УДК 621.391

СЕГМЕНТАЦИЯ РЕЧЕВЫХ СИГНАЛОВ НА ОСНОВЕ АНАЛИЗА РАСПРЕДЕЛЕНИЯ ЭНЕРГИИ ПО ЧАСТОТНЫМ ИНТЕРВАЛАМ

Е. Г. ЖИЛЯКОВ
Е. И. ПРОХОРЕНКО
А. В. БОЛДЫШЕВ
А. А. ФИРСОВА
М. В. ФАТОВА

Белгородский государственный национальный исследовательский университет

e-mail: Zhilyakov@bsu.edu.ru

В статье представлено описание некоторых алгоритмов сегментации речевых сигналов на основе анализа распределения энергии по частотным интервалам. Предложено несколько решающих функций для сегментации речевых сигналов, основанных на анализе распределений энергий по частотным интервалам.

Ключевые слова: речевой сигнал, анализ речевого сигнала, частотные представления, распределение энергии по частотным интервалам, сегментация речевого сигнала.

Речевые сообщения являются одной из естественных для человека форм информационного обмена. В связи с этим одним из основных направлений развития информационно-телекоммуникационных систем является использование речевых технологий. Одним из основных направлений исследований в этой области является распознавание и синтез речи с целью реализации речевого управления, идентификации по образцу речи, создание речевых запросно-ответных систем.

Вопросам распознавания речи уделяется большое внимание в современных информационно-телекоммуникационных системах [1,2,3,4,5,6,7,8]. Интерес к этой проблеме обусловлен тем, что ее решение позволяет сделать взаимодействие между человеком и компьютером более естественным. Это касается возможности преобразования речи в текст, в случае необходимости ведения письменных протоколов, лекций и т.д. Другим немаловажным преимуществом является предоставление возможности взаимодействия с компьютером людям с нарушением зрения и мелкой моторики рук.

Распознавание речи является сложной задачей, что обусловлено нестационарностью речевых сигналов, зависимостью их параметров от диктора, интонации, эмоционального состояния. Одним из важнейших этапов распознавания речи является сегментация речевых сигналов на участки, соответствующие одному звуку или их неразделимыми сочетаниями (фонемами). Сегментация – это процесс определения границ между участками сигналов, соответствующих разным звукам. От того, насколько точно определены границы между звуками, зависит эффективность и сложность следующего этапа алгоритма распознавания. Проблема осуществления точной сегментации связана с трудностью определения решающей функции. Звуки речи можно охарактеризовать распределением энергии по частотному диапазону. При этом каждый звук имеет свое особенное распределение энергии. При переходе от одного звука к другому распределение энергии по частотным интервалам меняется. Тем не менее, можно выделить участки, где это изменение незначительно. Такие участки называют фонемами. Переход от одной фонемы к другой не может происходить мгновенно. Это связано с особенностью речевого аппарата человека. Эта особенность может быть использована для принятия решения о наличии или отсутствии границы. Существуют различные алгоритмы сегментации речевых сигналов, основанные на анализе распределения энергии по частотному диапазону: по динамическим детекторам, по усредненному нормированному спектру, по корреляции между спектрами [1,2]. Вместе с тем, опыт и литературные источники показывают, что существующие методы сегмен-

тации не позволяют определить границы между некоторыми звуками, либо приводят к появлению дополнительных границ на участках, соответствующих одному звуку.

Основным недостатком метода сегментации по усредненному нормированному спектру является то, что он не позволяет обнаружить границу, если изменения происходят преимущественно в значении энергии сигнала. Для учета этих особенностей предлагается не производить нормировку спектра, но это приводит к появлению эффекта пересегментации из-за повышения чувствительности алгоритма. Экспериментальные исследования метода сегментации по динамическим детекторам показывает, что выбор достаточно большой величины порога приводит к пропуску большого количества границ, особенно между гласными и сонорными согласными. Уменьшение порогового значения приводит к появлению ложно установленных границ, особенно это проявляется для согласных звуков, имеющих неоднородное распределение энергии вдоль звука. Использование алгоритма сегментации по корреляции между спектрами имеет наилучшие показатели среди представленных алгоритмов. Но при этом важно также отметить тот факт, что наименьшая вероятность верного определения границы проявляется на участках между гласными и сонорными согласными. Уменьшение величины порога приводит к увеличению чувствительности и появлению ложно определенных границ.

Анализ решающих функций описанных алгоритмов показал, что они имеют неравномерный характер и зависят от выбора величины сдвига между началами анализируемых отрезков и длительности анализируемых отрезков.

