового тракта человека, различными эмоциональными состояниями диктора, сильным различием голосов разных людей. Два временны*х* представления одного и того же фрагмента речи даже для одного и того же человека, записанные в разное время, не будут совпадать. Необходимо искать такие параметры речевого сигнала, которые, с одной стороны, полностью бы его описывали (т.е. позволяли бы отличить одну речевую единицу от другой), и с другой стороны, нивелировали бы указанные выше вариации речи. Затем эти параметры должны сравниваться с образцами, причем это должно быть не простое сравнение на совпадение, а поиск наибольшего соответствия.

Таким образом, процедура распознавания речи должна основываться на использовании подходящей системы параметров (признаков) и выполняться с помощью разумных алгоритмов. В этой статье в сжатом виде описана *система распознавания целых слов*, позволяющая с довольно высокой точностью распознавать отдельно произнесенные слова из ограниченного словаря путём сравнения их с эталонными образцами. В качестве распознаваемых единиц речи в системе используются целые слова, каждое из которых рассматривается как неделимое целое.

1. Организация записи речевого сигнала, определение его начала и конца.

Звуковой (в частности, речевой) сигнал, оцифрованный звукозаписывающим устройством, представляет собой массив *отсчетов* (*сэмплов*) – замеренных с некоторым временным шагом значений напряжения на выходе микрофона. Если пренебречь погрешностью квантования и зависимостью получаемого цифрового сигнала от характеристик микрофона и звуковой карты, то можно рассматривать речевой фрагмент как дискретную функцию амплитуды сигнала от времени. Звуковой сигнал можно визуализировать, выведя график этой функции на монитор.

Произнесем в микрофон и запишем речевой фрагмент с частотой дискретизации 22050 Гц и разрядностью квантования 8 бит, так что его значения могут иметь $2^8 = 256$ градаций: от 0 до 255.

Прежде всего, нам нужно определить, на каком участке записанного сигнала начинается речь, и где она заканчивается; иными словами, отделить звучащую речь от «тишины» (паузы, фонового шума). Для этого воспользуемся величиной

$$V = \sum_{i=1}^{n} |x_{i+1} - x_i|, \qquad (1)$$

представляющей собой численный аналог *полной вариации функции* для дискретного случая. Здесь n – количество отсчетов на участке сигнала, x_i – значение i-го отсчета, $0 \le x_i \le 255$.

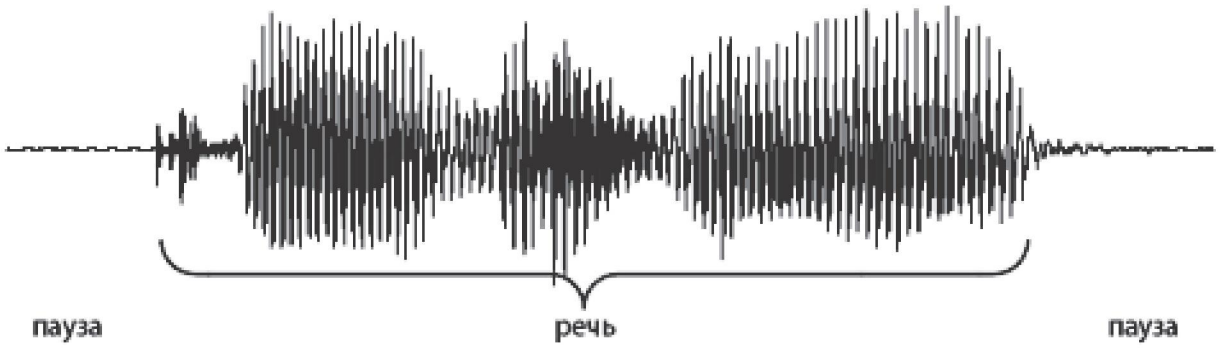

Рис.1 Определение начала и конца речевого сигнала

Назовем *точками постоянства* такие моменты времени, для которых в следующий момент величина сигнала остается неизменной:

$x_i = x_{i+1} \Rightarrow i$ – точка постоянства.

Остальные моменты времени назовем точками непостоянства.

И воспользуемся тем фактом, что для паузы, по сравнению с речью, характерно большое количество точек постоянства [1].

Запишем звук «тишины» – некоторое количество, например, 30000 отсчетов фонового шума – и проанализируем записанный сигнал. Разобьем его на последовательные отрезки по 256 отсчетов в каждом. Для каждого из них вычислим отношение

$$V/C, \qquad (2)$$

где $V = \sum_{i=1}^{256} |x_{i+1} - x_i|$ – численный аналог полной вариации, C – число точек постоянства.

Определим значение, наиболее часто встречающееся в массиве величин (2), как наиболее характерное значение для «тишины», с учетом используемой звуковой карты и акустических условий студии. Для уверенности увеличим это значение на 0,1 и запишем его в управляющий файл, например, под именем *Начальный Порог*.

