САНКТ – ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Математико-механический факультет

Кафедра системного программирования

Определение артикуляционных классов и точных границ сегментов слитной речи

Дипломная работа студента 545 группы

Меламуда Александра Евгеньевича

|  |  |  |
| --- | --- | --- |
| Научный руководитель | ……………… | к.т.н А.Е. Булашевич  |
| Рецензент | ……………… | д.ф.-м.н., проф. О.Н. Граничин |
| “Допустить к защите”,заведующий кафедрой | ……………… | д.ф.-м.н., проф. А.Н. Терехов  |

Санкт-Петербург

2011

SAINT PETERSBURG STATE UNIVERSITY

Mathematics & Mechanics Faculty

Software Engineering Department

DEFINITION OF ARTICULATION CLASSES AND STRICT BOUNDS OF CONTINUOUS SPEECH SEGMENTS

A graduate work by
Oleksandr Melamud

|  |  |  |
| --- | --- | --- |
|  Supervisor | ……………… | A.E. Bulashevich  |
|  Reviewer | ……………… | Professor O.N. Granichin |
| “Approved by”Head of Department | ……………… | Professor A.N. Terekhov |

Saint Petersburg

2011

**Оглавление**

[Введение 4](#_Toc293756941)

[Постановка задачи 6](#_Toc293756942)

[Анализ и выбор общей методологии сегментации 8](#_Toc293756943)

[Определение наличия голосового источника возбуждения 11](#_Toc293756944)

[Определение точных моментов прихода импульсов основного тона 16](#_Toc293756945)

[Разделение гласных и различных типов звонких согласных 22](#_Toc293756946)

[Разделение высокочастотного шума и паузоподобных участков 26](#_Toc293756947)

[Особенности реализации 28](#_Toc293756948)

[Результаты 32](#_Toc293756949)

[Список литературы 34](#_Toc293756950)

# Введение

Область применения речевых технологий охватывает широкий спектр задач и постоянно расширяется. Основные работы ведутся в направлении автоматизации распознавания человеческой речи и ее транскрибации. Этой задачей, как и задачей автоматического перевода начали заниматься чуть ли не с момента появления науки информатики как таковой. Однако на данный момент в общем случае задачу решить так и не удалось: автомата, способного распознавать слитную нелабораторную речь произвольного диктора с приемлемым качеством, до сих пор не создано. Основной проблемой речевого сигнала является его вариативность по большому набору параметров: длительность, темп, высота голоса, эмоциональные составляющие, искажения, вносимые каналом передачи информации. Следует заметить, что даже осциллограммы одинаковых звуковых фрагментов одного диктора, записанные в разное время с большой долей вероятности будут визуально различаться. Более того, подобное может наблюдаться даже в пределах одного звукового фрагмента с осциллограммами одного и того же слова.

Одними из первых попыток автоматической обработки речи были попытки распознавания изолированно произнесенных или выделенных из потока дикторской речи вручную гласных или изолированных слов, причем только по достаточно длинным стационарным участкам [1]. В качестве признаков использовались данные о формантах участка. Задача автоматического распознавания даже не ставилась.

В конце прошлого века интенсивность исследований в этом направлении значительно снизилась в виду широкого распространения статистических методов, основанных на применении скрытых марковских моделей (hidden markov model (HMM)). Применение подобного подхода позволило полностью решить задачу распознавания читаемой речи лабораторного качества. Ошибка на уровне слов не превышала нескольких процентов. Автоматы, построенные на основе HMM, приемлемо справляются с задачей распознавания дикторской речи в реальных условиях (офис, автомобиль), особенно при ограниченном словаре или при наличии настройки на конкретного диктора. В плюсы подобного подхода можно также отнести тот факт, что фонемы распознаются с учетом динамики их произнесения (обычно HMM состоит из трёх состояний, соответствующих стационарному участку фонемы (центральное состояние) и двум переходным (начальное и конечное состояния)).

Современные системы распознавания речи построены преимущественно по следующему принципу: исходный звуковой сигнал проходит стадию предобработки (нормализация, приведение к общему уровню громкости, иногда – подъем высоких частот), потом сигнал параметризируется (господствующая параметризация MFCC [9] и ее различные вариации), после этого сигнал передается на обработку автомату, реализующему HMM. В теории и диссертациях сетка кадров привязана на вокализованных участках к импульсам основного тона. На практике же господствует жесткая сетка кадров ввиду как отсутствия алгоритмов расстановки импульсов основного тона с требуемой надежностью, так и сравнительной сложности программной реализации гибкой сетки кадров.

На данный момент задачи распознавания и сегментации потока фонем решаются единовременно с помощью HMM. Хотя по своей природе это разные задачи. Границей между фонемами является переходный участок звука, где речевой тракт перестраивается. В сигнале это проявляется как сравнительно быстрое изменение характеристик сигнала. В рамках такого понимания границы между фонемами задача сегментации является независимой от конкретного языка (от набора фонем) [3], в то время как задача распознавания фонем, очевидно, зависит от языка. Проблема в том, что на данный момент подавляющее число исследований направлены на улучшение математико-статистического аппарата существующих решений. Практически отсутствуют исследования, которые бы называли достоверные корреляты, характерные той или иной группе фонем. Это выливается в то, что достоверного ответа на вопрос, контрасты каких признаков и главное в каких масштабах следует считать границами фонем (или групп фонем), особенно на слитной речи, особенно в телефонном канале, нет.

# Постановка задачи

Задачей дипломной работы является разработка механизма сегментации речевого потока на основании источника первичного возбуждения. Дополнительным ограничением является частота оцифровки сигнала до 8 кГц (частота оцифровки телефонного канала) и невозможность использования словаря. Последнее ограничение вводится исходя из того, что существует класс задач, где использование обучения в принципе невозможно. Данное исследование является составной частью разрабатываемого программного модуля по определению языка говорящего (подробнее про задачу определения языка см. [10], [11] ), отсутствие словаря, которым мог бы оперировать говорящий, и, более того, отсутствие даже набора фонем, из которых состоит речевой поток является характерной чертой этой задачи.

В соответствии с источником первичного возбуждения (голос, шум, взрыв) выделяются пять артикуляционных типов звукового сегмента:

* взрыв
* физическая пауза (глухая смычка)
* высокочастотный шум при отсутствии периодов частоты основного тона (глухая щель)
* гармонические колебания при наличии высокочастотного шума (звонкая щель)
* гармонические колебания при наличии формантной структуры (вокализованные участки)

Фонетисты выделяют [2] ещё один тип звукового сегмента, называемый звонкой смычкой. Он характеризуется наличием небольшого числа гармоник основного тона (обычно 2-3, редко 4) без формантной структуры. Взрыв занимает малое время (1-2 миллисекунды) и не имеет самостоятельного значения, так как взрыв всегда завершает смычку (и является, собственно говоря, раскрытием смычки).

