Санкт-Петербургский Государственный Университет

Математико-механический факультет

Кафедра системного программирования

Выделение импульсов основного тона на слитной речи с плохим соотношением «сигнал-шум»

Курсовая работа студента 445 группы

Такун Евгении Игоревны

Научный руководитель ……. А.Е. Булашевич

Кандидат технических наук

Санкт-Петербург

2010

Оглавление

История вопроса

Область применения речевых технологий постоянно расширяется. Особенно это касается автоматического распознавания речи. Поэтому данная задача не перестает быть актуальной в наши дни, несмотря на более чем пятидесятилетнюю историю проблемы. На сегодняшний день получены хорошие результаты для распознавания речи конкретного диктора, а также лабораторной речи. Это связано с тем, что большинство современных систем обработки речи основываются на статистическом подходе, который подразумевает под собой необходимость предоставлять системе большое количество обучающего материала. При переходе от лабораторных данных к реальной речи мы сталкиваемся с существенными проблемами, связанными с особенностями спонтанной речи. При разговоре большой акцент делается на вербальную составляющую общения (при общении человек не занимается четкой артикуляцией, так как по реакции собеседника можно легко понять, достаточно ли разборчиво произносятся слова и фразы). Спонтанная речь подразумевает под собой быструю скорость и нечёткость произношения (на минимальном уровне четкости, понятном собеседнику). Для обработки спонтанной речи применяются различные подходы, основная идея которых заключается в предобработке входного сигнала и предоставлении автомату некоторых «подсказок». Для предварительной обработки сигнала используются следующие механизмы:

Нормализация – отслеживание уровня громкости сигнала и приведение его к некоторому общему

Фильтрация – очистка сигнала от шумов, иногда – подъем высоких частот.

Параметризация – входной сигнал разбивает на кадры. Кадр параметризируется. Господствующая параметризация - MFCC и ее модификации (вычитание среднего, RASTA-фильтрация,…).

В теории и существующих научных работах сетка кадров привязывается к импульсам основного тона. Это связано с тем, что в результате деления на кадры от импульса к импульсу удаляется часть вариативности параметризованного сигнала, связанная со случайной фазой ОТ относительно начала кадра. Однако, на практике секта кадров жесткая. Это вызвано тем, что отсутствуют алгоритмы, которые могут размечать импульсы основного тона с требуемой точностью на всех участках, где основной тон проявляется. Существуют методы, которые размечают основной тон на гласных, но все

они сбиваются на участках, где уровень шума становится больше (например, на звонких согласных).

Целью данной курсовой работы является разработка алгоритма, который выделял бы импульсы основного тона на участках, где присутствует шумовое возбуждение. При этом следует различать внешние источники шума (помехи телефонного канала, фоновый шум) и внутренние источники шума (шум, порожденный турбулентностью речевого тракта). С точки зрения фонем мы должны выделить импульсы основного тона на гласных звуках и на звонких согласных звуках (то есть там, где присутствует основной тон).

Выделив импульсы основного тона, мы получим ещё один инструмент уточнения границ фонем (сегментация речи на фонемы должна быть согласована с приходами импульсов основного тона). Деление на кадры для параметризации следует делать по моментам импульса основного тона.

Обзор существующих решений

Проблемой определения основного тона начали заниматься в 60-х годах. В зависимости от конкретной ситуации и общей постановки задачи требуется либо выделить контур основного тона, либо сравнительно точно оценить частоту, либо сделать маркировку отдельных периодов. Выделение контура необходимо в том случае, если для нас важен общий характер фразы (повествование, вопрос, восклицание), а маркировка периодов требуется в том случае, если нам нужно знать, когда конкретно пришел импульс основного тона.

Рассмотрим существующие методы определения ОТ

Частотные методы

При вокализованном возбуждении речевого тракта в спектре сигнала присутствуют пики на частотах, кратных частоте основного тона. Строим дискретное преобразование Фурье с достаточно малым шагом и пытаемся в качестве оценки частоты основного тона использовать частоту, соответствующую максимальному значению энергии спектра. При использовании данного метода возникают следующие проблемы: часто возникает ситуация, когда в рассматриваемой полосе помимо ЧОТ лежит ее гармоника с большей энергией. Она и будет ошибочно принята за оценку ЧОТ[2]. Особенно часто данное явление возникает при выделении ОТ мужского голоса. С этим можно бороться путем оценки среднего значения ЧОТ на крупном сегмента речевого сигнала, однако проблема останется при обработке нелабораторной речи. Также особенно важно то, частотные методы не подходят для решения задачи маркировки периодов.

Автокорреляционный метод

Метод заключается в вычислении автокорреляционной функции (АКФ) речевого сигнала и определение аргумента, при котором АКФ максимальна[1]. Пусть речевой сигнал представляется в виде последовательности отсчётов: S[i], где i = 0,1.. Для вокализованных звуков верно то, что S[i] ≈ S[i + T], где T – период ОТ, выраженный в количестве отсчётов. Этот факт делает справедливым следующее предположение: оценка периода основного тона должна максимизировать функцию

$$R(k) = \sum_{i=0}^{N-1} S[n+i]S[n-k+i]$$

Данная функция называется автокорреляционной. За долгую историю данного подхода, он претерпел большое количество изменений и усовершенствований. Например, достоверное определение значения ОТ по одному сегменту иногда оказывается невозможным. Существенное повышение точности может быть достигнуто при учёте зависимостей значений периодов ОТ на соседних сегментах (обычно периоды ОТ на соседних сегментах близки, хотя бывают и исключения). Несмотря на широкую распространенность данного метода у него есть существенные недостатки: на звонких согласных метод часто сбивается. И как и в случае с частотными методами мы не получим фазы основного тона, 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 в данном случае всевозможные усреднения и зависимости от предыдущих периодов делают невозможным определение мгновенного импульса.