Для выявления причины нестабильности решающих функций было решено провести анализ изменения распределения энергии речевых сигналов при переходе от одного окна анализа к другому. Значения энергий, сосредоточенных в заданных частотных интервалах предлагается оценить с использованием выражения [9]:

$$P_r = \vec{x}^T A_r \vec{x}, \tag{1}$$

где: $\vec{x}$ – анализируемый отрезок сигнала;

r – номер частотного интервала, изменяющийся от 1 до R;

A_r – субполосная матрица, рассчитанная для r-го частотного интервала:

$$A_r = \{a_{ik}^r\}$$

$$a_{ik}^r = (\sin(v_{r+1}(i-k)) - \sin(v_r(i-k)))/(\pi(i-k)),\ i,k = 1,\ldots,N, \tag{2}$$

где v_r, v_{r+1} – границы r-ого частотного интервала, причем:

$$0 \le v_r < v_{r+1} \le \pi, \quad r=1,\ldots,R, \tag{3}$$

$$v_{r+1} - v_r = \pi / R, \tag{4}$$

где R – количество частотных интервалов, на которые разбивается частотная ось.

Для выявления особенностей изменения распределения энергии по частотным интервалам предлагается рассмотреть график изменения энергии в каждом частотном интервале при переходе от одного окна анализа к другому.

Исследования проводились для различных сигналов, соответствующих звукам и сочетаниям звуков русской речи, произнесенных разными дикторами, записанными с частотой дискретизации F_d=8000Гц и количеством бит на один отсчет 16. Речевой сигнал разбивался на окна одинаковой длины. При этом выбор отрезков анализа выбирался со сдвигом 1 отсчет относительно начала окна анализа. Длина окна анализа выбиралась достаточно большой, чтобы отразить периодичность звуков русской речи и достаточно малой, чтобы не превышала длины одного звука. В рамках данных исследований длина окна анализа выбиралась равной 64 и 128 отсчетов (что составляет 8мс и 16мс соответственно). Количество интервалов, на которые разбивалась частотная ось, выбиралось равным 16 и 32.

На рисунках 1-4 представлены фрагменты сигналов и графики изменения энергии при переходе от одного окна анализа к другому для двух одинаковых звуков, произнесенных в различных сочетаниях одним диктором.

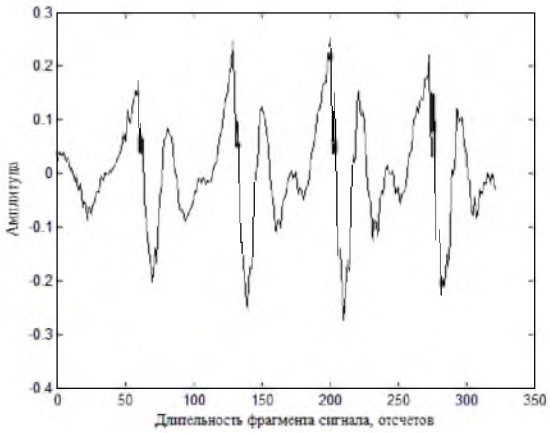

Рис. 1. Фрагмент сигнала, соответствующего первому звуку «е» в слове «ч**е**репАха»

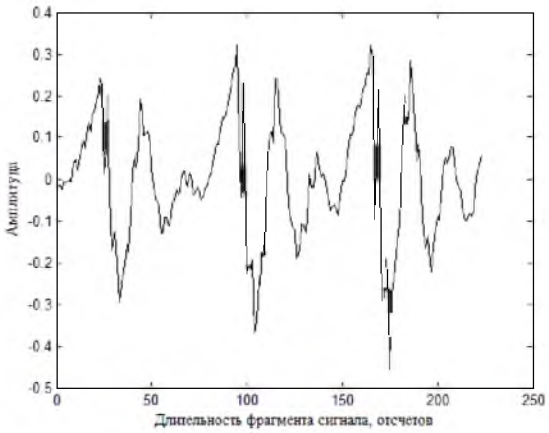

Рис. 2. Фрагмент сигнала, соответствующего второму звуку «е» в слове «чер**е**пАха»

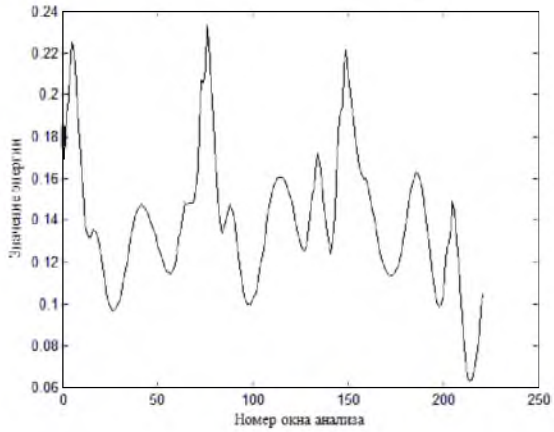

Рис. 3. Изменение энергии во втором частотном интервале сигнала, соответствующего первому звуку «е»в слове «ч**е**репАха» (*N*=64, *R*=16, шаг=1)

