Подходы для оценки периода основного тона речевого сигнала в зашумлённой среде

Имамвердиев Я.Н.,

Сухостат Л.В.

Высота голоса человека определяется периодом основного тона. Информация об основном тоне речевого сигнала очень важна в приложениях анализа и синтеза речи. В статье рассматриваются и анализируются существующие методы определения периода основного тона. Обсуждаются временные, частотные и гибридные методы, которые тестируются на речевых базах данных Keele и CSTR. Приводятся сильные и слабые стороны каждого из рассматриваемых методов. Каждый из них имеет свои особенности в условиях различного шума.

● *высота голоса* ● *вокализованный/невокализованный участок речевого сигнала* ● *период основного тона* ● *шум*

Pitch is determined by the fundamental frequency. It is the most important component in various speech processing systems. This paper discusses and analyzes the existing methods for pitch detection. Time-domain, frequency-domain and hybrid methods, which are tested on Keele and CSTR speech databases, are discussed. Advantages and disadvantages of each method are pointed out. Each of them has its own characteristics in terms of different noise conditions.

● *pitch* ● *voiced/unvoiced speech segment* ● *fundamental frequency* ● *noise*

Введение

Речевой сигнал представляет собой совокупность колебаний разных частот, источник которых – органы речеобразования человека. В процессе артикуляции воздух из лёгких проходит через трахею и голосовые связки, которые смыкаются и размыкаются, модулируя воздушный поток, вследствие чего он приобретает вид последовательности импульсов сложной формы. Частота колебания голосовых связок определяет такую характеристику, как высота голоса человека. Частота, с которой голосовые связки вибрируют, является частотой основного тона, а соответствующий ей звук определённой высоты – основным тоном. Чем больше период основного тона, тем звук громче, и наоборот. Частота основного тона является важным атрибутом вокализованной речи.

Помимо предоставления ценной информации о природе источника возбуждения для речеобразования, информация об основном тоне высказывания очень важна в задачах распознавания диктора, в определении его эмоционального состояния, обнаружении голосовой активности, обучении речи слабослышащих, а также в приложениях анализа и синтеза речи [1].

Причина отличия в частоте основного тона между дикторами состоит в различии размеров, массы и натяжения гортанного тракта, который включает голосовые складки и голосовую щель. В детстве частота основного тона составляет около 250 Гц, а длина голосовых складок – 10,4 мм. С возрастом длина голосовых складок у мужчин возрастает примерно до 15–25 мм, а у женщин – 13–15 мм. Эти изменения в размерах соотносятся с уменьшением частот, поступающих от голосовых складок. Женщины имеют более высокий диапазон частот, чем мужчины, потому что размер гортани у них меньше.

Тем не менее, точное и надёжное измерение частоты (или обратной ей величины – периода) основного тона речевого сигнала часто чрезвычайно трудно по нескольким причинам. Одна из трудностей состоит во взаимодействии между вокальным трактом и голосовым источником. В некоторых случаях форманты речевого тракта могут значительно изменить квазипериодическую структуру речевого сигнала таким образом, что фактический период основного тона трудно обнаружить. Ещё одна трудность может возникнуть в практической ситуации, когда речевой сигнал зашумлен. Это приводит к искажениям сигнала и потерям его компонент в спектральной и временной областях, и как следствие – к неправильному измерению периода основного тона.

Для решения проблем, связанных с измерением основного тона, были разработаны самые разнообразные методы и проведено их сравнение [2, 3].

Вычисление периода основного тона происходит в два этапа: сначала принимается решение вокализованный/невокализованный участок речевого сигнала, после чего на вокализованных участках вычисляется период основного тона. Следует отметить, что во многих алгоритмах принятие решения о вокализованности является частью процесса вычисления, а не отдельным этапом.

В последние годы были предложены альтернативы к более традиционным подходам в оценке периода основного тона [4–7].

В данной работе в первую очередь мы даём обзор существующих подходов для определения периода основного тона с целью выявления их сильных и слабых сторон. А затем проводим их детальное сравнение.

Эта статья организована следующим образом. В разделе 2 описываются методы обнаружения высоты во временной и частотной областях, а также гибридные подходы.

В разделе 3 приводятся меры оценки методов и даётся краткое описание применяемых баз данных. Раздел 4 содержит результаты и обсуждение рассматриваемых подходов, и последний раздел даёт некоторые заключительные замечания.

