УДК 004.522

© 2018 г. **А.Ю. Якимук**,
А.А. Конев, канд. техн. наук
(Томский государственный университет систем управления и радиоэлектроники)

АЛГОРИТМ СЕГМЕНТАЦИИ РЕЧЕВОГО СИГНАЛА НА ОСНОВЕ ЗНАЧЕНИЙ МИНИМАЛЬНОЙ МЕРЫ РАЗЛИЧИЯ*

В работе предложен способ повышения качества работы алгоритма сегментации речевого сигнала на вокализованные и невокализованные участки, основанного на модели периферической части слуховой системы человека. Усовершенствование заключается в использовании значений минимальной меры различия. Проведенное исследование показало: применение сглаживания повышает качество сегментации.

Ключевые слова: речевой сигнал, речевые технологии, вокализованные участки, минимальной меры различия, частота основного тона, сегментация.

DOI: 10.22250/isu.2018.56.108-121

Введение

Методы обработки речевых сигналов в настоящее время интенсивно развиваются, но по-прежнему в основе обработки речевой информации лежит сегментация. Огромное значение сегментация имеет в идентификации дикторов и распознавании их речи [1, 2]. Особенно она важна в условиях малого количества данных или высокого шума, наложенного на полезный сигнал [3, 4]. Распознавание естественно произнесенной речи намного труднее, чем отдельно сказанных слов. Это связано с тем, что границы отдельных слов определены нечетко и их произношение сильно искажено смазыванием произносимых звуков. В связи с этим в системах автоматического распознавания речи важной задачей является ее сегментация в соответствии с фонетической транскрипцией языка. Эта операция имеет большое значение не только при распознавании речи или выделении характеристик признаков голоса, но и при обратных задачах [5]: например, провести

* Работа выполнена при финансовой поддержке Министерства образования и науки РФ в рамках базовой части государственного задания ТУСУР на 2017-2019 гг. (проект №2.8172.2017/8.9).

точную проверку принадлежности к заданной модели диктора можно только после качественной сегментации [6].

В исследовательских системах и на этапе предварительной разработки возможно использование ручной сегментации. Однако она требует значительных затрат сил и времени (как утверждается в [7], обработка даже 30-секундой записи может потребовать около часа работы), и практически невозможно точно воспроизвести результаты ручной сегментации вследствие субъективности человеческого слухового и зрительного восприятия.

Подобных проблем не возникает при автоматической сегментации, которая, конечно, небезошибочна, но дает воспроизводимые результаты. Возникает необходимость разработки алгоритма сегментации, работающего с любыми языками и дикторами [8] и способного осуществлять сегментацию, близкую по результатам к ручной.

В настоящей статье рассматривается усовершенствование алгоритма сегментации речевого сигнала на вокализованные и невокализованные участки на основе значений минимальной меры различия.

Сегментация на основе значений меры различия

Используемая в исследованиях система фильтров основана на модели периферической части слуховой системы человека [9]. В ней учитывается эффект одновременной маскировки [10], возникающий в том случае, когда рядом расположенные нейроны воспринимают две или более компоненты, частоты которых находятся недалеко друг от друга. При этом частота с более высокой амплитудой подавляет частоту с более низкой амплитудой, вплоть до того, что вторая частота может вообще не восприниматься.

На основе модели был создан алгоритм автоматической сегментации сигнала на вокализованные и невокализованные участки [11]. При этом вокализованные участки включают звонкие согласные и сонорные звуки, а невокализованные – глухие согласные и «тишину».

Алгоритм основан на анализе в каждый дискретный момент времени частотной области, включающей две гармоники речевого сигнала (два непрерывных интервала единиц определенной длины, разделенных интервалом нолей). Для этого создается набор шаблонов, с которыми сравнивается структура сигнала в текущий момент времени. Шаблоны включают в себя первую и вторую гармоники основного тона. После прохождения сигнала через систему фильтров [12, 13] на каждом временном отрезке производится его свертка с частотной маской.

Алгоритм сегментации состоит из двух этапов:

1) определение вокализованности текущего временного отсчета;

2) сегментация речевого сигнала на вокализованные и невокализованные участки.

На первом этапе используются три способа вычисления меры различия. Для первого способа считается количество отличающихся каналов в шаблоне и берется их сумма; для второго – считается сумма количества отличающихся каналов в шаблоне и делится на общее количество каналов, которое было в самом шаблоне; для третьего – вместо отличающихся считаются совпавшие каналы в шаблоне.

Минимумы меры различия находятся в точках совпадения с маской, имеющей такую же частоту основного тона. Величина минимума различается для каждой маски, но остается постоянной для любого речевого сигнала. Если в какой-то момент времени эта величина больше заданного порога *min*, то данный участок признается невокализованным. На рис. 1 представлен алгоритм определения вокализованности текущего временного отсчета [14].

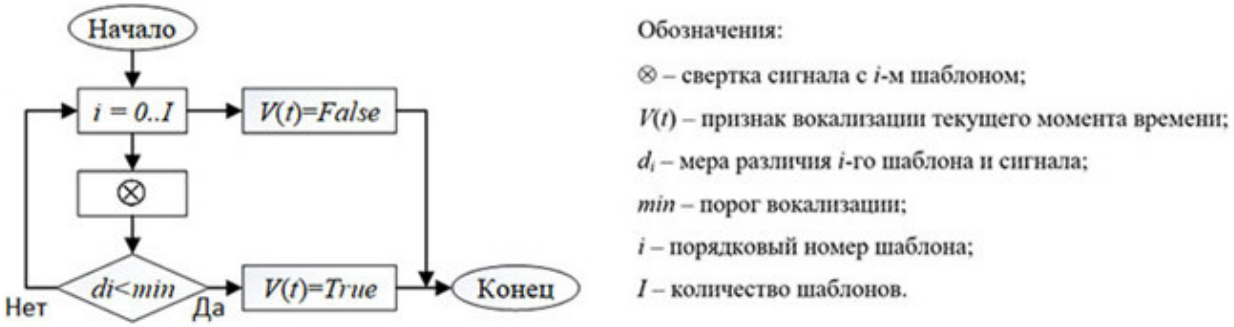


Рис. 1. Базовый алгоритм определения вокализованности временного отсчета.

Исследование алгоритмов проводилось с помощью программного комплекса SpeechSoft [15]. Ниже приводятся результаты работы исследуемого алгоритма сегментации (рис. 2), в котором еще не реализовано сглаживание минимального значения меры различия.