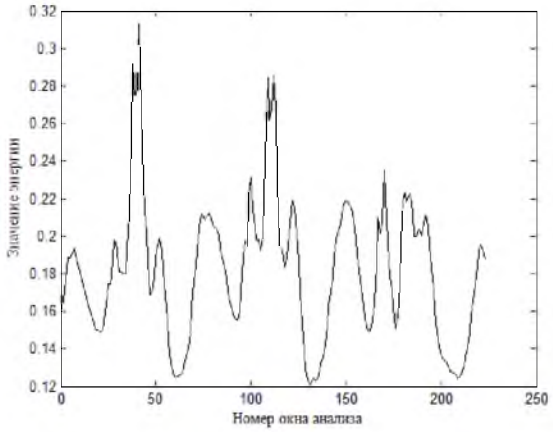

Рис. 4. Изменение энергии во втором частотном интервале сигнала, соответствующего второму звуку «е»в слове «чер**е**пАха» (*N*=64, *R*=16, шаг=1)

Выбор второго частотного интервала для данных звуков обусловлен тем, что их основная энергия сосредоточена во втором и третьем частотных интервалах. Анализ рисунков 3-4 показывает, что несмотря на то, что энергия оценивалась для одного и того же звука, ее значение отличается. Также можно отметить, что при изменении энергии в частотном интервале при переходе от одного окна анализа к другому наблюдается периодичность. Для данных фрагментов сигнала длительность периода составляет порядка 76 отсчетов. Аналогичные изменения для данных звуков наблюдаются и в третьем частотном интервале. Было рассмотрено также изменение энергии для других звуков речи. Выявлено, что подобная периодичность проявляется для всех гласных, а также звонких и сонорных согласных звуков. Эта нестационарность в изменении распределения энергии при переходе от одного окна анализа к другому приводит к нестационарности решающих функций. Анализ распределения энергии по частотным интервалам функции изменения энергии в заданном частотном интервале при переходе от одного окна анализа к другому показал, что основная энергии этого распределения сосредоточена в интервале [0, π/16]. Для устранения периодичности функции изменения энергии в заданном частотном интервале при переходе от одного

окна анализа к другому предлагается применить оптимальную фильтрацию к этой функции в полосе [0, $\pi/16$] [9]:

$$\vec{P}_{fr} = A_1 \vec{P}_r , \qquad (5)$$

где $\vec{P}_r$ – функция изменения энергии в r-ом частотном интервале,

$A1$ – субполосная матрица для интервала [0, $\pi/16$],

$\vec{P}_{fr}$ – результат фильтрации функции изменения энергии в r-ом частотном интервале.

На рис. 5-6 представлен сигнал, соответствующий слову «черепаха» после удаления пауз, и функция изменения энергии в 3-м частотном интервале до и после фильтрации.

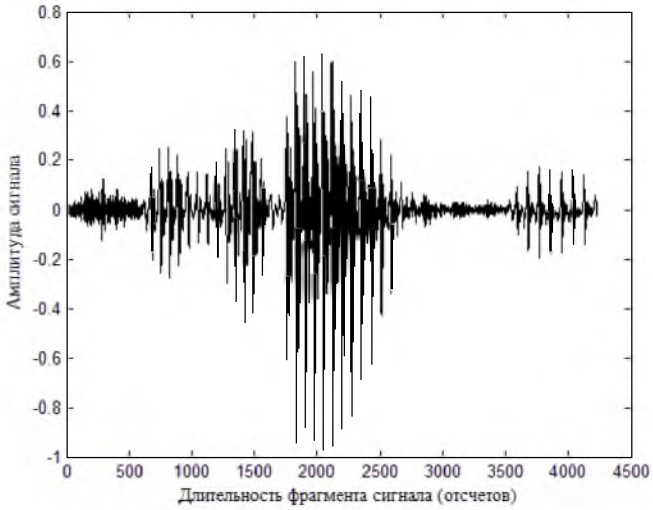

Рис. 5. Фрагмент речевого сигнала, соответствующего слову «черепаха»