1. Методы обнаружения частоты основного тона

Оценивание частоты (или периода) основного тона является одной из наиболее важных задач в обработке речи для получения просодических характеристик [8–10]. Существующие методы оценки частоты основного тона [2] можно условно разделить на три группы:

1. С использованием преимущественно свойств временной области речевых сигналов.
2. С использованием преимущественно свойств частотной области речевых сигналов.
3. Гибридный подход, объединяющий как свойства временной, так и свойства частотной области речевого сигнала.

Методы во временной области выполняют анализ амплитудно-временного представления речевого сигнала. Для этого типа детекторов периода основного тона [11–16] пики, долины (valley), частота переходов через нуль и измерения автокорреляции очень важны. Измерения временной области обеспечивают хорошую оценку периода основного тона, если квазипериодический сигнал был соответствующим образом обработан для минимизации последствий формантной структуры.

Идея методов из этой категории заключается в рассмотрении входного сигнала как колебания амплитуды во временной области и нахождении повторяющихся образов в звуковой волне, которые определяют её периодичность [17, 18].

Известными детекторами частоты основного тона во временной области являются метод параллельной обработки (Parallel Processing Time-Domain Method, PPROC) [13, 19] и метод редукции данных (Data Reduction Method, DARD) Миллера [11, 14], где частота ошибок не зависит от уровня шума. Эти два метода имеют несколько меньшее разрешение, чем другие методы, из-за чувствительности пиков звуковой волны, долин и частоты переходов через нуль к изменениям формант, шуму, искажениям и т.д. Добавление алгоритма сглаживания улучшает их производительность.

Основная трудность использования этих методов возникает при определении основного тона для мужских голосов: окно анализа имеет фиксированную длину, которая составляет 30–40 мс и, как правило, не подходит для оценки периода основного тона в области низких частот.

Детекторы основного тона в частотной области [20–23] проводят типичный анализ частотного спектра, который состоит в разбиении сигнала на небольшие фреймы, их умножении на оконную функцию и получении кратковременного преобразования Фурье для каждого фрейма. Если сигнал является квазипериодическим, преобразование Фурье покажет несколько пиков в области низких частот, соответствующих периоду основного тона. Поскольку человеческое восприятие высоты тона, свя-

занное с частотой звука, логарифмично, это означает, что низкие частоты могут быть отслежены менее точно, чем высокие.

К методам анализа сигнала в частотной области можно отнести произведение гармоник спектра (Harmonic Product Spectrum, HPS) [24], кепстральный метод [25] и метод линейного предсказания (Linear Predictive Coding, LPC) [26]. Преимущества метода HPS заключаются в том, что он вычислительно менее сложен, устойчив к шуму и помехам [27]. Метод LPC [28] с помощью обратной фильтрации отделяет сигнал возбуждения от вокального тракта и использует реальную часть кепстра для обнаружения частоты основного тона.

Гибридный детектор высоты анализирует речевой сигнал как во временной области, так и в частотной. Например, гибридный детектор высоты может использовать методы частотной области, чтобы обеспечить спектрально выровненную временную звуковую волну, а затем использовать измерения автокорреляции для оценки периода основного тона. Методы этой категории выполняют анализ во временной области выходов банка полосовых фильтров для вычисления множества кандидатов частоты основного тона на каждом фрейме сигнала.

Среди различных гибридных алгоритмов обнаружения периода (частоты) основного тона можно выделить следующие: модифицированный автокорреляционный метод (Modified Autocorrelation Method, AUTOC) [2], кепстральный метод, многополосная агрегация коррелограмм (Multi-Band Summary Correlogram, MBSC) [6], BaNa [7], YIN [4], YAAPT [5], среднее значение разностной функции (Average magnitude difference function, AMDF) [2], SWIPE' [29] и метод оценки высоты на основе амплитудного сжатия (Pitch estimation Filter with Amplitude Compression, PEFAC) [30].

Рассмотрим более подробно методы обнаружения периода основного тона всех трёх категорий.

1.1. Методы во временной области

1.1.1. Модифицированный автокорреляционный метод

Автокорреляционный подход является наиболее широко используемым методом во временной области для оценки периода основного тона речевого сигнала [1]. Этот метод [15] основан на выявлении наиболее высоких значений автокорреляционной функции в интересующих нас областях.

Функция автокорреляции речевого сигнала $\{s(n), n = 0, 1, ..., N-1\}$ вычисляется как

$$R(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} s(n)s(n+k) \quad k = 0,1,\ldots,N-1 . \qquad (1)$$

Согласно (1) ищется пик, расположение которого определяет значение высоты.

