Подход к решению задачи членения слитной речи на речевые единицы

И.А. Андреев[а], А.И. Армер[а], Н.А. Крашенинникова[б], В.С. Мошкин[а]

[а] *Ульяновский государственный технический университет, 432027, ул. Северный венец, 32, Ульяновск, Россия*
[б] *Ульяновский государственный университет, 432970, ул. Льва Толстого, 42, Ульяновск, Россия*

Аннотация

В статье описывается алгоритм членения речевого сигнала на речевые единицы – фонемы, их сочетания и паузы. Алгоритм основан на преобразовании речевого сигнала в особое двумерное изображение – автокорреляционный портрет. Для определения границ речевых единиц производится совмещение портретов анализируемого сигнала и эталонных портретов каждой речевой единицы. При совмещении используется метод динамического программирования, позволяющий получить оптимальное расстояние между портретами.

Ключевые слова: речевой сигнал; сегментация; автокорреляционный портрет; речевые единицы; дискретное динамическое программирование.

1. Введение

В настоящее время довольно востребованы алгоритмы членения слитной речи на составляющие речевые единицы – фонемы, их сочетания и паузы. Эта задача возникает, например, при создании систем для исследования, обработки, моделирования и автоматического распознавания речи. Для возможности применения таких систем в различных акустических условиях к ним предъявляются повышенные требования по устойчивости к акустическим шумам и искажениям речевого сигнала. В статье предлагается способ определения границ речевых пауз и речевых единиц, соответствующих таблице SAMPA+ русского языка [1]. Используемые в предлагаемом способе алгоритмы преобразования и обработки речевого сигнала дают возможность предъявлять к нему повышенные требования по устойчивости к акустическим шумам и искажениям речевого сигнала.

2. Объект исследования

Задача членения или сегментации речевого сигнала на составляющие является чрезвычайно сложной, и в настоящее время не найдено простого решения для общего случая. В литературе [2] отмечается, что существуют отдельные случаи, для которых точная сегментация затруднительна.

Для членения речевого сигнала, содержащего слитную речь, используют различные подходы [2, 3, 4], среди которых выделяются основанные на анализе спектральных характеристик, траектории энергии сигнала, логарифма энергии, числа переходов через ноль и статистических параметрах речевых единиц. Известные подходы дают хорошие результаты в хороших акустических условиях, а при наличии шумов результаты ухудшаются. Кроме того, длительность речевого сигнала изменяется от произнесения к произнесению, что также затрудняет его членение на составляющие. В статье предлагается для повышения устойчивости к шумам определения границ речевых единиц использовать автокорреляционное преобразование [5] речевого сигнала в двумерное изображение и соответствующие способы совмещения изображений. Автокорреляционное преобразование имеет ряд свойств, делающих его в некоторой степени помехоустойчивым [6], это даёт возможность предполагать меньшую зависимость предлагаемого способа членения речевого сигнала от акустических условий произнесения. Использование при совмещении двумерных изображений речевых сигналов метода дискретного динамического программирования [7] позволяет повысить устойчивость предлагаемого способа к изменениям длительности произнесения речевых единиц.

3. Алгоритм определения границ речевых единиц

1.1. Общий алгоритм

Алгоритм определения границ речевых единиц в участке слитной речи следующий: речевой сигнал, содержащий фрагмент слитной речи, анализируемой для определения в ней границ речевых единиц, представлен в виде цифровых отсчётов. Также в виде цифровых отсчётов представлены эталоны каждой речевой единицы. Для подготовки эталонов каждый пример соответствующей речевой единицы из таблицы SAMPA+ произносится диктором, затем в нём на слух определяются границы, и речевая единица образует эталон. Цифровые отсчёты анализируемого участка слитной речи и отсчёты каждой из эталонных речевых единиц с помощью автокорреляционного преобразования преобразуются в особые двумерные изображения - автокорреляционные портреты (АКП). Портреты анализируемого участка речи и каждой эталонной речевой единицы для возможности дальнейшего совмещения имеют одинаковую длину строки.