Временные методы

Автокорреляционный метод и спектральный метод фактически пользуются одними и теми же свойствами сигнала. Однако есть чисто временные методы, основанные на анализе непосредственно осциллограммы сигнала. Как правило, они изучают распределение пиков и переходов через ноль. Эти методы потенциально являются самыми точными. Особенно в условиях плохого сигнала, но требуют тщательной настройки на входной сигнал. Что ограничивает область применения этих методов интерактивной работой с участием специалиста. Широко распространены в практике криминалистической экспертизы.

Teager operator (далее оператор Тигра)

В 1990 году Джеймс Кайзер опубликовал статью «On a simple algorithm to calculate the energy of a signal» [6], в которой ввел функцию для измерения моментальной энергии дискретного синусоидального сигнала. Используя уравнение общей энергии источника и аппроксимацию sin(x) = x для малых значений x, Кайзер выводит следующую формулу:

$$T(i) = (s(i)')^2 - s(i) \bullet s(i)''$$

где s(i) – это значение сигнала на i-ом отсчете.

Данный нелинейный оператор получил название «оператор Тигра». Он обладает рядом привлекательных черт, таких как простота, эффективность и хорошая восприимчивость к изменению сигнала (в том числе и резкому изменению)[5].

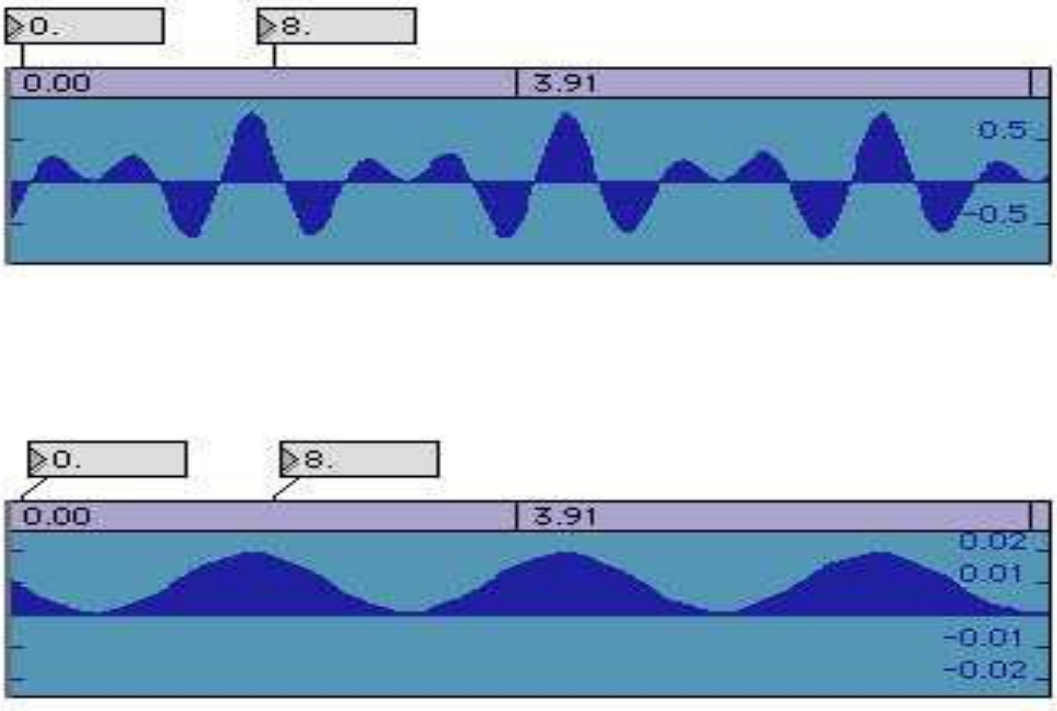

Рис 1

На рисунке выше можно легко видеть преимущества оператора Тигра с точки зрения слежения за изменением сигнала. Оператор Тигра давит синусоидальные колебания и оставляет только изменение амплитуды.

