УДК 621.391

В.Я. Жуйков, д-р техн. наук, Н.Н. Кузнецов, А.Н. Харченко

Алгоритм автоматической классификации сегментов речи на основе автокорреляционных и энергетических характеристик

В статье предложен алгоритм сегментации речевого сигнала по признакам вокализации, основанный на особенностях автокорреляционной функции и распределении энергии различных звуков по частотному спектру. Показано, что классификационные характеристики предложенного алгоритма обеспечивают достаточно высокое качество сегментации не зависимо от характеристик речевого сигнала, что демонстрирует его преимущество перед алгоритмами, ориентированными на сигналы с определенными характеристиками. Приведены результаты работы алгоритма для различных типов звуков мужских и женских голосов.

The article is devoted to the speech segmentation algorithm by vocal features, based on specifics of autocorrelation function and energy distribution over frequency domain. The algorithm's classification characteristics are high enough and independent of definite speech base, what demonstrates the proposed algorithm advantage compared to algorithm made for processing of voice with definite characteristics. The operational results with various male and female utterances are considered.

Введение

Предварительная сегментация речевого сигнала на различные фонетические группы используется во многих алгоритмах обработки и кодирования речи [1,2,3,4,5]. Обработка сигнала, учитывающая его характеристики, позволяет улучшить качество звука в устройствах кодирования и декодирования.

Существует множество алгоритмов разделения речи на различные классы звуков[1,2,3,4,5]. Их общей особенностью является зависимость от характеристик речевого сигнала, для обработки которого они предназначены, и, при применении их к сигналам с отличными характеристиками, качество разделения обычно ухудшается [4]. Алгоритмы подразделяются на: линейные, использующие сравнение результатов критериев классификации со статическими либо адаптивными порогами и основанные на нейронных сетях [4]. В алгоритмах разделения в качестве критериев применяются: 1) коэффициенты автокорреляционной функции; 2) распределенные по частотным поддиапазонам значения энергии; 3) коэффициенты оператора Тигера; 4) величина энтропии; 5) кепстральные коэффициенты; 6) значения формантных частот[4,5]. Наибольшее распространение получили три первых критерия [1,3,6,7], как такие, что отражают наличие периодической составляющей сигнала, ее частотную локализацию и величину энергии.

Предложенный алгоритм классификации сегментов речи основан на использовании критериев разделения речевого фрагмента на вокализованные и невокализованные группы звуков и включает: 1) три вышеназванных критерия; 2) функцию адаптации порогов; 3) функцию окончательного принятия решения о принадлежности сегмента речи к вокализованным или невокализованным звукам, что позволяет повысить вероятность правильного разделения по сравнению с независимым использованием отдельных критериев. Количество фонетических групп, используемых при классификации, минимально, поэтому для синтеза алгоритма выбрана простая линейная структура.

Критерии классификации

Автокорреляционная функция широко применяется в качестве критерия классификации [3,4,6], так как позволяет непосредственно оценивать квазипериодические свойства сигнала. В частности, при использовании такого критерия, оценкой степени вокализованости сегмента речевого сигнала выступает значение ближайшего локального максимума этой функции [3], а расстояние от ее начала до этого максимума близко к периоду основного тона сегмента [3,6]. Однако такая оценка существенно зависит от условий записи речи и индивидуальных характеристик говорящего, так как не является нормированной, и чувствительна к случайным выбросам. Для улучшения оценки свойств сигнала предлагается *относительный интервальный корреляционный критерий*, который определяется как отношение максимального по модулю значения автокорреляционной функции (АКФ) Ka исходного сегмента $x(n)$, расположенного не ближе некоторого интервала от начального отсчета функции $n=1$, к ее максимальному

значению в начале интервала. Интервал в определении критерия вводится для уменьшения влияния нестационарности сигнала. Приведенный критерий является нормированным.