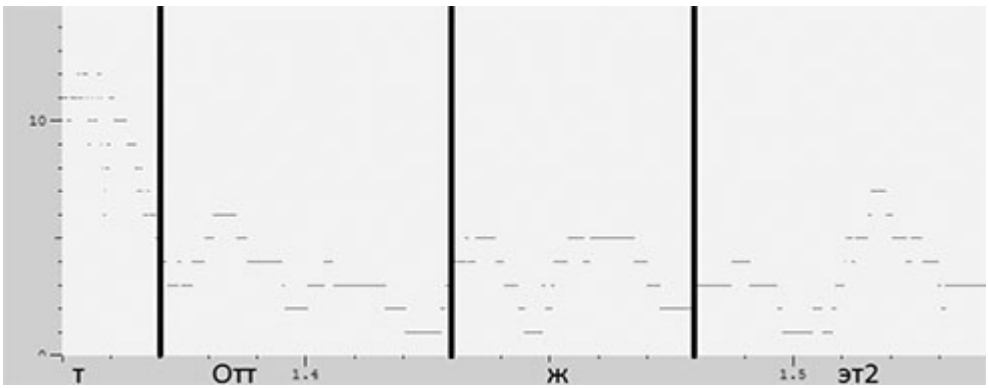


Рис. 2. Динамика изменения меры различия по способу №1.

На рис. 2 представлены границы звуков в рамках произнесенного диктором-мужчиной вокализованного сегмента «Тоже» – [т Отт ж эт2], где [т] и [ж] – твердые согласные звуки; [Отт] – ударный звук [О], расположенный между двумя твердыми согласными; [эт2] – безударный звук [э], расположенный после твердой

согласной в заударном слоге. По оси абсцисс отложено время в секундах, а по оси ординат – минимальное значение меры различия. Пунктирными линиями обозначены значения минимальной меры различия. Как правило, на невокализованных звуках она находится выше по сравнению с вокализованными. На рис. 3 представлены результаты оценки той же аудиозаписи, но по способу №2.

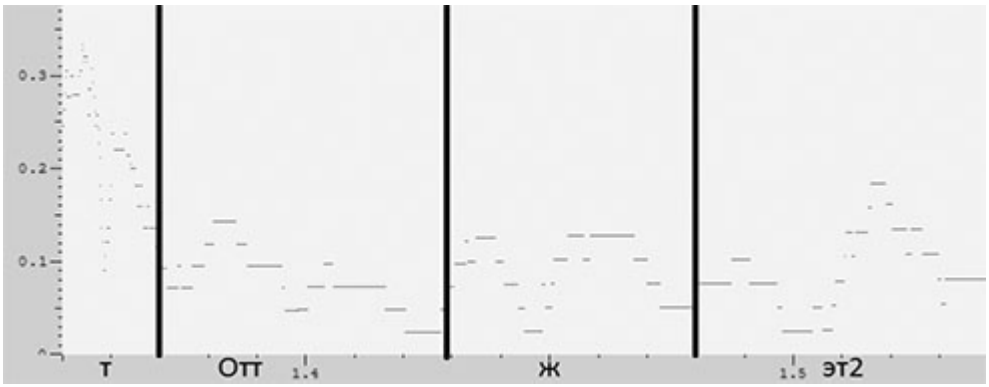


Рис. 3. Динамика изменения меры различия по способу №2.

При исследовании, является ли отсчет вокализованным или нет, учитывается наличие определенного значения порога. Выбросы, представленные выше на рисунках, на вокализованных участках часто превышают этот порог. На рис. 2 звуки [Отт] и [Ж] не выше 6, а звук [эт2] находится на отметке 7. Если значение порога задать равным 6, то участок на звуке [эт2] определится как невокализованный из-за случайного выброса, что в итоге может привести к дополнительным ошибкам в сегментации. Таким образом, задача данной работы заключается в том, чтобы "сгладить", т.е убрать эти случайные выбросы.

На рис. 4 представлен усовершенствованный алгоритм сегментации речевого сигнала на вокализованные и невокализованные участки на основе сглаживания значений минимальной меры различия.

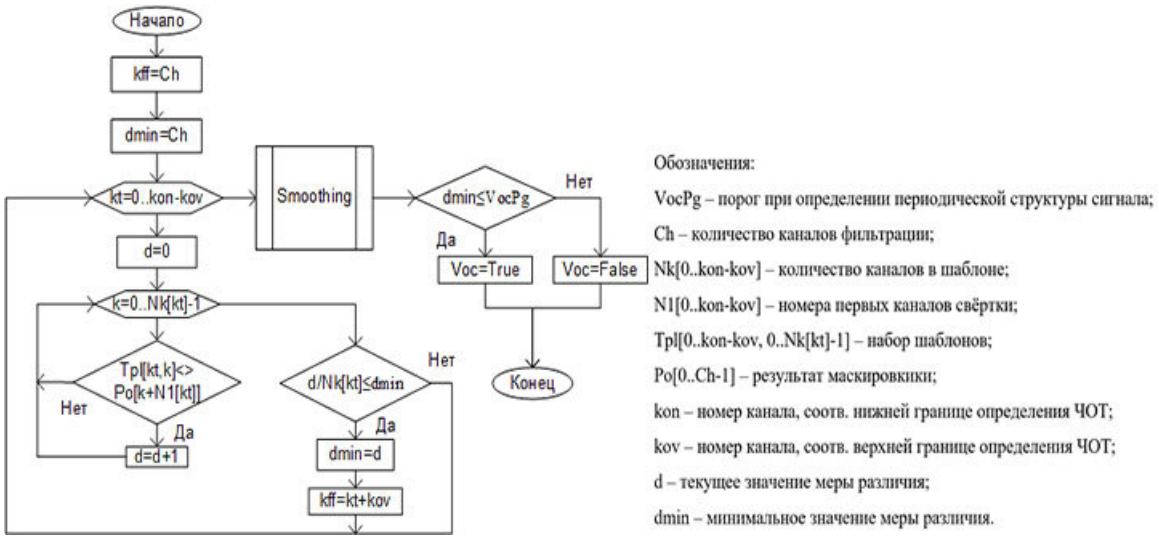


Рис. 4. Модифицированный алгоритм определения вокализованности временного отсчета с учетом получения минимального значения меры различия.