Рассмотрим поведение оператора Тигра на гласных и звонких согласных звуках. Оператору Тигра на вход подается сигнал, отфильтрованный по полосе 80-450 Гц (полоса, в которой может присутствовать основной тон)

Гласные звуки (для примера фонема [а])

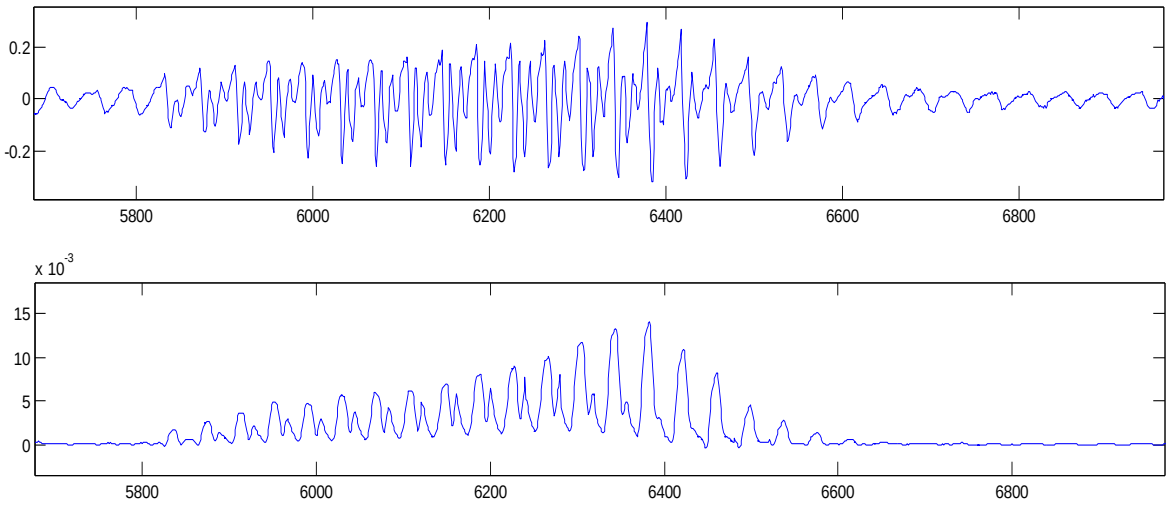

Рис 2.

На данном рисунке с отсчёта 5900 по отсчёт 6500 находится гласная [а]. Нетрудно заметить, что оператор Тигра очень хорошо отслеживает моменты прихода импульса основного тона.

Звонкие согласные (для примера фонема [в])

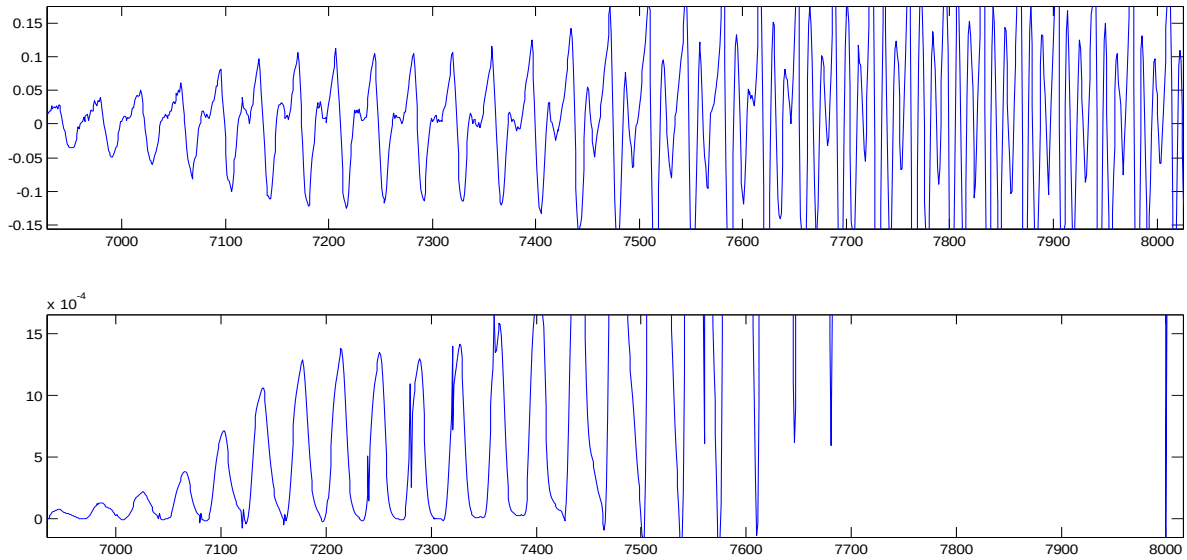

Рис 3

На данном рисунке с отсчёта 7100 по отсчет 7450 находится звонкий согласный [в]. Легко видеть, что оператор Тигра адекватно выделяет импульсы основного тона.

Звонкие согласные (для примера фонема [н])

