УДК 621.391

ОБ ЭФФЕКТИВНОСТИ РАЗЛИЧНЫХ ПОДХОДОВ К СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ НА ОСНОВЕ ОБНАРУЖЕНИЯ ПАУЗ[1]

Е.Г. ЖИЛЯКОВ
С.П. БЕЛОВ
А.С. БЕЛОВ
А.А. ФИРСОВА
А.В. ГЛУШАК

Белгородский государственный университет

e-mail: Zhilyakov@bsu.edu.ru

В статье приведена сравнительная оценка эффективности ряда существующих методов сегментации речевых сигналов на основе обнаружения пауз и метода, основанного на принципе учета отличий распределения энергии речевого сигнала по частотному диапазону, соответствующего звуку, по сравнению с распределением энергии сигнала в паузе.

Ключевые слова: речевой сигнал, анализ речевого сигнала, модель VAD, алгоритм обнаружения пауз, частотные представления.

Одним из этапов обработки речевых сигналов в современных информационно-телекоммуникационных системах, широко используемых в различных приложениях, является их сегментация на основе обнаружения пауз [1].

При этом в качестве основного критерия эффективности применяемых методов достаточно часто используется мера достоверности принятия решения о наличии паузы в обрабатываемом речевом сигнале, которая количественно может быть оценена вероятностями «ложной тревоги» или «пропуска цели».

В статье, на основе указанного критерия, проводится сравнительная оценка эффективности ряда существующих методов обнаружения пауз и метода, основанного на принципе учета отличий распределения энергии речевого сигнала по частотному диапазону, соответствующего звуку, по сравнению с распределением энергии сигнала в паузе.

Широкое применение в информационно-телекоммуникационных системах нашли алгоритмы Voice Activity Detector (VAD). Реализация алгоритмов VAD основана на различиях речевого сигнала и шума. При этом основное внимание уделяется следующим особенностям:

1) речь является нестационарным сигналом;

2) фоновый шум стационарен на более длинном отрезке времени по сравнению с речью;

3) уровень речевого сигнала обычно выше уровня фонового шума.

Одной из простейших реализаций VAD является принятие решение о наличии или отсутствии полезного сигнала на основе сравнения уровня энергии фрагмента сигнала с энергетическим порогом. Но такая реализация алгоритма целесообразна лишь в том случае, когда уровень фонового шума низкий по сравнению с уровнем сигнала, порождаемого звуками речи. Уровень фонового шума может меняться в течение времени. В этом случае используются более сложные алгоритмы. В системах GSM и IP-телефонии в алгоритмах VAD обработка сигнала осуществляется в частотной области. При приятии решения о наличии или отсутствии полезного сигнала используются особенности спектральных характеристик речи и шума [2, 3, 4].

В алгоритмах VAD широко используются коэффициенты автокорреляции для определения энергетического уровня сигнала и его стационарности. Решение о нали-

[1] Исследования выполнены при финансовой поддержке гранта РФФИ № 10-07-00326-а

чии речевого сигнала принимается в том случае, если энергия сигнала превышает пороговое значение и сигнал является нестационарным.

Для определения, является ли сигнал стационарным или нет, средний спектр, представленный усредненными автокорреляционными LPC-параметрами *A*, сравнивается со средними значениями автокорреляции сигнала, вычисленными в текущем фрейме с использованием уравнения:

$$Df_n = A_n(0)r_n(0) + 2\sum_{i=1}^{p} A_n(i)\frac{r_n(i)}{r_n(0)}, \quad (1)$$

$$r(i) = \sum_{k=0}^{N-i} x(k)x(k+i), \quad (2)$$

$$A(i) = \sum_{k=0}^{p-i} a(k)a(k+i), \quad (3)$$

где Df_n – средние значения автокорреляции сигнала, вычисленные для *n*-го отрезка;

$r_n(i)$ – коэффициенты автокорреляции *n*-го отрезка входного сигнала;

$A_n(i)$ – коэффициенты автокорреляции средних LPC-параметров *n*-го отрезка;

p – порядок модели;

i – изменяется от 0 до *p*;

N – длина окна анализа;

x – анализируемый сигнал;

a – средние LPC-параметры, рассчитываемые на основе средних коэффициентов автокорреляции с использованием алгоритма Дурбина.