Для женских голосов, имеющих изначально более высокий тон, чем у мужских, метод взаимной корреляции и метод автокорреляции [22] дают близкие результаты. Тем не менее, оба метода недостаточно устойчивы, особенно для речи с очень низкой частотой и быстро меняющейся тональностью.

Ниже приводится алгоритм работы метода AUTOC.

Вход: вектор отсчётов речевого сигнала $x \in R^{1xN}$; частота дискретизации F_S ; длина фрейма L_m ; сдвиг фрейма R_m .

Выход: значения периода основного тона.

1. Низкочастотная фильтрация до 900 Гц.
2. Преобразование L_m и R_m

$$L=L_mF_S/1000,\ R=R_mF_S/1000.$$

3. Нахождение первого $I_{pk}1$ и последнего $I_{pk}2$ образцов для каждого фрейма.
4. Определение уровня амплитудного ограничения $C_L=\alpha\ min(I_{pk}1, I_{pk}2)$.
5. Вычисление функции автокорреляции (1).
6. Нахождение максимального автокорреляционного пика.
7. Принятие решения вокализованный/невокализованный.
8. Медианная фильтрация.

Основным ограничением оценки частоты основного тона методом автокорреляции является наличие пиков, превышающих пики, соответствующие периоду основного тона. В результате может быть «собирание» пиков и, следовательно, неправильная оценка частоты.

1.1.2. Среднее значение разностной функции

Алгоритм обнаружения частоты основного тона методом AMDF [11] имеет относительно низкую вычислительную стоимость и лёгок в применении [1]. Принцип обнаружения частоты основного тона речевых сигналов с использованием AMDF основан на кратковременной функции, в которой вычисляется разность между исходной функцией речевого сигнала и функцией, сдвинутой по оси времени.

AMDF используется для измерения периода основного тона для вокализованных участков речевого сигнала (период основного тона не определён для невокализованных сигналов, поскольку в них нет периодического возбуждения (periodical excitation)). Сигнал разбивается на фреймы длиной N. На каждом из них вычисляются значения AMDF-функции [6] следующим образом

$$A(k)=\frac{1}{N}\sum_{n=0}^{N-1-k}\left|s(n)-s(n+k)\right|,\ k=0,1,\ldots,N-1, \qquad (2)$$

где $s(n)$ – отсчёты анализируемого речевого сигнала, $s(n+k)$ – отсчёты с временным смещением на k. Функция разницы будет иметь сильный локальный минимум, если сдвиг k равен или очень близок к периоду основного тона.

В отличие от автокорреляционной функции, расчёты AMDF не требуют операций умножения. Это хорошее свойство для приложений реального времени. После сегментации, сигнал предварительно обрабатывается, чтобы удалить эффекты фонового шума путём низкочастотной фильтрации. В дополнение к оценке периода основного тона, получаем соотно-

шение между максимальным и минимальным значениями AMDF. Это измерение совместно с энергией сигнала используется для выделения вокализованных/невокализованных участков, в которых могут возникнуть некоторые ошибки определения частоты основного тона F_0. К таким ошибкам относится удвоение F_0. Поэтому в основе AMDF используется медианная фильтрация. Расположение минимума AMDF-функции соответствует значению периода основного тона.

Вычисление AMDF включает следующие шаги:

Вход: вектор отсчётов речевого сигнала $x \in R^{1xN}$; частота дискретизации F_S ; длина фрейма L_m ; сдвиг фрейма M_m ; *pdlow* и *pdhigh* – наименьший и наибольший допустимые периоды основного тона соответственно; порог T_h.

Выход: значения периодов основного тона для каждого фрейма.

1. Низкочастотная фильтрация до 900 Гц.
2. Измерение частоты переходов через нуль

$$ZCR = \frac{1}{2}\sum_{n=1}^{N-1} \left| \mathrm{sgn}(x[n]) - \mathrm{sgn}(x[n-1]) \right| .$$

3. Измерение энергии

$$E = 10\log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]\right).$$

4. Вычисление функции AMDF согласно (2).
5. Принятие решения вокализованный/невокализованный, используя порог *Th*.

Методы, основанные на автокорреляции и AMDF, более популярны и широко используются [2].