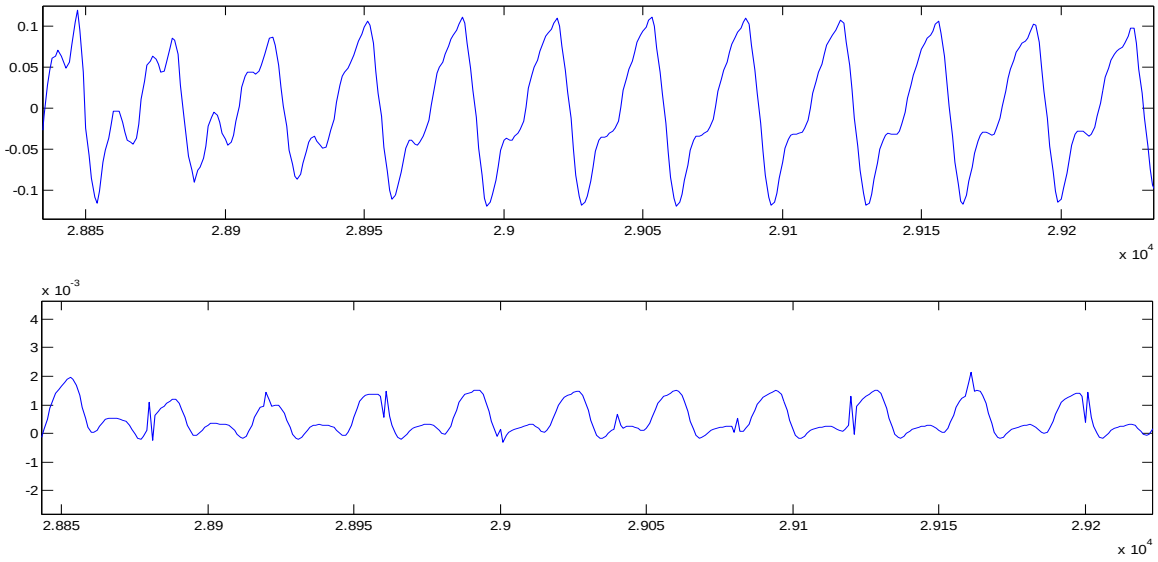

Рис 4

В данном случае оператор Тигра ведет себя неустойчиво. На рисунке 4 можно видеть, что с отсчета 28900 по отсчёт 29200 пики оператора Тигра становятся неочевидными. Верхушка смазывается, появляется несколько максимумов, расположенных рядом друг с другом. Определить в данном случае, какой из них значимый, довольно трудно.

Недостатки оператора Тигра

Оператор Тигра содержит производные (как первую, так и вторую). Нужно учитывать, что дифференцирование подчёркивает высокие частоты. То есть взятие производной ухудшает соотношение «сигнал-шум», а соответственно взятие второй производной усиливает этот эффект. Также следует помнить о той погрешности оценки мгновенного значения производной, которую мы получаем за счет дискретизации сигнала. Если мы будем рассматривать сигнал, оцифрованный на 16 кГц, то оператор Тигра будет вести себя абсолютно адекватно. На практике чаще всего имеется сигнал, оцифрованный на 8 кГц (телефонный канал). Теоретически, путем нетривиального усреднения можно добиться улучшения результата, но тогда теряется временное разрешение.

По результатам данного обзора было принято решение предложить метод, который сочетал бы в себе достоинства оператора Тигра, такие как простота и быстрая реакция на изменение энергии сигнала, и учитывал бы его недостатки.

Предлагаемое решение

Гетеродинирование

Вокализованный звук во временной области представляется суммой нескольких колебаний на формантных частотах, являющихся частотами резонаторов, сформированных артикуляционными органами. Известно, что для различения гласных звуков достаточно двух первых формант. Амплитуда колебаний на участке между импульсами основного тона плавно затухает, а в момент импульса основного тона при открытой щели голосовых связок происходит сравнительно быстрая по сравнению с затуханием подкачка энергии в резонаторы, сопровождающаяся скачкообразным нарастанием амплитуды колебаний. Рассмотренный выше оператор Тигра и предназначен для оценки мгновенной мощности колебаний с последующим объявлением момента импульса основного тона момента пиковой мгновенной мощности.

Таким образом перед нами стоит хорошо известная в технике задача – оценка огибающей синусоидального колебания. Неприятным отличием от хорошо изученного в технике амплитудно-модулированного сигнала является малое отношение частоты форманты, соответствующей несущей частоте, к полосе частот огибающей (к характерному времени затухания, если смотреть с временной точки зрения). Еще одним неприятным отличием является то, что в вокализованных звуках нет никакой отдельно генерируемой «несущей», формантные чатсоты извлекаются резонаторами из широкополосного спектра возбуждающих голосовых импульсов. Из этого следует, что в момент прихода очередного голосового импульса фаза остаточного колебания в резонаторе случайна – голосовые связки «не знают» о конфигурации артикуляционных органов.

Тем не менее в основу предлагаемого решения лег широко применяемый в технике способ обработки АМ сигнала – гетеродинирование.

Гетеродинирование основано на тригонометрическом равенстве

$$sin\,\alpha\,sin\,\beta = \frac{1}{2}cos(\alpha-\beta) - \frac{1}{2}cos(\alpha+\beta)$$

Используя это равенство, результат умножения двух гармоник $sin(2\pi f_1 t)$ и $sin(2\pi f_2 t)$ может быть выражен следующим образом:

$$sin(2\pi f_1 t)\, sin(2\pi f_2 t) = \frac{1}{2} cos2\pi(f_1 - f_2)t - \frac{1}{2} cos2\pi(f_1 + f_2)t$$

В результате получаем два сигнала частот $f_1 + f_2$ и $f_1 - f_2$, высокочастотный сигнал подавляется усреднением, а низкочастотный сигнал остается. Гетеродинирование переносит спектр АМ сигнала вниз по оси частот на частоту гетеродина.

В нашем случае играющая роль несущей частота форманты нам неизвестна, но эту трудность можно обойти.

Идея предлагаемого решения

Сигнал представляется в комплексном виде: $\cos x = \frac{e^{it} + e^{-it}}{2}$

Если закрыть глаза на отрицательную частоту, то получаем то, что нам нужно.