Если абсолютное значение разности между значениями *Df* текущего и предыдущего фреймов больше, чем установка порога, текущий фрейм считается нестационарным, иначе – стационарным [2, 3, 5].

Так как речевой сигнал может быть спектрально стационарным длительное время, для различения речи и фонового шума в качестве индикатора используется периодичность речи. Значения задержек LTP сравниваются с наименьшим значением задержки. Если оставшиеся задержки очень близки к минимальной задержке, фрейм считается периодическим, в противном случае – апериодическим [2].

Для определения энергии сигнала возбуждения также могут использоваться коэффициенты автокорреляции:

$$E = A(0)r(0) + 2\sum_{i=1}^{p} A(i)r(i), \quad (4)$$

где *E* – остаточная энергия;

r(i) – коэффициенты автокорреляции входного сигнала;

A(i) – коэффициенты автокорреляции средних LPC-параметров;

p – порядок модели.

Пороговые значения энергии и разницы между значениями *Df* текущего и предыдущего фреймов определялись на основе анализа обучающей выборки сигнала, относящегося к шуму. Для определения пороговых значений отрезок шума разбивался на фрагменты одинаковой длины *N* (64, 128 отсчетов) со сдвигом 5 отсчетов (всего для анализа использовалось 400 фрагментов). Для каждого фрагмента вычислялись значения остаточной энергии *E* (4) и средние значения автокорреляции *Df* (1). В качестве энергетического порога выбиралось максимальное значение остаточной энергии среди фрагментов шума, используемых на этапе обучения. В качестве порога для принятия решения о стационарности выбиралось максимальное значение из полученных на этапе обучения абсолютных величин разностей между *Df* соседних фрагментов.

Исследование эффективности работы метода проводилось для различных значений порядка модели предсказания *p*=2?30. Решение об отсутствии паузы принимается в том случае, если рассчитанное значение остаточной энергии сигнала *E* (4)

и абсолютное значение разности между *Df* текущего и предыдущего фреймов больше пороговых значений.

Оценка эффективности работы алгоритма осуществлялась на основе определения вероятностей ошибок первого и второго рода. При этом за основную принималась гипотеза о наличии паузы. В этом случае $P_{л.т.}$ – вероятность ошибки «ложная тревога» (когда основная гипотеза о наличии паузы ошибочно отвергается), а $P_{п.ц.}$ – вероятность ошибки «пропуск цели» (когда основная гипотеза о наличии паузы ошибочно принимается).

Вероятность принятия ошибочного решения определялась в два этапа. На первом этапе анализировался фрагмент сигнала, относящийся к паузе, отличающийся от обучающей выборки. Вероятность ошибки «ложная тревога» определялась как:

$$P_{л.т.} = 1 - N_o / N_n, \quad (5)$$

где $N_{o.}$ – количество отрезков, отнесенных к паузе,

N_n – количество отрезков паузы.

На втором этапе анализировался фрагмент сигнала, относящийся к речи. Вероятность ошибки «пропуск цели» определялась как:

$$P_{п.ц.} = N_o / N_p, \quad (6)$$

где N_o – количество отрезков, отнесенных к паузе,

N_p – количество отрезков речевого сигнала.

Для определения значения вероятности $P_{л.т.}$ анализировалось 3992 отрезка. Для определения значения вероятности $P_{п.ц.}$ анализировалось 3843 отрезка. В табл. 1 представлены результаты исследования работы алгоритма VAD при различных значениях длины окна анализа для значения порядка фильтра равного 8, которое наиболее часто используется в фильтрах линейного предсказания [2].

Таблица 1.

Оценка вероятности принятия ошибочного решения алгоритма VAD

Параметры	$P_{л.т.}$		$P_{п.ц.}$	
	N=64	N=128	N=64	N=128
1	2	3	4	5
p=8	0,16	0,15	0,00	0,00

Основную опасность при обработке сигнала представляют ошибки «пропуск цели», поэтому при разработке алгоритма VAD главным является, чтобы вероятность $P_{п.ц.}$ была минимальна, при этом вероятность $P_{л.т.}$, чаще всего выбирается достаточно большой.