1.1.3. Метод YIN

Алгоритм YIN был предложен в [4] и основан на автокорреляции и разностной функции, которая, подобно автокорреляции, стремится минимизировать разницу между функцией речевого сигнала $x(t)$ и функцией, сдвинутой по оси времени.

Алгоритм YIN после разбиения сигнала на фреймы включает 6 шагов:

1. Вычисление на каждом фрейме автокорреляционной функции [31] от $x(t)$

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j \cdot x_{j+\tau}$$

где τ – задержка, t – начальный индекс времени и W – размер фрейма.

2. Нахождение разностной функции для уменьшения гармонических и субгармонических ошибок

$$d_t(\tau) = r_t(0) - r_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2$$

3. Вычисление интегральной средней нормализованной функции:

$$d'_t(\tau) = \begin{cases} 1, & \tau = 0 \\ \dfrac{d_t(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} d_t(j)}, & \text{в противном случае.} \end{cases} \quad (3)$$

4. Применение порога для выбора решений, направленных на минимизацию субгармонических ошибок. Порог определяет список кандидатов.

5. Интерполяция $d'_t(\tau)$ для расчёта истинной частоты от целого периода с использованием кубического сплайна.

6. Для каждого индекса времени t ищется минимум в (3) на интервале $[t-T_{max}/2, t+T_{max}/2]$, где T_{max} является наибольшим ожидаемым периодом.

Точность метода YIN ограничена целочисленными временными задержками и зависит от частоты дискретизации.

Хотя YIN является очень успешным и стабильным во временной области алгоритмом оценки частоты основного тона, он не имеет стадии пост-обработки для коррекции полученного результата.

1.2. Методы в частотной области

1.2.1. Кепстральный метод обнаружения высоты

Кепстр определяется как обратное преобразование Фурье от логарифма амплитуды спектра. Медленно меняющиеся компоненты при этом представляют собой огибающую, соответствующую вокальному тракту, а быстро изменяющиеся компоненты – источнику возбуждения.

Кепстр вокализованных участков речевого сигнала имеет сильный пик, соответствующий периоду основного тона [25]. Предполагается, что последовательность отсчётов вокализованных участков речевого сигнала $s(n)$ может быть представлена в виде свёртки

$$s(n) = e(n)\ h(n),$$

где $e(n)$ – сигнал возбуждения в виде случайного шума;
$h(n)$ – дискретная импульсная характеристика речевого тракта.

В частотной области это отношение принимает вид:

$$S(\omega) = E(\omega)\ H(\omega), \quad (4)$$

где $S(\omega) = F\{s(n)\}$,
$E(\omega) = F\{e(n)\}$,
$H(\omega) = F\{h(n)\}$.

Символ F обозначает дискретное преобразование Фурье.

Далее функция (4) может быть представлена в виде:

$$F^{-1}\left(\log\left[S(\omega)\right]\right) = F^{-1}\left(\log\left[E(\omega)\right]\right) + F^{-1}\left(\log\left[H(\omega)\right]\right) \quad (5)$$

Можно отделить часть кепстра, которая представляет исходный сигнал, и найти период основного тона. Именно поэтому, в общем, результаты кепстрального анализа по определению основного тона являются более точными, чем при использовании автокорреляции [1]. Для определения периода основного тона достаточно реальной части кепстра. Логарифмическая магнитуда спектра $s(n)$ определяется как

$$S(k)=\log\left\{\sum_{n=1}^{N-1} s(n)\cdot e^{-j\frac{2\pi}{N}nk}\right\}$$

Рассмотрим алгоритм работы кепстрального метода.

Вход: вектор отсчётов речевого сигнала $x \epsilon R^{1xN}$; частота дискретизации F_s; длина фрейма L_m; сдвиг фрейма M_m.

Выход: значения периода основного тона для каждого фрейма.

1. Измерение частоты переходов через нуль

$$ZCR=\frac{1}{2}\sum_{n=1}^{N-1}\left|\mathrm{sgn}(x[n])-\mathrm{sgn}(x[n-1])\right|.$$

2. Применение к сигналу оконной функции Хэмминга

$$w[n]=\begin{cases}0.54-0.46\cdot\cos\left(2\pi\frac{n}{N-1}\right), & 0\le n\le N-1\\ 0, & \text{в противном случае}\end{cases}.$$

3. Нахождение магнитуды $FFT|S_\omega|$ данного временного ряда, где $\omega = 1, ..., N$.
4. Вычисление $F^{-1}\{Log(S_\omega)\}$.
5. Получение значений реального кепстра

$$C(m)=\frac{1}{N}\left\{\sum_{k=1}^{N-1} S(k)\cdot e^{-j\frac{2\pi}{N}mk}\right\}$$

6. Нахождение максимального пика F_{0max}.
7. Принятие решения вокализованный/невокализованный.