Умножаем сигнал на сопряженный задержанный к нему

$S_k \times S_{k-1}^*$

В итоге получаем квадрат амплитуды и фазу, но фаза нас в данном конкретном случае не интересует.

Соотношение «сигнал-шум» в данном случае ухудшается вдвое, но только один раз по сравнению с оператором Тигра. Причем степень ухудшения зависит от фазы, в которой пришли сигнал и шум.

Данный подход лишен основного недостатка оператора Тигра – необходимости дифференцирования в дискретном времени. По сути методическая погрешность сосредоточена в преобразовании в комплексный сигнал и преселекторе.

Алгоритм выделение импульсов основного тона

В качестве предварительной обработки сигнала помимо фильтрации следует выделить те участки, на которых присутствует основной тон. За основу был принят алгоритм, предложенный Кочаровым Д.А. в работе «Автоматическая интерпретация звуков речи». Суть алгоритма заключается в следующем:

вычисляется мгновенная энергия E(n) отфильтрованного по полосе 80-900Гц сигнала с небольшим шагом. Далее на крупном сегменте речевого сигнала вычисляется среднее

арифметическое значение мгновенной энергии E_m. Вычисляется отношение между мгновенной энергией и средней мгновенной энергией.

ОТ присутствует, если $E(n) < T \times E_m$

ОТ отсутствует, если $E(n) > T \times E_m$

где T – это коэффициент, определяющий пороговое значение.

Было вычислено, что при значении T = 0.05 алгоритм работает корректно.

После проведения предварительной обработки сигнала и применения предложенного алгоритма мы получаем значение квадрата амплитуды для каждого отсчёта. Теперь нам необходимо выделить максимумы амплитуд. Проблема заключается в том, что для нас важны максимумы, вызванные ударами основного тона. Максимумы, вызванные пиками второй и последующих формант, нас не интересуют, поэтому необходимо их корректно обработать и отфильтровать. Это делается в три этапа.

Первый этап

Необходимо избавиться от расколов пиков. При обнаружении подобного явления алгоритм убирает два «расколотых» максимума и образует один, равный максимальному.

Второй этап

Необходимо избавиться от максимумов, порожденных пиками второй и последующих формант. Для этого несколько раз (эмпирическим путем установлено, что достаточно двух) проделывается следующая операция: проходится массив уже выделенных максимумов, и в случае, если между двумя пиками находится третий, который меньше левого и правого, то он убирается из массива максимумов.

Третий этап

Трассировка. По итогам второго этапа может возникнуть ситуация, когда алгоритмом будут помечены как ложные на самом деле имеющие место максимумы. Для исправления ошибок в таком случае следует применить алгоритм трассировки. Он основывается на соображении, что в пределах соседних периодов длина периода основного тона не может значительно измениться. Исходя из этого, следует провести проверку по длинам периодов и в случае, если значение резко изменилось, отметить незаслуженно пропущенный максимум заново.

После всех вышеперечисленных действий мы получаем массив отсчётов, на которых теоретически должен быть импульс основного тона. Дальше мы можем использовать данную информацию для привязки сетки кадров при параметризации и корректировки границ фонем

Выделим импульсы основного тона для предложенного алгоритма и оператора Тигра и рассмотрим результаты на различных фонемах

Гласные (для примера фонема [а])