Таким образом, рассмотренный метод имеет достаточно большое значение $P_{л.т.}$, что не позволяет минимизировать объем передаваемых данных и приводит к тому, что сегментация не является достоверной.

Исследования тонкой структуры энергетического спектра речевого сигнала в частотной области позволили установить, что энергия звуков речи распределена неравномерно и, сосредоточена в достаточно узких частотных интервалах, в то время как энергия отрезка сигнала, принадлежащего паузе, распределена равномерно во всем анализируемом частотном диапазоне. В связи с этим, в работе предлагается в качестве процедуры обнаружения пауз использовать метод, основанный на принципе учета отличий распределения энергии речевого сигнала по частотному диапазону, соответствующего звуку, по сравнению с распределением энергии сигнала в паузе.

Для анализа особенностей речевых сигналов можно использовать метод вычисления точных значений долей энергии, попадающих в заданный частотный интервал [6].

Полный набор долей энергии отрезка сигнала можно определить следующим образом:

$$P_r = \vec{x}^T A_r \vec{x}, \quad (7)$$

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:

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

$$a_{ik}^r = (\sin(v_{r+1}(i-k)) - \sin(v_r(i-k)))/(\pi(i-k)), \; i,k = 1,...,N, \quad (8)$$

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

$$0 \le v_r < v_{r+1} \le \pi, \quad r=1,...,R, \quad (9)$$

$$v_{r+1} - v_r = \pi / R, \quad (10)$$

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

Для принятия решения о наличии или отсутствии паузы вычисляется решающая функция для проверки гипотезы о том, что анализируемый отрезок сигнала соответствует паузе между звуками речи (основная гипотеза) [7]:

$$W_{NR} = f_{NR}^m / R, \quad (11)$$

где f_{NR}^m – минимальное количество частотных интервалов (частотная концентрация), в которых сосредоточена заданная доля энергии m звукового отрезка, т.е.:

$$f_{NR}^m = \min d_{NR}^m \quad (12)$$

Здесь выполняется неравенство:

$$\sum_{k=1}^{d_{NR}^m} P_{(k),N} \ge m\left\|\vec{x}_N\right\|^2 = m\sum_{i=1}^{N} x_i^2 \quad (13)$$

где $\vec{x}_N$ – анализируемый отрезок сигнала,

m – заданное значение доли энергии сигнала,

$P_{(k),N}$ – упорядоченные по убыванию доли энергий сигнала, попадающих в заданные частотные интервалы, т.е.:

$$P_{(k),N} \in \left\{P_{rN}, r = 1,...,R\right\} \; P_{(k+1),N} \le P_{(k),N}, \; k=1,...,R \quad (14)$$

где P_{rN} – доли энергий сигнала, попадающих в заданные частотные интервалы, определяемые с помощью (7).

Если выполняется неравенство:

$$W_{NR} < w_{\text{пор}}, \quad (15)$$

то основная гипотеза отвергается, в противном случае принимается решение о наличии паузы.

$w_{\text{пор}}$ в (15) – пороговое значение, которое выбирается на основе анализа особенностей распределения долей энергии звуков речи и шума [7]. Анализ особенностей распределения энергии по частотным интервалам звуков русской речи показал, что все звуки речи имеют различное распределение долей энергии по частотным интервалам, при этом основная энергия сигнала сосредоточена в узком частотном диапазоне. В данной работе представлены результаты экспериментов для пороговых значений $w_{\text{пор}}$=0,4 и $w_{\text{пор}}$=0,5.

Для оценки эффективности метода анализировались отрезки одинаковой длины N (64, 128 отсчетов). В данной работе проводились эксперименты при различных значениях количества частотных интервалов, на которые разбивается частотная ось R: 16, 32, 64; и значения заданной доли энергии m=0,80? 0,99.

Оценка вероятностей $P_{л.т.}$ (когда основная гипотеза о наличии паузы ошибочно отвергается) и $P_{п.ц.}$ (когда основная гипотеза о наличии паузы ошибочно принимается) осуществлялась, так же как и при исследовании эффективности алгоритма VAD (5), (6).