Модификация заключается в сглаживании минимального значения меры различия, а также в возможности изменения ширины окна. На рис. 5 представлена подпрограмма Smoothing, отвечающая за сглаживание.

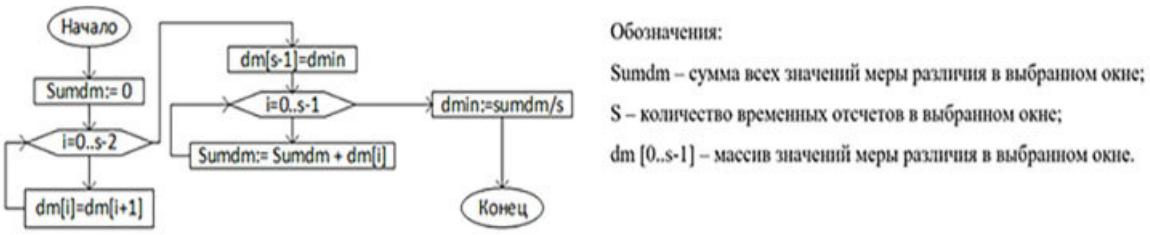
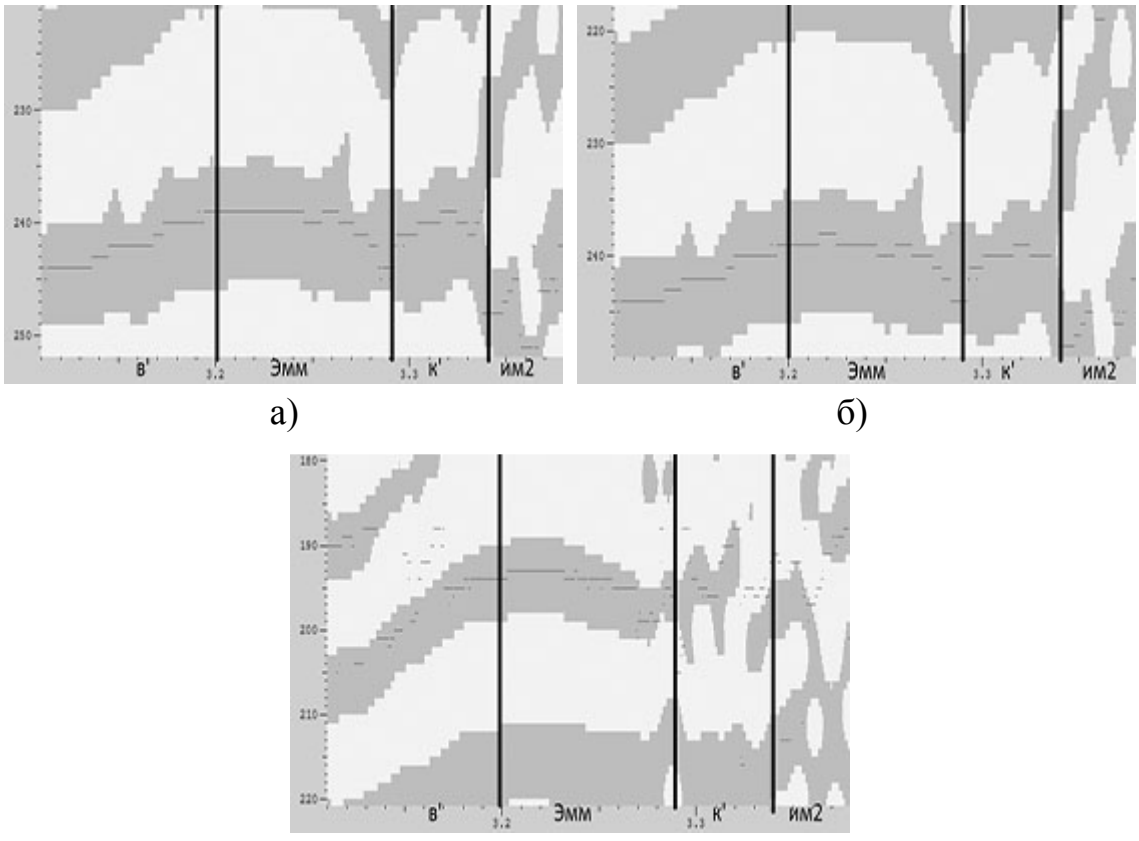


Рис. 5. Блок-схема подпрограммы Smoothing.

На рис. 6 приведена динамика изменения номера канала основного тона (далее – ОТ), соответствующая трем перечисленным выше способам, в рамках вокализованного сегмента «Веки», произнесенного диктором-мужчиной – [в’Эмм к’им2], где[в’] и [к’] – мягкие согласные звуки; [Эмм] – ударный звук [Э], расположенный между двумя мягкими согласными; [им2] – безударный звук [и], расположенный после мягкой согласной в заударном слоге. По оси абсцисс отложено время в секундах, а по оси ординат – номера частотных каналов.



а) б)

в)

Рис. 6. Динамика изменения номера канала ОТ, определяемая способами: а) №1; б) №2; в) №3.

На данных графиках белым цветом выделены компоненты, не воспринимаемые слуховой системой и представленные значением на частотном канале равном нулю. Темным выделены компоненты, воспринимаемые слуховой систе-

мой человека, отображающие гармоническую структуру сигнала. На рис. 6а и 6б номер канала ОТ определен на первой гармонике, т.е на основном тоне, это свидетельствует, что способы №1 и №2 сработали верно, а на рис. 6в номер канала ОТ определен на второй гармонике, а не на первой, как должно быть.

Полученные результаты позволяют сделать вывод, что способ №3 сработал хуже других и для дальнейшего исследования непригоден.

На всех рисунках границы звуков проставлены вручную, исходя из эталонной сегментации, которая составлялась экспертом без привлечения сторонних программ. Необходимо сравнение границ в ручной сегментации с границами, которые получились автоматически, в зависимости от входных параметров. Это позволит определить набор входных данных, при котором возможно максимальное качество сегментации.

