ЗД

УДК 681.142.66

Дорохин О.А., Старушко Д.Г., Федоров Е.Е., Шелепов В.Ю.
Институт проблем искусственного интеллекта

Сегментация речевого сигнала

Авторы предлагают несколько методов деления речевого фрагмента на части, отвечающие отдельным аллофонам.

The authors propose some methods for partition of the speech signal on pieces corresponding individual allophones.

В работе [1] предложена схема распознавателя с автоматическим построением эталонов, которая требует предварительного обучения (создания) лишь пофонемной кодовой книги. Разработка пофонемного распознавателя в прямом смысле слова, то есть системы, распознающей цепочку фонем, требует прежде всего решения задачи сегментации речевого сигнала - разбиения его на участки, отвечающие отдельным аллофонам. Авторами предлагаются некоторые подхода к решению этой задачи.

Прежде всего отметим, что сегментация сама по себе является некоторой задачей распознавания. Ее особенность состоит в том, что она может решаться для каждого речевого сигнала отдельно. Это значит, что мы должны распознать границы конкретных объектов – аллофонов данного речевого сигнала. При этом нам не нужно распознавать сами аллофоны, не нужно заботиться о том, чтобы одинаковые аллофоны одинаково идентифицировались в различных словах, при произнесении различными дикторами и т.д. Для сегментации может использоваться только информация, содержащаяся в данном сигнале. Это при удачном выборе алгоритмов позволяет сделать их не зависящими от каких-либо априори задаваемых пороговых величин. Такие величины должны формироваться непосредственно в процессе сегментации.

Первый из описываемых методов сегментации является, по сути своей, амплитудным. Для краткости будем в дальнейшем называть его «амплитудной обработкой». Записанный речевой фрагмент разбивается на отрезки по 300 отсчетов. Если в конце образуется отрезок меньшей длины, то он отбрасывается. Для каждого из этих отрезков вычисляется величина

$$p_i = \frac{1}{300}\sum_{k=0}^{299}\left|x_{ik} - 128\right| \ . \qquad (1)$$

Здесь i - номер отрезка $(0 \le i \le N)$, x_{ik} - значение сигнала на k-ом отсчете i-го отрезка Величины x_{ik} при 8-битной записи, как известно, принимают целочисленные значения от 1 до 256, так что если иметь в виду визуализацию сигнала, то под знаком суммы стоят отклонения от средней линии. Затем вычисляется среднее величин p_i:

$$p = \frac{p_0 + p_1 + \ldots + p_N}{N} \ .$$

Наконец, весь сигнал разбивается на участки, состоящие из отрезков, для которых

$$p_i < p$$

(будем называть их низкоамплитудными), и участки, состоящие из отрезков, для которых

$$p_i \geq p .$$

(будем называть их высокоамплитудными). Границы между ними принимаются за искомые границы сегментации.

Этот метод с высокой надежностью позволяет выделить в исходном сигнале участки, отвечающие звукам «с», «ш», «щ», «ц», «ч», «ф», «х», «б», «г», «д», «п», «к», «т». Он может выделять также другие согласные, особенно при отсутствии в слове вышеперечисленных звуков. В целом, этот простой, надежный и чрезвычайно быстро работающий метод может, как показывает опыт, с успехом служить как часть более сложной процедуры сегментации речевого сигнала.

Второй подход реализует аналогичный алгоритм, но в формуле (1) величина, выражаемая суммой, заменяется числом локальных максимумов на соответствующем отрезке. Этот метод (назовем его обработкой с помощью функции максимумов) позволяет отделять шипящие или свистящие звуки от соседних взрывных «п, к, т».

Третий подход тесно связан с методами построения кодовой книги. Привязавшись к прежнему разбиению сигнала на отрезки по 300 отсчетов, для каждого из них вычислим традиционный для нас 29-мерный вектор признаков, использующий относительные частоты длин полных колебаний [2]. Получим множество векторов

$$\mathrm{a}_0, \mathrm{a}_1, ..., \mathrm{a}_N .$$

Выберем число $2 \leq k \leq N$ и построим вектора

$$\mathrm{b}_k = \frac{\mathrm{a}_1 + \mathrm{a}_2 + ... + \mathrm{a}_{k-1}}{k-1} , \qquad \mathrm{c}_k = \frac{\mathrm{a}_k + ... + \mathrm{a}_N}{N-(k-1)} .$$