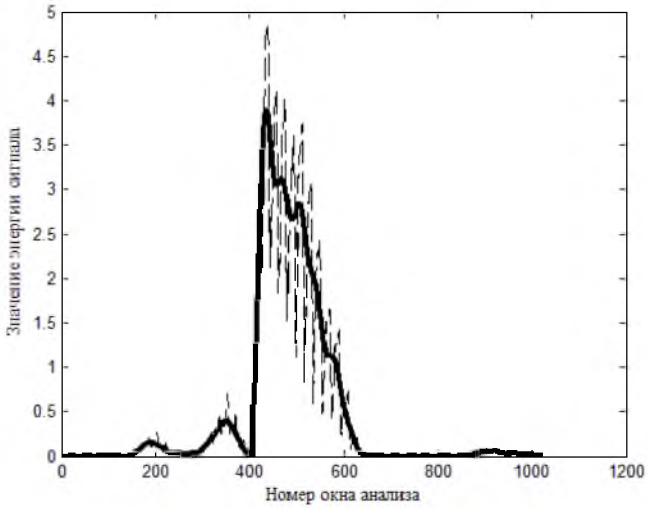

Рис. 6. Распределение энергии в 3-м частотном интервале при переходе от одного окна анализа к другому до и после фильтрации (N=64, шаг=4, υ_1=0; υ_2=$\pi/16$)

На рисунке 6 пунктиром обозначена функция изменения энергии в 3-м частотном интервале до фильтрации, а сплошной линией – после фильтрации. Выбор третьего частотного интервала обусловлен тем, что основная энергия рассматриваемого сигнала сосредоточена именно в третьем частотном интервале. Аналогичные ре-

зультаты получаются и при анализе других частотных интервалов. Анализ рисунка 6 показывает, что использование фильтрации позволяет устранить колебания функции.

На рис. 7-10 представлены фрагменты сигналов и распределение энергий этих отрезков до и после фильтрации.

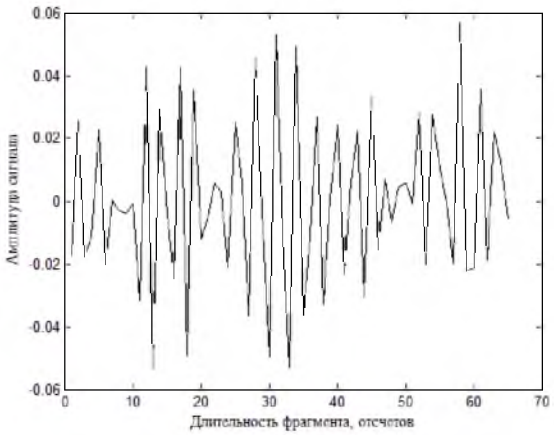

Рис. 7. Фрагмент сигнала 1, соответствующего первому звуку «е» в слове «чер*е*п*А*ха»

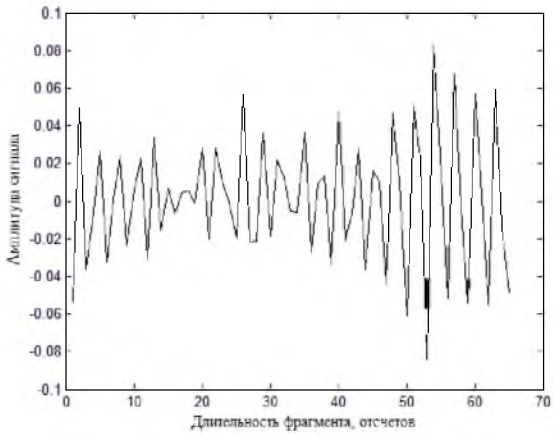

Рис. 8. Фрагмент сигнала 2, сдвинутого на 32 отсчета относительно сигнала 1

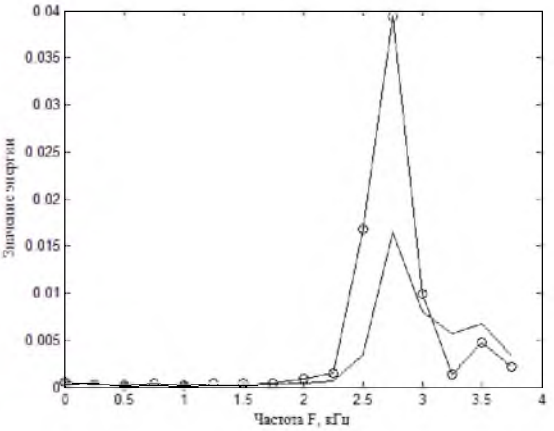

Рис. 9. Распределение энергий по частотным интервалам анализируемых отрезков сигналов до фильтрации:
–o – фрагмента сигнала 1;
– – фрагмента сигнала 2

