УДК 004.934

АЛГОРИТМ ОЦЕНКИ МГНОВЕННОЙ ЧАСТОТЫ ОСНОВНОГО ТОНА РЕЧЕВОГО СИГНАЛА

Азаров И.С., *к.т.н., Белорусский государственный университет информатики и радиоэлектроники, e-mail: azarov@bsuir.by*
Вашкевич М.И., *аспирант кафедры электронных вычислительных средств Белорусского государственного университета информатики и радиоэлектроники, e-mail: vashkevich@bsuir.by*
Петровский А.А., *д.т.н., профессор, зав. кафедрой электронных вычислительных средств Белорусского государственного университета информатики и радиоэлектроники, e-mail: palex@bsuir.by*

Ключевые слова: оценка основного тона, мгновенная частота, синусоидальная модель, алгоритм слежения, кросс-корреляционная функция.

Предлагается способ оценки мгновенной частоты основного тона на основе устойчивого к ошибкам алгоритма слежения за основным тоном RAPT (robust algorithm for pitch tracking). В отличие от RAPT, который выполняет оценку частоты, относящуюся к фрейму анализа, предлагаемый метод выполняет оценку, относящуюся к заданному моменту времени. Другая особенность метода - низкая чувствительность точности оценки к модуляциям частоты основного тона. Перечисленные свойства достигаются за счет использования специальной функции оценки периодичности, которая аналогична нормированной кросс-корреляционной функции, используемой в RAPT, однако вычисляется на основе мгновенных гармонических параметров синусоидальной модели сигнала. Предложенный алгоритм сравнивается с другими современными алгоритмами при помощи искусственных и натуральных речевых сигналов. В случае значительных частотных модуляций основного тона предложенный метод обеспечивает ошибку оценки в несколько раз меньшую по сравнению с ближайшим конкурентом, о чем свидетельствуют результаты анализа синтетических сигналов с известными значениями мгновенной частоты основного тона.

Введение

Параметрическое представление речи часто подразумевает использование частоты основного тона (ЧОТ) в качестве параметра модели [1]. Выбор определенного алгоритма для оценки частоты основного тона зависит от целевого приложения и всегда представляет собой некоторый компромисс между частотно-временным разрешением, устойчивостью к ошибкам, алгоритмической задержкой и вычислительной сложностью. Настоящая работа направлена на поиск алгоритма оценки основного тона для приложений, где необходима максимальная точность оценки, и частота основного тона рассматривается как непрерывная функция от времени. К таким приложениям относиться широкий класс систем обработки речи, использующих детерминистическую/стохастическую декомпозицию сигнала. Точностью оценки основного тона определяется насколько хорошо можно разделить сигнал на детерминистическую и стохастическую составляющие, от нее зависит также число разделяемых гармоник, которые можно описать отдельными наборами параметров. Точность оценки частоты основного тона определяется двумя основными характеристиками: 1 – временное разрешение т.е. как быстро алгоритм оценки реагирует на изменения частоты, 2 – частотное разрешение т.е. насколько малые изменения частоты алгоритм может определить. Обе характеристики чувствительны к модуляциям основного тона и степени зашумленности сигнала (интенсивности шума как фонового так и обусловленного смешанным возбуждением речевого тракта).

В настоящее время предложено большое число разнообразных алгоритмов оценки основного тона, основные из них описаны в работах [1-6]. Сегодня наиболее популярными алгоритмами оценки ЧОТ являются RAPT [7], YIN [8] и SWIPE' [9]. Популярность данных алгоритмов обусловлена хорошей функциональностью, низким процентом грубых ошибок и наличием свободно распространяемых версий их реализаций. Тем не менее, как будет показано ниже, возможность этих алгоритмов оценивать мгновенную частоту существенно ограничена. Ограничение обусловлено периодической (стационарной) моделью сигнала, лежащей в их основе, которая подразумевает точное повторение периода основного тона и не допускает его изменения на протяжении анализируемого фрейма. При появлении модуляций (изменений частоты основного тона) точность оценок существенно снижается.

Для оценки мгновенной частоты основного тона в последнее время было предложено несколько оригинальных методов [10-13], которые имеют хорошее теоретическое основание, однако не имеют свободных программных реализаций, доступных для использования и тестирования. Потому сложно объективно оценить данные методы и на практике убедиться в их применимости к тем или иным приложениям.

Авторы настоящей работы видят своей целью создание алгоритма, который: 1) обеспечивает оценку мгновенной ЧОТ, что позволит повысить качество параметрического синтеза речи (в частности в задаче конверсии голоса [14]); 2) устойчив к частотным модуляциям основного тона; 3) имеет достаточно низкую вычислительную сложность для работы в реальном времени; 4) имеет алгоритмическую задержку не более 100мс; 5) применим

в практических задачах обработки речи (обладает достаточно высокой робастностью); 6) имеет свободную программную реализацию[1].

Предложенный в настоящей работе алгоритм основывается на RAPT [7]. Из RAPT заимствован общий «каркас» алгоритма и отдельные элементы, в частности используется функция оценки периодичности сходная с нормированной кросс-корреляционной функцией (НККФ). За основу можно было бы взять любой другой алгоритм, использующий корреляционные функции в качестве генератора кандидатов периода основного тона, тем не менее в пользу RAPT можно привести следующие аргументы:

– RAPT широко распространенный алгоритм с хорошо изученными преимуществами и недостатками;

– RAPT имеет относительно низкую алгоритмическую задержку, низкую вычислительную сложность и обеспечивает хорошую устойчивость к ошибкам в условиях зашумленности;

– практические эксперименты показывают, что RAPT в большинстве случаев более других алгоритмов устойчив к влиянию частотных модуляций основного тона;

– RAPT широко используется в речевых приложениях и имеет несколько доступных реализаций с открытым исходным кодом.

Идея алгоритма, предложенного в данной работе, заключается в том, что можно существенно улучшить характеристики RAPT, используя функцию оценки периодичности на основе мгновенных гармонических параметров синусоидальной модели вместо НККФ. Дополнительное улучшение точности можно достичь на этапе постобработки путем оценки параметров сигнала, масштабированного во времени.