На рис 1. показаны примеры АКФ периодического а) и шумового б) сигналов частотой дискретизации 8 кГц длительностью 30 мс (240 отсчетов). АКФ периодического сигнала изменяется с частотой, близкой к частоте исходного сигнала. Амплитудное значение АКФ $Ka\left(x_{П}(1)\right)$ в точке $n=1$, незначительно превышает последующие локальные максимумы. АКФ шумового сигнала затухает с большей скоростью, чем АКФ периодического сигнала, $Ka\left(x_{Ш}(1)\right)$ гораздо больше остальных последующих значений. Заштрихованная область соответствует минимальному интервалу, который исключается из рассмотрения для устранения влияния возможной нестационарности сигнала на значение определяемого критерия, что повышает адекватность оценки вокализованности сигнала.

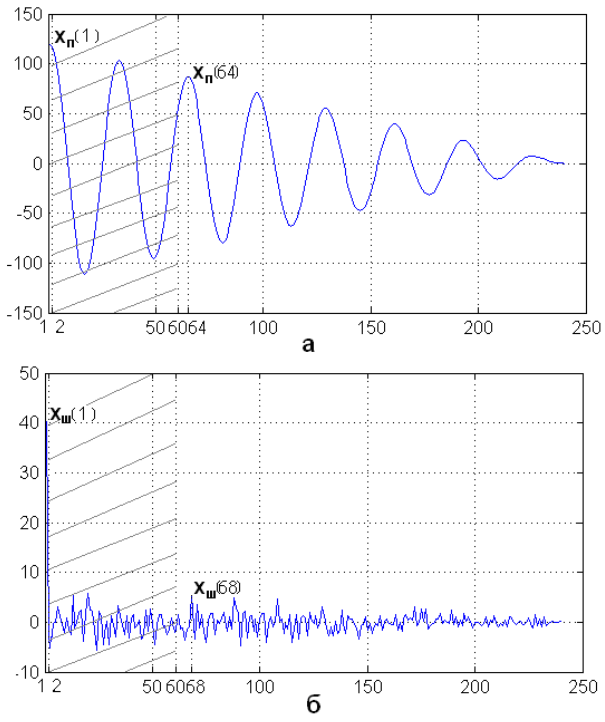

Рис. 1. АКФ периодического $x_{П}(n)$ *(а)* и шумового $x_{Ш}(n)$ *(б)* сигналов

Одним из вариантов анализа распределения энергии по спектру сегмента сигнала как оценки степени вокализованности, является отношение усредненных энергий между коэффициентами аппроксимации и/или детализации вейвлет декомпозиции сигнала в различных комбинациях [4]. При этом уровни разложения и соответствующие им частотные поддиапазоны не всегда соответствуют значимым формантным частотам речи. Для характеристики распределения энергии сегмента сигнала между низкочастотными (0…2 кГц) и высокочастотными (более 2 кГц) поддиапазонами используется *относительный полосный энергетический критерий.* Так как частота основного тона лежит в пределах 60-450 Гц, а вторая формантная частота не превышает 2кГц [8; 9], то в части спектра 0…2 кГц вокализованных сегментов следует ожидать большего сосредоточения энергии [5]. Критерием выступает значение отношения суммы энергий поддиапазонов с частотами менее 2 кГц к сумме энергий поддиапазонов с частотами 2…4 кГц. В этом критерии энергия рассматривается как сумма квадратов амплитуд последовательности отсчетов частотных поддиапазонов, при этом информация о частоте в пределах одного поддиапазона не учитывается.