Ниже (рис. 7 и 8) показаны результаты одновременной маскировки фразы «Он целует», произнесенной диктором-мужчиной – [Отт н ц Эт1 л Утм й’эм3 т], где [Отт] – ударный звук [О], расположенный в начале слова перед твердой согласной; [н], [ц], [л] и [т] – твердые согласные звуки; [Эт1] – безударный звук [э], расположенный после твердой согласной в предударном слоге; [Утм] – ударный звук [У], расположенный между твердой и мягкой согласными; [й’] – мягкий согласный звук; [эм3] – безударный звук [э], расположенный после мягкой согласной в заударном слоге в абсолютном конце слова. Присутствие твердой согласной перед ударной гласной не влияет на ее звучание в отличие от мягкой, что аналогично паузе или тишине. Чтобы не вводить дополнительное обозначение, будем считать, что стоящая в начале слова ударная гласная располагается после твердого согласного звука. Рассматриваться будут только первые две гармоники.

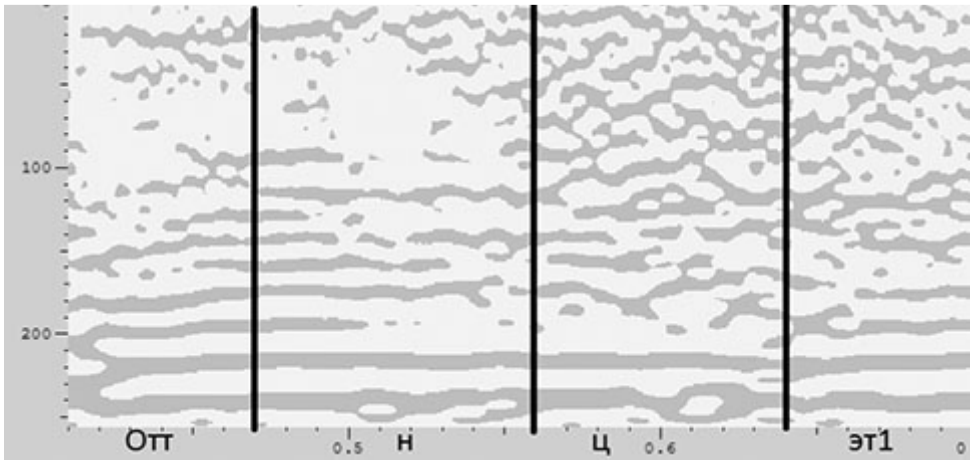


Рис. 7. Результат маскировки сегмента «Он це» – [Отт н ц Эт1].

На представленных рисунках приведены результаты одновременной маскировки. По оси абсцисс отложено время в секундах, по оси ординат – номера частотных каналов. Темным цветом выделены компоненты, которые воспринимают-

ся слуховой системой человека, белым цветом – компоненты, не воспринимаемые слуховой системой. На сегментах рис. 8, соответствующих звукам [л] и [Утм], четко просматривается «полосатая» гармоническая структура сигнала. На сегменте, соответствующем звуку «т», гармоническая структура отсутствует.

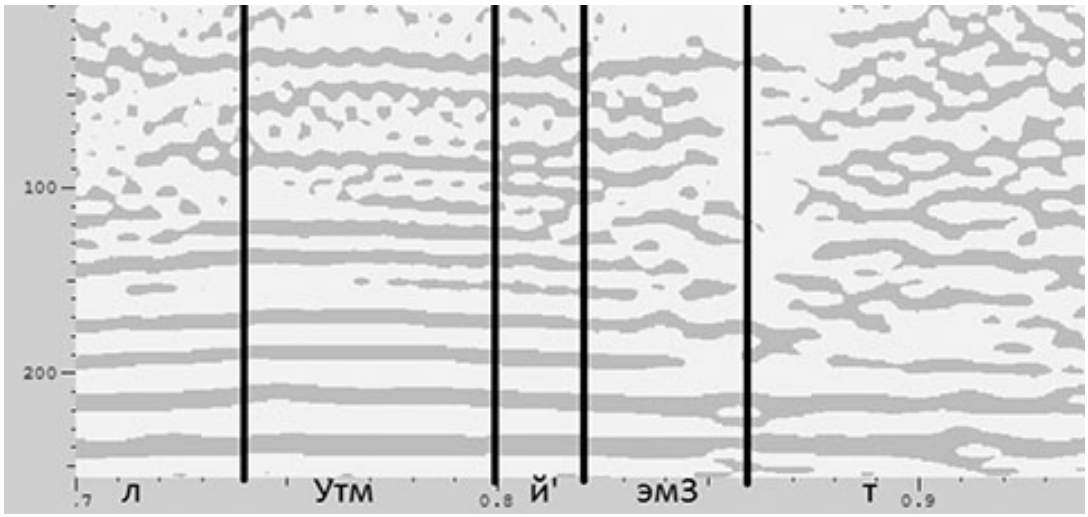
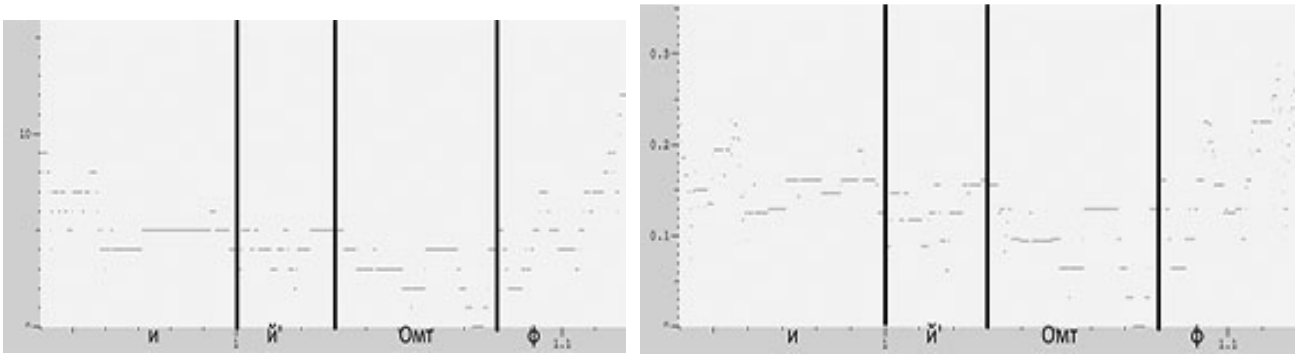


Рис. 8. Результат маскировки сегмента «Лует» – [л Утм й’эм3 т].