Следует отметить, что кепстральный детектор основного тона использует весь речевой сигнал для обработки. Для каждого фрейма определяется пиковое кепстральное значение и его расположение. Если значение этого пика превышает фиксированный порог, фрейм помечается как вокализованный и период основного тона соответствует расположению пика. Если пик не превышает порог, то выполняется вычисление частоты переходов через нуль для данного блока. В свою очередь, если количество переходов через нуль превышает заданный порог, то фрейм помечается как невокализованный.

Кепстр не обеспечивает хорошие результаты для зашумленной речи, поскольку кепстральные коэффициенты чувствительны к шумам и искажениям сигнала. В условиях шума трудно найти пик, соответствующий периоду основного тона, даже при использовании порога, адаптированного под шум.

1.2.2. Метод SWIPE'

Метод SWIPE для оценки основного тона был предложен в [29]. Он похож на автокорреляционные методы тем, что выполняет интегральное преобразование спектра с помощью косинусного ядра (теорема Винера-Хинчина). Тем не менее, вместо того, чтобы использовать квадрат магнитуды спектра, SWIPE рассматривает его квадратный корень. Кроме того, метод содержит некоторые изменения в косинусном ядре во избежание некоторых проблем, возникающих при автокорреляции.

Существенное улучшение достигается в алгоритме SWIPE' путём анализа спектра только на его первой и простых гармониках. Это выполняется путём удаления из ядра непростых пиков (положительных косинусных лепестков) и их соседних долин (отрицательных косинусных лепестков). Что и является главным отличием SWIPE от SWIPE'.

1.3. Гибридные методы

1.3.1. Метод оценки высоты на основе амплитудного сжатия

Алгоритм PEFAC [30] оценивает период основного тона каждого фрейма путём свёртки его спектральной плотности мощности (power spectral density) в области логарифмических частот с помощью фильтра, который суммирует энергию гармоник высоты, в то же время отбрасывая аддитивный шум, который имеет плавно изменяющийся спектр мощности. Сжатие амплитуды применено до фильтрации для ослабления узкополосных шумовых составляющих.

Для идеального периодического источника на частоте f_0 модель сигнала в момент времени t в области спектральной плотности мощности имеет вид

$$Y_t(f)=\sum_{k=1}^{K} a_{k,t}\,\delta(f-kf_0)+N_t(f),$$

где $N_t(f)$ представляет собой спектральную плотность мощности нежелательных шумов и $a_{k,t}$ – мощность k-й гармоники. В логарифмической области модель сигнала может быть выражена как

$$Y_t(f)=\sum_{k=1}^{K} a_{k,t}\,\delta(q-\log k-\log f_0)+N_t(f),$$

где $q = \log f$. В этой области расстояние между гармониками не зависит от f_0 и, следовательно, их энергии могут быть объединены путём свёртки $Y_t(q)$ с помощью фильтра с импульсной характеристикой

$$h(q)=\sum_{k=1}^{K}\delta(q-\log k)$$

Хотя шум с плавно меняющимся спектром будет подавлен с помощью фильтра $h(q)$, некоторые источники шума содержат высокоамплитудные узкополосные компоненты, которые могут доминировать на выходе фильтра. Для этого используется сжатие спектра каждого временного фрейма перед свёрткой с помощью $h(q)$, установив

$$Y_t'(q)=Y_t(q)^{\alpha_t(q)}$$

где t – временной индекс, а $\alpha_t(q)$ – показатель сжатия.

Алгоритм PEFAC надёжен при отрицательном соотношении сигнал-шум (signal-to-noise ratio, SNR). Алгоритм превзошёл другие широко используемые методы, даже те, которые имеют временные ограничения на непрерывность.

1.3.2. Метод YAAPT

YAAPT от «yet another algorithm for pitch tracking» («ещё один алгоритм обнаружения основного тона»), который был впервые предложен в работе [5], основан на рассмотрении как временной области, так и частотной. Ключевым компонентом метода является нормализованная кросскорреляционная функция (Normalized Cross Correlation Function, NCCF) [32], которая также используется в методе RAPT [33].