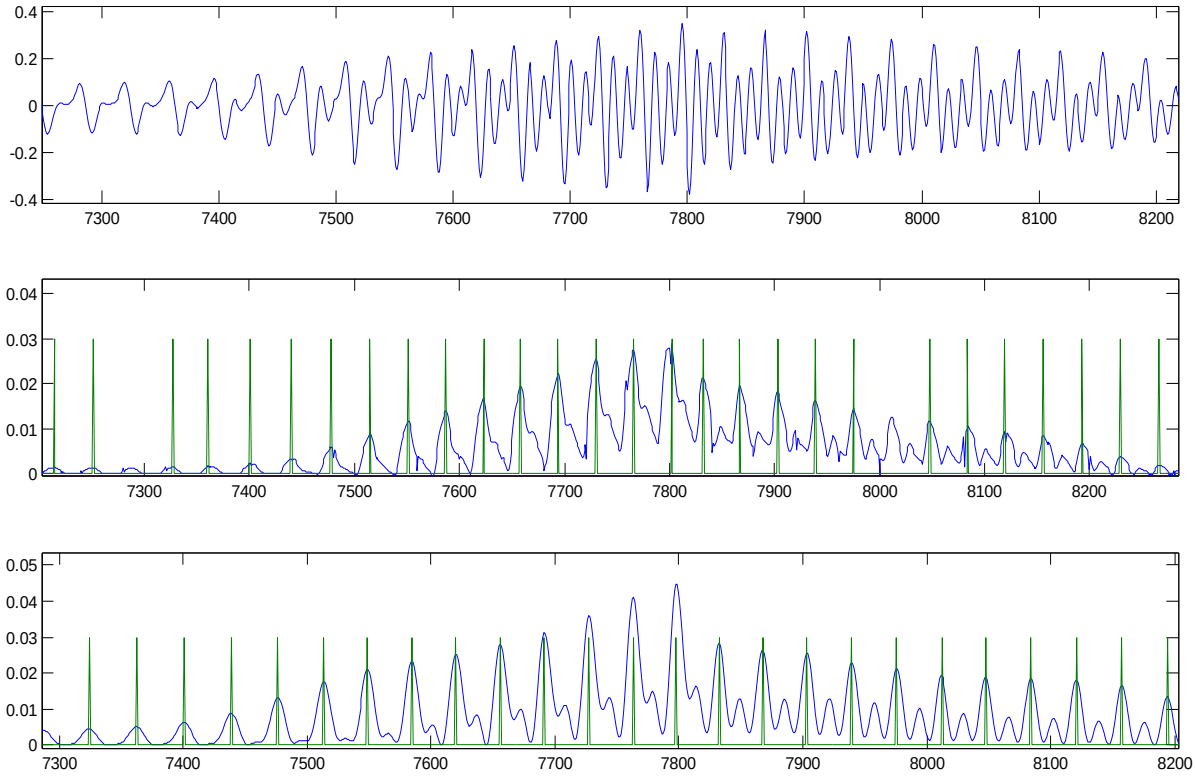

Рис 5

На рисунке 5 мы видим результат работы оператора Тигра и предложенного алгоритма на гласных звуках (с отсчёта 7450 по отсчёт 8200 находится фонема [а]). Легко заметить, что на гласных звуках результаты практически идентичны. И оператор Тигра, и предложенный алгоритм успешно выделили импульсы основного тона. К примеру, анализируя сигнал можно определить, что на отсчете 7800 пришел импульс основного тона. Оба алгоритма корректно отреагировали на это и пометили данный отсчет, как

интересующий нас. Разницы в результатах нет, так как на гласных звуках практически отсутствует шумовое возбуждение (мы используем уже отфильтрованный по полосе 80-900 Гц сигнал).

Звонкие согласные (для примера фонема [в])

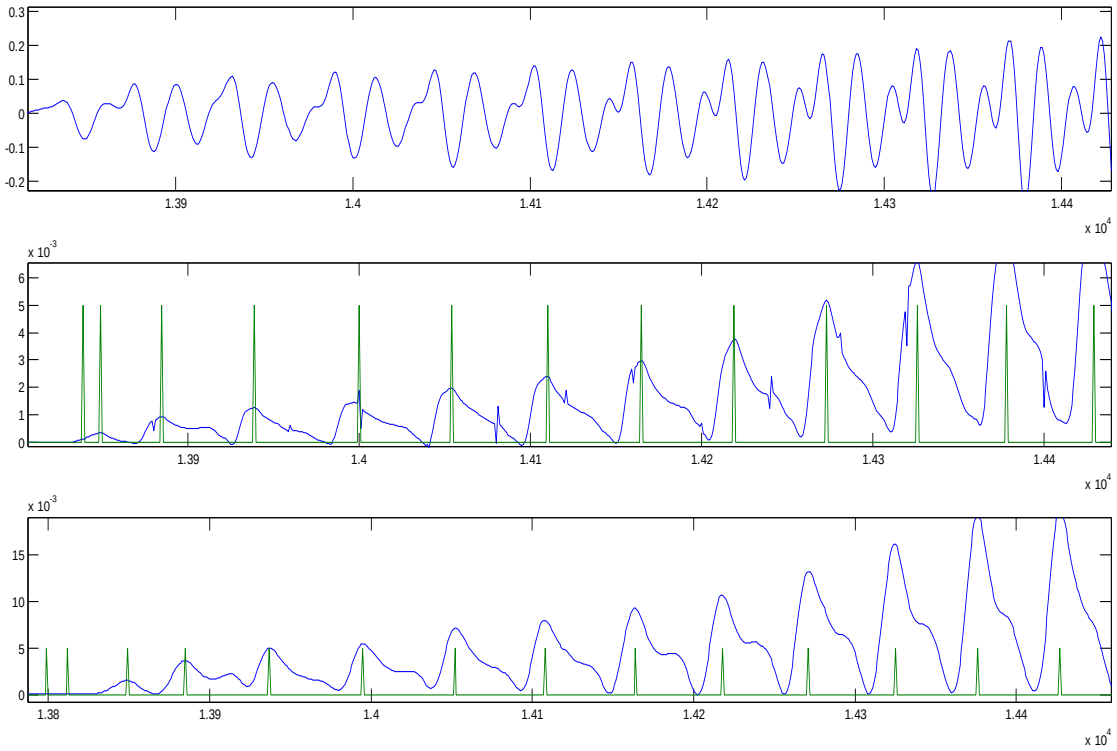

Рис 6

На рисунке 6 мы видим результат работы оператора Тигра и предложенного алгоритма на звонких согласных звуках. В качестве примера был выбран согласный звук [в]. Можно заметить, что в данном случае оператор Тигра начал давать несколько искаженный результат. Это связано с тем, что на звонких согласных присутствует естественное шумовое возбуждение. Двойное дифференцирование подчеркнуло шумовой эффект, что привело к появлению дополнительных пиков, одна алгоритм смог отличить ложные пики от существенных и отработал корректно. Импульсы основного тона были верно.

Далее рассмотрим результат работы алгоритма на звонкой согласной [н]. Она интересна тем, что шумовое возбуждение в данном случае сильнее.