Учитывая, что согласные звуки имеют длительность от 10 мс, а гласные от 30 мс [2], мы можем считать началом звучания речи тот участок звукового сигнала, на котором величина (2) впервые превысит *Начальный Порог*, скажем, не менее 5-ти раз подряд. (5 отрезков по 256 отсчетов при частоте дискретизации 22050 гц будут длиться около 60 мс, т.е. достаточно долго, чтобы отличить звук речи от случайного кратковременного шума.)

Кроме того, нам нужно также определить и *Конечный Порог*, т.е. значение величины (2), определяющее конец звучания речи. Речевой сигнал затухает долго, ему может сопутствовать «послезвучие», поэтому конец его звучания определить гораздо сложнее, чем начало [1]. Для упрощения этой задачи положим *Конечный Порог* в 10 раз бо́льшим, чем *Начальный Порог*.

Теперь организуем запись речевого сигнала следующим образом.

По нажатии кнопки записи система начинает записывать звуковой сигнал, поступающий с микрофона, и вычислять для последовательных отрезков по 256 отсчетов величину (2). Система фиксирует момент, после которого эта величина впервые не менее 5-ти раз подряд превышает *НачальныйПорог*. Начиная с этого момента, система заносит отсчеты в массив M вплоть до момента, после которого на протяжении 10000 отсчетов величина (2) окажется меньше, чем *Конечный Порог*. После этого запись останавливается. Сформированный массив отсчетов M и есть цифровое представление нашего речевого фрагмента, с которым мы продолжим работать (рис.1).

2. Предварительная обработка речевого сигнала

Так как мы собираемся распознавать отдельные слова из ограниченного словаря, оценим их длительность. Общеупотребительные слова короткой и средней длины (например, числительные 0-9 или голосовые команды из одного слова), произнесенные в среднем темпе речи, укладываются в полсекунды звучания [3].

Пусть речевой фрагмент, вводимый с микрофона в течение 0,5 секунд, оцифрован с частотой дискретизации 22050 Гц. Массив M содержит 10000 сэмплов:

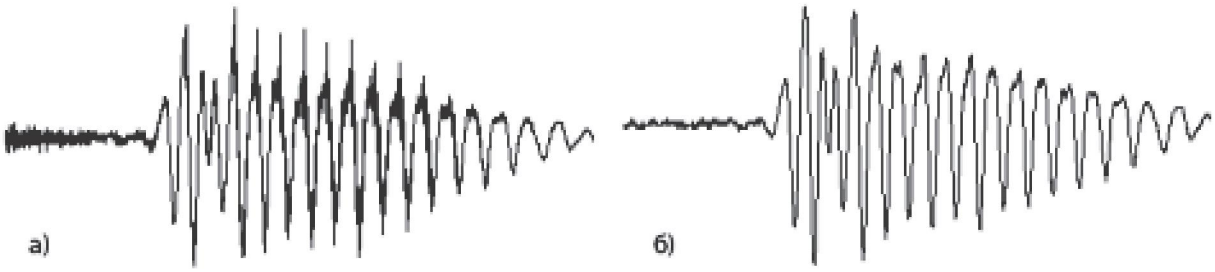

Рис.2 Речевой сигнал до (*а*) и после (*б*) обработки «сглаживающим» фильтром

$$x_1, x_2, \ldots, x_{10000} \qquad (3)$$

Назовем *сигналом* функцию

$$x(i) = x_i \qquad (4)$$

Чтобы избавиться от неинформативных при распознавании высокочастотных составляющих, используем для «сглаживания» сигнала линейный 3-точечный фильтр:

$$x_i = \frac{x_{i-1} + x_i + x_{i+1}}{3}, \; i = 2,3,\ldots,9999 \qquad (5)$$

Подвергнем сигнал последовательному 10-кратному сглаживанию. Если теперь сравнить его с исходным (рис. 2), можно отметить, что в «сглаженном» сигнале разборчивость речи сохранилась, однако он в значительной степени очистился от индивидуального тембра диктора (что можно считать шагом в направлении дикторонезависимости системы распознавания) [3].

В дальнейшем под *сигналом* мы будем подразумевать именно этот 10-кратно «сглаженный» сигнал.

3. Построение системы признаков, представление слова.

Как было сказано в начале, распознавание речи должно основываться на использовании подходящей системы признаков и выполняться с помощью разумных алгоритмов. Построение удачной, эффективной, сбалансированной системы признаков – ключевой момент в создании любой системы распознавания речи. В основе нашей системы признаков лежит понятие *полного колебания*.

Пусть l – число отсчетов между двумя соседними локальными максимумами сигнала (рис.3).