Для оценки энергии сигнала, требуемой для его генерации, с учетом его частотных характеристик удобно использовать оператор Тигера $\Psi_n\left[x(n)\right]$ [10], дискретная форма записи которого имеет вид [11]:

$$\Psi_n\left[x(n)\right] = x^2(n) - x(n-1)x(n+1). \qquad (1)$$

Оператор Тигера является дифференциальным, и для оценки энергии поддиапазона его коэффициенты предлагается усреднить. Значения этого оператора существенно зависят от частотных составляющих исходного сигнала [10], поэтому для независимой оценки отдельных составляющих оператор целесообразно применять к частотным поддиапазонам, образованным, например, коэффициентами разложения дискретного вейвлет-преобразования [4,5]. Для расчета *селективного частотного - оператора Тигера критерия* предлагается использовать усредненное значение коэффициентов оператора Тигера $\Psi_n\left[x(n)\right]$ для поддиапазонов с частотами от 100 до 500 Гц. Для возможности сравнения различных по уровням амплитуд сигналов с использованием приведенного критерия, исходный сегмент сигнала необходимо нормировать по амплитуде.

Примеры использования критериев

Исходный фрагмент звукового сигнала (частота дискретизации 8 кГц) разбивается на сегменты $x(n)$ длительностью 30 мс, что соответствует $N=240$ отсчетам. Указанная длительность сегмента выбрана из условия стационарности речевого сигнала [3,12]. В *относительном интервальном корреляционном критерии*

величина интервала составляет 7.5 мс, что соответствует одной четверти длины сегмента. Такое значение интервала не содержит информации о вокализованности, так как начальный фрагмент сегмента длительностью менее 10 мс не обладает свойствами стационарности [12]. Критерий рассчитывается по выражению:

$$K_1 = \frac{\max_{n=N/4\ldots N}\left(Ka\left(\left|x(n)\right|\right)\right)}{Ka\left(\left|x(n)\right|\right)_{n=1}}, \qquad (2)$$

где $x(n)$ – исходный решетчатый сегмент сигнала; $n = \varepsilon\left[t_{сиг} f_{\partial}\right], n = 1,2\ldots240$; ε – целая часть $[.]$; $t_{сиг}$ – временная координата сегмента; f_{∂} – частота дискретизации.

Относительный полосный энергетический критерий и *селективный частотный - оператора Тигера* критерии предполагают разделение исходного сегмента сигнала на частотные поддиапазоны. Для такого разделения удобно применять дискретное вейвлет-преобразование, так как, например, при 5-ти уровнях разложения коэффициенты декомпозиции образуют поддиапазоны, соответствующие локализации основных формантных частот речи [8,9]: 1) 0…125; 2) 125..250; 3) 250…500 Гц; 4) 0.5…1; 5) 1…2; 6) 2…4 кГц. Первых три поддиапазона содержат частоту основного тона; третий, четвертый и пятый – первую и вторую форманты; шестой – высшие формантные частоты. При этом в отличие от фурье-анализа, в образованных поддиапазонах сохраняется информация о временной локализации частотных характеристик сигнала [13]. Так как энергия поддиапазонов с первого по пятый в большей степени зависит от вокализованности сегмента, чем энергия шестого поддиапазона, а второй и третий поддиапазоны локализуют частоту основного тона, то выражения для относительного полосного энергетического критерия K_2 и селективного частотного - оператора Тигера критерия K_3 соответственно имеют вид:

$$K_2 = \frac{\sum_{j=1}^{5}\left(\frac{1}{L_j}\sqrt{\sum_{n=1}^{L_j} C_j(n)^2}\right)}{\frac{1}{L_6}\sqrt{\sum_{n=1}^{L_6} C_6(n)^2}}, \qquad (3)$$

где $C_j(n)$ – коэффициенты j частотного поддиапазона; L_j – количество коэффициентов j поддиапазона; и

$$K_3 = \frac{1}{2}\sum_{j=2}^{3}\frac{1}{L_j}\sum_{n=1}^{L_j}\Psi_n\left(C_j(n)\right). \qquad (4)$$

Показателем T эффективности критериев разделения речи по типам звуков, служит вероятность правильного разделения, представляющая собой отношение количества правильно классифицированных сегментов $W_{П}$ к их общему числу [4]:

$$T = \frac{W_{П}}{W_{П} + W_{O}}, \qquad (5)$$

где W_O - количество ошибочно классифицированных сегментов.