В результате аналогичного исследования качества определения номера канала ОТ с помощью способа №3 была установлена неприменимость данного подхода для дальнейшей работы. Номер канала должен находиться на первой гармонике, т.е на основном тоне. Однако алгоритм определил их на второй гармонике. Поскольку номера каналов определены неправильно, значения меры различия тоже будут определены с ошибкой. Существенной разницы в результатах между способами №1 и №2 на данном этапе не обнаружено.

Ниже приведены рисунки динамики изменения меры различия без сглаживания на сегменте «Ее в» – [и й’ Омт ф], где [и] – безударный гласный звук в начале слова; [й’] – мягкий согласный звук; [Омт] – ударный звук [О], расположенный между твердой и мягкой согласными; [ф] – твердый согласный звук.



а б

Рис. 9. Динамика изменения значения меры различия без сглаживания:
а) по способу №1; б) по способу №2.

Далее представлены результаты динамики изменения меры различия, определяющиеся по способу №1, со сглаживанием. На рис. 10 представлено сглаживание с шириной окна в 5 мс.

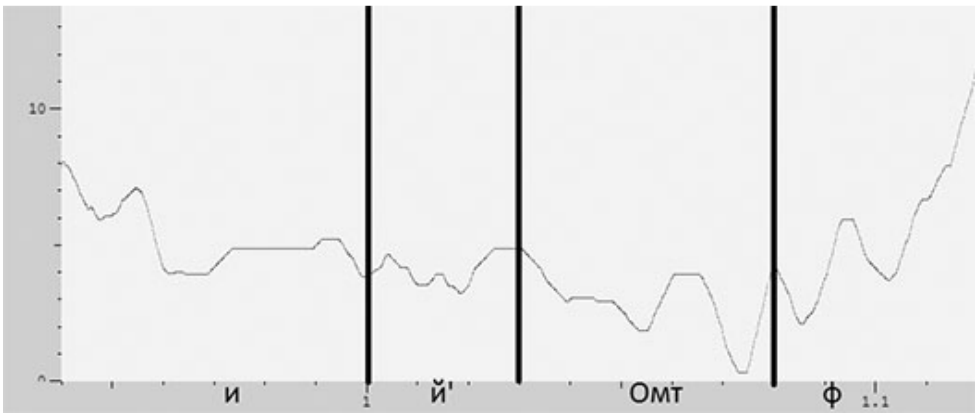
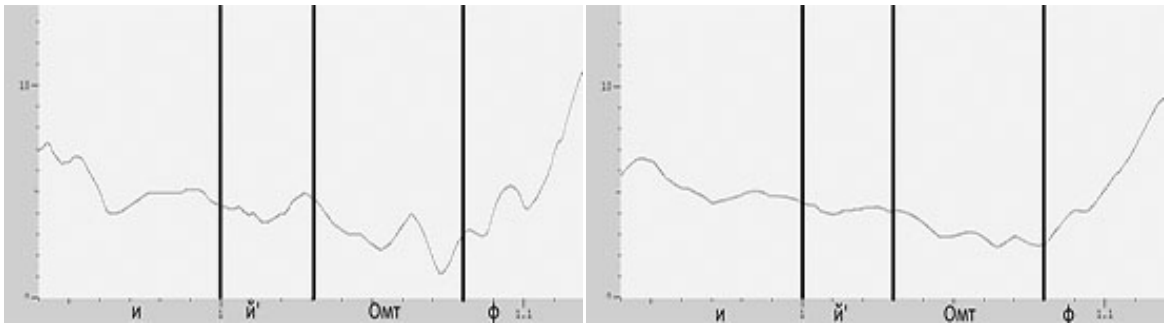


Рис. 10. Динамика изменения значения меры различия (ширина окна 5мс).

На рис. 11 представлены аналогичные графики по отслеживанию динамики изменения значения меры различия при увеличении ширины окна способом №1.

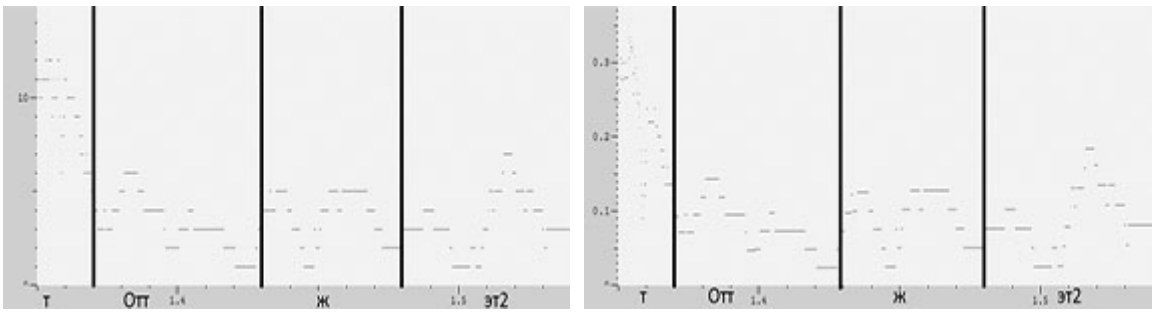


а б

Рис. 11. Динамика изменения значения меры различия при ширине окна:
а) 10 мс, б) 25 мс.

По приведенным рисункам можно сделать вывод: чем больше ширина окна, тем более сглажено значение меры различия. Следовательно, и выбросов становится меньше.

Ниже представлены результаты границ вокализованных и невокализованных звуков, без сглаживания значения меры различия.

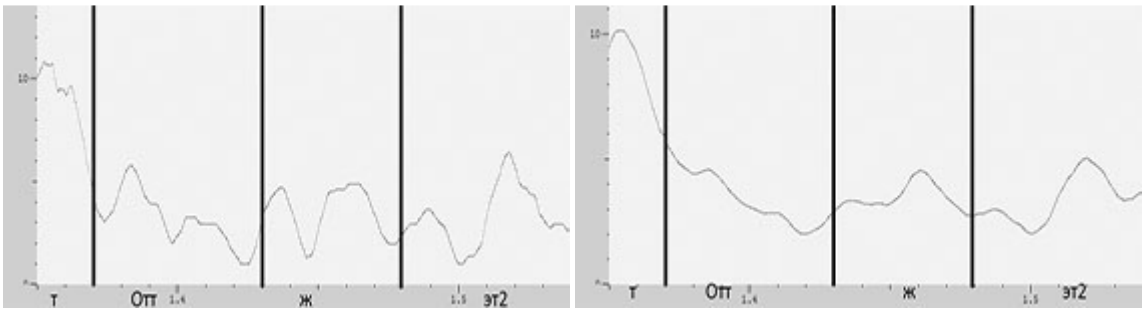


а б

Рис. 12. Динамика изменения значения меры различия:
а) по способу №1; б) по способу №2.