В работе кратко излагаются основы RAPT, теоретически обосновывается преимущества основных предлагаемых модификаций, приводится описание нового алгоритма и его экспериментальное сравнение с существующими алгоритмами.

Оценка основного тона речевого сигнала

Если допустить, что анализируемый сигнал является строго периодичным, то частота основного тона (F_0) может быть определена как величина обратная длине его периода. Период в свою очередь определяется как минимальный временной сдвиг, сохраняющий исходный сигнал. Почти все сигналы, с обработкой которых встречаются в практических приложениях, не являются строго периодичными.

Определение ЧОТ в различных приложениях имеет различный смысл. Например, в обработке вокализованной речи обычно считают, что ЧОТ соответствует частоте колебаний голосовых связок. Предполагается, что хотя их колебания и не являются строго периодичными, то, во всяком случае, на некотором непродолжительном временном интервале можно наблюдать почти повторяющиеся фрагменты. Процесс речеобразования является сложным и нестационарным. Изменение параметров голосовых связок, голосового тракта, интонации произношения, участие в процессе нескольких источников возбуждения делает оценку ЧОТ достаточно сложной задачей. Несмотря на разнообразие уже предложенных методов решения, эта задача продолжает привлекать пристальное внимание и множество усилий со стороны современных исследователей. Среди опубликованных подходов можно выделить использование аудиторных моделей [15], нейронных сетей [15,16] и специфических моделей сигнала [9,17]. Отдельно можно выделить направление, посвященное оценке мгновенной ЧОТ [11-13]. Данное направление представляет особый интерес, поскольку рассматривает сигнал как непрерывный процесс и позволяет его интерпретировать в виде непрерывных параметрических моделей. В этом случае к обычным характеристикам, важным для всех оценщиков без исключения (например устойчивость к грубым ошибкам, устойчивость к шуму, вычислительная сложность и т.д.), добавляются и играют ключевую роль такие характеристики как частотно/временное разрешение и устойчивость к модуляциям основного тона. Эти требования повышают сложность задачи, однако позволяют получить новую, более детальную информацию о сигнале. Оценщики мгновенной ЧОТ применимы не только в традиционных приложениях обработки речи, но и в специальных задачах, среди которых диагностика заболеваний речевого тракта, оценка эмоционального состояния диктора, измерение оборотной частоты вращающихся механизмов по звуковому сигналу, диагностика состояния турбин, шестерен и т.д.

Краткое описание RAPT

Основной задачей при разработке RAPT являлось достижение максимальной точности оценки и устойчивости к ошибкам. Настройки алгоритма позволяют адаптировать его к различным требованиям конечного приложения, особенностям голоса диктора или условиям записи.

Большинство оценщиков основного тона (и RAPT в том числе) состоят из трех основных компонентов: 1 – предобработка или приведение сигнала к требуемым характеристикам, 2 – генератор кандидатов действительного искомого периода основного тона, 3 – постобработка или выбор наилучшего кандидата с последующим уточнением значения частоты.

Вид предобработки определяется видом последующей функции генератора кандидатов периода основного тона. Основной целью предобработки является очистка сигнала от его составляющих, которые могут негативно повлиять на точность оценки (например, акустические шумы, состояние речевого тракта, постоянное смещение и т.д.). В алгоритме RAPT не применяется специальной предобработки такого рода. Однако, для уменьшения вычислительных затрат используется понижение частоты дискретизации.

В качестве функции, определяющий кандидатов пе-

[1] MATLAB реализация предложенного алгоритма доступна по адресу http://dsp.tut.su/irapt.html

риода основного тона используется НККФ (ее формальное определение будет дано ниже), которая позволяет оценить степень периодичности сигнала в зависимости от задержки сигнала в отсчетах.

Подразумевается, что анализируемый сигнал обладает следующими свойствами, характерными для речи [7]:

- локальный максимум НККФ, соответствующий действительному периоду основного тона вокализованной речи (исключая нулевую задержку), обычно является наибольшим и близким к единице;
- в случае, когда имеется несколько локальных максимумов НККФ близких к единице, то правильным будет выбрать тот, который соответствует наименьшему периоду;
- максимумы НККФ смежных фреймов расположены близко друг к другу, поскольку частота основного тона изменяется медленно;
- действительная частота основного тона иногда может резко увеличиться или уменьшиться в два раза;
- изменение состояния вокализованности происходит редко;
- для невокализованной речи значения НККФ (за исключением нулевой задержки) значительно ниже единицы;
- кратковременный спектр вокализованных и невокализованных фреймов сигнала обычно существенно различны;
- при переходе от невокализованного фрейма к вокализованному амплитуда сигнала обычно увеличивается, и, наоборот, при переходе от вокализованного к невокализованному – уменьшается.

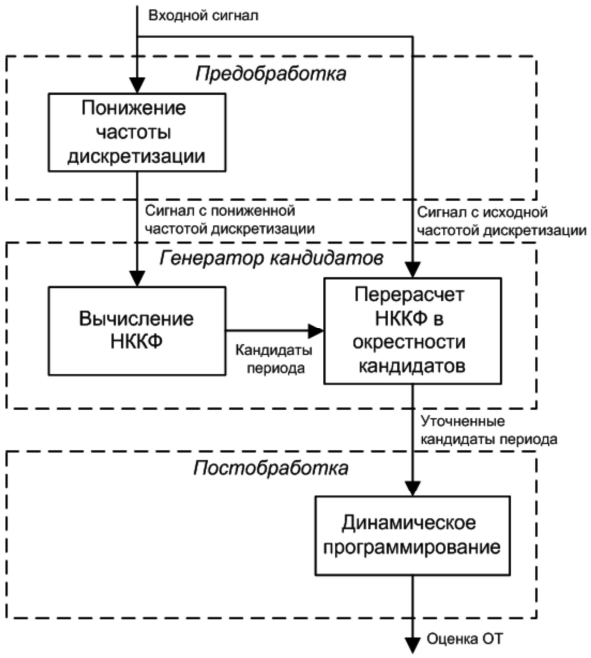


Рис. 1. Схема RAPT

На стадии постобработки выполняется поиск контура основного тона, объединяющего локальные максимумы НККФ, при этом накладывается ограничение, что частота основного тона изменяется медленно и, таким образом, значения ЧОТ смежных фреймов не должны сильно отличаться.