Таким образом ставится задача обнаружения границ четырех типов звуковых сегментов: глухая смычка, глухая щель, вокализованные участки, звонкая щель/звонкая смычка с по возможности точным определением границ сегментов. Заметим, что при физически осмысленном подходе к сегментации проявляется близость этой задачи к задаче собственно распознавания, так как различным артикуляционным типам сегмента соответствуют чётко определенные группы фонем (гласные, звонкие согласные, глухие согласные и паузы). Так же следует заметить, что по своей сути задача распадается на несколько изолированных задач, последовательно использующих результаты предыдущих:

1. Определение наличия голоса: по этому критерию фонемы однозначно делятся на глухие и звонкие.
2. Разделение на звонкие щелевые и чисто вокализованные участки на интервалах наличия голоса.
3. Разделение глухих смычных и щелевых

Ввиду необходимости определения точных границ на вокализованных участках ставится вспомогательная задача определения точных моментов открытия голосовых связок на звонких участках (моментов прихода импульсов основного тона): именно к ним будут в последствии привязаны границы участков. Необходимо заметить и самостоятельную ценность этой задачи: при успешном, достоверном, и главное относительно точном, определении моментов прихода импульсов основного тона появляется возможность на практике при распознавании звонких участков оперировать гибкой сеткой кадров, привязанной как раз к моментам прихода импульсов основного тона.

Результатом данной работы должен стать прототип сегментирующего автомата, способный с точностью, сопоставимой с точностью работы автомата, реализующего HMM, решать поставленные выше задачи. При успешной реализации независимого сегментирующего автомата, появится возможность предобработки сигнала перед передачей его на распознающий автомат. Учитывая довольно большой процент спутывания фонем из разных классов, отсутствие больших обучающих выборок, появление предобработки в виде языконезависимой сегментации на группы фонем сможет повысить процент распознавания. Так же, учитывая кардинальные различия в физике образования различных классов фонем, появляется возможность применять качественно различные алгоритмы при распознавании фонем из одного класса.

# Анализ и выбор общей методологии сегментации

Существует несколько общих подходов при проведении независимой сегментации речи: анализ спектра сигнала, анализ временных статистик, статистико-вероятностные методы.

**Вероятностные методы**

Суть вероятностных подходов заключается в том, чтобы сконструировать автомат, в основе которого будет лежать HMM, способная выделять из речевого потока и распознавать 4 литерала (которые и соответствовали бы звуковым сегментам). Был поставлен следующий эксперимент: распознающему русский язык автомату подавались на вход речевые сигналы. После распознавания фонемы, принадлежащие одному классу по источнику первичного возбуждения, сливались в один класс. Подобный автомат сегментации показал точность порядка 70%. Причем каких-либо методов существенно улучшить полученный результат (кроме существенного увеличения обучающей выборки) не известно. Так же в минусы этого подхода можно записать то, что для успешной работы алгоритма необходима значительная по объему обучающая выборка (не менее 3-4 часов), размеченная вручную. В то же время следует отметить, что корректная ручная разметка и распознавание 1 минуты сигнала занимает порядка 8-10 часов работы квалифицированного лингвиста. Более того в задачах, где применение обучающих выборок невозможно или существенно ограниченно, применение этого подхода практически невозможно.

**Анализ сигнала в спектральной области**

Суть метода заключается в анализе амплитудного спектра сигнала. Анализируется изменение градиента амплитудного спектра с течением времени. Моменты значительного изменения градиента как раз и принимают за переходные участки. Проводились эксперименты по сегментации, которые анализировали сигнал в спектральной области. Например, [4]*.* Были получены результаты на уровне 60-70% правильных границ, причем граница считалась правильной, если находилась в пределах 15 мс от границы, поставленной при ручной сегментации. При наличии в потоке речи среднего темпа фонем с длительностью порядка 25-30 мс такая точность определения места границы представляется явно недостаточной.

**Анализ временных статистик**

Суть метода заключается в том, чтобы по изменению временных характеристик распознать текущее состояние речевого тракта, а так же наличие или отсутствие того или иного источника первичного возбуждения. Преимуществом этого подхода является очень хорошее временное разрешение. В то время как спектральные методы подобного преимущества лишены. Сложности заключаются в том, что активное внедрение HMM в начале 90-х годов, решившее проблему распознавания лабораторной речи, стало причиной того, что исследования проблемы во временной области практически прекратились. Практически отсутствуют исследования, которые бы называли более ли менее достоверные акустические корреляты тех или иных фонетических признаков, характерные той или иной группе фонем. Достоверного ответа на вопрос, контрасты каких признаков и главное в каких масштабах следует считать границами фонем (или групп фонем), нет.

**Влияние телефонного канала**

Аналоговый телефонный канал должен иметь полосу пропускания от 300Гц до 3400 Гц. Разборчивость речи на уровне слогов и слов при этом практически не отличается от разборчивости непосредственно записанной речи в широкой полосе. А на уровне фонем разборчивость ощутимо падает, так как многие согласные звуки имеют максимум спектра на высоких частотах (более 10 кГц у звука «с»). При восприятии слитной речи такие фонемы восстанавливаются человеком по контексту. Речь остается разборчивой на уровне слов при значительном сужении полосы пропускания канала, поэтому в реальности часто встречаются случаи завала АЧХ аналогового телефонного канала на высоких частотах (обычно выше 2 кГц) – за этим просто никто внимательно не следит.

Цифровой вокодерный канал, характерный для мобильной связи, обеспечивает разборчивость речи, сохранение интонации и узнаваемости голоса, однако довольно сильно изменяет спектральные характеристики сигнала. Более того, в ходе одного и того же разговора по мобильному телефону может меняться используемый вокодер при смене помехо-сигнальной обстановки. Разумеется, это существенно усложняет задачу.

Поэтому задача работы по вокодерному сигналу, несмотря на её очевидную практическую важность, в дипломной работе не ставится.

**Выбор методологии решения**

Учитывая ограничения, накладываемые постановкой задачи, использование вероятностных методов представляется затруднительным. Дополнительным аргументов является и тот факт, что проведенные эксперименты на русской речи показали точность порядка 70%. Учитывая структуру алгоритмов, реализующих HMM, качественное улучшение результата видится бесперспективным, возможно только увеличение обучающей выборки, что так же крайне трудно осуществимо. Чисто спектральные методы не способны обеспечить нужное временное разрешение, что так же ставит под вопрос возможность их использования.

Наиболее перспективным представляется решение данной задачи путем исследования временных закономерностей и зависимостей сигнала. Единственным его существенным недостатком является слабая теоретическая база. В то же время, используя этот подход, может быть достигнуто несравнимое со спектральными методами временное разрешение.

# Определение наличия голосового источника возбуждения

**Голосовой и шумовой источник возбуждения**

Рис. 1 До 20650 отсчета присутствует только шумовой источник возбуждения, после появляется голосовой источник возбуждения

Одним из наиболее характерных артикуляционных признаков разделения групп фонем является наличие или отсутствие голосового источника возбуждения [2]. Поток воздуха, мощность которого модулируется квазипериодическими (с частотой основного тона) сокращениями голосовых связок возбуждает систему резонаторов, которые и составляют речевой тракт. Почти все резонаторы включены последовательно, только носовой – параллельно. Открытыми голосовые связки находятся непродолжительное время (младшие единицы миллисекунд). После закрытия голосовых связок, резонаторы, возбужденные импульсом, порождают затухающие синусоидальные колебания, на сумму которых и приходится практически вся энергия звонких участков речевого сигнала. Для звонких участков характерна линейчатая структура спектра (заметны гармоники, следуемые с частотой основного тона), на осциллограмме легко заметить как моменты прихода импульсов основного тона, так и псевдопериодичность порождаемой осциллограммы. Вторым источником возбуждения является шумовой, основу которого составляют шумы обтекания потока воздуха, образующиеся вокруг препятствий в речевом тракте (язык, зубы, губы и т.п.). Как правило, порождаются высокочастотные негармонические колебания. Следует отметить, что шумовой источник возбуждения может присутствовать и на звонких участках (этим, в частности, и объясняется наличие гласных и звонких согласных: в первых шумовой источник отсутствует полностью), однако энергия шумовых формант значительно уступает энергии голоса. Более того энергия порождаемого шума так же модулируется ударами основного тона. Кардинально другая картина наблюдается на глухих участках, где присутствует только и исключительно шумовой источник возбуждения. Подобные участки, как правило, высокочастотны (если это только не паузы), наблюдается относительно плоский, равномерный спектр, характерный для шума. Так же следует отметить, что звонкие участки в среднем более энергетичны. На осциллограмме это проявляется в значительно большей амплитуде.