Далее для определения границ речевых единиц происходит совмещение портрета анализируемого участка речи с каждым из портретов эталонных речевых единиц. Для этого в скользящем окне размером в количество строк портрета соответствующей речевой единицы производится вычисление расстояния [8]. В процессе вычисления расстояние между окнами оптимизируется с использованием метода дискретного динамического программирования. Для каждой речевой единицы определяется массив расстояний вдоль портрета анализируемого участка речи. Расстояния, соответствующие одинаковым фрагментам портрета анализируемого участка речи, сравниваются между собой. В результате, портреты речевых единиц, давшие наименьшее расстояние своими размерами образуют искомые границы. Если наименьшее расстояние дают портреты одинаковых речевых единиц, следующие друг за другом, то они объединяются в границы одной речевой единицы.

1.2. Автокорреляционные портреты речевых сигналов

Из-за того, что автокорреляционные связи достаточно информативны, то есть отражают характерные свойства речевых сигналов, АКП индивидуальны для каждой речевой единицы, это определяет хорошие результаты при оконном поиске границ речевых единиц в участке слитной речи. В [9] АКП строятся следующим образом. Пусть $s(i)$ — -й отсчет оцифрованного речевого сигнала; $s(i+k)$ — отсчет, отстоящий от $s(i)$ на k отсчетов. Степень зависимости этих отсчетов выражается выборочным коэффициентом корреляции:

$$R_s(k) = R[s(i), s(i+k)] = \frac{\mathrm{cov}[s(i),s(i+k)]}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2(i) - m^2_{s(i)}}\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2(i+k) - m^2_{s(i+k)}}},$$

$$\mathrm{cov}[s(i), s(i+k)] = \frac{1}{N}\sum_{i=1}^{N} s(i)s(i+k) - \left[\frac{1}{N}\sum_{i=1}^{N} s(i)\right]\frac{1}{N}\sum_{i=1}^{N} s(i+k), \quad (1)$$

где N — число отсчетов отрезка, на котором ищется зависимость; $\mathrm{cov}[s(i), s(i+k)]$ — выборочная ковариация $s(i)$ и $s(i+k)$ при $i = 1..N$; $m_{s(i)}$ — выборочное среднее $s(i)$ при $i = 1..N$; $m_{s(i+k)}$ - выборочное среднее $s(i+k)$ при $i = 1..N$. Функция, определенная выборочными коэффициентами корреляции по (1), является автокорреляционной функцией (АКФ) сигнала. С помощью вычисления АКФ осуществляется преобразование отсчетов РС $s(i)$ $i = 1..M$ (M — число отсчетов в речевом сигнале) в двумерное изображение. Для этого $s(i)$ разбивается на отрезки по $N < M$ отсчетов, далее, в каждом j −м $(j = 1, N, 2N, \dots, M - 2N)$ отрезке ищется локальный максимум сигнала $i_m^j = \max|s|$. Будем считать, что M делится на N нацело, иначе отбросим оставшиеся отсчеты от конца РС. Далее с помощью соотношения (1) вычисляем элементы соответствующей строки АКП начиная от $i_m^j (j = 1, N, 2N, \dots, M - 2N)$ и составляем из них строки АКП:

$$R\left[s\left(i_m^j\right), s\left(i_m^j + k\right)\right]_{j=1,N,2N,\dots,M-2N}^{k=1..N},$$
$$X(j,k) = R. \quad (2)$$

Полученное по формуле (2) двумерное изображение $X(j,k)$, где j − номер строки, k − номер столбца, является АКП речевого сигнала $s(i)$ с размерами $N \times \left(\frac{M}{N} - 2\right)$, построенным по локальным максимумам речевого сигнала. Отметим, что АКП, построенные по локальным максимумам индивидуальны для каждой речевой единицы, и в силу привязки к локальным максимумам речевого сигнала менее подвержены геометрическим искажениям связанным с изменчивостью речи. На рисунке 1 представлены АКП речевых единиц [“a], [o], [n`:], [f] (SAMPA+).