Основные шаги RAPT перечислены ниже [7] (рис. 1):

- создаются две версии анализируемого сигнала: одна с исходной частотой дискретизации, другая с существенно пониженной;
- вычисляется НККФ для всех фреймов сигнала с пониженной частотой дискретизации и всех задержек из допустимого диапазона периода основного тона; выполняется поиск и сохранение локальных максимумов полученных значений НККФ;
- вычисляется НККФ для всех фреймов сигнала с исходной частотой дискретизации в окрестностях локальных максимумов, найденных на предыдущем шаге; выполняется поиск и сохраняются локальные максимумы полученных значений НККФ с высоким разрешением;
- каждый из найденных максимумов является кандидатом периода основного тона для соответствующего фрейма;
- при помощи динамического программирования выполняется поиск контура частоты основного тона, соединяющего найденные кандидаты периода;

Краткое описание предлагаемого алгоритма

В предлагаемом алгоритме оценки частоты основного тона так же используется понижение частоты дискретизации (до 6кГц). При этом предполагается, что основная доля энергии вокализованного речевого сигнала приходиться на нижнюю часть спектра (до 3кГц), и таким образом оценки мгновенной частоты каждой гармоники из этого диапазона достаточно для получения точного значения основного тона. При анализе других сигналов (отличных от речевых, например звука работающего двигателя) для получения более точных оценок частота дискретизации может не понижаться. Для оценки НККФ используются мгновенные гармонические параметры синусоидальной модели, характеризующие периодичность сигнала в каждый момент времени.

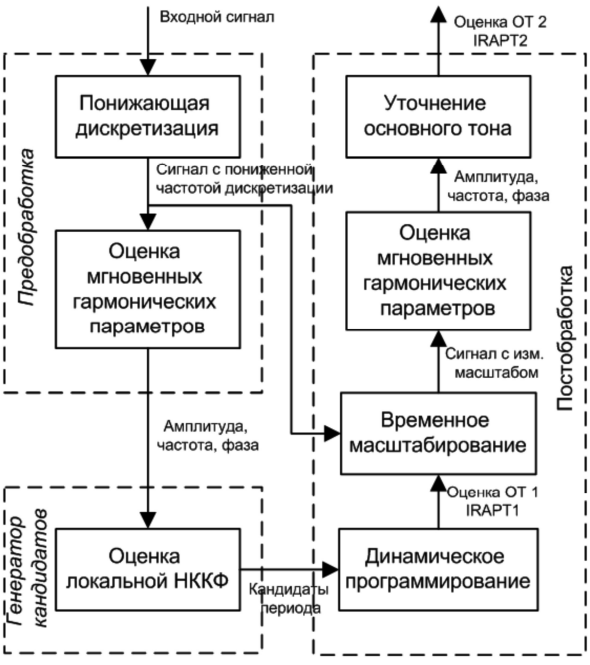


Рис. 2. Схема предлагаемого алгоритма оценки

Основные шаги предлагаемого алгоритма перечислены ниже [18] (рис. 2):

- понижение частоты дискретизации (как и в RAPT это выполняется для уменьшения числа требуемых операций); при обработке речевого сигнала новая частота дискретизации составляет примерно 6кГц; в отличие от RAPT точность оценки кандидатов не так сильно зависит от частоты дискретизации (в данном случае потеря точности обусловлена только потерей верхних гармоник), и потому оценка основного тона можно выполнить, используя всего один сигнал;
- вычисляются мгновенные параметры синусоидальной модели сигнала;
- вычисляется НККФ, используя полученные мгновенные параметры; значения НККФ сохраняются;
- каждый из локальных максимумов НККФ является кандидатом периода основного тона для соответствующего момента времени;
- при помощи метода динамического программирования выполняется поиск контура частоты основного тона, соединяющий найденных кандидатов периода; получаемая в результате оценка мгновенной частоты основного тона обозначена на схеме «IRAPT1»;
- используя контур полученной частоты основного тона сигнал масштабируется во временной области для того, чтобы обеспечить его стационарность;
- вычисляются мгновенные параметры синусоидальной модели масштабированного сигнала;
- на основе полученных параметров вычисляется уточненная оценка основного тона, которая обозначена на схеме «IRAPT2».

Оценка мгновенных гармонических параметров

Понятие мгновенных гармонических параметров возникает из предположения синусоидальной модели сигнала [19], которая представляет действительный сигнал *s(m)* в виде суммы синусоид или действительной части комплексных экспонент с непрерывной амплитудой, частотой и фазой:

$$s(m) = \sum_{p}^{P} A_p(m)\cos\varphi_k(m) = \mathrm{Re}\left[\sum_{p}^{P} A_p(m)e^{\varphi_p(m)j}\right]$$

где P – число синусоид (комплексных экспонент), $A_p(m)$ – мгновенная амплитуда *p*-ой синусоиды, $\varphi_p(m)$ – мгновенная фаза *p*-ой синусоиды. Мгновенная частота $F_p(m)$, находящаяся в интервале $[0,\pi]$ (π соответствует частоте Найквиста), является производной от мгновенной фазы. Предполагается, что амплитуда $A_p(m)$ и частота $F_p(m)$ изменяются медленно, что означает ограничение полосы каждой из составляющих.

Представим сигнал *s(m)* в виде суммы узкополосных аналитических сигналов $S_{F_\Delta,F_C^p}(m)$ [20], где $2F_\Delta$ соответствует ширине полосы, а F_C^p - центру полосы с индексом *p*, тогда

$$s(m) = \mathrm{Re}\left[\sum_{p}^{P} S_{F_\Delta,F_C^p}\right] =$$

$$= \sum_{p}^{P} A_p(m)\cos\phi_k(m) = \mathrm{Re}\left[\sum_{p}^{P} A_p(m)e^{\phi_k(m)j}\right],$$

$$A_p(m) = \sqrt{R_p^2(m) + I_p^2(m)},$$

$$\varphi_p(m) = \arctan\left(\frac{-I(m)}{R(m)}\right),$$

$$F_p(m) = \varphi'_p(m)$$

где $R(m)$ и $I(m)$ являются действительной и мнимой частью $S_{F_\Delta,F_C^p}(m)$ соответственно. При вычислении $F_p(m)$, для того чтобы избежать разрывов, используются значения развернутой фазы из диапазона $[-\pi,\pi]$.