Рис2. Явление остаточного озвончения

До 8150 отсчета – гласный ‘a’, 8150 – 8900 – глухой смычный ‘t’, 8900 – 9250 – глухой смычный ‘п’
8150 – 8600 – остаточное озвончение после гласной. На спектрограмме явно видно шлейф первой гармоники, который затухает довольно медленно. Остальные гармоники затухают довольно быстро

Следует заметить, что фонемы практически однозначно делятся на звонкие и глухие. Гласные, звонкие согласные и звонкие смычные так или иначе образованы вследствие работы голосового источника, в глухих смычных и шипящих (глухая щель) голос отсутствует в принципе. Следует отдельно остановится на явлении остаточного озвончения (рис. 2), которое несколько выбивается из выше названного разбиения. Суть явления заключается в том, что в момент, когда голосовой источник перестает работать, резонаторы не затухают мгновенно. Какое-то время они продолжают порождать затухающие низкочастотные псевдопериодические колебания, которые в чем-то похожи на тихую звонкую смычку. Отличием является явно выраженное достаточно быстрое затухание. Обычно при ручной сегментации подобные участки относят к глухим, несмотря на наблюдаемую периодичность картинки.

**Определение вокализованных участков на основании энергии**

В рамках этой работы предпринимались попытки определить интервалы наличия/отсутствия голоса, анализируя изменения энергии сигнала. Подобный подход является довольно распространенным ввиду своей относительной простоты и относительно хорошего результата. За основу принят метод, предложенный Кочаровым Д.А. в своей диссертационной работе [5]. Необходимость его модификации и изменения некоторых параметров было вызвано тем фактом, что Кочаров работал с более чистым и лучше оцифрованным сигналом. Следует отметить, что в своей работе он не ставил целью определить точные границы всех вокализованных участков. Его целью было определение достоверных интервалов, на которых он бы смог оценить частоту основного тона.

Суть метода заключается в предположении, что в среднем по звуковому фрагменту энергия низкочастотных колебаний звонких участков значительно превосходит энергию глухих.

**Описание предлагаемого алгоритма для выделения звонких фрагментов**: наличие вокализации в исходном сигнале S(t) вычисляется на основе вычисления мгновенной энергии предварительно отфильтрованного сигнала SF(t). Для фильтрации применялся полосный фильтра, который оставляет полосу 400-700 Гц. Это делается для того, чтобы по возможности заглушить все частоты, на которых возможно шумовое возбуждение (как низкочастотные шумы паузы, так и высокочастотные шумы шумового возбуждения речи). На озвонченных участках остается мощная составляющая. На неозвонченных участках энергия будет практически равна нулю. После фильтрации вычислялась энергия фильтрованного сигнала по небольшому участку (по окну длиной около 10 мс с шагом в 1 мс). Далее по большому участку сигнала (в пределах 2 с) вычислялось среднее арифметическое значение мгновенной энергии. Решение о том, вокализовано ли окно (кадр) или нет принималось на основании вычисления отношения между мгновенной энергией и средней мгновенной энергией.

Окно (кадр) считается вокализованным, если отношение энергии по короткому к энергии по окну усреднения было больше некоего порога, и невокализованным иначе.

**Выявленные проблемы:** проведенные эксперименты выявили довольно сильную вариативность оптимального порогового значения (от 0.01 до 0.1). Возможным объяснением подобного феномена может служить следующее: перед началом сегментации сигнал приводится к одному уровню громкости. Учитывая то, что уровень паузы и так сильно варьируется от файла к файлу, приведение к общему уровню только усиливает эту вариативность. Для вокализованных участков, которые, как правило, являются высокоэнергетичными, подъем уровня энергии паузы существенного влияния не оказывает, а вот на участки, где голоса нет (и соответственно общий уровень энергии не сильно велик) подъем уровня энергии паузы заметен. Так же остаются вопросы и в определении длины большого участка усреднения: существуют сегменты, где звонкие непрерывные участки сопоставимы по длине с окном усреднения. Как следствие под уровнем порога оказываются значительные промежутки тихих звонких. В то же время дальнейшее увеличения окна является нецелесообразным из-за того, что громкость в пределах звукового фрагмента может существенно варьироваться и есть так же возможность пропустить тихие вокализованные участки. Так же при использовании вышеописанного подхода в поставленных экспериментах точность сегментации подобным методом не превышала 85-88%., что ставит под сомнение возможность использования только подобного механизма при первичном разделении.

**Определение звонкости с помощью линейного предсказания**

Методы линейного предсказания начали применяться при распознавании речи довольно давно [6]. Суть метода заключается в том, что мы пытаемся подобрать набор коэффициентов {аn}, линейная комбинация которых с M-предыдущих отсчетов (количество отсчетов M называется порядком модели линейного предсказания) используется для предсказания M+1отсчета. Так как основу звонких участков составляют псевдопериодические синусоидальные колебания, то предполагается, что отношение энергии линейно предсказанной части сигнала, к энергии всего сигнала будет значительно выше, нежели на глухих. Следует заметить, что существуют участки, размечаемые вручную как глухие, на которых доля линейно предсказанной части будет так же велика: это участки остаточного озвончения и паузоподобные участки. Однако они довольно легко отфильтровываются энергетически.

**Описание предлагаемого алгоритма для выделения звонких фрагментов**: сигнал фильтруется в полосе 150-4000 Гц. Фильтрация нижних частот производится для того, чтобы по возможности подавить низкочастотные шумы канала связи, однако оставить низкочастотные составляющие голоса. Вычисляются коэффициенты линейного предсказания путем минимизации ошибки на основании метода наименьших квадратов в 3-х отчетной модели. Порядок модели выбирается заведомо не сильно большими, чтобы участки, где случайным образом присутствует нечто похожее на гармоничность (хотя заведомо не является голосом), по возможности предсказывались плохо. Вычисляется сигнал предсказания на основании полученных коэффициентов. В кадрах небольшой длины (порядка 4-5 мс) считается отношение энергии линейно-предсказанной части к энергии самого сигнала. Учитывая, что в сигнале также присутствуют паузообразные фрагменты, которые также предсказываются хорошо, решение о наличии или отсутствии вокализации предлагается принимать на основании совокупности 2-х параметров: отношения энергии линейно-предсказанной части к энергии самого фрагмента (StatLin) и отношения энергии сигнала в полосе до 1000 Гц к средней энергии всего сигнала в данной полосе(StatEner). Полоса до 1000 Гц выбрана исходя из того, что практически во всех вокализованных звуках в данной полосе присутствует высокоэнергетическая составляющая.