На рис. 12а видно, что большая часть звука вокализованного находится в районе 5. Если взять значение порога 5, то на участках [Отт] и [эт2] получаются небольшие выбросы. Это может привести к ухудшению сегментации на вокализованные и невокализованные участки, поскольку звуки [Отт] и [эт2] определятся как невокализованные. Задача заключается в том, чтобы убрать эти случайные выбросы путем сглаживания.

На рис. 13 представлены результаты измерения динамики изменения меры различия, определяющиеся по способу №1, со сглаживанием при изменении ширины окна от 5 до 20 мс. Со сглаживанием в 20мс у звука [Отт] больше 5 нет ничего, у [эт2] участок выброса стал намного меньше.



а б

Рис. 13. Динамика изменения значения меры различия при повышении ширины окна: а) ширина 5 мс; б) ширина 20 мс.

В дальнейшем оценивание качества сегментации проводилось с помощью алгоритма сегментации речевого сигнала, описанного в [16]. Для оценки автоматической сегментации были выбраны следующие критерии качества: количество поставленных несуществующих границ и количество правильно определенных временных значений границ. Правильно определенным временным значением границы принималось отличие автоматической сегментации от ручной не более чем в 0,01 сек.

Далее, в табл. 1 – 5, приведены результаты исследований по параметрам: P_0 – коэффициент определения временного значения границы без погрешности; P_1 – коэффициент определения временного значения границы с погрешностью 0,01 сек.; P_2 – коэффициент определения временного значения границы с погрешностью 0,02 сек.; $P_{>3}$ – коэффициент определения временного значения границы с погрешностью равной или более 0,03 сек.; P_- – коэффициент пропуска существующей границы; P_+ – коэффициент определения несуществующей границы.

По итогам полученной статистики осуществлен поиск соотношения данных: величина порога, ширина окна, способ определения меры различия, при котором получится наилучший результат сегментации речевого сигнала на вокализованные и невокализованные участки. Чем выше коэффициент правильно определенного временного значения границы и чем ниже коэффициент определения

несуществующей границы, тем более качественными получаются результаты.

В табл. 1 приведены окончательные результаты отношения суммы найденных границ (по 8 дикторам) к сумме эталонных границ ($Э_{гр}$ = 296) в процентном соотношении при ширине окна 0 мс. Аналогичные исследования были приведены с увеличением ширины окна до 25 мс с шагом в 5 мс.

Таблица 1

	P_0		P_1		P_2		$P_{>3}$		P_-		P_+		
Порог 1	1	2	1	2	1	2	1	2	1	2	1	2	Порог 2
6	23	34	8	8	1	2	40	41	27	12	34	36	0,09
7	23	34	8	8	1	2	40	41	26	12	34	36	0,1
8	23	34	8	8	1	2	40	41	26	12	34	36	0,11
9	23	34	8	8	1	2	40	41	26	12	34	36	0,12
10	23	34	8	8	1	2	40	41	26	12	34	36	0,13
11	23	34	8	8	1	2	40	41	26	12	34	36	0,14

В табл. 2 приведены результаты суммирования коэффициентов P0+P1, определявшиеся по способу №1, для всех значений меры различия и порога.

Таблица 2

Порог	D = 0	D = 5	D = 10	D = 15	D = 20	D = 25
6	31	67	74	71	75	75
7	31	76	77	78	77	77
8	31	70	75	75	67	75
9	31	67	61	65	58	68
10	31	67	39	67	67	67
11	31	64	67	67	65	65

По данным из табл. 2 построен график (рис. 14), на абсциссе которого обозначена ширина окна (0, 5, 10, 15, 20, 25), на оси ординат – сумма коэффициентов P_0 и P_1, на оси аппликат – величина порога (6, 7, 8, 9, 10, 11).

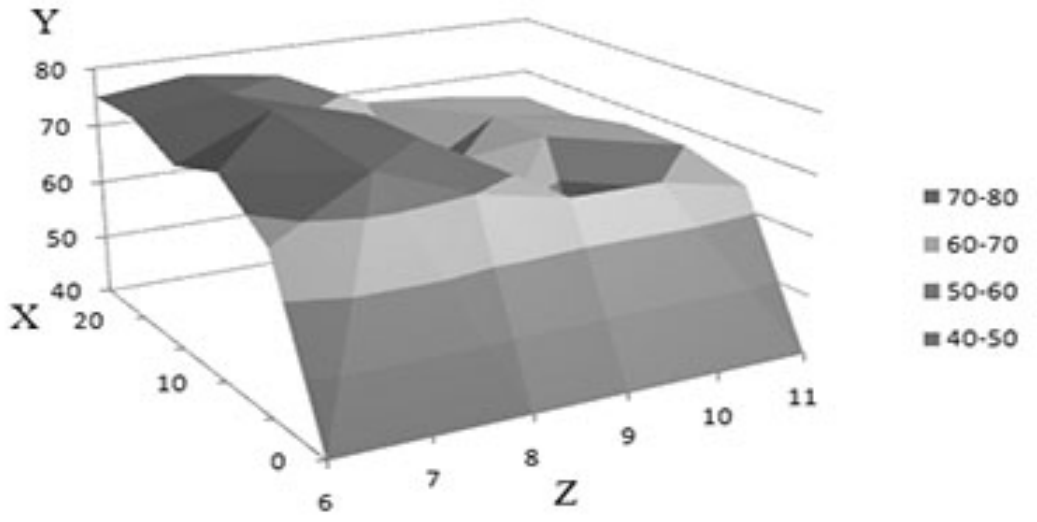


Рис. 14. Сумма правильно определенных границ P_0+P_1, определяющаяся по способу №1.

В табл. 3 представлены лишние границы P_+ для всех значений меры различия и порога.

Таблица 3

Порог	D = 0	D = 5	D = 10	D = 15	D = 20	D = 25
6	34	17	15	15	14	14
7	34	15	15	14	14	14
8	34	19	18	17	18	16
9	34	19	21	21	21	20
10	34	19	23	20	19	19

На рис. 15 представлен график, на осях абсцисс и аппликат обозначены те же значения, что и на рис. 14, а на оси ординат – лишние границы P_+.