Нужно отметить, что приведенные выше соотношения имеют практический смысл только в том случае, если субполосные сигналы $S_{F_\Delta,F_C^p}(m)$ являются однокомпонентными [20], т.е. содержат не более одной гармоники основного тона. Для этого ширина полосы $2F_\Delta$ должна быть не больше минимально возможной частоты основного тона. Это условие достаточно легко удовлетворить в случае речевого сигнала, для которого принято считать, что основной тон не опускается ниже 50 Гц.

Требуемые аналитические сигналы $S_{F_\Delta,F_C^p}(m)$ могут быть получены путем фильтрации, используя следующую импульсную характеристику фильтров

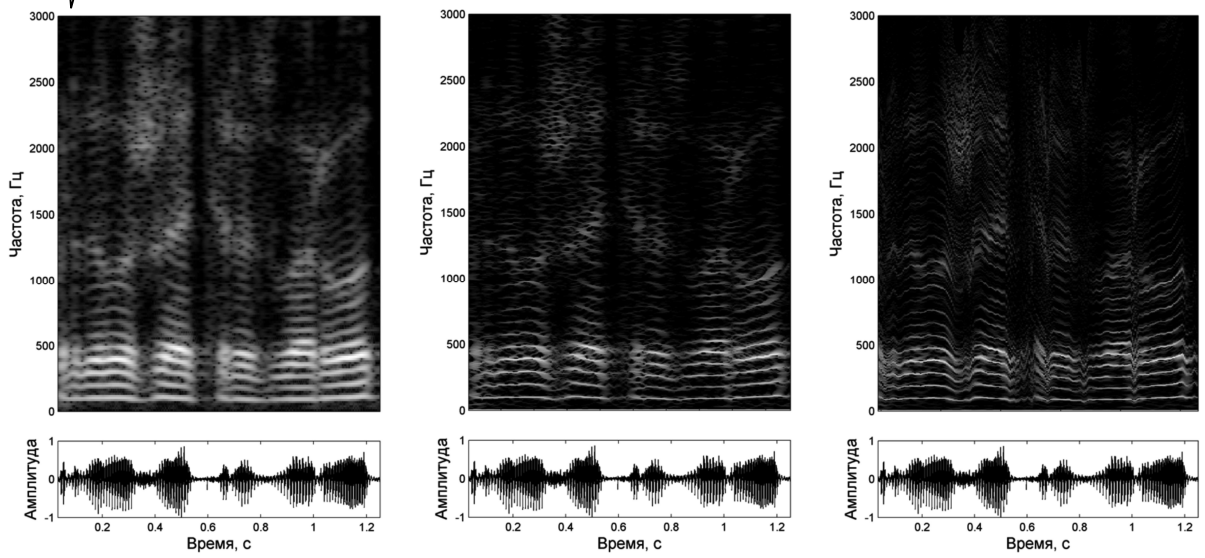
$$h_p(n) = 2\frac{\sin(F_\Delta n)}{n\pi}w(n)e^{-jF_C^p n}$$

где $w(n)$ – четная оконная функция. Данная импульсная характеристика может быть синтезированна оконным методом [20]. Выходом фильтра является требуемый узкополосный аналитический сигнал, который представляет собой свертку входного сигнала $s(m)$ с импульсной характеристикой $h_p(n)$:

$$S_{F_\Delta,F_C^p}(m) = \sum_{n=-\infty}^{\infty} h_p(n)s(m-n) \qquad (1)$$

При условии, что оконная функция $w(n)$ имеет конечную длину (равна нулю за пределами некоторого конечного интервала) и что значения частоты F_C^p распределены равномерно во всем частотном диапазоне, последнее выражение может быть вычислено при помощи быстрого преобразования Фурье (БПФ) или банка фильтров.

Изложенный метод оценки параметров синусоидальной модели позволяет более концентрированно распределить энергию квазипериодического сигнала в частотной области по сравнению с дискретным преобразованием Фурье (ДПФ). На рис. 3,а-б приведен пример представления речевого сигнала при помощи дискретного преобразования (ДПФ) и синусоидальной модели.



а) *б)* *в)*

Рис. 3. Представление речевого сигнала в частотной области различными методами: а) ДПФ, б) оценка параметров синусоидальной модели, в) оценка параметров синусоидальной модели с временным масштабированием сигнала

В случае быстрых изменений частоты гармоники невозможно получить точной оценки ее параметров из-за указанного выше ограничения ширины полосы аналитических сигналов ($2F_\Delta$). Избежать влияния модуляций основного тона на оценку параметров можно путем использования частотно-модулированной импульсной характеристики фильтра [20]:

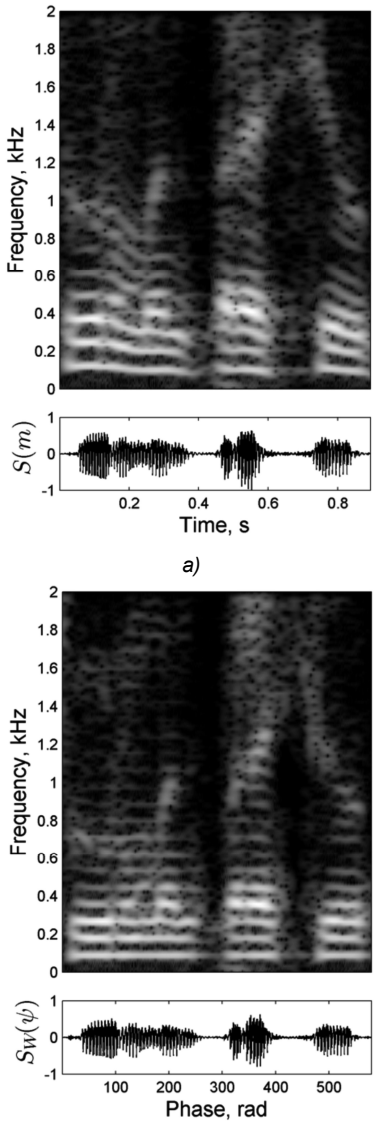
$$S_{F_\Delta, F_C^p}(m) = 2\sum_{n=-\infty}^{\infty} \frac{\sin(F_\Delta n)}{n\pi} w(n)s(m+n)e^{-j\varphi_C(n)} \qquad (2)$$

где $\varphi_C(n) = \sum_{l=0}^{n} F_C(l)$.