Путем исследования гистограмм распределения вышеназванных статистик в зависимости от артикуляционных классов были выбраны следующие пороговые значения, на основании которых и принималось решение о наличии вокализации во фрагменте:

Фрагмент считается вокализованным, если:

(StatLin > 0.8 && StatEner> 0.01) || (StatLin > 0.7 && StatEner> 0.1) || (StatLin > 0.6 && StatEner> 0.2)

Автомат, реализующий вышеописанный алгоритм показал точность порядка 95-96%, что уже является сопоставимым с точностью ручной сегментации. Так же следует отметить, что большинство ошибочных отсчетов сконцентрировано на переходных участках между классами, которые однозначно отнести к тому или иному классу не представляется возможным.

# Определение точных моментов прихода импульсов основного тона

Источником голосовых колебаний являются ритмические смыкания и размыкания голосовых связок человека. Частота, с которой голосовые связки размыкаются называется частотой основного тона, а сами моменты, когда связки открыты (как правило младшие единицы миллисекунды) – импульсы основного тона. Название обычно объясняется тем, что именно в этот момент происходит скачкообразный рост энергии воздушного потока, которым возбуждаются резонаторы голосового тракта. Учитывая отсутствие других источников возбуждения резонаторов и неспособность человеческого голосового тракта быстро менять свою конфигурацию, можно считать 1 период основного тона стационарным в смысле изменения основных характеристик речевого потока (таких как спектр, динамика затухания и т.д). Характерной особенностью является тот факт, что в эти моменты происходит резкое увеличение энергии сигнала и сбрасываются фазы синусоидальных сигналов, порождаемые резонаторами в 0, что приводит к краткосрочным аномалиям в целом периодического сигнала. В зависимости от рода задач обычно либо выделяют контур поведения основного тона, либо оценивают его частоту на определенном промежутке. Подобные оценки используются, например, при эмоциональном анализе фразы. Для некоторых же задач подобных усредненных оценок недостаточно. В теории и диссертациях [6] уже давно сложилось понимание, что границы кадров обработки следует привязывать к моментам прихода импульсов основного тона, но отсутствие механизма, способного с достаточной точностью расставлять импульсы, мешало воплощению этим планам на практике. Так же подобная трассировка пригодилась бы в задачах сегментации, т.к тогда снимается вопрос, к каким моментам привязывать точную границу сегментов.

**Обзор существующих подходов к определению и оценке параметров основного тона**

**Частотные методы**

Основой метода является наблюдение факта, что в спектре сигнала, в местах, где присутствует голосовое возбуждение, различимы относительно четкие локальные максимумы на частотах кратных частоте основного тона (рис 3.). Какую-то оценку основного тона можно получить, используя вышеописанное наблюдение. Как вариант, можно пытаться оценить частоту основного тона следующим образом: дискретным преобразованием Фурье считаем амплитудный спектр сигнала, предварительно отфильтровав сигнал в полосе до 400-450 Гц, за оценку основного тона принимаем частоту наиболее энергонесущей гармоники спектра. При использовании этого подхода может возникать следующая проблема: часто в вышеназванной полосе находятся так же и гармоники, имеющие большую энергию (например, если в фонеме присутствует высокоэнергетичная низкая первая форманта, видно на рис. 3). В таком случае оценка будет неверной. Так же можно пытаться оценивать основной тон через спектр спектра сигнала. Учитывая, что гармоники идут как раз с частотой основного тона, амплитудный спектр амплитудного спектра (особенно, если после первого вычисления спектра в нем поднять энергию дальних гармоник, например, логарифмированием) сигнала покажет характерно выраженный всплеск как раз на частоте основного тона.

Следует отметить, что общим недостатком этой группы методов является необходимость выделения относительно длинного (порядка 250 отсчетов – в этом случае можно говорить о приемлемом спектральном разрешении) стационарного (при резких изменениях сигналах на промежутке анализа спектр будет смазан) участка вокализованной речи. Так же при подобном подходе о точной разметке импульсов говорить не приходится

Рис. 3. Спектр звука ‘а’. Высокоэнергетичные гармоники следуют с частотой основного тона вплоть до старших частот. Энергия гармоники на 250 Гц выше энергии гармоники на 150 Гц – частоты основного тона.

**Автокорреляционный метод**

метод заключается в максимизации автокорреляционной функции. Суть метода заключается в том, что для речевого сигнала, который близок к периодическому, верно следующее: S[i] ≈ S[i + T], где T – период основного тона, выраженный в количестве отсчётов. Этот факт делает справедливым следующее предположение: оценка периода основного тона должна максимизировать функцию (\*), которая называется автокорреляционной [7]

$$R\left(k\right)= \sum\_{i=0}^{N-1}S\left[n+i\right]S\left[n-k+i\right] (\*)$$

Несмотря на широкую распространенность этого метода у него есть существенные недостатки: на плохом соотношении «сигнал-шум» метод часто сбивается. И как и в случае с частотными методами мы получаем всего лишь оценку частоты, что делает этот алгоритм непригодным для решения задачи маркировки периодов. Также можно привести ещё несколько примеров подобных методов и их модификаций, однако все они не подходят для решения задачи выделения импульсов основного тона, так как всевозможные усреднения и зависимости от предыдущих периодов делают невозможным определение мгновенного импульса.

**Анализ осциллограммы сигнала**

Эта группа методов работает напрямую с осциллограммой сигнала, анализируя ее во временной области. Анализу подвергаются статистические данные о количестве переходов через ноль и/или анализе локальных экстремумов сигнала. Учитывая, что в момент прихода импульса основного тона происходит резкое увеличение энергии и как следствие амплитуды сигнала, довольно часто величина экстремумов в окрестности импульса больше, нежели в других частях. Однако часто наблюдается ситуация, когда в окрестности импульса находятся несколько локальных экстремумов, приблизительно равных по величине. Однозначно определить, какой из них соответствует моменту прихода импульса основного тона часто оказывается неразрешимой задачей.

**Анализ изменения мгновенной энергии сигнала**

**Teager operator (оператор Тигра):** была опубликована серия статей (например, [12]) посвященных оценке мгновенной энергии сигнала с помощью так называемого «оператора Тигра». Его формула: T(i) = (s(i)′)2 – s(i)\*s(i)′′, где s(i) – значение сигнала на i-м отсчете. Из положительных особенностей можно так же отметить простоту вычисления и хорошее временное разрешение (в дискретном случае достаточно 3-х отсчетов, чтобы оценить первую и вторую производную сигнала в точке). Главным же достоинством оператора Тигра является его способность полностью подавлять синусоидальные сигналы, оставляя только энергию: становятся видны моменты резкого увеличения энергии в момент прихода импульса основного тона (видно на рис.4). Из недостатков следует выделить необходимость двойного дифференцирования сигнала. Учитывая, что дифференцирование является фильтром высоких частот, на участках сигнала с плохим соотношением «сигнал-шум» на сигнале, получаемом из оператора тигра наблюдается значительная зашумленность. Так же дает о себе знать и необходимость дифференцирования дискретного сигнала. Если на 16 кГц сигналах тигр ведет себя вполне приемлемо, то при понижении частоты оцифровки до 8 кГц появляется значительное число ложных локальных экстремумов.