Сравнение результатов работы алгоритма показывает, что при наименьшей вероятности $P_{п.ц.}$ меньшее значение вероятности $P_{л.т.}$ достигается при N=128, R=32, $w_{пор}$=0,5, m=0,96. В табл. 2 представлены результаты экспериментов при некоторых параметрах модели.

Таблица 2

Оценка вероятности принятия ошибочного решения алгоритма без обучения при N=128 R=32

Параметры	$P_{л.т.}$		$P_{п.ц.}$	
	$w_{пор}$ =0,4	$w_{пор}$ =0,5	$w_{пор}$ =0,4	$w_{пор}$ =0,5
1	2	3	4	5
m=0.96	0,02	0,15	0,06	0,00

Сравнение работы рассмотренного метода с работой алгоритма VAD показывает, что на различных участках сигнала рассмотренный алгоритм может работать с меньшим значением вероятности ошибки. Но этот метод существенно зависит от типа шума и особенностей речевого аппарата диктора, и на некоторых участках он работает хуже алгоритма VAD. Для анализируемого фрагмента вероятность $P_{л.т.}$ для $w_{пор}$ =0,5, m=0,96 ($P_{п.ц.}$ ≈0, $P_{л.т.}$ ≈0,15) такая же, как и вероятность $P_{л.т.}$ алгоритма VAD ($P_{п.ц.}$ ≈0, $P_{л.т.}$≈0,15).

Другой способ обнаружения пауз заключается в использовании процедуры обучения на основе анализа особенностей распределения долей энергии по частотным интервалам в паузе.

На этапе обучения для отрезков сигнала, заведомо относящихся к шуму, оцениваются характеристики вида [6]:

$$P_r^{П} = \sum_{k=1}^{N_y} (P_r)_k^{П} / N_y \,, \tag{16}$$

где N_y – количество отрезков сигнала в паузе, которые используются для усреднения (обучения), что соответствует оцениванию математических ожиданий вычисляемых долей энергий в соответствующих частотных интервалах;

$(P_r)_k^{П}$ – доли энергий в соответствующих частотных интервалах для N_y отрезков обучающей выборки.

В данном случае решающая функция имеет вид:

$$S = \max\left(P_r / P_r^{П}\right), \forall r = 1,\ldots,R \,, \tag{17}$$

где P_r – доли энергий, попадающих в заданные частотные интервалы (7);

$P_r^{П}$ – результаты предварительного усреднения по достаточно большому количеству отрезков сигнала, заведомо относящихся к паузам, долей энергий, попадающих в заданный частотный интервал (16):

Если выполняется неравенство:

$$S > h_\alpha, \tag{18}$$

где h_α – порог, обеспечивающий заданный уровень вероятности ложной тревоги α на обучающей выборке,

то основная гипотеза о наличии паузы отвергается, в противном случае принимается решение о наличии паузы.

Для определения значения порога используется обучающая выборка относящихся к паузе данных. При этом после вычислений оценок математических ожида-

ний вида (17) вычисляются оценки математического ожидания и дисперсии решающей функции [6]:

$$\overline{S}_{П} = \sum_{k=1}^{N_y} (S_k^{П}) / N_y \,, \tag{19}$$

$$D_{П}^2 = \sum_{k=1}^{N_y} (S_k^{П})^2 / N_y - \overline{S}_{П}^2 \,, \tag{20}$$

где $S_k^{П}$ – значение решающей функции на *k*-ом анализируемом отрезке заведомо относящихся к паузе данных;

N_y – количество отрезков сигнала обучающей выборки заведомо относящихся к паузе.

Пороговое значение, обеспечивающее заданный уровень вероятности ложной тревоги α на обучающей выборке, определяется на основе неравенства:

$$h_\alpha \le \overline{S}_{П} + D_{П} / a_m \sqrt{\alpha} \,, \tag{21}$$

где α – вероятность ложной тревоги, задаваемая на этапе обучения;

$\overline{S}_{П}$ – математическое ожидание решающей функции;

$D_{П}$ – дисперсия решающей функции;

a_m – коэффициент, превышающий значение 2 и определяемый в процессе обучения [7].