Метод YAAPT включает следующие шаги:

1. **Нелинейная обработка** применяется к речевому сигналу для восстановления периода основного тона.
2. **Расчёт контура F_0 из спектрограммы:** спектральный контур F_0 оценивается с помощью спектральных гармоник корреляции (Spectral Harmonics Correlation, SHC) из спектрограммы нелинейного обрабатываемого сигнала.
3. **Оценка F_0 кандидата:** кандидаты извлекаются на основе NCCF во временной области.
4. **Вычисление конечного F_0:** динамическое программирование применяется для получения конечного F_0 контура.
5. **Определение вокализованных и невокализованных участков** на основании нормализованного низкочастотного коэффициента энергии (Normalized Low Frequency Energy Ratio, NLFER).

1.3.3. Метод многополосной агрегации коррелограмм

MBSC [6] является ещё одним алгоритмом обнаружения основного тона в зашумлённой речи. Он включает применение гребенчатого фильтра для получения потока отдельных поддиапазонов агрегации коррелограмма (summary correlogram, SC) и надёжное взвешивание потока для объединения этих SC в единый MBSC.

MBSC-детектор основного тона включает 5 основных этапов обработки сигналов:

1. **Частотное разложение с использованием широкополосных КИХ-фильтров:** входной речевой сигнал разлагают с помощью четырёх КИХ-фильтров.
2. **Обнаружение огибающей для каждого поддиапазона:** огибающая Гильберта [34] в каждом поддиапазоне извлекается путём вычисления магнитуды аналитического сигнала.
3. **Многоканальная гребенчатая фильтрация:** гребенчатая фильтрация выполняется в частотной области для каждого потока поддиапазона, гребенчатые фильтры захватывают гармоническую энергию сигнала.
4. **Выбор канала на основе HPS и вычисление SC для каждого потока:** вычисление HPS направлено на улучшение оценки периода основного тона и обнаружение вокализованности.
5. **Оценка основного тона и обнаружение вокализованных участков:** поток SC далее объединяется в MBSC. Обнаружение вокализованности выполняется на основании порогового значения.

1.3.4. Метод BaNa

BaNa [7] является гибридным подходом для обнаружения основного тона в зашумлённом сигнале, который включает использование гармонических частот и кепстрального метода.

Алгоритм BaNa включает следующие шаги:

1) предобработка;
2) поиск гармонических пиков;
3) выбор F0 -кандидатов;
4) оценка конечного F0.

2. Оценка методов и экспериментальные базы данных

Для проведения экспериментов были рассмотрены речевые базы данных, содержащие, помимо речевых образцов, также и записи с помощью ларингографа (laryngograph), что облегчает вычисление периода основного тона. Базы данных также содержат эталонные значения частот основного тона (ground truth).

База данных CSTR для оценки алгоритмов определения основного тона [35, 36] была собрана в Университете Эдинбурга и включает образцы мужских и женских голосов. Она доступна на сайте http://www.cstr.ed.ac.uk/research/projects/fda. Речевые сигналы были записаны с частотой дискретизации 20 кГц. Период основного тона был вычислен путём оценки расположения импульсов голосового источника в данных ларингографа и взятия обратного расстояния между каждой парой последовательных импульсов.

Другая база данных Keele [37] была создана в Кильском университете. Она содержит речевые сигналы с глубиной квантования 16 бит при частоте дискретизации 20 кГц. Период основного тона был оценен с помощью автокорреляции с окном 25,6 мс и сдвигом в 10 мс.

Обе эти базы данных, CSTR и Keele, как сообщается в [4], содержат ошибки, особенно в конце предложений, где значения энергии речевого сигнала понижаются. Корпус содержит записи, прочитанные 5 женщинами и 5 мужчинами. Записи длятся около 30 секунд. Эталонные значения периода основного тона были извлечены из ларингографа с помощью алгоритма автокорреляции.

Для тестирования устойчивости к внешним шумам рассматриваемых алгоритмов определения основного тона, к акустическим сигналам добавляются 3 вида шума с различными уровнями SNR, которые были выбраны из базы данных шумов NOISEX-92 [38] и включали в себя: «бульканье» (babble), Volvo (car) и белый шум (white) (табл. 1). Для генерации зашумлённой речи с определённым значением SNR, энергия сигнала вычисляется только на вокализованных участках речевого сигнала, и шум усиливается или ослабевает до определённого уровня, чтобы удовлетворить значению целевого SNR.