Рис. 4. На верхнем рисунке показана осциллограмма гласного звука ‘a’. На нижнем сигнал, получаемый после обработки оператором Тигра. Места увеличения мощности видны намного более явно

**Гетеродинирование:** в рамках курсовой работы 2010 года Такун Евгении [8] был предложен подход, который так же анализирует мгновенную энергию сигнала. Предлагаемый метод сочетает в себе такие преимущества Тигра как хорошее временное разрешение и простоту реализации, но заметно лучше себя ведет на зашумленных сигналах. Суть метода заключается в том, что сигнал представляется в комплексном виде (например, $\cos(x= \frac{e^{it}+ e^{-it}}{2})$). После этого каждый отсчет комплексного сигнала умножаем на сопряженный к нему. Одним из результатов подобного действия как раз и является достаточно точная оценка мгновенной мощности (которая равна квадрату амплитуды). Этот подход лишен основного недостатка тигра: необходимости дифференцирования в дискретном времени. Ошибка сосредоточена только в процедуре перевода сигнала в комплексную форму. Как результат, получается менее зашумленная оценка мгновенной энергии с явно различимыми всплесками в момент прихода импульсов основного тона в том числе и на фрагментах сигнала с плохим соотношением «сигнал-шум»

**Предлагаемое решение**

Импульсы основного тона образуются при смыкании и размыкании голосовых связок. Очевидно, что при сомкнутых голосовых связках энергия колебаний резонаторов убывает (полезно излучается вовне и бесполезно затухает, уходя в деформацию стенок). При открытых связках происходит «накачка» энергии в резонаторы.

Предлагаемый алгоритм ищет в выходном сигнале импульсы основного тона, отмечаемые как моменты смыкания голосовых связок. Основной признак – максимум энергии первого резонатора (резонатора первой форманты), отмечаемый в речевом сигнале как максимум амплитуды колебания первого (наиболее энергонесущего) резонатора.

Осложняющие факторы:

* Наличие в сигнале иных составляющих, кроме сигнала первой форманты, а именно – второй форманты и шумового возбуждения.
* Существенная нестационарность периода основного тона, который может резко меняться на ударных гласных. Женские голоса также используют изменение основного для формирования требуемого значения частоты первой форманты.

**Идея:** Для достижения цели: нахождения моментов прихода импульсов основного тона, используется следующее наблюдение: момент работы голосового источника характеризуется значительной «накачкой» энергии в речевой сигнал. Как следствие в сигнале в большинстве мест наличия голоса прослеживаются характерные «всплески» энергии (видно, например, прекращение экспоненциального затухания амплитуды огибающей, скачек, после которого возобновляется затухание), которые и являются признаком прихода очередного импульса. Достижение цели осложняется тем, что априори мы не знаем ни средних значение периодов основного тона, ни тех мест, в которых нам следует его искать. Разумным кажется следующий подход: выделить в частотной области из сигнала составляющую, на которой четко бы прослеживались «удары», выделить в сигнале места, где наличие голоса очевидно. На этих интервалах попытаться получить приблизительную оценку периода основного тона. Имея какую-то оценку основного тона мы можем попытаться построить первичные трассы импульсов основного тона, состоящие из моментов, в которых наблюдается максимум энергии, которые расположены с тем периодом ОТ, оценку которого мы смогли получить. Содержательным является вопрос, каким образом оценивать энергию сигнала. После анализа существующих решений и подходов наилучшим был признан метод гетеродинирования. Он позволяет с отличным временным разрешением получать довольно точную оценку изменения энергии сигнала. Применение этого подхода позволяет существенно ограничить число мест, подозрительных на наличие импульса основного тона. После определения первичных траекторий мы знаем достаточно точные оценки периодов основного тона по всему звуковому сигналу. Используя эти данные мы можем пытаться искать периодические траектории среди тех максимумов, которые есть в промежутках между найденными траекториями. Следует заметить, что для поиска нужных максимумов можно использовать тот факт, что максимумы в анализируемом сигнале в момент прихода импульса основного тона локально самые большие среди всех остальных.

Вышеописанная идея может быть трансформирована в следующий **алгоритм**:

1. Определить интервалы наличия голоса с помощью алгоритма, описанного в предыдущей главе.
2. На каждом из «длинных» (более 250 отсчетов) интервалов попытаться оценить основной тон спектральными методами (например, спектром спектра)
3. Попытаться провести трассы импульсов основного тона в пределах интервала, используя в качестве первого приближения частоты основного тона оценку, полученную в пункте 2, таким образом, чтобы в совокупности они покрывали до 90% длительности интервала, а период основного тона в пределах трассы изменялся не больше чем на пороговое значение (эмпирическим путем было получено значение в 25% предыдущего периода). Если такие трассы построить не удалось, это значит, что первичная оценка была получена с ошибкой, построенные трассы с большой долей вероятности неверны.
4. После выполнения пункта 3 на некоторой части интервалов наличия голоса получены достоверная разметка импульсов. Это позволяет произвести оценку частоты основного тона на всех оставшихся интервалах. В ходе экспериментов использовалось усредненное значение периодов основного тона предыдущей и последующей траекторий импульсов.
5. Пользуя оценками, полученными в пункте 4, провести трассы импульсов основного тона в интервалах, где таких трасс еще нет.

# Разделение гласных и различных типов звонких согласных

**Гласные и звонкие согласные**

Звонкие участки речевого потока делятся лингвистами на два непересекающихся класса: гласные и звонкие согласные. На артикуляционном уровне различия между этими классами заключаются в следующем:

* Во время произнесения гласных работает только голосовой источник возбуждения. В комбинации с чистым речевым трактом (полностью отсутствуют препятствия в виде резких сужений (например, щель между зубами, препятствие в виде языка и т.п.)) резонаторами порождаются мощные квазипериодические колебания, спектр вокализован вплоть до старших частот (все форманты гармонические), просматривается четкая формантная структура.
* Считается [2], что при звукообразовании согласных звуков (и звонких согласных в том числе), активное участие принимает шумовой источник возбуждения. В речевом тракте создаются шумообразующие преграды (сужение в районе зубов, положение языка, сужения в районе гортани), которые и становятся местом образования турбулентных шумов обтекания этих преград потоками воздуха. Порождаются, как правило, высокочастотные (от 2 кГц) колебания. Если параллельно с шумовым источником возбуждения работает голосовой источник, то подобные звуки относят к звонким согласным. Следует отметить, что класс звонких согласных отличается сильной неоднородностью, однако все они характеризуются:
	+ наличием низкочастотных квазипериодических колебаний: следствие работы голосового источника
	+ наличием высокочастотных шумовых формант (наблюдаются на частотах свыше 2500 Гц, следствие работы шумового источника возбуждения) – лингвистами считается одним из основных отличительных признаков
	+ неразвитой второй вокалической формантой: сужения голосового тракта затрудняют прохождение высоких частот, порождаемых резонаторами, как следствие в звонких согласных часто наблюдается заметное спадание энергии в полосе частот второй форманты: 900 – 2000 Гц
	+ меньшей энергией (следствие того, что часть мощных высокоэнергетичных колебаний голоса тухнет в речевом тракте)

Следует отдельно остановится на сложностях, связанных с частотой оцифровки сигнала в 8 кГц. Значительная часть энергии некоторых звуков (например, ‘р’) сосредоточена на частотах выше 4 кГц, которые в телефонном канале вырезаются. Очень часто часть звуков (например, то же ‘р’) произносятся как нечто, похожее на паузу. Такие звуки «додумываются» человеком. Дополнительной проблемой является так же и тот факт, что энергия шума неоднородна: она так же модулируется ударами основного тона. Так же можно отметить общую слабую изученность вопроса качественного разделения гласных и звонких согласных, отсутствие работ, которые бы указывали набор параметров и, главное, методологию подбора пороговых значений, которые позволяли бы с большой точностью разделять звонкие участки слитной телефонной речи.