Для использования данного фильтра необходимо приближенно знать частотный контур каждой гармоники $F_C(l)$, который можно получить, имея приближенную оценку основного тона. В результате становится возможным оценить параметры гармоник более высоких порядков и еще больше локализовать энергию в частотной области – рис. 3 в.

Выражение (2) является неэффективным с вычислительной точки зрения и потому в практической реализации алгоритма не используется. Аналогичный результат можно получить, применяя выражение (1) к масштабированному сигналу. Временное масштабирование сигнала *s(m)* заключается в его передискретизации с переменной частотой, согласованной с частотой основного тона, в результате чего частотный диапазон каждой гармоники сужается – рис. 4.

Операция временного масштабирования может существенно повысить точность оценки гармонических параметров для сигналов с быстрым изменением основного тона, однако она подразумевает увеличение алгоритмической задержки и вычислительной сложности. Контур частоты $F_C(l)$ не известен с самого начала, и это означает что данная техника может применяться только после предварительной (более грубой) оценки основного тона. В предлагаемом алгоритме данная процедура уточнения частоты основного тона является опциональной и выполняется на стадии постобработки.



а)

б)

Рис. 4. Временное масштабирование сигнала, согласованное с частотой основного тона: а) исходный речевой сигнал; б) масштабированный сигнал

Функции оценки периодичности
Начальная оценка ЧОТ (IRAPT1)

Одним из традиционных способом генерации кандидатов периода основного тона является автокорреляционная функция [7]. Пусть *s(m)* – анализируемый сигнал, *z* – величина шага в отсчетах и *n* – размер окна, тогда автокорреляционная функция $R(x,k)$ для K отсчетов, задержки k и анализируемого фрейма x определяется как

$$R(x,k) = \sum_{i=m}^{m+n-k-1} s(i)s(i+k),$$

$$k = 0, K-1; m = xz; x = 0, M-1.$$

Благодаря относительной устойчивости к шуму автокорреляционная функция с успехом используется во многих алгоритмах оценки ЧОТ. Тем не менее, она имеет ряд недостатков, которые ограничивают ее использование в качестве функции генератора кандидатов периода. Основным из недостатков является необходимость использовать продолжительные окна анализа для того чтобы оценить периодичность сигнала во всем интересующем диапазоне. В результате резкие изменения ЧОТ приводят к потере четких пиков $R(x,k)$ в точке, соответствующей действительному периоду [7]. Другим недостатком является неодинаковое число отсчетов, участвующих в оценке $R(x,k)$ для разных задержек k. Это приводит к тому, что устойчивость автокорреляционной функции к шумам так же зависит от задержки и приводит к тому, что если для больших значений k окно анализа достаточно по длине, то для малых оно избыточно.

В RAPT периодичность фрагмента сигнала определяется при помощи НККФ $\phi(x,k)$, в которой недостатки автокорреляционной функции менее выражены. НККФ определяется как

$$\phi(x,k) = \frac{\sum_{i=m}^{m+n-1} s(i)s(i+k)}{\sqrt{e_m e_{m+k}}},$$

$$k = 0, K-1; m = xz; x = 0, M-1,$$

где $e_i = \sum_{l=i}^{i+n-1} s_l^2$.

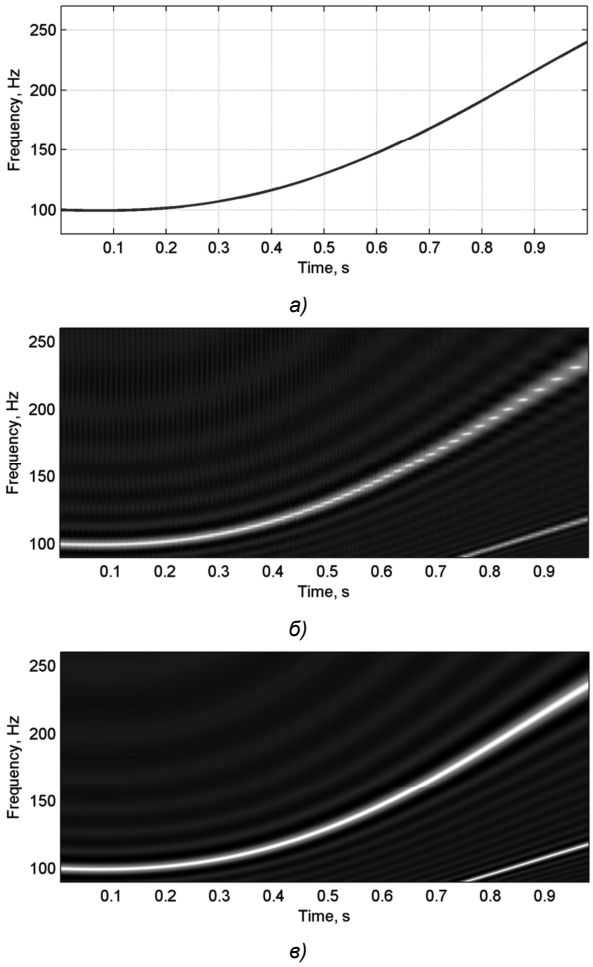
Следует отметить, что значения $\phi(x,k)$ находятся в диапазоне от -1 до +1, причем функция приближается к верхнему пределу для задержек, кратных действительному периоду основного тона вне зависимости от амплитуды анализируемого сигнала. Допустимый диапазон периода основного тона не зависит от продолжительности окна анализа. Если анализируемый сигнал является белым шумом, то $\phi(x,k)$ будет приближаться к нулю для всех $k > 0$ при увеличении длины окна анализа.