В качестве обучающей выборки использовалось 400 отрезков сигнала, соответствующего паузе. Отрезки были получены в результате разбиения сигнала на окна одинаковой длины *N* (64, 128 отсчетов) с шагом 5 отсчетов.

Для оценки эффективности метода анализировались отрезки одинаковой длины *N* (64, 128 отсчетов). В данной работе проводились эксперименты при различных значениях количества частотных интервалов, на которые разбивается частотная ось *R*: 16, 32, 64.

Оценка вероятностей $P_{л.т.}$ (когда основная гипотеза о наличии паузы ошибочно отвергается) и $P_{п.ц.}$ (когда основная гипотеза о наличии паузы ошибочно принимается) осуществлялась, так же как и при исследовании эффективности алгоритма VAD (5), (6).

В табл. 3 представлены результаты экспериментальной оценки вероятностей ошибок «ложная тревога» и «пропуск цели».

Таблица 3

Оценка вероятности принятия ошибочного решения алгоритма с обучением N=128 R=32

Параметры	$P_{л.т.}$	$P_{п.ц.}$
1	2	3
α=0,00002	0,02	0,00

Сравнение результатов работы алгоритма VAD, алгоритма без обучения и алгоритма с обучением показало, алгоритм обнаружения пауз с обучением дает наименьшее значение вероятности $P_{л.т.}$ при условии, что вероятность $P_{п.ц.}$ для всех исследованных алгоритмов одинакова. Так для алгоритма с обучением $P_{л.т.} \approx 0{,}02$, а для алгоритма без обучения и алгоритма VAD $P_{л.т.} \approx 0{,}15$. Таким образом, легко видеть, что применение алгоритма обнаружения пауз с обучением позволяет точнее определять участки отсутствия звука в фрагменте сигнала.

Литература

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

2. Шелухин, О.И. Цифровая обработка и передача речи [Текст] /О.И. Шелухин, Н.Ф.Лукьянцев; под ред. О.И. Шелухина. – М.: Радио и связь, 2000. – 456 с.: ил.

3. Герасимов, А.В. Применение метода модифицированного линейного предсказания к задачам выделения акустических признаков речевых сигналов [Текст] /А.В.Герасимов, О.А. Морозов, В.Р. Фидельман // Радиотехника и Электроника. – 2005. – том 50. №10. – С. 1287-1292.

4. Рабинер, Л. Теория и применение цифровой обработки сигналов [Текст] / Л. Рабинер, Г. Голд. – М.: Мир, 1988. – 512 с.

5. Коротаев, Г.А. Некоторые аспекты линейного предсказания при анализе речевого сигнала [Текст] /Г.А. Коротаев // Зарубежная радиоэлектроника. – 1991. – № 7. – С.13-31.

6. Жиляков Е.Г. Методы обработки речевых данных в информационно-телекоммуникационных системах на основе частотных представлений / Е.Г. Жиляков, С.П. Белов, Е.И. Прохоренко. – Белгород, 2007. – 136 с.

7. Белов, А.С. Разработка математических моделей и алгоритмов анализа и синтеза звуковых сигналов в цифровых слуховых аппаратах: автореферат диссертации на соискание ученой степени кандидата технических наук // Белгород, 2009. – 22 с.

ABOUT EFFECTIVENESS DIFFERENT APPROACHES TO SEGMENTATION OF SPEECH SIGNALS BASED DETECTION OF PAUSE

E.G. ZHILYAKOV
S.P. BELOV
A.S. BELOV
A.A. FIRSOVA
A.V. GLUSHAK

Belgorod state university

e-mail: Zhilyakov@bsu.edu.ru

The article presents a comparative evaluation of the effectiveness of several existing methods for the segmentation of speech signals based on the detection of breaks and a method based on the principle of taking into account differences in the energy distribution of the speech signal in the frequency range corresponding to the sound, as compared with the distribution of signal energy in a pause.

Key words: speech signal, speech signal analysis, a model of VAD, pause detection algorithm, the frequency representation.