Определим теперь функцию

$$\phi(k) = \sum_{i=1}^{k-1} \left| \mathrm{a}_i - \mathrm{b}_k \right| + \sum_{i=k}^{N} \left| \mathrm{a}_i - \mathrm{c}_k \right|, \quad 2 \leq k \leq N .$$

Если наименьшее значение этой функции достигается при $k = k_0$, то правый конец отрезка с номером k_0 принимается за границу аллофона. Далее отрезки по одну сторону найденной границы отбрасываются, к оставшейся части сигнала применяется вышеописанная процедура и т.д. Этот метод в отличие от первого позволяет весьма надежно разделять две соседние гласные фонемы или гласную и соседствующую с ней звонкую согласную.

Наконец, четвертый из опробованных нами методов использует спектральное разложение речевого сигнала. Он состоит в следующем. Речевой сигнал, представленный отсчетами $x_i, i = \overline{0, N}$, разбивается на перекрывающиеся фраг-

менты объемом 256 отсчетов с шагом S, т. е. формируется множество векторов $\nu = \{\overline{v}_k\}_{k=0}^{M}$, где

$$M = \frac{N - 256}{S}.$$

Для каждого вектора вычисляются спектральные коэффициенты дискретного преобразования Фурье, которые могут быть определены по формуле

$$X_{km} = \frac{1}{256}\sum_{l=0}^{255} x_{kS+l} e^{-j\frac{2\pi lm}{256}}.$$

Затем формируется множество векторов, представляющих собой мгновенные логарифмические спектральные срезы речевого сигнала: $\vartheta = \{\overline{\vartheta}_k\}_{k=0}^{M}$, где

$$\overline{\vartheta}_k = \left(\log|X_{ko}|, \cdots, \log|X_{k255}|\right).$$

Далее в полученном множестве векторов выделяются смежные группы, состоящие из D векторов, для каждой из которых определяется среднее значение логарифмического спектра мощности в соответствии с формулой

$$\widetilde{\vartheta}_h = \frac{1}{D}\sum_{q=0}^{D-1} \overline{\vartheta}_{h+q}, \text{ где } h = \overline{0, M - D - 1}.$$

Затем вводится функция однородности речевого сигнала, определяемая выражением

$$\Re_g = e^{-\left\|\widetilde{\vartheta}_g - \widetilde{\vartheta}_{g+1}\right\|},$$

где $g = \overline{0, M - D - 2}$. Полученные значения функции сглаживаются методом скользящего среднего и выполняется поиск локальных минимумов функции однородности, соответствующих границам однородных участков. Найденные границы принимаются за границы между фонемами. Этот метод в большинстве случаев сегментирует достаточно качественно. Однако он может строить лишние границы на длинных шипящих. Кроме того, он пока работает много дольше каждого из ранее описанных методов, несмотря на то, что вместо быстрого преобразования Фурье нами применено более экономное по времени преобразование Хартли.

В заключение мы опишем способ сегментации речевого сигнала, который состоит в сочетании методов фильтрации и обработки функцией максимума с последующей амплитудной обработкой. Он является быстрым и надежно сегментирует большой класс слов, а именно: требуется, чтобы рядом стоящие гласные и звонкие согласные звуки последовательно чередовались. При этом пока мы можем гарантировать успех только для случая гласных «а, о, э». Опишем этот способ на примере сегментации слова «ДОШТАМПОВАЛ». Вот визуализация соответствующего речевого сигнала при 8-битной записи с частотой 22050 Гц:

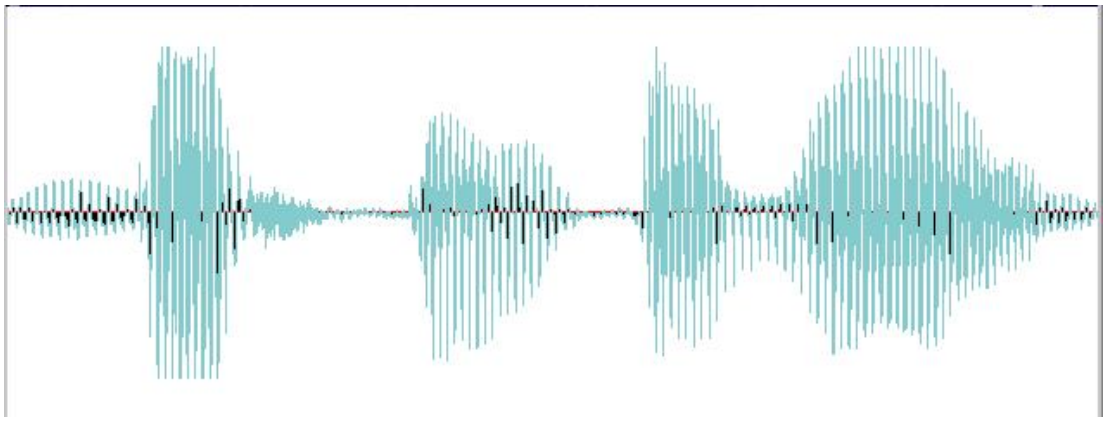

Рис. 1

Прежде всего, этот файл обрабатывается фильтром низких частот с частотой среза 500 Гц. Получаем файл следующего вида:

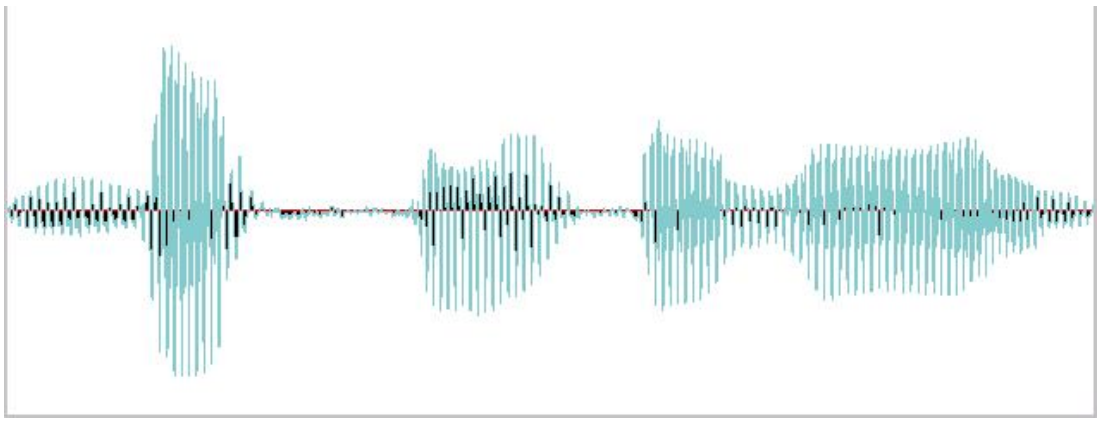

Рис. 2

Результатом проделанной обработки является то, что часть сигнала, соответствующая шипящему звуку «ш», стала малой по амплитуде. Этот шаг необходим при дальнейшей амплитудной обработке потому, что первоначально участки шипящих могут быть сравнимы по амплитуде с участками голосовых звуков или даже превышать их.

К полученному файлу применяется амплитудная обработка с заменой величины p величиной $0{,}3\,p$. В результате получаем файл, визуализация которого имеет вид:

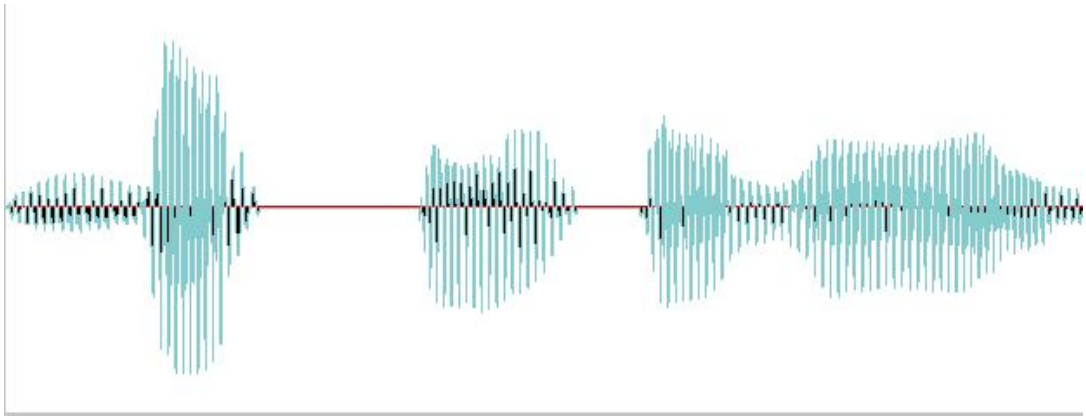

Рис. 3

Это дает предварительную сегментацию, в результате которой получаются следующие файлы, соответствующие частям исходного сигнала:

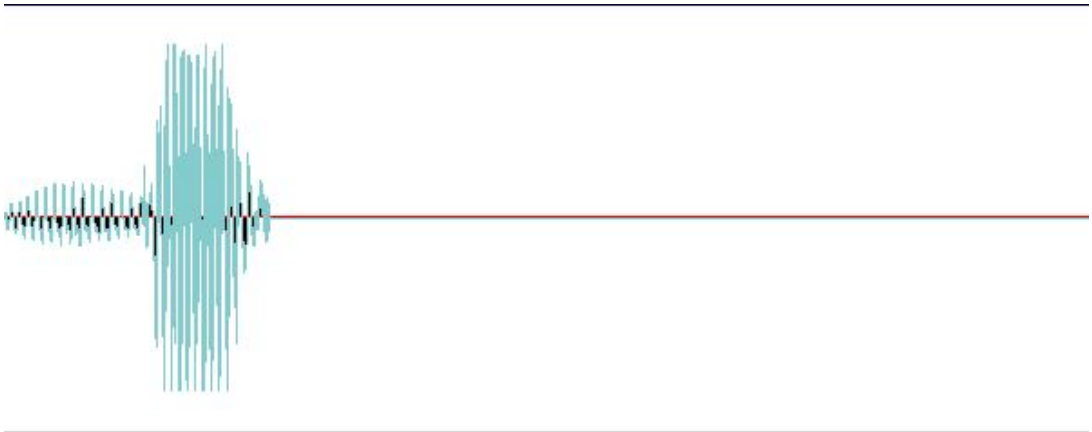

Рис. 4. Файл 0H

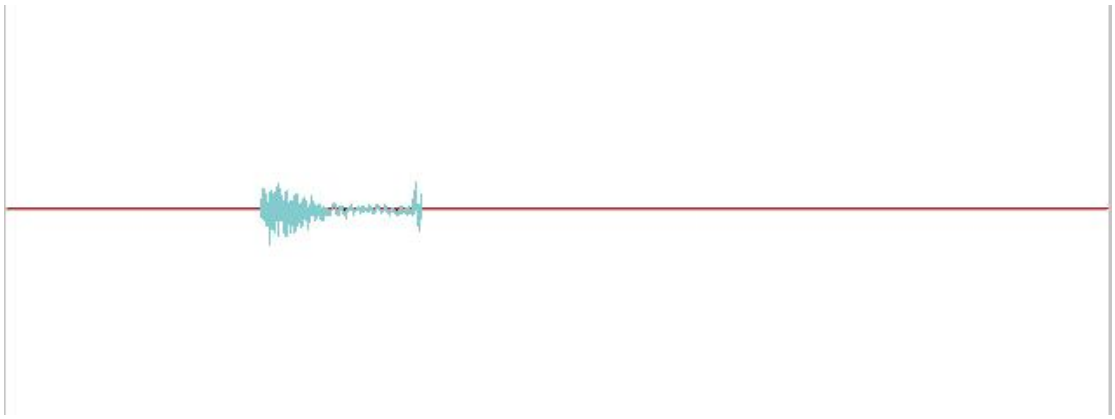

Рис. 5. Файл 1L

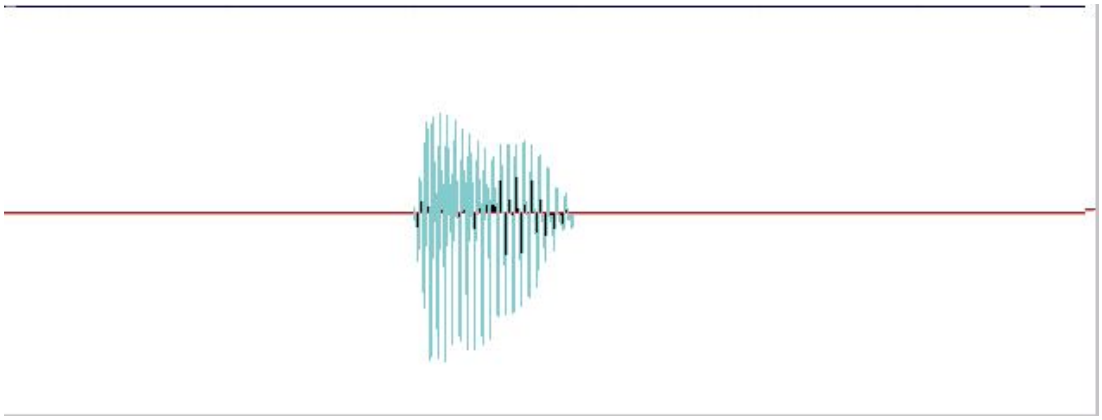

Рис. 6. Файл 2H

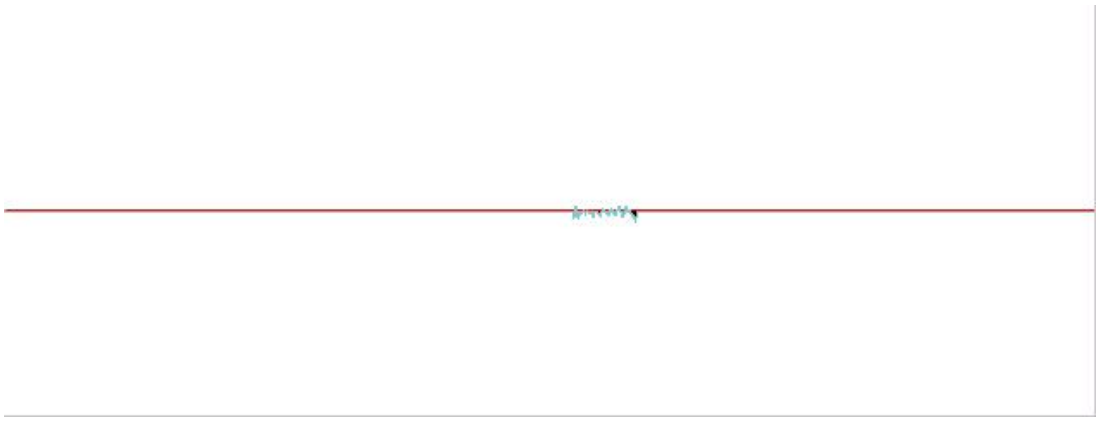

Рис. 7. Файл 3L

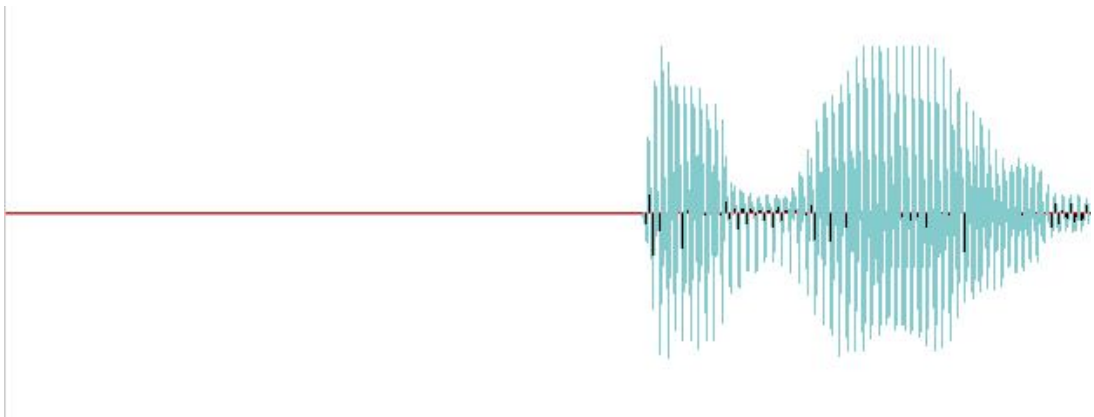

Рис. 8. Файл 4H

В названиях этих рисунков цифра обозначает порядковый номер файла. Буква «H» означает, что при первоначальной сегментации соответствующая часть сигнала классифицирована как высокоамплитудная: буква «L» , что она классифицирована как низкоамплитудная. Высокоамплитудная часть соответствует некоторой последовательности голосовых звуков, низкоамплитуд-

ная – последовательности шипящих и глухих взрывных. Для последних большая часть места занята паузой (задержка воздуха в начале глухого взрывного).

Для того чтобы в низкоамплитудных частях отделить шипящие от пауз, соответствующие файлы обрабатываются «функцией максимумов» с заменой величины p величиной $0{,}3\,p$.