Как было сказано выше, в предлагаемом алгоритме оценки ЧОТ функция $\phi(x,k)$ оценивается при помощи мгновенных параметров синусоидальной модели сигнала. Параметрическое представление каждого отсчета *s(m)*, определяемое синусоидальной моделью, может быть использовано для вычисления мгновенной автокорреляционной функции $R_{inst}(m,k)$ используя теорему Винера-Хинчина:

$$R_{inst}(m,k) = \frac{1}{2}\sum_{p=1}^{P} A_p^2(m)\cos(F_p(m)k) \cdot$$

$R_{inst}(m,k)$ соответствует автокорреляционной функции, вычисленной для периодического сигнала бесконечной длины с постоянными значениями A_p и F_p. Поскольку окно анализа в данном случае бесконечно, то не будет разницы между нормированной автокорреляционной функцией и НККФ. Следовательно, НККФ можно оценить через мгновенные параметры синусоидальной модели следующим образом:

$$\phi_{inst}(m,k) = \frac{\sum_{p=1}^{P} A_p^2(m)\cos(F_p(m)k)}{\sum_{p=1}^{P} A_p^2(m)} \quad (3)$$



а)

б)

в)

Рис. 5. Функции генерации кандидатов искомого периода основного тона: а) действительный контур частоты основного тона, б) НККФ (RAPT), в) НККФ на основе синусоидальной модели сигнала (IRAPT)

Особенностью этой функции является то, что в отличие от НККФ, задержка k не обязательно должна быть целой и, таким образом, можно получить оценку периодичности для любого вещественного периода. Вторым важным отличием является то, что предлагаемая функция нечувствительна к любым изменениям частоты ос-

новного тона в окрестности отсчета m при условии, что полученные гармонические параметры являются достаточно точными. На рис. 5 показано, что для частотно-модулированного сигнала традиционная НККФ подвержена «эффекту ступенек», в то время как НККФ на основе синусоидальной модели формирует непрерывный контур кандидатов искомого периода основного тона.

Уточнение оценки ЧОТ (IRAPT2)

Учитывая то, что после первоначальной оценки основного тона и выполнения временного масштабирования каждый узкополосный аналитический сигнал соответствует одной гармонике основного тона, уточнение частоты основного тона может быть выполнено при помощи взвешенного среднего:

$$F_0(m) = \sum_{p=1}^{P} \frac{F_p(m) A_p(m)}{p \sum_{j=1}^{P} A_j(m)}.$$

Детали реализации алгоритма

Аппроксимация НККФ

Вычисление мгновенной функции генерации кандидатов периода основного тона $\phi_{inst}(m,k)$ с высоким частотным разрешением непосредственно из приведенного выражения требует большого количества операций. На практике целесообразно использовать ее аппроксимацию $\hat{\phi}_{inst}(m,k)$, которая может быть получена при помощи следующих шагов:

1. Аппроксимация мгновенной частоты фиксированной равномерной шкалой $G = 0, \frac{\pi}{H}, \frac{2\pi}{H}, ..., \frac{(H-1)\pi}{H}, \pi$. В результате чего вместо исходного вектора $F(m)$ мгновенной частоты формируется вектор $\hat{F}(m)$, содержащий элементы из фиксированного набора значений G. Число элементов шкалы $H+1$ определяет частотное разрешение аппроксимации.

2. Формирование амплитудного вектора $\hat{A}(m)$ размерности $2H$

$$\hat{A}_h(m) = \begin{cases} 0, G_{I(h)} \notin F(m) \\ \dfrac{A_p^2(m)}{\sum_{p=1}^{P} A_p^2(m)}, G_{I(h)} \in F(m), причем \quad G_{I(h)} = F_p(m) \end{cases},$$

где $I(h) = H + 1 - |H + 1 - h|$, $h = 1,2,...,2H$.

3. Вычисление быстрого обратного преобразования Фурье от вектора $\hat{A}(m)$.

4. Sinc-интерполяция полученных значений в нужном диапазоне.

Снижение вычислительной сложности происходит за счет использования быстрого обратного преобразования Фурье вместо многократного вычисления функции косинуса в выражении (3). В результате аппроксимация $\hat{\phi}_{inst}(m,k)$ вычисляется для некоторых фиксированных дискретных значений k. Потеря точности определяется числом элементов частотной шкалы $H+1$ и шагом интерполяции.

Аппроксимация отсчетов масштабированного сигнала

Операция временного масштабирования может быть определена функцией масштабирования $W(\psi) = m$, которая ставит в соответствие отсчетам сигнала $s(m)$ фазу основного тона ψ. Масштабированный сигнал $s_W(\psi)$ вычисляется в моменты времени, соответствующие равному шагу фазы основного тона: $s_W(\psi) = s(W(\psi))$. Для вычисления сигнала в произвольный момент времени, соответствующий дробному значению m, используется теорема Котельникова, утверждающая, что непрерывный сигнал, соответствующий дискретному сигналу с ограниченным спектром можно интерполировать при помощи свертки дискретного сигнала с sinc-функцией. Для того, чтобы не вычислять sinc-функцию для каждого момента времени в процессе работы алгоритма используется заранее сформированная таблица sinc-функций со смещениями от -0.5 до 0.5 отсчетов и шагом 10^{-3}. Использование дискретного шага приводит к незначительному ухудшению точности, однако существенно сокращает число требуемых вычислительных операций.

Вычислительная сложность алгоритма и алгоритмическая задержка

Вычислительная сложность понижения частоты дискретизации зависит линейно от частоты дискретизации исходного сигнала. Сложность оценки мгновенных параметров синусоидальной модели при помощи ДПФ-модулированного банка фильтров составляет $O(N + M \log_2 M)$, где N – порядок фильтра прототипа и M – число каналов банка. Сложность вычисления аппроксимации НККФ при помощи мгновенных параметров составляет $O(B \log_2 B)$, где B – формат быстрого обратного преобразования Фурье.

Первоначальная оценка частоты основного тона (IRAPT1) доступна с алгоритмической задержкой в 50мс, уточненная оценка (IRAPT2) требует дополнительной задержки в 43 мс.