Звонкие согласные (для примера фонема [н])

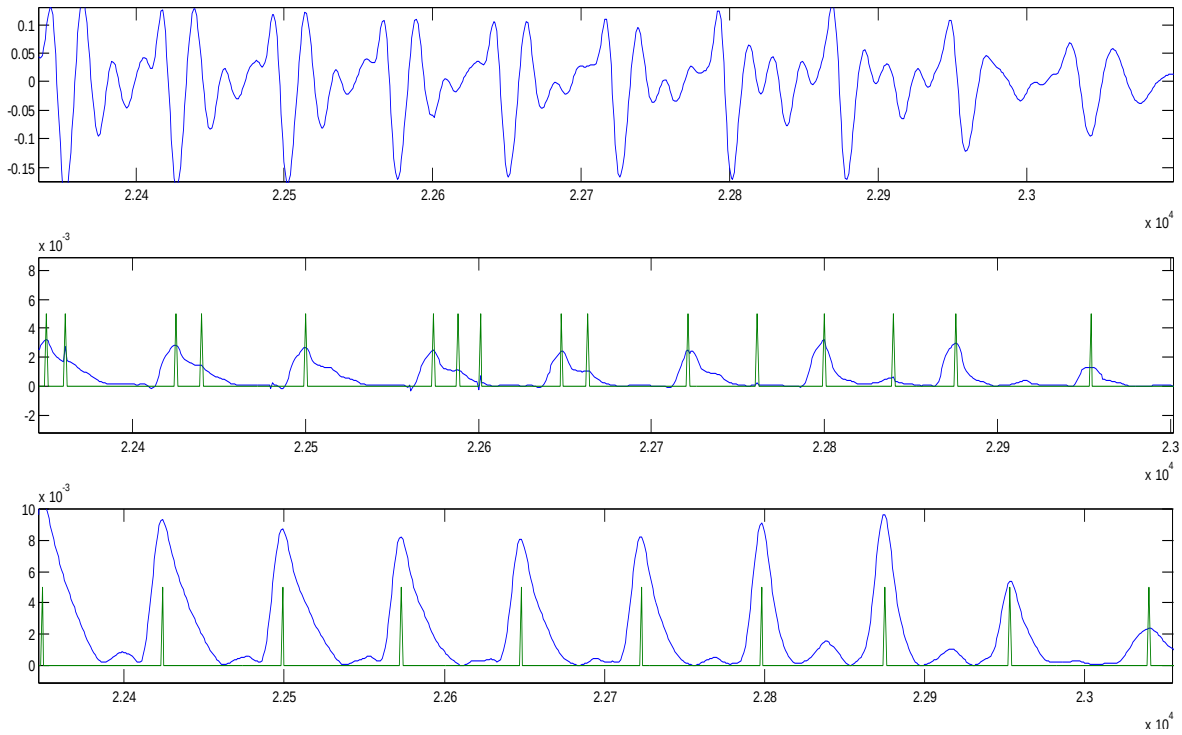

Рис 7

Звонкий согласный [н] имеет более мощную шумовую составляющую, чем рассмотренный выше звонкий согласный [в]. Поэтому в данном случае результат, получаемый при использовании оператора Тигра, существенно искажен. Предложенный алгоритм продолжает работать корректно, так как соотношение «сигнал-шум» ухудшилось незначительно.

Для оценки общего результата рассмотрим результат работы алгоритма на более длинном речевом сегменте. Для наглядности и чистоты эксперимента используем речевой сегмент, на котором присутствуют звонкие согласные звуки и гласные звуки, то есть имеются сегменты, на которых присутствует шумовое возбуждение и голосовое и на которых присутствует только голосовое возбуждение.

В качестве примера используется сочетание фонем [ано’]