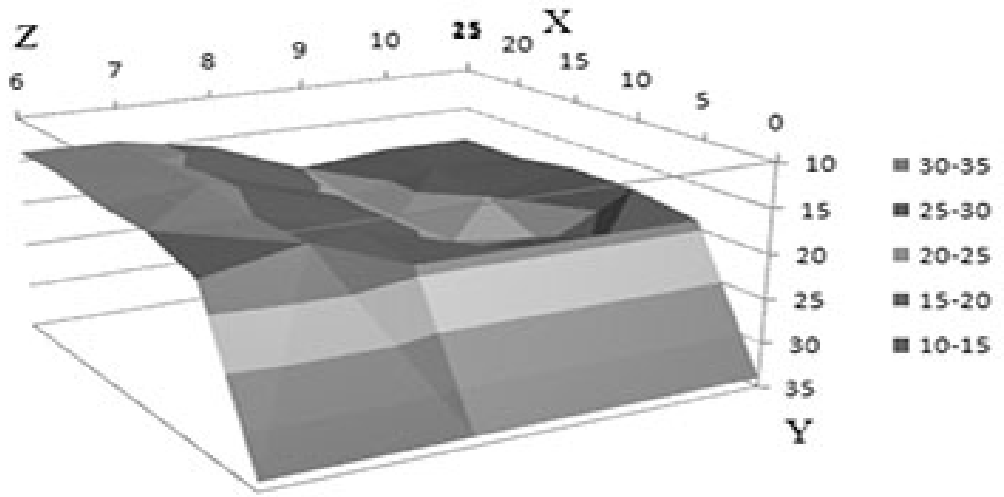


Рис. 15. Количество лишних границ P_+ , определяющееся по способу №1.

В табл. 4 показаны суммы коэффициентов P_0+P_1 для всех значений меры различия и порога, на рис. 16 – графики суммы правильно определенных границ P_0+P_1 и количества лишних границ, определяющиеся по способу №2. Ось аппликат – величина порога (0,09, 0,1, 0,11, 0,12, 0,13, 0,14).

Таблица 4

Порог	D = 0	D = 5	D = 10	D = 15	D = 20	D = 25
0,09	42	56	60	63	63	60
0,1	42	65	61	65	62	66
0,11	42	68	70	71	60	69
0,12	42	71	73	74	71	71
0,13	42	59	75	74	74	74

Аналогичным образом было проведено исследование количества лишних границ P_+, определяющихся по способу №2, результаты представлены в табл. 4 и на рис. 17.

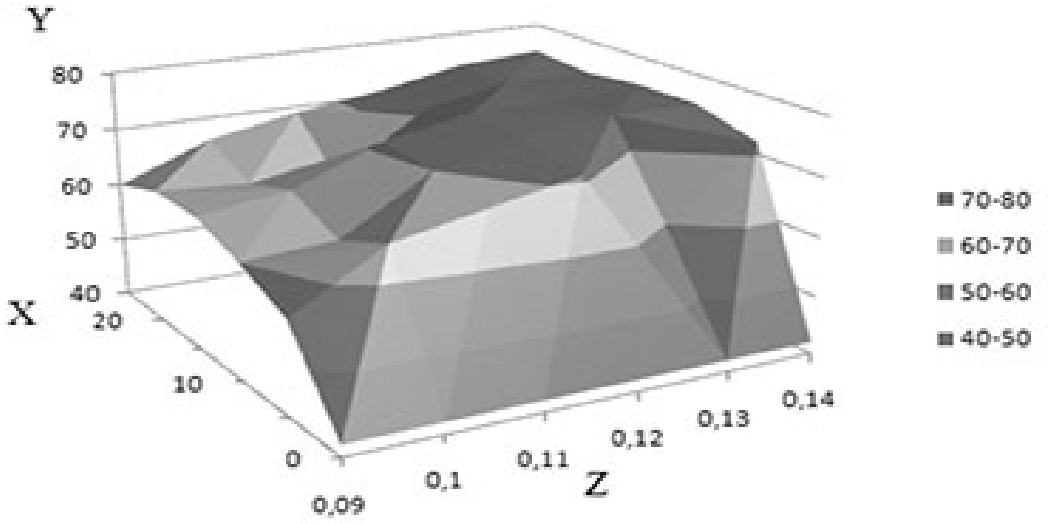


Рис. 16. Сумма правильно определенных границ P_0+P_1, определяющаяся по способу №2.

Таблица 5

Порог	D = 0	D = 5	D = 10	D = 15	D = 20	D = 25
0,09	36	18	17	17	16	16
0,1	36	17	14	16	16	15
0,11	36	14	14	15	16	14
0,12	36	15	15	15	15	14
0,13	36	18	15	16	14	14

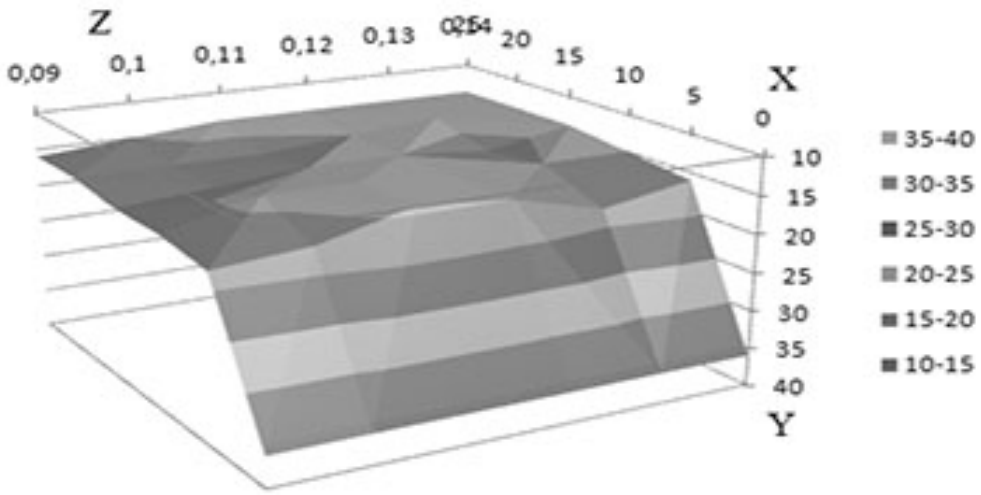


Рис. 17. Количество лишних границ P_+, определяющееся по способу №2.