Экспериментальная оценка точности алгоритма

Для оценки точности предложенного алгоритма используется набор искусственных, синтетических сигналов с заранее известной мгновенной частотой основного тона. Скорость изменения частоты основного тона сигналов изменяется от 0 до 2 Гц/мс. Значения тона находятся в пределах от 100 до 350Гц. Частота дискретизации сигналов – 44.1 кГц. К чистому тональному сигналу добавляется белый шум различной интенсивности для того, чтобы оценить устойчивость алгоритма к аддитивным шумам. Интенсивность шума определяется соотношением гармоники/шум (*HNR*)

$$HNR = 10 \lg \sigma_H^2 / \sigma_N^2,$$

где σ_H^2 - энергия гармонического сигнала и σ_N^2 - энергия шума. Диапазон HNR изменяется от 25дБ до 5дБ. Нижняя граница в 5дБ обусловлена тем, что фреймы с большим содержанием шума часто классифицируются RAPT как невокализованные.

Сравнивается пять различных алгоритмов: RAPT [7], YIN [8], SWIPE' [9] и две версии предложенного алгоритма оценки основного тона – одна без уточнения частоты основного тона (IRAPT 1) и вторая с уточнением частоты основного тона путем временного масштабирования сигнала (IRAPT 2).

Результат работы алгоритмов сравнивается в терминах: 1) процент грубых ошибок (gross pitch error - GPE) и 2) средний процент мелких ошибок (mean fine pitch error - MFPE).

Процент грубых ошибок вычисляется как

$$GPE(\%) = \frac{N_{GPE}}{N_V} \times 100,$$

где N_{GPE} - число фреймов с отклонением полученной оценки более чем на $\pm 20\%$ от настоящего значения основного тона, N_V - общее число вокализованных фреймов.

Средний процент мелких ошибок вычисляется для вокализованных фреймов без грубых ошибок

$$MFPE(\%) = \frac{1}{N_{FPE}} \sum_{n=1}^{N_{FPE}} \frac{\left| F_0^{true}(n) - F_0^{est}(n) \right|}{F_0^{true}(n)} \times 100,$$

где N_{FPE} - число вокализованных фреймов без грубых ошибок, $F_0^{true}(n)$ - действительные значения основного тона и $F_0^{est}(n)$ - оценочные значения основного тона.

Результаты тестирования алгоритмов с использованием синтетических сигналов приведены в табл. 1.

Приведенные результаты экспериментов показывают, что все алгоритмы имеют низкие показатели GPE и MFPE в случае неизменной частоты основного тона и преимущество IRAPT 1-2 становится заметным с увеличением частотных модуляций – рис. 6.

При наличии белого шума высокой интенсивности предлагаемый алгоритм сохраняет свое преимущество, однако при низких значениях HNR версия IRAPT 1 может быть предпочтительнее чем IRAPT 2.

Работа алгоритмов сравнивается с использованием натуральной речи при помощи речевой базы данных PTDB-TUG [21]. База данных содержит 2342 предложения, взятых из речевого корпуса TIMIT, прочитанных 10 дикторами мужчинами и 10 дикторами женщинами. База данных включает контрольные сигналы, полученные при помощи ларингографа и их оценочные значения частоты основного тона. Данные значения не могут рассматриваться как мгновенные, поэтому нельзя сравнить алгоритмы так же достоверно как в случае с синтетическими сигналами, однако эксперимент позволяет оценить применимость предложенного алгоритма к обработке настоящих речевых сигналов. Полученные результаты приведены в табл. 2.

Таблица 1. Сравнение алгоритмов оценки основного тона с использованием синтетических сигналов

		Скорость изменения частоты основного тона Гц/мс				
		0	0.5	1	1.5	2
HNR 25dB						
RAPT	GPE	**0**	**0**	**0**	7.90	18.42
	MFPE	0.037	0.103	0.219	0.405	0.778
YIN	GPE	**0**	**0**	**0**	**0**	5.36
	MFPE	**0.002**	0.156	0.778	2.136	3.905
SWIPE'	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.09	0.150	0.337	0.607	1.206
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.111	0.094	0.100	0.104	0.255
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.013	**0.050**	**0.051**	**0.060**	**0.114**
HNR 15dB						
RAPT	GPE	**0**	**0**	**0**	7.90	18.42
	MFPE	0.053	0.108	0.217	0.415	0.778
YIN	GPE	**0**	**0**	**0**	**0**	5.16
	MFPE	**0.004**	0.154	0.785	2.103	3.803
SWIPE'	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.165	0.193	0.347	0.632	1.194
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.113	0.094	0.102	0.111	0.273
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.049	**0.056**	**0.065**	**0.074**	**0.148**
HNR 5dB						
RAPT	GPE	**0**	**0**	**0**	10.52	18.42
	MFPE	0.161	0.205	0.268	0.506	0.871
YIN	GPE	**0**	**0**	**0**	**0**	4.33
	MFPE	**0.019**	0.151	0.813	1.948	3.524
SWIPE'	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.316	0.253	0.373	0.706	1.307
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.143	**0.099**	**0.115**	**0.147**	0.356
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.162	0.131	0.145	0.164	**.256**

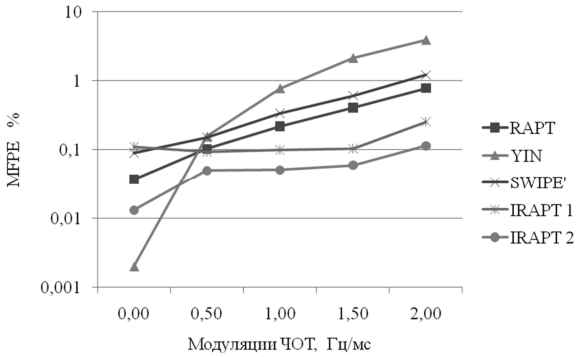


Рис. 6. Изменение точности оценки основного тона с увеличением частотных модуляций

Таблица 2. Сравнение алгоритмов оценки частоты основного тона с использованием речевых сигналов

	Мужской голос		Женский голос	
	GPE	MFPE	GPE	MFPE
RAPT	3.687	1.737	6.068	1.184
YIN	3.184	**1.389**	3.960	0.835
SWIPE'	**0.783**	1.507	4.273	**0.800**
IRAPT 1	1.625	1.608	**3.777**	0.977
IRAPT 2	1.571	1.565	**3.777**	1.054

Для натуральных речевых сигналов предложенный алгоритм показывает близкий результат к другим алгоритмам оценки, что говорит о его применимости в реальных приложениях обработки речи.