Рассмотрим три варианта использования данных критериев:

Вариант 1. Предполагается, что все критерии равнозначны и однозначно относят сегмент к вокализованным либо невокализованным звукам. Совместное значение критериев Y_1 определяется выражением:

$$Y_1 = \frac{1}{3}\sum_{i=1}^{3} y_i', \qquad (6)$$

где $y_i' = \mathrm{sgn}(K_i - A_i)$ - промежуточный параметр; $\mathrm{sgn}(.)$ - знаковая функция; A_i – порог разделения i-го критерия на вокализованные и невокализованные сегменты (рис. 2).

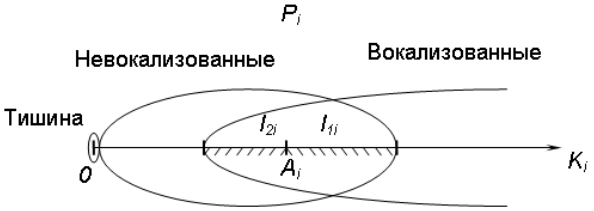

Рис 2. Параметры и пороги, используемые в различных вариантах построения алгоритма

Представленный вариант достаточно прост, но и обладает невысокой точностью, так как учитывает наименьшее количество информации о критериях.

Вариант 2. В данном варианте учитывается неоднозначность отнесения сегмента к вокализованным или невокализованным звукам и предполагается, что все критерии равнозначны.

Совместное значение Y_2 этого варианта вычисляется по выражению:

$$Y_2 = \frac{1}{3}\sum_{i=1}^{3} y_i'', \qquad (7)$$

где y_i'' – промежуточный параметр, учитывающий неоднозначность отнесения сегмента к вокализованным и невокализованным звукам,

$$y_i'' = \begin{cases} -1, \dfrac{K_i - A_i}{l_{2i}} \le -1 \\ \dfrac{K_i - A_i}{l_{2i}}, -1 < \dfrac{K_i - A_i}{l_{2i}} < 0 \\ 0, K_i - A_i = 0 \\ \dfrac{K_i - A_i}{l_{1i}}, 0 < \dfrac{K_i - A_i}{l_{1i}} < 1 \\ 1, \dfrac{K_i - A_i}{l_{1i}} \ge 1 \end{cases};$$

l_{1i} – диапазон значений, в котором невокализованные сегменты определяются i-м критерием как вокализованные; l_{2i} – диапазон значений, в котором вокализованные сегменты определяются i-м критерием как невокализованные (рис. 2).

Данный вариант более точный, чем предыдущий, так как в нем учитываются диапазоны получения возможных ошибочных значений l_{1i} и l_{2i}; $i = 1,2,3$, а промежуточный параметр y_i'' представляет собой вероятность принадлежности классифицируемого сегмента к вокализованным или невокализованным звукам согласно результатам расчета i-го критерия.

Вариант 3 является модификацией предыдущего варианта, учитывающей как неоднозначное отнесение сегмента к вокализованным и невокализованным звукам, так и неравнозначность критериев. Отличие от варианта 2 заключается во введении субъективного параметра адекватности критериев P_i. Выражение для совместного значения Y_3 этого варианта:

$$Y_3 = \frac{1}{3}\sum_{i=1}^{3} y_i''', \qquad (8)$$

где $y_i''' = \begin{cases} -1, P_i \dfrac{K_i - A_i}{l_{2i}} \le -1 \\ P_i \dfrac{K_i - A_i}{l_{2i}}, -1 < P_i \dfrac{K_i - A_i}{l_{2i}} < 0 \\ 0, K_i - A_i = 0 \\ P_i \dfrac{K_i - A_i}{l_{1i}}, 0 < P_i \dfrac{K_i - A_i}{l_{1i}} < 1 \\ 1, P_i \dfrac{K_i - A_i}{l_{1i}} \ge 1 \end{cases}$

- промежуточный параметр.