По представленным рисункам можно сделать вывод, что без сглаживания, т.е при ширине окна 0, результаты получаются гораздо хуже, чем со сглаживанием.

Наилучшие результаты получились при сглаживании с шириной окна 15мс и значением порога 7, соответствующие первому способу определения меры различия.

Заключение

В ходе проведенного исследования был рассмотрен подход к сегментации речевого сигнала, исследована структура речевого сигнала на основе значений минимальной меры различия, протестирован алгоритм сегментации речевого сигнала на вокализованные и невокализованные участки. Изучение способов определения меры различия и величины порога дало следующие результаты: от 6 до11 – для первого способа, от 0,09 до 0,14 – для второго. Третий способ был исключен из исследования, так как неверно определял номер канала основного тона.

Было проведено исследование по сбору статистических данных о качестве сегментации на вокализованные и невокализованные участки, в ходе которого выявлено, что без сглаживания результаты сегментации получаются гораздо хуже, чем со сглаживанием. Максимальное качество сегментации получилось при таком сочетании входных данных: первый способ определения меры различия, величина порога 7, ширина окна 15мс ($P_0 = 63$, $P_1 = 15$, $P_2 = 1$, $P_{\geq 3} = 15$, $P_- = 5$, $P_+ = 14$, $P_0 + P_1 = 78$).

Результатом проделанной работы стала разработка алгоритма сегментации речевого сигнала на вокализованные и невокализованные участки путем сглаживания значений минимальной меры различия. Разработанный алгоритм был включен в программный комплекс автоматического исследования речевых сигналов. Поскольку сегментация является ключевой задачей в вопросах обработки вокального исполнения, что можно заметить по работам [17, 18], данный алгоритм может быть включен в программный комплекс по распознаванию нот [19].

ЛИТЕРАТУРА

1. *Benati N., Bahi H.* Spoken term detection based on acoustic speech segmentation // 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications. SETIT 2016. – 2017. – P. 267-271.
2. *Kamper H., Jansen A., Goldwater S.* A segmental framework for fully-unsupervised large-vocabulary speech recognition // Computer Speech and Language. –2017. – Vol. 46. – P.154-174.
3. *Pakoci E., Popovic B., Jakovljevic N., Pekar D., Yassa F.* A Phonetic Segmentation Procedure Based on Hidden Markov Models // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2016. – Vol. 9811. – P. 67-74.
4. *Biagetti G., Crippa P., Falaschetti L., Orcioni S., Turchetti C.* Speaker identification in noisy conditions using short sequences of speech frames // Smart Innovation, Systems and Technologies. – 2017. – Vol. 73. – P. 43-52.
5. *Рабинер Р.Л., Шавер Р.В.* Цифровая обработка речевых сигналов. – М.: Радио и связь, 1981.
6. *Рахманенко И.А.* Программный комплекс для идентификации диктора по голосу с применением параллельных вычислений на центральном и графическом процессорах // Доклады ТУСУР. – 2017. – Т. 20, № 1. – С. 70-74

7. *Brognaux, S., Roekhaut, S., Drugman, T., Beaufort, R.* Train&Align: a new online tool for automatic phonetic alignment. // IEEE Signal Processing Society. Spoken Language Technology Workshop (SLT). – 2012. – P. 416-421.
8. *Вишнякова О.А., Лавров Д.Н.* Автоматическая сегментация речевого сигнала на базе дискретного вейвлет-преобразования // Математические структуры и моделирование. – 2011. – Вып. 23. – С. 43-48
9. *Bondarenko V.P., Moor V.R., Chabanets A.N.* The analysis of speech perception mechanisms on the models of auditory system // Proceedings XIth ICPhS. Tallinn. – 1987. – V. 2. – P. 77-80.
10. Слуховая система / под ред. Я.А. Альтмана. – Л.: Наука, 1990.
11. *Конев А.А., Мещеряков Р.В., Жевуров С.В., Хлебников В.С.* Сегментация вокализованных участков речевого сигнала // Сборник трудов XXII сессии Российского акустического общества. – 2010. – Т. III. – С. 45-48.
12. *Бондаренко В.П., Коцубинский В.П., Мещеряков Р.В.* Адаптивный анализ голосового сигнала // Интеллектуальные системы в управлении, конструировании и образовании. – Томск: STT, 2004. – С.58-61.
13. *Бондаренко В.П., Пономарев А.А., Рогозинская Е.А.* Модель одновременной маскировки // Интеллектуальные системы в управлении, конструировании и образовании – Томск: STT, 2004. – С. 167-174.
14. *Бондаренко В.П., Конев А.А., Мещеряков Р.В.* Обработка речевых сигналов в задачах идентификации // Изв. вузов. Физика. – 2006. – Т. 49, №9. – С. 207-210.
15. *Егошин Н.С., Конев А.А., Якимук А.Ю.* Идентификация параметров речевого сигнала // Электронные средства и системы управления. – 2015. – № 1-2. – С. 147-150.
16. *Черных Д.В., Конев А.А., Мещеряков Р.В.* Элементы программного комплекса для оценки биометрических параметров в защищенных системах // Электронные средства и системы управления: Материалы докладов Международной научно-практической конференции. – Томск: В-Спектр, 2011. – С. 188-190.
17. *Kokkinidis K., Stergiaki A., Tsagaris A.* Error prooving and sensorimotor feedback for singing voice // ACM International Conference Proceeding Series. 3rd International Symposium on Movement and Computing. MOCO 2016. – Vol. 05-06-July-2016.
18. *Marxer R., Purwins H.* Unsupervised incremental online learning and prediction of musical au-dio signals // IEEE/ACM Transactions on Audio Speech and Language Processing. – 2016. – Vol. 24(5). – P. 863-874.
19. *Конев А.А., Онищенко А.А., Костюченко Е.Ю., Якимук А.Ю.* Автоматическое распознавание музыкальных нот // Научный вестник Новосибирского государственного технического университета. – 2015. – № 3(60). – С.32-47

Статья представлена к публикации членом редколлегии А.А. Шелупановым.

E-mail:
Якимук Алексей Юрьевич – yay@keva.tusur.ru
Конев Антон Александрович – kaa1@ keva.tusur.ru.