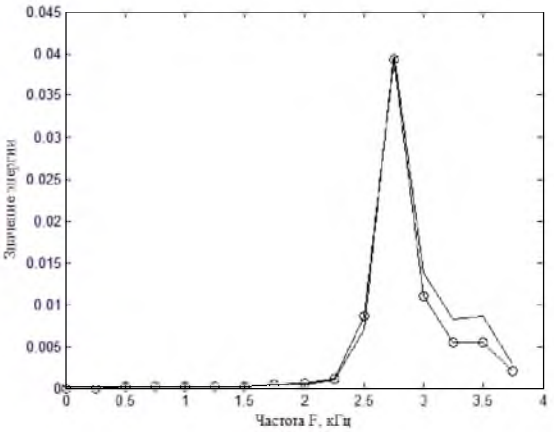

Рис. 10. Распределение энергий по частотным интервалам анализируемых отрезков сигналов после фильтрации:
–o – фрагмента сигнала 1;
– – фрагмента сигнала 2

Анализ рисунков 9 и 10 показывает, что использование фильтрации вида (5) приводит к устранению существенных различий в распределении энергий по частотным интервалам для фрагментов сигналов, соответствующих одному звуку речи.

Использование предложенного алгоритма позволит уменьшить колебания решающих функций при сегментации речевых сигналов, что приведет к уменьшению вероятности неправильного определения границ сегментов.

В рамках данной статьи рассматривается несколько решающих функций для принятия решения о наличии или отсутствии границы. Для принятия решения о наличии или отсутствии границ между двумя отрезками сигнала рассматриваются свойства анализируемых отрезков. В качестве сопоставляемых характеристик предлагается использовать распределение энергий по частотным интервалам вида (1). При этом важно отметить, что данные характеристики предлагается использовать после фильт-

рации вида (5). Если анализируемые отрезки принадлежат одному и тому же звуку речи, то их характеристики должны отличаться незначительно.

Пусть нулевая гипотеза *Ho* звучит следующим образом: сопоставляемые отрезки сигналов порождены одним и тем же звуком речи. В идеальном случае для сопоставляемых отрезков должно выполняться:

$$P_{r1} = P_{r2}, \quad r = 1,\ldots,R, \tag{6}$$

где P_{r1} – значение энергии в r-ом частотном интервале первого отрезка,

P_{r2} – значение энергии в r-ом частотном интервале второго отрезка.

Для оценки шансов выполнения гипотезы *Ho* может быть использована характеристика вида:

$$S_1 = 2\frac{IntG_{12}}{IntR_{11} + IntR_{12}} \le 1, \tag{7}$$

где *IntR*11, *IntR*12 – мощность множеств *R*11 и *R*12 соответственно, где:

$$\frac{\sum_{r\in R_{11}} P_{r1}}{\left\|\vec{x}_1\right\|^2} = \frac{\sum_{r\in R_{12}} P_{r2}}{\left\|\vec{x}_2\right\|^2} \approx m, \tag{8}$$

где *R*11, *R*12 – наименьшее количество частотных интервалов, в которых сосредоточена заданная доля энергии m соответственно для первого и второго отрезка анализа,

m – доля энергии, выбираемая порядка 0,9,

P_{r1}, P_{r2} – значение энергии в r-ом частотном интервале соответственно для первого и второго отрезка анализа,

$\vec{x}_1$, $\vec{x}_2$ – анализируемые отрезки сигнала,

*IntG*12 – мощность множества $G_{12} = R_{11} \cap R_{12}$ – пересечения множеств *R*11 и *R*12.

Чем большая доля частотных интервалов совпала при анализе двух отрезков, тем больше функция *S*1, а, следовательно, тем больше вероятность того, что гипотеза *Ho* верна.

На рисунках 11, 12 представлены фрагмент анализируемого сигнала и функция *S*1 для этого фрагмента.