При этом файл 1L, например, превращается в файл:

Рис. 9

и машина строит метки.

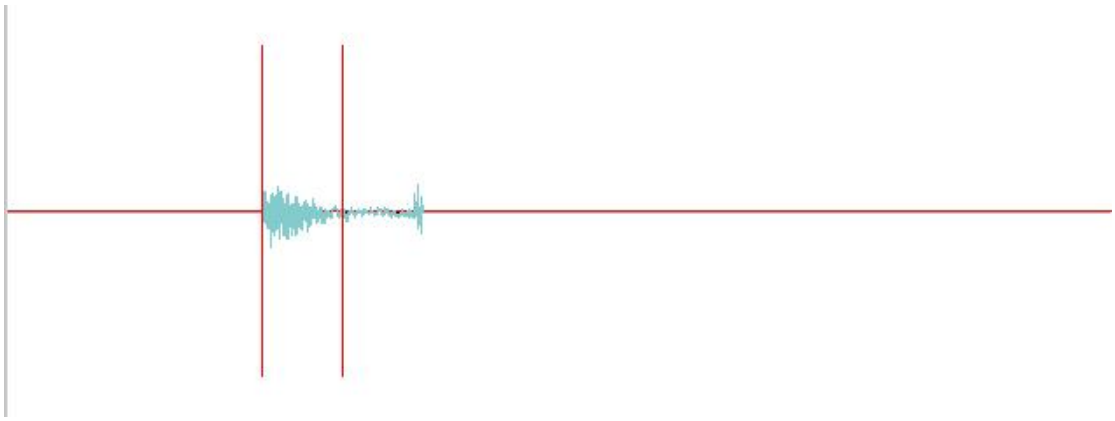

Рис. 10

Файлы, соответствующие высокоамплитудным частям, обрабатываются фильтром высоких частот с частотой среза 500 Гц. Это позволяет разделить по амплитуде гласные и согласные звуки. Так, например, файл 2H при этом превращается в файл:

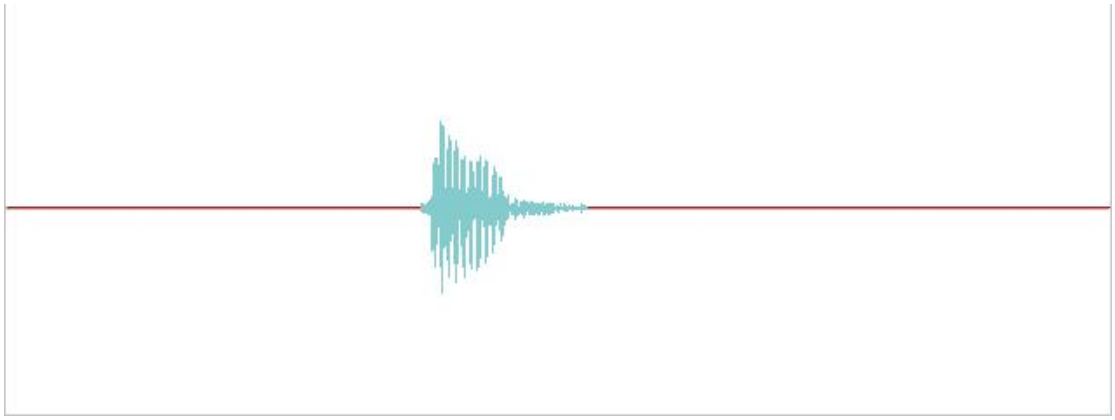

Рис. 11

После этого амплитудная обработка позволяет получить метки

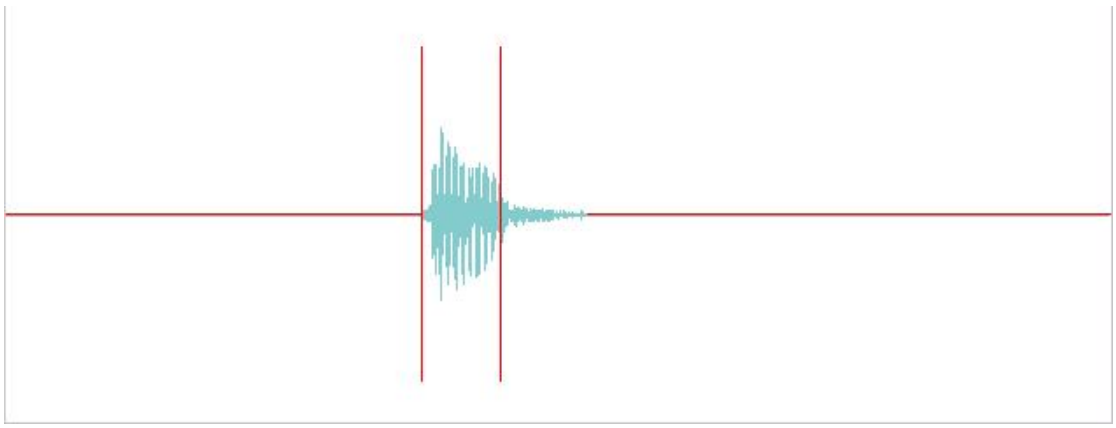

Рис. 12

В процессе сегментации частей возможны отдельные ошибки: некоторые кусочки высокоамплитудных частей машина классифицирует как низкоамплитудные и наоборот. То же возможно и при применении функции максимумов. Такие посторонние вкрапления, если их длина не превышает 600 отсчетов, убираются программой. Далее множество меток, полученных при сегментации отдельных частей, собирается воедино. При этом на стыке соседних частей может образоваться по две рядом стоящих метки вместо одной. Тогда они заменяются одной меткой посредине. В результате всех этих операций, которые современная машина выполняет очень быстро, мы получаем следующую сегментацию исходного сигнала:

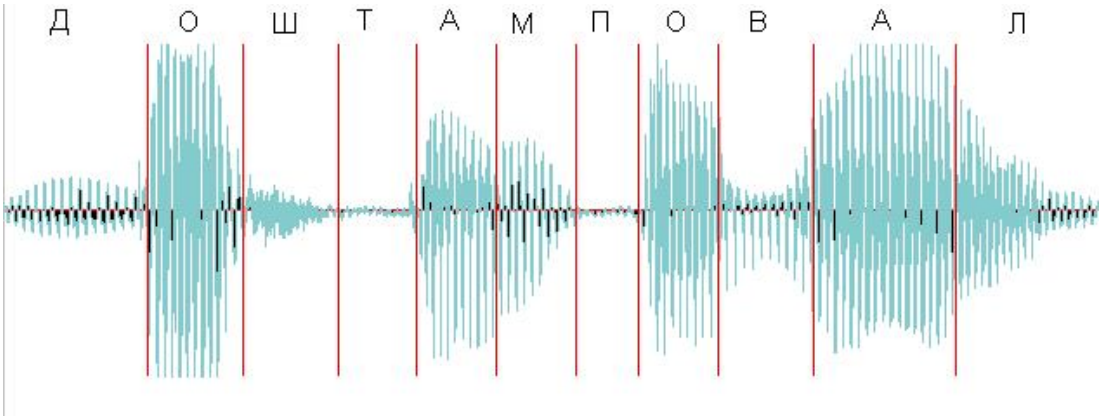

Рис. 13

Все описанные выше методы позволяют создавать реальные программы, в известной степени сегментирующие речевой сигнал. Каждый имеет свои достоинства и недостатки. Наибольшими возможностями на сегодняшний день, по нашему мнению, обладает только что изложенный подход, который состоит в сочетании методов фильтрации и обработки функцией максимума с последующей амплитудной обработкой. Вообще же, вероятно, качественная программа сегментации может быть создана на основе объединения всех этих методов так, чтобы они дополняли и контролировали друг друга.

Литература

1. Грабовая В.А., Федоров Е.Е., Шелепов В.Ю. О системе компьютерного распознавания русской речи с автоматическим построением эталонов // Искусственный интеллект. – 2000. – №№1. – С. 76-81.
2. Дорохин О.А., Засыпкин А.В., Червин Н.А., Шелепов В.Ю. О некоторых подходах к проблеме компьютерного распознавания устной русской речи // Международная конференция «Знания - Диалог – Решение». Сборник научных трудов. – Ялта, 1997. – Т.1. – С. 234-240.
3. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 262 с.
4. Rabiner L.R., Juang B.-H. Fundamentals of Speech Recognition. Prentice Hall PTR, 1993. – 507 p.
5. Бабий Л.В., Винцюк Т.К. Робастные алгоритмы распознавания речи // Автоматическое разпознавание слуховых образов. Тезисы докладов 16-го всесоюзного семинара (АРСО-16). – Москва, 1991.