Заключение

В работе предложен алгоритм оценки мгновенной частоты основного тона. Основными достоинствами алгоритма являются высокое частотно/временное разрешение и устойчивость к частотным модуляциям основного тона. Алгоритм использует RAPT в качестве архитектурной основы, однако содержит ряд существенных модификаций: 1) генератор кандидатов искомого периода основного тона вычисляется на основе мгновенных гармонических параметров синусоидальной модели сигнала, что позволяет оценить периодичность применительно к отдельному моменту времени, а не к целому фрейму сигнала; 2) функция генератора кандидатов нечувствительна к частотным модуляциям основного тона; 3) на стадии постобработки выполняется уточнение основного тона путем временного масштабирования сигнала и более точной оценки параметров модели. Тестирование алгоритма выполнено с использованием синтетических сигналов с известной мгновенной частотой основного тона и различным соотношением гармоники/шум. Показано, что точность оценок предложенного алгоритма снижается медленнее чем у RAPT, YIN и SWIPE' при повышении скорости изменения частоты основного тона. Алгоритм является устойчивым к аддитивному шуму. Эксперименты с натуральной речью показали, что предложенный алгоритм применим к приложениям обработки речи. Алгоритм может быть использован в реальном масштабе времени в приложениях, где допускается постоянная алгоритмическая задержка в 50-90мс.

Литература

1. Hess, W. J. «Pitch and voicing determination», in Advances in Speech Signal Processing, edited by S. Furui and M. M. Sohndi (**Marcel Dekker, New York), 1992, pp. 3–48.

2. Hermes, D. J. «Pitch analysis,» in Visual Representations of Speech Signals, edited by M. Cooke, S. Beet, and M. Crawford Wiley, ** New York, 1993, pp. 3–25.

3. D. Gerhard. Pitch Extraction and Fundamental Frequency: History and Current Techniques, technical report, Dept. of Computer Science, University of Regina, 2003.

4. V. Sercov, A. Petrovsky «The method of pitch frequency detection on the base of turning to its harmonics» / Sercov V., Petrovsky A. //Proc. of the European Signal Processing Conference (EUSIPCO-98), pp. 1137-1140, Island of Rhodes, Greece, 8-11 September, 1998.

5. A. Pavlovets and A. Petrovsky, «Robust HNR-based closed-loop pitch and harmonic parameters estimation» / Pavlovets A., Petrovsky A. // Proc. the 12th Annual Conference of the International Speech Communication Association (Interspeech-2011), Italy, Florence, 27-31 August 2011.

6. P. Zubrycki and A. Petrovsky, «Quasi-periodic signal analysis using harmonic transform with application to voiced speech processing»/ Zubrycki P., and Petrovsky A. // ISCAS 2010: 2374-2377.

7. D. Talkin, «A Robust Algorithm for Pitch Tracking (RAPT)» in «Speech Coding & Synthesis», W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.

8. A. Cheveigné and H. Kawahara «YIN, a fundamental frequency estimator for speech and music», *Journal Acoust. Soc. Am.*, vol. 111, no. 4, pp 1917-1930, Apr. 2002.

9. A. Camacho and J. G. Harris, «A sawtooth waveform inspired pitch estimator for speech and music», *Journal Acoust. Soc. Am.*, vol. 123, no. 4, pp 1638-1652, Sep. 2008.

10. J.O. Hong and P.J. Wolfe, «Model-based estimation of instantaneous pitch in noisy speech» in Proceedings of INTERSPEECH, 2009.

11. B. Resch, M. Nilsson, A. Ekman and W. B. Klejin «Estimation of the Instantaneous Pitch of Speech», IEEE Trans. on Audio, Speech, and Lang. Process., 2007, vol. 15, no. 3, pp. 813-822.

12. Kobayashi, T. Fundamental frequency estimation based on instantaneous frequency amplitude spectrum / T. Kobayashi, D. Arifianto, T. Masuko // Proc of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). – 2002. – P. I-329-I-332.

13. Kawahara, H., Masuda-Katsuse, I., and de Cheveig-ne., A. «Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,» Speech Commun. 1999. 27, P. 187–207.

14. E. Azarov, and A. Petrovsky, «Real-time voice conversion based on instantaneous harmonic parameters» / E. Azarov, and A. Petrovsky // Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2011), pp. 5140 - 5143, Prague, Czech Republic, May 22-27, 2011.

15. Hajime Sano and B. Keith Jenkins. A neural network model for pitch perception. Computer Music Journal, 13(3):41–48, Fall 1989.

16. Barnard, E., Cole, R. A., Vea, M. P., and Alleva, F. A. «Pitch detection with a neural-net classifier», IEEE Trans. Signal Process. 39, 1991, pp. 298–307.

17. Shahnaz, C., Zhu W.P., Ahmad M. O. «Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme», IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 322-335, January 2012.

18. Azarov, E. Instantaneous Pitch Estimation Based on RAPT Framework / E. Azarov, M. Vashkevich and A. Petrovsky // Proc. of the 20th European Signal Process. Conf. (EUSIPCO-2012). – 2012. – P. 2787–2791.

19. Abe, T. and Honda, M. «Sinusoidal model based on instantaneous frequency attractors», IEEE Transactions on Audio, Speech, and Language Processing, Volume: 14, Issue 4, pp. 1292 – 1300, July 2006.

20. Al. Petrovsky, E. Azarov and A. Petrovsky, «Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding», Signal Processing, vol. 91, Issue 6, Fourier Related Transforms for Non-Stationary Signals, pp. 1489-1504, June 2011.

21. G. Pirker, M. Wohlmayr, S. Petrik and F. Pernkopf, «A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario», in Proceedings of INTERSPEECH, 2011, p. 1509-1512.