1.3. Совмещение автокорреляционных портретов с использованием метода дискретного динамического программирования

Из-за большой степени изменчивости речевого сигнала автокорреляционные портреты одной речевой единицы, произнесённой в разные моменты времени отличаются. На рисунке 2 изображены АКП речевой единицы «безударная [а]», один из которых (а) был получен из произнесения слова «Вера», другой (б) из «сопутствующие». Видно, что портреты отличаются количеством строк. При этом несколько строк портрета а) могут соответствовать одной строке портрета б).

Расстояние между соответствующими строками АКП определяется для i − й строки портрета X и j − й строки портрета Yпо формуле

$$\rho_{i,j} = \sum_{k=1}^{N}\left(X(i,k) - Y(j,k)\right)^2. \quad (3)$$

Для определения меры соответствия АКП применен метод дискретного динамического программирования [7], который позволяет минимизировать функционал $\rho = \min_{\Omega}\sqrt{\sum \rho_{i,j}}$, характеризующий близость АКП. Множество Ω задаёт допустимые соответствия строк портретов, которые определяются исходя из следующих ограничений. 1. Количество строк в АКП может быть разным. 2. Каждая строка одного АКП не может соответствовать строке другого, отстоящей от предыдущей строки с найденным соответствием, более чем на cстрок. 3. Порядок соответствия строк

сохраняется, то есть если -я строка одного АКП соответствует j-й строке другого, то $(i+1)$-я строка не может соответствовать $j-l, l=1,2,..4$. Общее расстояние между АКП произнесений одинаковых речевых единиц, складывающееся из расстояний между соответствующими строками, по второму свойству метрики должно быть минимально с условием ограничений, описанных в 1)-3).

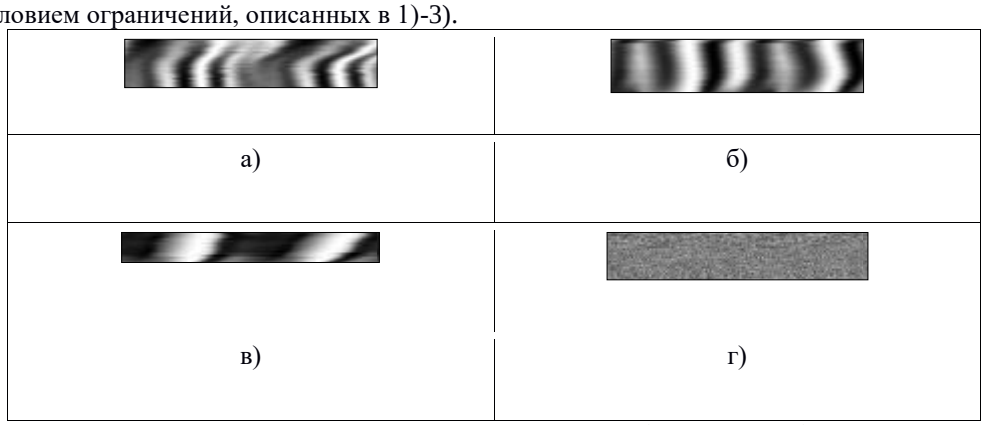

Рис. 1. АКП произнесений речевых единиц а) [“а], б) [о], в) [n`:], г) [f].

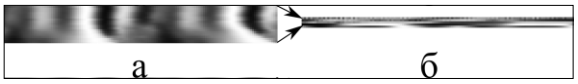

Рис. 2. АКП произнесений речевой единицы “безударная [а]”: а) эталон, б) в составе произнесения слова “сопутствующие”.