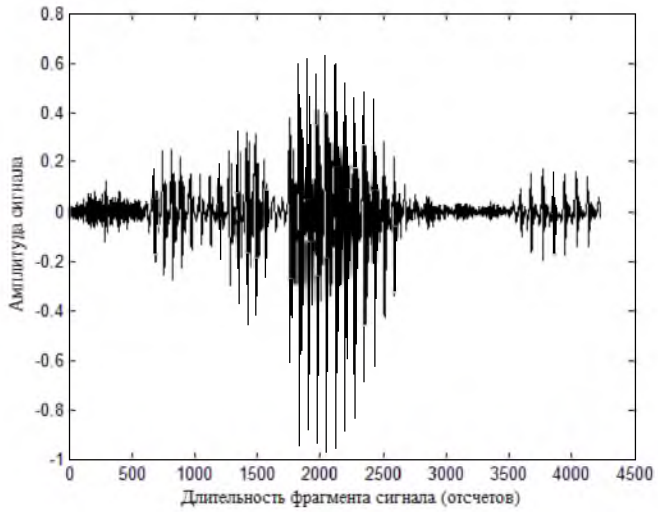

Рис. 11. Фрагмент речевого сигнала, соответствующего слову «черепаха»

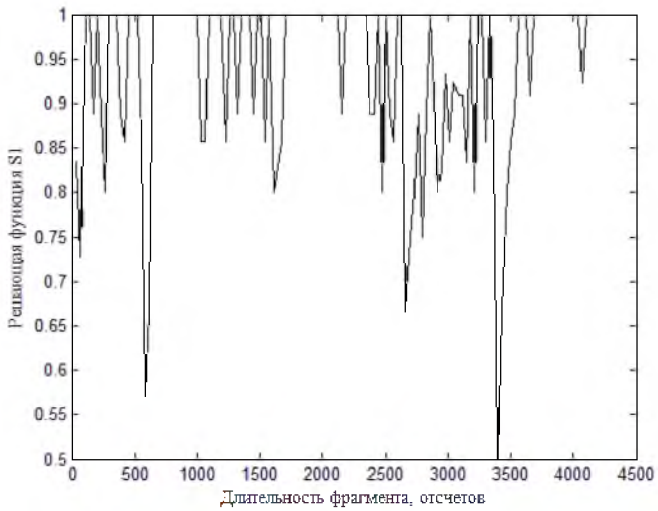

Рис. 12. Решающая функция $S1$ (N=64, R=16, m=0,9, шаг=32)

Анализ рисунков 11 и 12 показывает, что наименьшее значение решающая функция $S1$ принимает на границе между звуками «ч» и «е», «а» и «х», а также «х» и «а». В то же время значение решающей функции на участках, соответствующих сочетаниям «ере», а также звукам «ч», «а» и «х», изменяются незначительно в диапазоне 0,75÷1. Незначительное изменение функции $S1$ на участке «ере» (0,80÷1) обусловлено тем, что данные звуки имеют похожее распределение энергии и основное изменение связано со значением энергии в этих интервалах. Учет этих особенностей может быть осуществлен при использовании решающей функции вида:

$$S_2 = \begin{cases} \max\limits_{r \in G_{12}} \left(\max\left(\dfrac{P_{r1}}{P_{r2}}, \dfrac{P_{r2}}{P_{r1}} \right) \right) \geq 1, & \text{при} \quad S_1 > 0 \\ >> 1, & \text{при} \quad S_1 = 0 \end{cases}, \quad \forall r \in G_{12} \qquad (9)$$

где P_{r1}, P_{r2} – значение энергии в r-ом частотном интервале соответственно для первого и второго отрезка анализа,

$G_{12} = R_{11} \cap R_{12}$ – пересечения множеств $R11$ и $R12$.

Чем больше анализируемые фрагменты отличаются друг от друга, тем больше значение решающей функции $S2$, а, следовательно, тем меньше вероятность того, что гипотеза $H0$ верна.

На рисунке 13 представлена функция S2 для фрагмента сигнала, соответствующего слову «черепаха», представленного на рисунке 11.

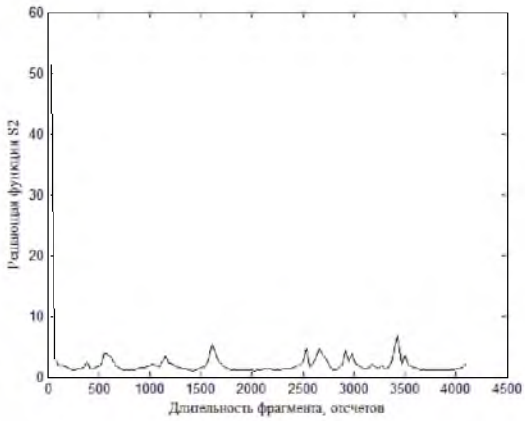

Рис. 13. Решающая функция $S2$ (N=64, R=16, m=0,9, шаг=32)

Анализ рисунка 13 показывает, что наблюдаются всплески на участках между «ч» и «е», «е» и «р», окончания второго звука «е», «а» и «х», «х» и «а». При этом важно отметить, что невозможно выбрать однозначного порога, так как наблюдаются всплески решающей функции *S*2 на фрагментах сигналов, соответствующих звукам «ч» и «х». Это связано с особенностями воспроизведения этих звуков, в частности, их неоднородностью.