Определим величину z:

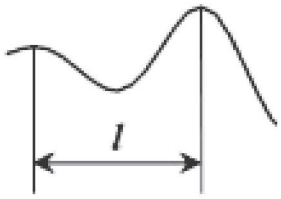

Рис. 3. Полное колебание

$$\begin{cases} z = l, 2 \le l < 20, \\ z = 20 + \dfrac{l-20}{6}, 20 \le l < 50, \\ z = 25 + \dfrac{l-50}{10}, 50 \le l < 90, \\ z = 29, l \ge 90. \end{cases} \qquad (6)$$

Ближайшее целое число, не превосходящее z, назовем *длиной* соответствующего *полного колебания*. Таким образом, длина полного колебания учитывается тем более точно, чем оно короче.

Выделим произвольный участок сигнала и обозначим: N – общее число полных колебаний на этом участке; N_1 – число полных колебаний длины 2 отсчета; N_2 – число полных колебаний длины 3 отсчета; N_{28} – число полных колебаний длины 29 отсчетов.

Поставим в соответствие выделенному участку *вектор признаков*

$$(y_1, y_2, \ldots, y_{28}, \varepsilon), \qquad (7)$$

где $y_k = \dfrac{N_k}{N}$, k = 1, 2, …, 28; ε – отношение амплитуды (разности наибольшего и наименьшего значений) рассматриваемого участка сигнала к амплитуде всего сигнала. Величина ε вводится

для того, чтобы надежно отделить паузу, возникающую внутри слова перед взрывными согласными, от значащей части сигнала, а нормировка ее нужна, чтобы отвлечься от громкости произносимого [3].

Теперь используем описанную систему признаков для представления нашего записанного «сглаженного» сигнала в 10000 отсчетов. Разобьем сигнал на отрезки по 368 отсчетов в каждом (368 отсчетов при частоте дискретизации 22050 гц – приблизительно столько составляет удвоенный квазипериод основного тона для мужского голоса средней высоты). Для каждого из 27-ми полных отрезков вычислим вектор (7). Последний неполный отрезок отбросим. В результате мы получим последовательность 27-ми–29-мерных векторов:

$$A = (a_1, a_2, ..., a_{27}).$$

Эта последовательность представляет слово в нашей системе.

4. Распознавание слов по эталонам. Алгоритм DTW

Распознавание «чистых», произнесенных изолированно фонем мало что дает для распознавания слов. Это связано с тем, что по сути дела фонема – это, с акустической точки зрения, некая абстракция. Артикуляторные органы человека обладают инерцией. Их положение в данный момент в значительной степени определяется их конфигурацией в предшествующий и последующий моменты времени. Поэтому реализация фонемы в конкретной речевой ситуации очень сильно зависит от ее окружения. Отсюда следует, что перспективен подход к распознаванию слова как целого. Вместе с тем, пофонемное распознавание, которое является более сложным, все-таки возможно, а при распознавании больших словарей и необходимо [3]. Однако в этой статье мы рассматриваем задачу распознавания слова как целого.

Пусть некоторая реализация слова принимается за эталон. Как изложено выше, мы представляем ее в виде набора 27-ми – 29-мерных векторов:

$$E = (e_1, e_2, ..., e_{27}) \qquad (8)$$

Создадим такой эталон для каждого слова из распознаваемого словаря. Для словаря из k слов получим систему эталонов $E_1, E_2, ..., E_k$.

Пусть

$$A = (a_1, a_2, ..., a_{27}) \qquad (9)$$

– представление слова, которое подлежит распознаванию.

Теперь нам нужно разумным образом определить *расстояние* между двумя наборами вида (8), (9), с тем чтобы, вычислив расстояние от распознаваемого слова до всех эталонов, объявить результатом распознавания то слово из словаря, эталон которого является ближайшим.

Кажется естественным определить расстояние между наборами A и E как сумму расстояний между их соответствующими векторами. При этом за расстояние между векторами можно выбрать, например, сумму модулей разностей соответствующих координат (-метрика).

Однако это нецелесообразно по следующей причине. Расстояние между двумя реализациями одного и того же слова должно быть минимальным. Чтобы так оно и было, при вычислении расстояния нужно было бы сравнивать между собой вектора, относящиеся к одинаковым фонемам. Но темп произнесения слова может быть различным. Более того, он может меняться на протяжении слова (можно сказать «мама», а можно сказать «ма-ама»).

Разрешить эту трудность позволяет известный *алгоритм DTW* (Dynamic Time Warping, «динамическое выравнивание времени»).