Что касается вопроса определения точных границ сегментов, то, наиболее логичным кажется сегментация, привязанная к ударам основного тона: учитывая непродолжительность периодов основного тона (до 8 мс) и невозможность человеческого речевого тракта быстро менять свою конфигурацию, один период основного тона с большой долей достоверности можно считать промежутком стационарности основных параметров речевой волны. Учитывая, что метод, описанный в предыдущей главе позволяет с приемлемой точностью получить разметку ударов основного тона, предлагается распознавание участков (в дальнейшем кадров), равных одному периоду основного тона, отнесение каждого такого участка на основании какого-то признака к тому или другому классу, склейка из таких кадров более крупных фрагментов, постобработка, основанная на том факте, что средняя длительность вокализованной фонемы редко бывает короче 3-4 периодов основного тона.

**Разделение на основании определения наличия шумового источника возбуждения на высоких частотах**

Подход строится на предположении, что главным отличительным признаком звонких согласных является наличие шума на высоких частотах. Суть метода заключается в попытках различить характер частот выше 2,5 кГц и установить, являются ли они вокализованными или шумовыми. В рамках данной работы были опробованы следующие методы и подходы.

* **Анализ локальных максимумов анализируемого кадра:** из кадра вырезаются интервалы в начале и в конце (чтобы исключить моменты накачки энергии в момент удара импульса основного тона), в обрезанном кадре считаются моменты, в которых у сигнала наблюдается локальный максимум, считаются расстояния от максимума до следующего максимума. После этого считается дисперсия получившейся выборки. Кадры, в которых полученное значение выше некоторого порога должны относиться к согласным, остальные к гласным. Слежение за локальными экстремумами дает нам представление о высоких частотах в сигнале (ибо как раз ними и будет порождаться значительная часть локальных максимумов), анализ дисперсии в теории должен выявить вокалический характер старших частот (дисперсия расстояний между максимумами у гармонических колебаний, очевидно, должна быть значительно меньше, нежели у тех, которые являются шумом). На практике же при анализе гистограмм распределения значений определить пороговое значение оказалось невозможно, более того оказалось, что в около 20% случаев дисперсия согласных меньше дисперсии гласных.
* **Анализ на основании линейного предсказания:** смысл метода заключается в том, гармонические колебания должны предсказываться заметно лучше (в смысле энергии сигнала-остатка). В рамках дипломной работы была предпринята попытка сделать следующее: в кадре считался сигнал на основании метода линейного предсказания. Анализировалось отношение энергии сигнала-остатка к энергии самого сигнала. Ожидалось, что подобное отношение будет на гласных заметно ниже, нежели на согласных. В поставленных экспериментах подобного не наблюдалось. Существенного различия между классами замечено не было
* **Анализ на основании прямого синтеза 4-х синусов:** наиболее ресурсоемкие, но в то же время потенциально наиболее точный способ. Смысл метода заключает в следующем. Сигнал в кадре за исключением моментов ударов основного тона представляется в модели суммы 4-х экспоненциально затухающих синусов, находящихся в произвольной фазе. Идея метода заключается в том, что если нам удалось разложить сигнал в сумму подобных синусов (в смысле минимизации энергии сигнала остатка), то сигнал весь гармонический (тогда каждый синус, в том числе и высокочастотный, соответствует определенной форманте гласного), иначе в нем присутствует шум, который как раз и препятствует синтезировать сигнал в описанной модели. В рамках поставленных экспериментов существенного различия в отношении энергии сигнала остатка между классами замечено не было.

Возможным объяснением неудачного использования методов, пытающихся различить классы на основании «канонического» признака наличия высокочастотного шума у согласных является работа в условии 8 кГц телефонного канала. Учитывая, что шумовые форманты должны начинаться в районе 2,5-3 кГц, при передаче их через телефонный канал, значительная часть их могла оказаться выше полосы пропускания канала. Так же вызывает сомнение способность определить гармоничность сигнала, который представлен несколькими отсчетами на период. Этим возможно объяснить тот факт, что вышеперечисленные подходы, пытающиеся качественно различными методами определить наличие шума на высоких частотах, не смогли подтвердить или опровергнуть его присутствие.

**Разделение на основании анализа отношения энергии двух первых формант**

Из-за того, что на основании «канонического» признака разделить классы не удалось, начали проводиться изыскания других признаков, способных более ли менее достоверно разделить классы. Учитывая специфику канала передачи, перспективным начало казаться слежение за распределением энергии в полосах 1-й и 2-й формант (частоты от 300-400 до 1500-1800). Естественным кажется следующее: в кадре выделяются частоты, соответствующие формантам, и смотрится энергия в этих полосах. Проблема заключается в том, что надежного алгоритма трассирования формант (определения зависимости формантной центральной частоты от времени) на данный момент не известно.

Был предложен следующий метод: из кадра равного одному периоду основного тона вырезается фрагмент приблизительно в 20 отсчетов (такой период имеют колебания с частотой около 400 Гц – т.е приблизительно частота первой форманты), фрагмент нормируется. После этого считается вариация. При наличии развитой второй форманты, чья частота заведомо меньше первой, вариация должна быть ощутимо больше. В проведенных экспериментах, анализируя гистограммы распределения значений вариации, посчитанной вышеназванным способом, удалось подобрать пороговые значения, на основании которых появилась возможность провести сегментацию с точностью около 70%.

# Разделение высокочастотного шума и паузоподобных участков

Противоположностью вокалических участков являются глухие участки. Главной отличительной особенностью является наличие только одного источника возбуждения: шумового. Определение наличия этого источника возбуждение (или же практически полного его отсутствие) и является предметом анализа.

Глухие шумовые участки речи делятся на следующие два класса фонем [2]:

* Глухие щелевые: отличаются активной работой шумового источника, в осциллограмме наблюдаются ясноразличимые шумовые колебания на протяжении всей длительности фонемы, спектр шумовой, плоский, центр тяжести находится на высоких частотах, начиная от 2,5 кГц, очень часто выше 5-6 кГц.
* Глухие смычные: большая часть длительности глухих смычных приходится на так называемую смычку, по сути, паузу. В этот момент губы сомкнуты, в ротовой полости происходит нарастание давления воздуха, но никаких колебаний не производится. После смычки происходит короткая фаза взрыва (несколько мс, рис 5.), которая, по сути, является раскрытием смычки. Для взрыва характерен резкий рост энергии сигнала, энергия более ли менее равномерно распределена по всему спектру. Довольно часто после фазы взрыва наблюдается фаза аспиративного шума, продолжаться, в зависимости от темпа речи, может до 10 мс или даже больше (рис 5.). По своей структуре она очень похожа на глухие щелевые и может быть отличена только определением предшествующего ей взрыва. Учитывая, что в рамкой данной работы определение взрыва не предполагается, при оценке сегментирующего автомата подобные фрагменты будут относиться к шумным, хотя при сегментации по фонемам однозначно будут относиться к смычным.