Для определения меры соответствия АКП речевого сигнала (в двумерном скользящем окне) АКП речевой единицы получен следующий алгоритм. Создается матрица D размером $m \times m$ элементов, где m – количество строк АКП скользящего окна X, количество строк АКП речевой единицы Y то же. Для примера установим величину $c=3$. На первом этапе находятся расстояния между $Y(1)$ и $X(1), X(2), X(3)$, эти расстояния сохраняются в D,

$$D_{1,i} = \rho\big(Y(1), X(i)\big), i = 1..3. \quad (4)$$

На втором этапе находятся расстояния между $Y(2)$ и $X(1), X(2), X(3), X(4), X(5)$, с учетом положения строки $Y(1)$, т. е. если $Y(1)$соответствует $X(2)$, то $Y(2)$ может сравниваться только с $X(2), X(3), X(4)$. Запоминаем в каждом случае номер строки портрета X, заполняем матрицу $D_{2,i} = D_{1,i} + \rho\big(Y(2), X(j)\big), j = i..i+2$, причем каждый элемент D в связи с пересечением возможных положений строк может заполняться несколько раз, в этом случае сохраняется минимальное значение (рисунок 3):

$$D_{k,i} = min\big[D_{k,j}, D_{k-1,j} + \rho\big(Y(k), X(j)\big)\big], j = i..i+2. \quad (5)$$

На следующих этапах таким же образом по формуле (5) находятся остальные элементы матрицы D, i на каждом этапе изменяется от 1 до $I+2$, где I – максимальное значение i на предыдущем этапе, для первого этапа $I=1$. Алгоритм останавливается когда матрица D окончательно заполнится. Минимальному расстоянию между X и Y соответствует минимальный элемент из -й строки и m-ого столбца матрицы.

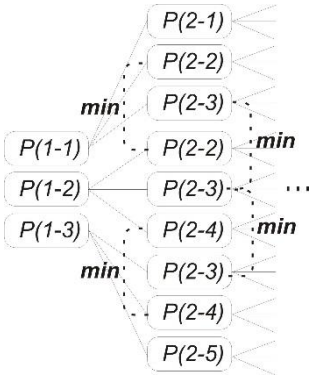

Рис. 3. Схема распределения строк сравниваемых АКП. P(i-j) – расстояние между i-й строкой одного АКП и j-й другого. Значком min обозначено, что из возможных одинаковых сравнений на разных этапах программирования выбирается сравнение с минимальным **расстоянием**.

4. Эксперименты

Предложенный алгоритм определения границ речевых единиц в участке слитной речи был проверен экспериментально. На рисунке 4 изображены границы речевых единиц в фрагменте речи, содержащем произнесение слова «основного». Например, участок произнесения речевой единицы [а], начинающей слово «основного», был верно определён в границах от 800 до 4800 цифровых отсчётов речевого сигнала, речевой единицы [s] в границах от 2400 до

5600 отсчётов, речевой единицы [n] в границах от 5600 до 9200 отсчётов, речевой единицы [a] в границах от 9600 до 11200 отсчётов, речевой единицы [v] в границах от 11200 до 16000 отсчётов, речевой единицы [n] в границах от 16000 до 17200 отсчётов, речевой единицы [“o] в границах от 17200 до 26400 отсчётов, речевой единицы [v] в границах от 26400 до 28000 отсчётов и последней в речевом сигнале единицы [a] в границах от 28000 до конца сигнала. Сравнение с экспертными границами не производилось, однако при визуальном сравнении определенных границ с действительными наблюдается их близость. Эксперименты показывают практическую применимость алгоритма определения границ речевых единиц в участке слитной речи.