Таблица 1

Характеристики шумов

Тип шума	Спектральные характеристики
babble	Нестационарный шум
car	Стационарный низкочастотный шум
white	Стационарный шум

Перечень рассмотренных алгоритмов, точность которых была оценена на речевых базах данных Keele и CSTR, представлен в табл. 2. Эти алгоритмы выбраны, потому что в них реализованы наиболее популярные подходы для определения параметров основного тона. Для этих алгоритмов параметры установлены по умолчанию. Диапазон частот установлен в пределах от 50 до 600 Гц.

Таблица 2

Популярные алгоритмы определения частоты (периода) основного тона

Алгоритм	URL-адрес
AUTOC (Rabiner, 1977)	www.mathworks.co.uk/matlabcentral/fileexchange/45309-autocorrelation-pitch-detector
AMDF (Ross и др., 1974)	www.mathworks.com/matlabcentral/fileexchange/45274-amdf
Кепстральный метод (Noll, 1967)	www.ece.rochester.edu/projects/wcng/project_bridge.html
YIN (Cheveigné и Kawahara, 2002)	http://audition.ens.fr/adc/sw/yin.zip
PEFAC (Gonzalez и Brookes, 2011)	www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/fxpefac.html
YAAPT (Kasi и Zahorian, 2002)	www.ws.binghamton.edu/zahorian/yaapt.htm
MBSC) (Tan и Alwan, 2013),	www.ee.ucla.edu/%7Espapl/code/MBSC_matlab.zip
BaNa (Ba и Yang, 2012)	www.ece.rochester.edu/projects/wcng/project_bridge.html
SWIPE' (Camacho и Harris, 2008)	www.cise.ufl.edu/~acamacho/publications/swipep.m

На этапе тестирования для сравнения производительности методов нахождения основного тона используются следующие стандартные метрики ошибок [39]:

а) ошибка принятия решения о вокализованности (Voicing Decision Error, VDE) определяет число фреймов, для которых принято неправильное решение вокализованный / невокализованный

$$VDE = \frac{N_{VU} + N_{UV}}{N} \times 100\% ,$$

где *Nvu* – ошибка принятия вокализованного фрейма как невокализованного, *Nuv* – ошибка принятия невокализованного фрейма как вокализованного, *N* – общее число фреймов;

б) процент грубых ошибок (Gross Pitch Error, GPE) определяет соотношение фреймов вокализованных участков, на которых относительная погрешность рассчитанного значения основного тона выше, чем определённый порог (обычно 20%) от эталонных значений [40].

3. Результаты и обсуждение

В этом разделе основное внимание уделяется оценке эффективности работы 9 алгоритмов по определению периода основного тона в условиях шума. Для искажения сигналов был применён дополнительный шум. Он наиболее известен как «фоновый». Отношение SNR включало –5, 0, 15 и 20 дБ для представления всего спектра зашумленного речевого сигнала. Было проведено сравнение качества работы временных, частотных и гибридных методов.

Таблицы 3 и 4 показывают результаты оценки речевых образцов для баз данных CSTR и Keele. В них приводятся VDE-значения для каждого из рассматриваемых методов при различных видах шума.

Таблица 3

Сравнение методов для базы данных Keele на основе VDE

Алгоритм	babble	car	white	Среднее значение
BANA	4.37	4.37	4.37	4.37
YIN	46.56	46.56	46.56	46.56
PEFAC	46.45	46.45	46.45	46.45
YAAPT	25.02	15.80	16.25	19.02
MBSC	29.96	19.15	21.22	23.44
SWIPE’	31.23	26.88	24.82	27.64
Cepst	46.56	46.56	46.56	46.56
AMDF	27.36	26.94	26.38	26.89
AUTOC	38.19	29.23	39.78	35.73

На рисунках 1–3 показаны GPE-значения каждой группы методов в зависимости от соотношения сигнал/шум. Как видно из рис. 1, среди временных методов наилучшим оказался YIN. При сравнении кепстрального метода (Cepst) и SWIPE’ первый показал наилучшие результаты (рис. 2). Анализируя гибридные методы, как видно из рис. 3, PEFAC лишь незначительно уступает BaNa в производительности при SNR 15 и 20 дБ.

Сравнение тестируемых методов BaNa, YIN и кепстрального подхода даёт надёжную оценку периода основного тона при разных типах шума во временной и частотной областях (рис. 4–6).