Рис.5. Фонемы шипящая‘tS’ (до 13200 отсчета) и глухая смычная ‘t’ (от 13200 и до конца). 13200-13600 – фаза смычки,
13600 – 13625 – взрыв, 13625 – 13800 – фаза аспирации. В данном случае занимает приблизительно 20 мс - до трети длительности фонемы ‘t’

Следует отметить, что наличие фазы аспирации может заметно повлиять при подсчете точности сегментации. Так как проверка осуществляется на файлах, рассегментированных вручную на фонемы, то логичным кажется отнесение фрагментов глухих смычных к паузоподобным. Однако длинные фазы аспирации практически наверняка будут распознаваться как фрагменты, где присутствует сильный широкополосный шум. Что касается вопросов определения точных границ, то, учитывая, что как правило переходные участки могут доходить до 10 мс, сегментация кадрами по 8-10 мс кажется вполне логичным решением.

**Предлагаемое решение**

Определять наличие шума на глухих участках с хорошим временным разрешением можно различными способами: можно анализировать общую энергию, собирать статистику частоты переходов через ноль в кадре, анализировать локальные экстремумы. После надлежащей предобработки все эти методы показывают вполне приемлемые результаты. В рамках дипломной работы предлагается метод, чьи результаты оказались немного лучше. Сигнал нарезается на кадры, сигнал в каждом кадре нормируется по амплитуде. После этого по отнормированному сигналу считается вариация. На основании анализа гистограмм распределения значений вариации в кадре было определено пороговое значение, на основании которого удалось удачно произвести сегментацию. В терминах глухая щелевая/глухая смычная точность составила порядка 80-85%, в терминах высокоэнергетичный шум/паузоподобные участки – в районе 90%.

# Особенности реализации

Прототипы автоматов реализовывались в среде разработка MatLab. Выбор объясняется огромным количеством реализованных математических функций, большим набором специализированных тулбоксов (в этой работе активно используется Signal Processing Toolbox). В связи с этим значительно экономится время разработки по сравнению с другими языками. Следует отметить, что и по быстродействию (особенно в последних версиях Matlab) программные реализации не сильно уступают, аналогичным программам, реализованным не на Matlab-е.

В рамках дипломной работы в этой среде были реализованы несколько автоматов, в основе которых лежат алгоритмы и идеи, описанные выше. Именно на этих прототипах проверялась жизнеспособность предложенных решений.

**Автомат разделения голоса и шума**

Автомат построен на основе реализованной функции вычисления линейно предсказанной части и сигнала остатка (ошибки линейного предсказания). По исходному сигналу нарезается жесткая сетка кадров, причем кадры следуют с перекрытием равным половине длины кадра. В каждом кадре с помощью матлабовской реализации вычисления LPC формируется фильтр. Пропуская через этот фильтр сигнал в кадре, мы получаем линейно предсказуемую в рамках данной модели часть сигнала. После этого сигнал из четных и нечетных кадров склеивается с косинусным весом.

С помощью вышеописанной процедуры считается сигнал линейного предсказания в рамках модели LPC 3-го порядка на кадрах в 92 отсчета на сигнале в полосе выше 150 Гц. Учитывая способ построения сетки кадров при вычислении линейно предсказанной части у нас появляется возможность работать на кадрах вдвое меньшей длины, что уже не превышает среднестатистических длин переходных участков между классами. В каждом кадре длиной 46 отсчетов считается отношение энергии в линейно предсказанной части к энергии сигнала (statLin) и энергия сигнала в кадре в полосе до 1000 Гц(statEner). Следует заметить, что границы полос в 150 Гц и 1000 Гц не являются точными, незначительное изменения этих границ существенного влияния на результат не оказывает, поэтому необходимости построения аккуратных фильтров с минимальным временем переходного процесса и полосой пропускания аккуратно возле вышеназванных частот не возникает.

После получения массивов двух статистик, они подвергаются медианной фильтрации 7-го порядка. Кадр считается звонким, если: (StatLin > 0.8 && StatEner> 0.01) || (StatLin > 0.7 && StatEner> 0.1) || (StatLin > 0.6 && StatEner> 0.2). Результатом работы автомата является сигнал, равный по длине исходному сигналу, где в отсчеты, соответствующие звонким участкам записывается 1, глухим – 0.

**Определение трасс основного тона**

С помощью автомата, реализация которого описана в предыдущем пункте, считаются интервалы наличия голоса. В дальнейшем работа производится только и исключительно на этих интервалах. Предобработкой является слежение за мгновенной мощностью сигнала с помощью «гетеродинирования». Фильтруем сигнал в полосу «приблизительно первой форманты», например 200-700Гц. Следует отметить, что незначительное изменение как нижней, так и верхней границы интервалов значительного влияния на результат не оказывает (основные частоты, где характерно проявляются импульсы ОТ, полоса так или иначе захватывает, до шумовой форманты не дотягивает). Для отфильтрованного сигнала вычисляется соответствующий аналитический сигнал. Требуемое преобразование Гильберта реализуется в частотной области с косинусным окном и двойным перекрытием. Вычисляется квадрат мгновенной амплитуды аналитического сигнала (умножение отсчета на сопряженный к нему), которая и является промежуточным сигналом, на котором в дальнейшем ищутся достаточно большие локальные максимумы, которые могут быть кандидатами моментов прихода импульсов ОТ. В каждом звонком интервале производится первичная оценка частоты ОТ следующим способом:

1. Звонкий фрагмент должен быть длиннее 250 отсчетов. Эта цифра выбрана из следующих соображений: подобная длина дает достаточное спектральное разрешение, само число 250 позволяет иметь дело с целыми значениями частоты (кратно частоте оцифровки в 8000), в то же время в современных реализациях matlab проигрыш в производительности из-за некратности степеням 2 незначителен.
2. Из середины фрагмента вырезается кадр длинной 250 отсчетов. К нему применяется косинусное окно. С помощью матлабовской реализации преобразования Фурье считается спектр сигнала.
3. С помощью логарифмирования спектра поднимаем энергию дальних гармоник
4. Еще раз получаем спектр отлагорифмированного спектра.
5. Оценкой ОТ считаем частоту наиболее энергонесущей гармоники в частоте 90-400 Гц.

Если вышеописанным способом нам удалось получить оценку частоты ОТ, то пытаемся используя данную оценку провести траектории импульсов основного тона. Любая траектория проводится следующим образом: ищется локальный максимум с максимальным значением мгновенной мощности (используется предположение, что локально импульсы ОТ самые мощные – глобальный максимум обязан быть ударом ОТ). Влево и вправо ищутся удары, которые бы отстояли на оценочное значение периода основного тона от последнего обнаруженного удара, но могли бы отклоняться от ожидаемого места появления на пороговое значение (значение порога – 25% оценки интервала от последнего до предпоследнего, при поиске первых двух импульсов используется оценка, известная нам из каких-то источников априори). Если в интервале ожидания находится несколько локальных максимумов (хотя такое редкость, так как мы анализируем частоты до 600-700 Гц), то в качестве удара основного тона берем удар с большей мгновенной мощностью. Траектории, которые отстоят одна от другой на расстояние не больше одного периода ОТ, склеиваются в одну.

При первичном поиске траекторий с помощью спектральной оценки частоты траектории записываются в результат, если в совокупности они покрывают до 90% длины звонкого интервала, на котором они ищутся (только если нам удалось провести длинные траектории, мы можем быть уверены, что первичная спектральная оценка оказалась верной).