Обозначим расстояние между векторами e_i и a_j наборов (8), (9) через и для всех определим величину — расстояние между частью сигнала (8) от начала до i-го вектора включительно и частью сигнала (9) от начала до j-го вектора включительно:

$$\begin{aligned} C_{11} &= D_{11} \\ C_{i1} &= C_{i-1,1} + D_{i1} \\ C_{1j} &= C_{1,j-1} + D_{1j} \\ C_{ij} &= D_{ij} + \min(C_{i-1,j}, C_{i,j-1}, C_{i-1,j-1}) \end{aligned} \qquad (10)$$

Получается типичный алгоритм динамического программирования.

Тогда расстояние между полными сигналами (8) и (9) определяется как $C_{27,27}$. Для того, чтобы понять смысл этого определения, обозначим символом ~ отношение соответствия меж-

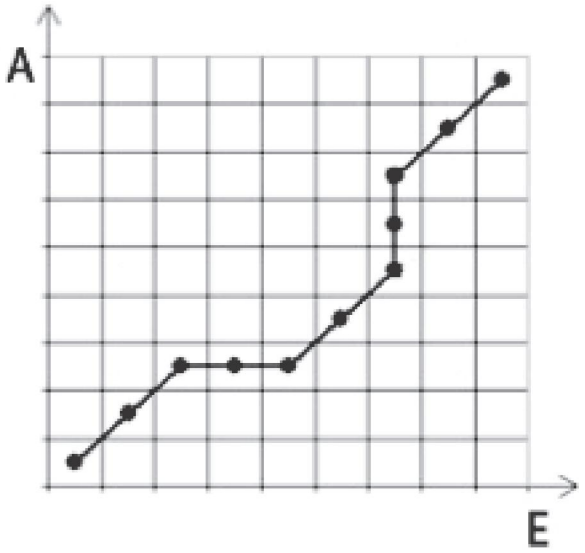

Рис. 4. Пример соответствия между векторами

ду вектором из набора (8) и вектором из набора (9), которое определяется следующим образом:

$$e_{27} \sim a_{27}.$$

Далее, если $e_i \sim e_j$, то в том случае, когда минимум в (10) есть $C_{i-1, j-1}$, полагаем

$$e_{i-1} \sim a_{j-1};$$

если минимум есть $C_{i, j-1}$, полагаем

$$e_i \sim a_{j-1};$$

если минимум есть $C_{i-1, j}$, полагаем

$$e_{i-1} \sim a_j.$$

На рис.4 приведен пример соответствия: центры квадратов, отвечающих соответствующим векторам, соединены прямолинейными отрезками. Наличию вертикального отрезка соответствует случай, когда несколько векторов набора A соответствуют одному вектору набора E. Наличию горизонтального отрезка соответствует случай, когда несколько векторов набора E соответствуют одному вектору набора A.

DTW-расстояние между наборами (8) и (9) определяется по формуле (10) при $i = j = 27$. При вычислении этого расстояния суммируются только расстояния между соответствующими векторами этих наборов.

Таким образом, алгоритм DTW обеспечивает выравнивание акустически наиболее близких кусков сигнала и их сравнение. Распознавание с помощью этого алгоритма сводится к вычислению DTW-расстояния от распознаваемого слова до каждого эталона, после чего результатом распознавания объявляется то слово из словаря, эталон которого оказывается ближайшим. Отметим, что при данном подходе в системе не предусмотрен отказ от распознавания, так как один из эталонов обязательно окажется ближайшим для распознаваемого слова.

Описанная система распознавания целых слов демонстрирует весьма удовлетворительные результаты для словаря размером до 1000 слов. При этом вся процедура подстройки ее под диктора заключается в единоразовом создании речевого эталона для каждого слова из словаря. Это позволяет использовать систему автономно, например, для распознавания голосовых команд, либо интегрировать ее отдельным модулем в более сложную систему пофонемного распознавания для повышения точности результатов.

ЛИТЕРАТУРА

1. *Федоров Е.Е., Шелепов В.Ю.* Автоматическое определение начала и конца записи речи // Материалы Междунар. конф. «Искусственный интеллект – 2002». – Т. 2. – Таганрог: Изд. ТРТУ. – 2002. – С. 44-47.

2. *Алдошина И.* Основы психоакустики // Звукорежиссер. – 2002. – № 5.

3. *Козлов А.В., Саввина Г.В., Шелепов В.Ю.* Система пофонемного распознавания отдельно произносимых слов // Искусственный интеллект. – 2003. – № 1. – С. 156-165.

Резюме

Бұл мақалада бүтін сөздерді тану жүйесі сипатталған. Ол шектелген сөздіктегі жеке айтылған сөздерді эталонды үлгілермен салыстыру арқылы жоғарғы дәлдікпен тануға мүмкіндік береді.

Summary

In the present article a system of isolated words recognition is described. It allows effective recognition of separately uttered words from a limited dictionary by comparison with patterns.

УДК 004.432.4

Евразийский национальный университет им. Л.Н. Гумилева *Поступила 23.09.09 г.*