В частности, в качестве P_i предлагается использовать вероятность T (выр. 5) правильного отнесения i-м критерием сегмента к вокализованным или невокализованным звукам.

В этом варианте допускается варьирование влияния отдельных критериев на конечный результат, обеспечивая наивысшую точность при удачном подборе параметров P_i, за счет которых так же обеспечивается дополнительная возможность адаптации алгоритма.

Описание алгоритма и результаты моделирования

На рис.3 показана блок-схема алгоритма автоматической классификации сегментов речи. На первом этапе исходный фрагмент речи передискретизируется до 8 кГц и клипируется по уровню 0.01 амплитудного значения. Такое преобразование необходимо для корректной работы критериев.

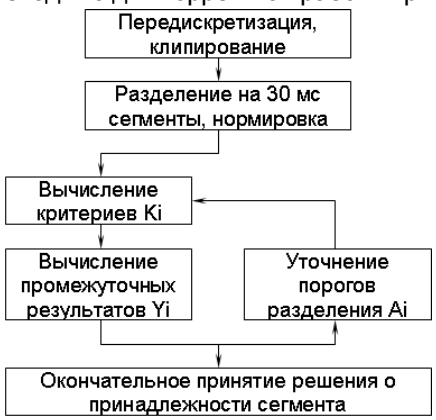

Рис 3. Блок-схема алгоритма автоматической классификации сегментов речи

В дальнейшем сигнал разделяется на 30 мс сегменты, которые нормируются по амплитуде, к каждому из них применяются критерии классификации K_i и вычисляется промежуточный результат одного из трех вариантов алгоритма Y_i. При $Y_i > 0$ сегмент считается вокализованным, при $Y_i < 0$ – невокализованным. $Y_i = 0$ – признак тишины. Так как пороги A_i в формулах (6)…(8) зависят от индивидуальных параметров говорящего и от особенностей записи речи, значения A_i уточняются в процессе работы алгоритма. Показателем адекватности величин A_i служит отношение количества вокализованных сегментов к количеству невокализованных. Окончательное решение о принадлежности сегмента к вокализованным или невокализованным звукам принимается на основании анализа значений Y_i для предыдущего, текущего и последующего сегментов. Так как частота артикуляции в среднем составляет 10-12 фонем в секунду [8], то при $Y_{i-1} > 0$ и $Y_{i+1} > 0$, величина Y_i принимает положительные значения в не зависимости от результата, рассчитанного по формулам (6) … (8). При $Y_{i-1} < 0$ и $Y_{i+1} < 0$, величина Y_i принимает отрицательные значения, при $Y_{i-1} = 0$ и $Y_{i+1} = 0$, $Y_i = 0$.

Для эксперимента использовалась база звуков английской речи, состоящая из 45 записей гласных из набора A, E, I, O, U; 28 записей согласных из набора Ch, F, H, K, P, S, Sh, T; 26 записей согласных вокализованных звуков из набора D, L, M, N, R, V. Длительность записей составляет от 30 до 270 мс, частота дискретизации – 8кГц. Звуки взяты из фраз, произносимых 2-мя мужчинами и 2-мя женщинами. Таким образом, тестовая выборка состоит из 140 гласных сегментов, 94 согласных сегмента, 50 согласных вокализованных сегментов.

Для определения среднего отношения количества вокализованных сегментов к количеству невокализованных использовались 24 записи длительностью 8 секунд фраз трех мужчин и трех женщин, записанных в тех же условиях, что и звуки речи в предыдущей базе.

Тестовая база состоит из 4-х записей фраз длительностью от 1 до 2 секунд, произнесенных на немецком языке двумя женщинами. Частота дискретизации сигнала – 16 кГц.