После нахождения первичных траекторий начинается обработка интервалов, где траектории найдены не были. В качестве первичной оценки частоты ОТ берется среднее значение средних значений частоты ОТ предыдущей и последующей траектории.

Результатом работы автомата является массив структур, каждая из которых представляет отдельную трассу отсчетов, в которых зафиксирован приход импульса основного тона.

**Разделение гласных и звонких согласных**

Анализируются только звонкие участки. Одним кадром анализа считается один период основного тона. Эта информация получена с помощью результатов работы автомата, описанного в предыдущем разделе. В каждом кадре в начале вырезается интервал в 20 отсчетов (соответствует частота в 400 Гц: приблизительно частота первой форманты), кадр нормируется, в отнормированном кадре считается оценка полной вариации. Полученный массив данных фильтруется медианным фильтром с порядком модели 3. После этого кадр считается звонкой согласной, если значение статистики <= 0.25 или >= 0.5. После этого происходит следующая постобработка: учитывая, что средняя длительность фонем превышает 3-4 удара основного тона, то интервалы одного класса такой длины считаются определенными неверно и становятся противоположным классом. Результатом работы автомата является сигнал, равный по длине исходному сигналу, где в отсчеты, соответствующие гласным записывается 1, звонким согласным – 2.

**Разделение высокоэнергетичного шума и паузоподобных участков**

Задача решается на жесткой сетке кадров с длиной кадра в 46 отсчетов только на глухих участках. Данные о том, какие участки глухие, получаются с помощью автомата разделения голоса и шума. В каждом кадре считается вариация отнормированного сигнала. После обработки полученной статистики медианным фильтром 7-го порядка, кадр считается глухой щелевой, если значение статистики >= 0.6, иначе – смычной. После этого происходит следующая постобработка: учитывая, что средняя длительность шумных фонем превышает 15-20 мс, то интервалы одного класса меньшей длины считаются определенными неверно и становятся противоположным классом. Результатом работы автомата является сигнал, равный по длине исходному сигналу, где в отсчеты, соответствующие глухим щелевым записывается 3, глухим смычным – 4.

# Результаты

В рамках дипломной работы были реализованы автоматы, совокупность которых смогла решить поставленные задачи.

С помощью реализованных автоматов были поставлены эксперименты по сегментации. Их результаты содержатся в Табл.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Эталон \ Автомат | Гласные | Зв. Согласные | Гл. Щелевые | Гл. Смычные |
| Гласные | **25,93** | 5,08 | 0,49 | 0,55 |
| Зв. Согласные | 10,69 | **19,11** | 0,75 | 1,39 |
| Гл. Щелевые | 0,52 | 0,27 | **10,25** | 1,18 |
| Гл. Смычные | 0,19 | 0,47 | 4,07 | **17,07** |

Табл. 1 «Матрица спутывания» работы автомата сегментации. Ячейка [n,m] означает, что k% отсчетов, размеченных вручную
как класс n, были распознаны автоматом как класс m. Переход класс в класс: 72.36% - сумма зеленых ячеек на диагонали. Красным отмечены ячейки с существенным процентом ошибок, которые в дальнейшем необходимо минимизировать, желтым – ошибки, которые в данной постановке задачи однозначно к ошибкам отнесены быть не могут.
 Представлены усредненные результаты обработки 2150 файлов от 10 дикторов, 5 мужчин и 5 женщин, приблизительно 4 часа речи речи.

Точность сегментации составила чуть больше 70%. Следует заметить, что часть ошибок в рамках постановки задачи строго говоря ошибками не являются. Характерным примером является спутывание глухих щелевых и глухих смычных. Учитывая, что эталонной разметки артикуляционных классов не существует, а имеются в наличии только разметка файлов на фонемы, то при сравнении с таким эталоном фазы аспирации смычных (большую часть которых занимает фаза смычки – по сути паузы) естественно не будут определяться смычными. Существенный процент ошибок сконцентрирован в гласных/звонких согласных. Следует отметить, что подобный результат был получен без привлечения обучения в каком-либо виде, используя только данные, что обрабатываемый сигнал – человеческая речь, оцифрованная на 8 кГц. Следует отметить, что эксперименты, поставленные с помощью HMM, показали сопоставимые результаты. Суть экспериментов сводилась к тому, что после распознавания речи существующим распознающим автоматом, фонемы соответствующих классов склеивались в один. Точность подобной сегментации так же оказалась чуть больше 70%. Следует отдельно остановиться на автомате, проводящем трассировку основного тона.

При визуальной проверке результатов работы специалистами по ручной разметке, всего в 3% случаев разметка была признана неточной, причем большинство таких мест было сконцентрировано в концах фраз. Учитывая, что под конец фразы в легких уже практически не остается воздуха, то и импульс основного тона существенно смазываются (а иногда одиночные импульсы просто пропадают). Характерным является тот факт, что содержательно ответить на вопрос, как следует провести трассу импульсов основного тона в подобных случаях, человек не смог. При сравнении реализации с существующими аналогами (в качестве сравнения использовался механизм по трассировке ОТ, поставляемый в рамках matlab, функция ‘fxrapt’ [13]), предлагаемое решение показало несравнимо большую устойчивость на зашумленных участках и значительно меньшее количество аномальных ошибок прерывания траекторий основного тона. Появление подобного автомата позволяет говорить о реальной возможности реализации гибкой сетки кадров при распознавании звонких участков.

# Список литературы

[1]. Фант Г. Акустическая теория речеобразования. – Издательство «Наука», 1964.

[2] Кодзасов С.В., Кривнова О.Ф. Общая фонетика. – М., 2001.

[3] Скрелин П.А. Сегментация и транскрипция. – СПб: Издательство СПбГУ, 1999.

[4] Сорокин В.Н., Цыплихин А.И. Сегментация и распознавание гласных // Информационные процессы. Том 4. № 2.

[5] Кочаров Д.А. Автоматическая интерпретация звуков речи, Диссертационная работа СПбГУ 2008.

[6] Маркел Дж. Д., Грей А.Х. Линейное предсказание речи. – М: «Связь», 1980.

[7] Баронин С.П. Автокорреляционный метод выделения основного тона речи. Пятьдесят лет спустя // Журнал «Речевые технологии» 2008

[8] Такун Е.И. Выделение импульсов основного тона на слитной речи с плохим соотношением «сигнал-шум». Курсовая работа СПбГУ, мат.-мех. 2010.

[9] ETSI Standard, ETSI ES 202 050 V1.1.3 (2003-11)

[10] Zissman M.A., Comparison of four approaches to automatic language identification of telephone speech // IEEE Trans. Speech and Audio Proc., SAP-4(1), Jan. 1996

[11] Adda-Decker M., Antoine F., Mareuil P.B., Vasilescu I., Lamel L., Vaissiere J.,
Geoffrois E., Lienard J.-S., Phonetic Knowledge, Phonotactics and Perceptual Validation for Automatic Language Identification // In Proceedings of the 15th International Congress of Phonetic Sciences, 2003.

[12] Kaiser J., On a simple algorithm to calculate the 'energy' of a signal // IEEE ICASSP 1990, pp 381-384.

[13] Описание алгоритм трассировки ОТ в среде MatLab <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/fxrapt.html>