Для оценки вероятности истинности гипотезы *H*0 может быть использована решающая функция вида:

$$S_3 = \begin{cases} \max\left\{ \dfrac{\sum\limits_{r \in G_{12}} P_{r1}}{\sum\limits_{r \in G_{12}} P_{r2}}, \dfrac{\sum\limits_{r \in G_{12}} P_{r2}}{\sum\limits_{r \in G_{12}} P_{r1}} \right\}, & \text{при } S_1 > 0 \\ >> 1, & \text{при } S_1 = 0 \end{cases}, \qquad (10)$$

где P_{r1}, P_{r2} – значение энергии в r-ом частотном интервале соответственно для первого и второго отрезка анализа,

$G_{12} = R_{11} \cap R_{12}$ – пересечения множеств *R*11 и *R*12.

Чем больше анализируемые фрагменты отличаются друг от друга, тем больше значение решающей функции *S*3, а, следовательно, тем меньше вероятность того, что гипотеза *H*0 верна.

На рисунке 14 представлена функция S3 для фрагмента сигнала, соответствующего слову «черепаха», представленного на рисунке 11.

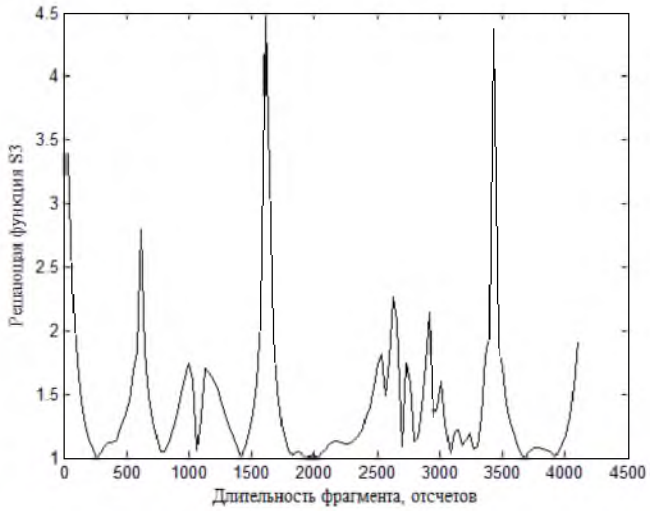

Рис. 14. Решающая функция *S*3 (*N*=64, *R*=16, *m*=0,9, шаг=32)

Анализ рисунка 14 показывает, что , как и для решающей функции *S*2 наблюдаются всплески на участках между «ч» и «е», «е» и «р», окончания второго звука «е», «а» и «х», «х» и «а». При этом в отличии от решающей функции *S*2 наблюдается всплеск на участке между звуками «р» и «е». Важно также отметить, что функция *S*3 имеет более яркие всплески в отличии от функции *S*2. Также как и при анализе функции *S*2 можно наблюдать значительные всплески на участке, соответствующем звуку «х». Таким образом, для решающей функции *S*3 сложно подобрать порог, который позволит обнаруживать границы между всеми звуками при условии, что не будет возникать участков с ложно определенными границами.

В качестве решающей функции может также использоваться сравнение долей энергий в пересекающихся частотных интервалах:

$$S_4 = \begin{cases} \dfrac{\sum\limits_{r\in G_{12}} P_{r1}}{\|\vec{x}_1\|^2} \cdot \dfrac{\sum\limits_{r\in G_{12}} P_{r2}}{\|\vec{x}_2\|^2}, & \text{при } S_1 > 0 \\ 0, & \text{при } S_1 = 0 \end{cases}, \quad (11)$$

где P_{r1}, P_{r2} – значение энергии в r-ом частотном интервале соответственно для первого и второго отрезка анализа,

$G_{12} = R_{11} \cap R_{12}$ – пересечения множеств $R11$ и $R12$,

$\vec{x}_1$, $\vec{x}_2$ – анализируемые отрезки сигнала.

Чем больше функция $S4$, тем больше вероятность того, что данные отрезки были порождены одним и тем звуков, т.е. больше вероятности того, что гипотеза $H0$ верна.

На рисунке 15 представлена функция S4 для фрагмента сигнала, соответствующего слову «черепаха», представленного на рисунке 11.