При определении критериев для вейвлет-анализа использовался вейвлет Добеши 3 (‘db3’) [13], позволяющий получить на данной длительности сегмента пять уровней разложения с выше указанными частотными параметрами.

Эксперимент состоит из четырех этапов.

Этап 1. Определение характеристик распределений значений критериев, и задание порогов и диапазонов. Результаты первого этапа эксперимента приведены в таблице 1.

Таблица 1. Характеристики распределений значений критериев

Критерий	Относительный дистанционный корреляционный критерий			Относительный полосный энергетический критерий			Селективный частотный - оператора Тигера критерий		
Тип звука	Гласные	Согласные	Согласные вокализованные	Гласные	Согласные	Согласные вокализованные	Гласные	Согласные	Согласные вокализованные
Математическое ожидание выборки	0,544	0,187	0,596	53,577	4,735	90,648	0,701	0,062	1,392
Среднеквадратическое отклонение	0,130	0,045	0,078	38,929	2,629	48,503	0,629	0,125	0,927
Дисперсия	0,017	0,002	0,006	1515,5	6,910	2352,6	0,396	0,016	0,859
Вероятность правильного отнесения сегмента	95%	95,74%	100%	100%	98,94%	100%	85%	86,17%	100%
Коэффициент корреляции с относительным дистанционным корреляционным критерием	1	1	1	0,398	-0,496	0,499	0,375	-0,269	0,231
Коэффициент корреляции с относительным полосным энергетическим критерием	0,398	-0,496	0,499	1	1	1	0,596	0,514	0,355
Коэффициент корреляции с селективным частотным - оператора Тигера критерием	0,375	-0,269	0,231	0,596	0,514	0,355	1	1	1
Порог A_i	0,27			12			0,1		
Параметр качества P_i	0,95			0,98			0,85		
Диапазон l_{1i}	0,05			4			0,3		
Диапазон l_{2i}	0,11			1			0,09		

Этап 2. Определялись вероятности правильного отнесения сегмента к вокализованным и невокализованным звукам каждого варианта алгоритма на основе полученных на предыдущем этапе порогов диапазонов и параметров. Результаты второго этапа эксперимента приведены в таблице 2.

Так как 3-й вариант обладает наилучшими характеристиками, он и был выбран для дальнейшего синтеза алгоритма.

Этап 3. На этом этапе определялось среднее отношение количества вокализованных сегментов к количеству невокализованных с использованием третьего варианта алгоритма. Результаты этого этапа эксперимента приведены в таблице 3.

Определенное среднее отношение количества вокализованных сегментов к количеству невокализованных, равное 1.6 (см. табл. 3), используется в алгоритме как нормальное значение отношения при корректировке порогов A_i .

Этап 4. Определение вероятности правильного отнесения сегмента к вокализованным и невокализованным звукам. Использовались записи из базы предыдущего этапа и из тестовой базы. Результаты приведены в таблице 4.

Как видно из таблицы 4, алгоритм автоматической классификации обладает лучшими характеристиками, чем отдельные критерии, что подтверждает целесообразность их совместного использования.

Вероятность правильного разделения T алгоритма составляет 91 % и несущественно зависит от характеристик записей фрагментов речи.

Таблица 2. Вероятности правильного отнесения сегмента к вокализованным и невокализованным звукам вариантов алгоритма

	Вероятность правильного отнесения сегмента к вокализованным звукам	Вероятность правильного отнесения сегмента к невокализованным звукам	Вероятность правильного отнесения сегмента к вокализованным звукам
Первый вариант	97,86 % (137 сегментов из 140)	97,86 % (137 сегментов из 140)	98,57 % (138 сегментов из 140)
Второй вариант	98,94 % (93 сегментов из 94)	98,94 % (93 сегментов из 94)	98,94 % (93 сегментов из 94)
Третий вариант	100 % (50 сегментов из 50)	100 % (50 сегментов из 50)	100 % (50 сегментов из 50)