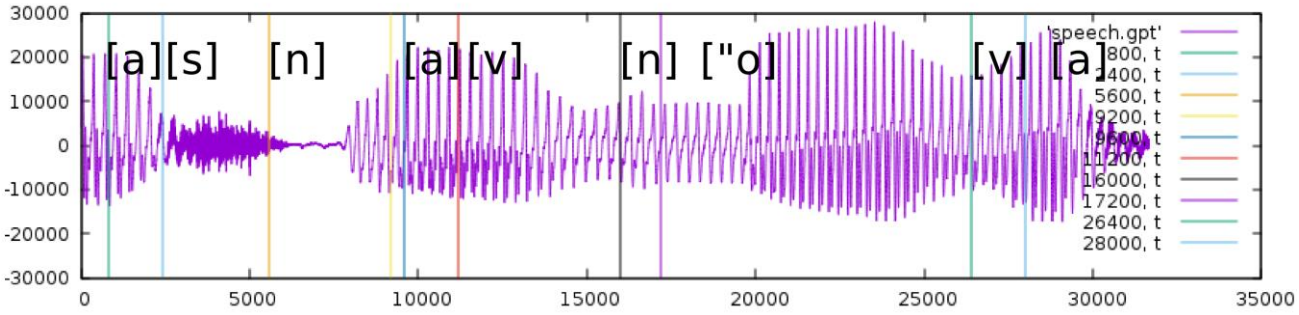

Рис. 4. Границы речевых единиц в фрагменте речи, содержащем произнесение слова «основного».

5. Заключение

Обнаруженные границы речевых единиц в дальнейшем предполагается использовать для более детального анализа речевого сигнала с целью идентификации речевых единиц. У авторов есть идея для этой задачи также использовать преобразование речевого сигнала в АКП. Однако параметры преобразования в АКП и метод совмещения портретов потребуются другие.

Благодарности

Работа выполнена при финансовой поддержке РФФИ. Проекты № 16-48-732046 и №16-48-730305.

Литература

[1] Galounov, V.I. Speech Database for the Russian Launguige / V.I. Galounov, H. Heuvel, J.L. Kochanina, A.V. Ostroukhov, H. Tropf, A.V. Vorontsova // SPEECOM 98, Proceedings of international workshop. – 1998. - SPb.

[2] Рабинер, Л. Р. Цифровая обработка речевых сигналов: пер. с англ. Под ред. М. В. Назарова и Ю. Н. Прохорова. / Л. Р. Рабинер, Р. В. Шафер. - М.: Радио и связь, 1981. - 496 с.

[3] Goldenthal, W. Statistical Trajectory Models for Phonetic Recognition / W. Goldenthal. PhD thesis. - M.I.T., Auguest 1994. - 170 p.

[4] Ostendorf, M. A stochastic segment model for phoneme-based continuous speech recognition / M. Ostendorf, S. A. Roukos // IEEE Transaction on Accoustics, Speech, and Signal Processing. – 1989. - Vol. 37, no. 12. - P. 1857-1869.

[5] Therrien, C. Probability and Random Processes for Electrical and Computer Engineers / C. Therrien, M. Tummala. — CRC Press, 2012. — P. 287.

[6] Крашенинников, В.Р. Некоторые задачи, связанные с распознаванием речевых команд на фоне интенсивных шумов / В.Р. Крашенинников, А.И. Армер, Н.А. Крашенинникова, В.В. Кузнецов, А.В. Хвостов // Инфокоммуникационные технологии. – Самара. – 2008. – Т. 1. – С. 72–75.

[7] Беллман, Р. Динамическое программирование / Р. Беллман. – М.: ИЛ, 1960. – 400 с.

[8] Крашенинников, В.Р. Autocorrelated Images and Search for Distance between them in Speech Commands Recognition / В.Р. Крашенинников, А.И. Армер, В.В. Кузнецов / Pattern Recognition and Image Analysisю – 2008. - Vol. 18, No. 4. – P. 663-666.

[9] Крашенинников, В.Р. Распознавание речевых команд на фоне интенсивных шумов с помощью автокорреляционных портретов / В.Р. Крашенинников, А.И. Армер, Н.А. Крашенинникова, А.В. Хвостов // Наукоемкие технологии. - 2007. - Т. 8, № 9. - С. 65-76.