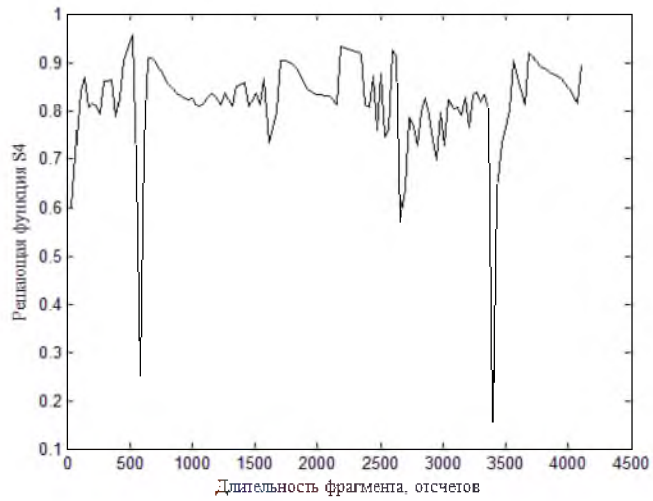

Рис. 15. Решающая функция $S4$ (N=64, R=16, m=0,9, шаг=32)

Анализ рисунка 15 показывает, что наименьшие значения решающей функции $S4$ имеют участки соответствующие переходу между звуками «ч» и «е», «а» и «х», «х» и «а», а также окончанию второго звука «е». Анализ представленной функции показывает, что также как и для предыдущих решающих функций, невозможно подобрать однозначный порог, так как имеются участки, где решающая функция $S4$ имеет относительно малые значения, несмотря на то, что эти участки принадлежат одному и тому же звуку речи.

Анализ рассмотренных решающих функций показывает, что наилучшие результаты показывают решающие функции $S2$ и $S3$, основанные на сравнении значений энергий в заданных частотных интервалах. При этом важно отметить, что выбор порога для представленных решающих функций является сложной задачей, требующий адаптивного подхода. Для повышения вероятности правильного обнаружения границ сегментов можно использовать комбинацию нескольких решающих функций.

Работа выполнена в рамках гранта РФФИ 10-07-00326-а.

Литература

1. Сорокин, В.Н. Сегментация речи на кардинальные элементы / В.Н. Сорокин, А.И. Цыплихин// Информационные процессы, 2006, Т. 6, № 3, с. 177-207.

2. Сорокин В.Н. Сегментация и распознавание гласных/В.Н. Сорокин, А.И. Цыплихин // Информационные процессы, Т. 4 2004. № 2 – С. 202-220.
3. Аграновский А.В. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов/ А.В. Аграновский, Д.А. Леднов – М.: Радио и связь, 2004. – 164с.
4. Жуйков В.Я. Алгоритм классификации сегментов речевого сигнала/В.Я. Жуйков, А.Н. Харченко//Электроника и Связь, тематический выпуск "Электроника и нанотехнологии", часть 1, № 2-3, 2009, стр. 130-137.
5. Жуйков, В.Я. Алгоритм автоматической классификации сегментов речи на основе автокорреляционных и энергетических характеристик /В.Я. Жуйков, Н.Н. Кузнецов, А.Н. Харченко// Электроника и связь 5' Тематический выпуск «Электроника и нанотехнологии», 2010, с. 83-89.
6. T. Van Pham. Wavelet analisys for robust speech processing and applications. – 2007. – 171 p.
7. Осин А.В. Сегментация речи с использованием вейвлет-преобразования / А.В. Осин, Р.Р. Ахметшин// Электротехнические и информационные комплексы и системы №2, т.2, 2006 г., с.30-32.
8. Федоров В.М. Сегментация сигналов на основе дискретного вейвлет-преобразования /В.М. Федоров, П.Ю. Юрков// Информационное противодействие угрозам терроризма, №12, 2009г. с. 138-146.
9. Жиляков Е.Г. Методы обработки речевых данных в информационно-телекоммуникационных системах на основе частотных представлений [Текст] / Е.Г. Жиляков, С.П. Белов, Е.И. Прохоренко – Белгород, 2007. – 136 с.

SEGMENTATION OF SPEECH SIGNALS BASED ON ANALYSIS OF ENERGY FOR FREQUENCY BAND

E. G. ZHILYAKOV
E. I. PROKHORENKO
A. V. BOLDYSHEV
A. A. FIRSOVA
M. V. FATOVA

Belgorod National Research University

e-mail: Zhilyakov@bsu.edu.ru

The article describes some algorithms for the segmentation of speech signals based on the analysis of energy distribution in frequency range. Proposed several crucial functions for the segmentation of speech signals based on an analysis of the energy distribution over frequency range.

Key words: speech signal, analysis of the speech signal, frequency representation, the distribution of energy over the frequency bands, the segmentation of the speech signal.