Таблица 3. Отношение количества вокализованных сегментов к количеству невокализованных

	Фразы мужчин	Фразы женщин	Смешанные
Количество фраз	12	12	24
Среднее отношение	1,53	1,66	1,60
Максимальное отношение	2,15	2,11	2,15
Минимальное отношение	0,89	1,40	0,89
Среднеквадратическое отклонение	0,40	0,25	0,33
Дисперсия	0,16	0,06	0,11

Таблица 4. Вероятности правильного отнесения сегмента к вокализованным и невокализованным звукам

		Общее количество ненулевых сегментов	Количество правильно определенных	Количество ошибочно определенных
	записи из базы 3-го этапа	853	782 (91,68 %)	71 (8,32 %)
	записи тестовой базы	211	192 (91,00 %)	19 (9,00 %)
записи тестовой базы	относительный дистанционный корреляционный критерий	211	86 (40,76 %)	125 (59,24 %)
	относительный полосный энергетический критерий	211	173 (81,99 %)	38 (18,01 %)
	селективный частотный - оператора Тигера критерий	211	185 (87,68 %)	26 (12,32 %)

Выводы

Учет диапазона возможных ошибочных значений критериев, адаптация параметров алгоритма по среднему отношению количества вокализованных сегментов к количеству невокализованных, комбинирование различных вариантов объединения критериев и величин параметров, позволило создать алгоритм, для которого вероятность правильного разделения сигнала на 30-мс фреймы по признакам вокализации не ниже 91%, сохраняется для широкого круга различных голосовых сигналов, что является его преимуществом перед алгоритмами, ориентированными на сигналы с определенными характеристиками.

Литература

1. *C. Lemyre, M. Jelinek, R. Lefebvre.* New approach to voiced onset detection in speech signal and its application for frame error concealment // ICASSP. – 2008. – Vol. 9. – P. 4757–4760.
2. *M. Kulesza, G. Szwoch, A. Czyżewski* High Quality Speech Coding using Combined Parametric and Perceptual Modules//PWASET. – 2006. – Vol. 13. – P. 244-249.
3. *Продеус А.Н.* Цифровое кодирование речи: моделирование вокодеров в среде Matlab// Электроника и связь, тематический выпуск "Проблемы электроники", ч.1, 2006, с.56-64.
4. *T. Van Pham*. Wavelet analisys for robust speech processing and applications. – 2007. – 171 p.
5. *Жуйков В.Я., Харченко А.Н.* Алгоритм классификации сегментов речевого сигнала// Электроника и Связь, тематический выпуск "Электроника и нанотехнологии", часть 1, № 2-3, 2009, стр. 130-137
6. *Орлов А.И.* Прикладная статистика, М.: Издательство «Экзамен», 2004
7. G. Szwoch, M. Kulesza, A. Czyżewski Transient Detection for Speech Coding Applications
8. *Калинцев Ю.К.* Разборчивость речи в цифровых вокодерах. – М.: Радио и связь, 1991.
9. *Джеймс Л. Фланаган.* Анализ, синтез и восприятие речи пер. под ред. А.А, Пирогова. – М.:Связь,1968.
10. *J.F Kaiser*. On a simple algorithm to calculate the ‘energy’ of a signal // IEEE. – 1990. – Vol. 2. – P. 381–384.
11. *P. Maragos, A. Potamianos.* Higher order differential energy operators // IEEE. – 1995. – Vol. 2 – No.8 – P.152-154.
12. *Дж.Д. Маркел, А.Х. Грэй.* Линейное предсказание речи. пер. под ред. Ю.Н. Прохорова и В.С. Звездина. – М.:Связь, 1980.
13. *Электронная* книга В. Хардле, Ж. Крекьячаряна, Д. Пикара и А. Цыбакова “Вейвлеты, аппроксимация и статистические приложения”, пер. К.А.Алексеева.