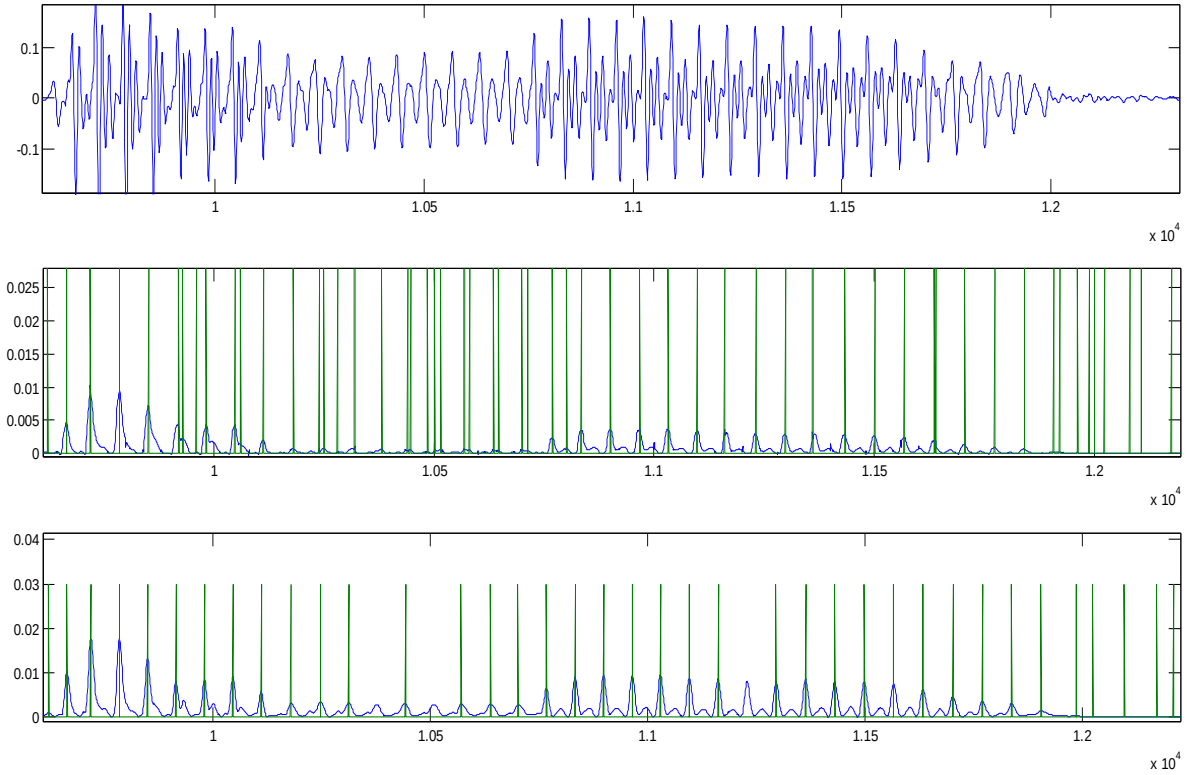

Рис 8

По рисунку можно видеть, что гласные звуки корректно обрабатываются обоими алгоритмами. Звонкие согласный звуки корректно обрабатываются предложенным алгоритмом, а использование тигрового оператора приводит к существенным искажениям. В итоге на нижнем графике мы имеем более адекватно расставленные импульсы основного тона.

Результаты

Проанализировав полученные данные, можно утверждать, что алгоритм выделения импульсов основного тона работает более корректно, если ему на вход подаётся квадрат амплитуды, полученный с помощью предложенного алгоритма, нежели с помощью оператора Тигра. Отказ от операции дифференцирования оказался полезным и привел к существенному улучшению результатов. При этом предложенный алгоритм сохранил такие преимущества оператора Тигра, как вычислительная простота и хорошее разрешение по времени. В следующей таблицы представлено сравнение результатов работы предложенного алгоритма и оператора Тигра на различных типах фонем. Указывается отношение верно определенных импульсов основного тона к общему количеству выделенных максимумов.

	Гласные звуки	Звонкие согласные ([м], [н],[л])	Звонкие согласные (прочие)	Итого
Оператор Тигра	96%	74%	76%	82%
Предложенный алгоритм	97%	93%	94%	95%

В итоге мы получили разметку импульсов основного тона с высокой точностью для гласных и звонких согласных звуков. Данная информация будет использоваться в задаче автоматической сегментации речи на фонемы, как средство уточнения границ фонемы. Это связано с тем, что границы вокализованных фонем должны привязаны к импульсам основного тона. Также информация об импульсах основного тона может быть использована для параметризации на гибкой сетке кадров, что позволит уменьшить структурный шум параметризации.

Дальнейшее направление исследования

Для более точного определения импульсов основного тона требуется улучшить алгоритм поиска максимумов. Предполагается основываться на том, что между двумя существенными максимумами должен находиться существенный минимум. Это позволит избежать необходимости фильтровать массив максимумов несколько раз. И как следствие упростит алгоритм трассировки. Следует заметить, что существует определенный произвол разработчика в распределении усилий между алгоритмом первичного обнаружения кандидатов в моменты импульса ОТ и алгоритмом трассировки. Например, более качественный преселектор уменьшит величину ложных пиков, порожденных старшими формантами, что упростит трассировку. Работа же в полной полосе делает пики более острыми, потенциально улучшая разрешение по времени за счет необходимости сравнительно интеллектуального алгоритма трассировки/отбраковки обнаруженных пиков.

Объектом тщательной проработки в будущем является преселектор. Дело в том, что преселектор приходится делать следящим вследствие того, что области первой и второй формант, вообще говоря, перекрываются (это иногда случается и в пределах одного диктора, а с учетом междикторской изменчивости встречается довольно часто).

Подлежит доисследованию вопрос о возможности применить обнаружение пиков мгновенной энергии сигнала как альтернативного способа обнаружить наличие голосового возбуждения.

Список литературы

1. Баронин С.П. Автокорреляционный метод выделения основного тона речи. Пятьдесят лет спустя. //Журнал «Речевые технологии» 2008, стр 3-13

2. Кочаров Д.А. Автоматическая интерпретация звуков речи // Диссертационная работа СПбГУ 2008

3. Г. Фант. Акустическая теория речеобразования // Издательство «Наука», 1964 г.

4. В.В. Митянок. Метод аппроксимации для определения числовых характеристик некоторых низкочастотных звуков человеческой речи // Электронный журнал «Техническая акустика» http://www.ejta.org 2008, 15

5. Rebecca Fiebrink. An Exploration of then Teager Operator// MUMT 605, 2004

6. Jim Kaiser. On a simple algorithm to calculate the 'energy' of a signal// IEEE ICASSP 1990, pp 381-384