Алгоритм YIN показал хорошие результаты в условиях белого шума, машины и «булькания». Метод SWIPE' имеет низкую производительность (рис. 5) и, таким образом, не может хорошо работать в шумных условиях. Алгоритм PEFAC хорошо работает в зашумленных условиях. Его производительность немного ниже, чем у BaNa (рис. 6).

Для баз данных CSTR и Keele алгоритм BaNa обладает наименьшим GPE по сравнению с другими методами.

Таблица 4

Сравнение методов для базы данных CSTR на основе VDE

Алгоритм	babble	car	white	Среднее значение
BANA	0	0	0	0
YIN	47.87	47.87	47.87	47.87
PEFAC	46.64	46.64	46.64	46.64
YAAPT	26.98	15.93	13.51	18.81
MBSC	26.28	20.34	21.47	22.73
SWIPE'	38.29	28.91	27.92	31.71
Cepst	47.87	47.87	47.87	47.87
AMDF	30.44	36.14	26.66	31.08
AUTOC	39.38	32.54	40.25	37.39

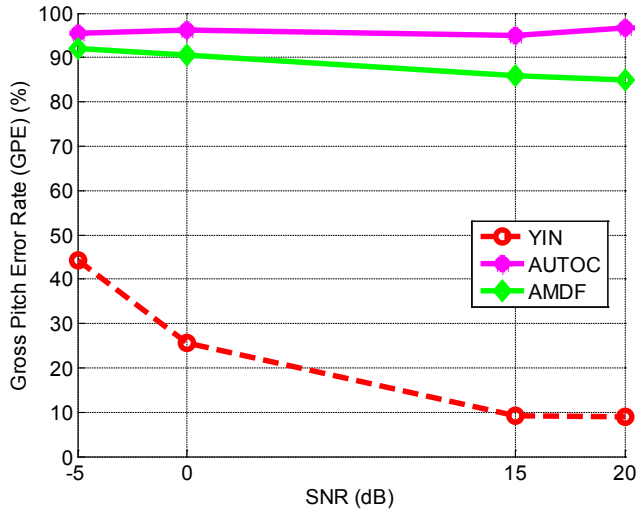

Рис. 1. GPE временных методов для CSTR

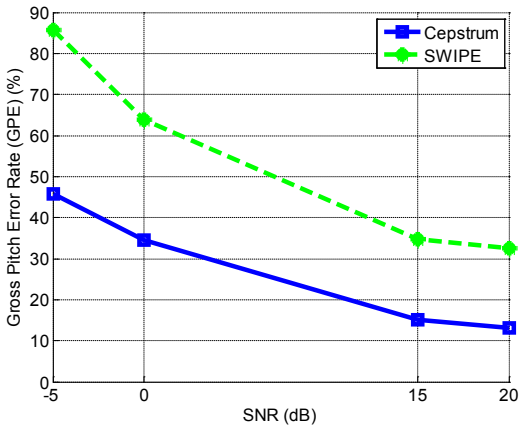

Рис. 2. GPE частотных методов для CSTR

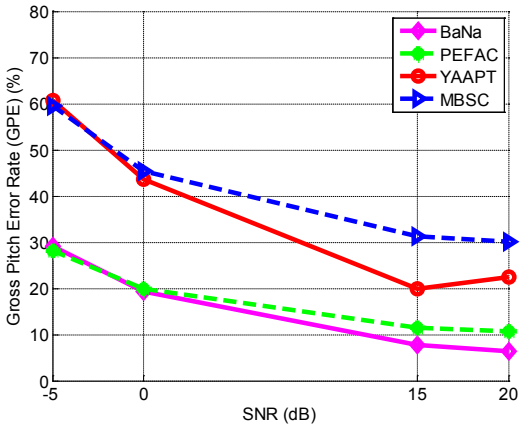

Рис. 3. GPE гибридных методов для CSTR

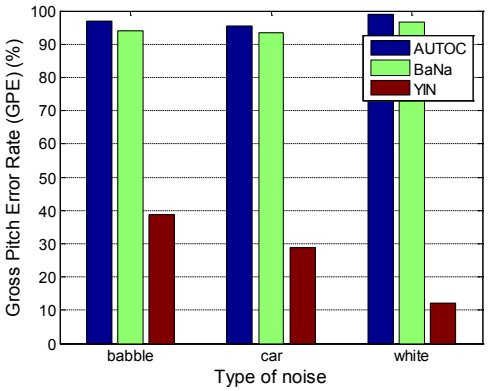

Рис. 4. GPE временных методов при различных типах шума (Keele)

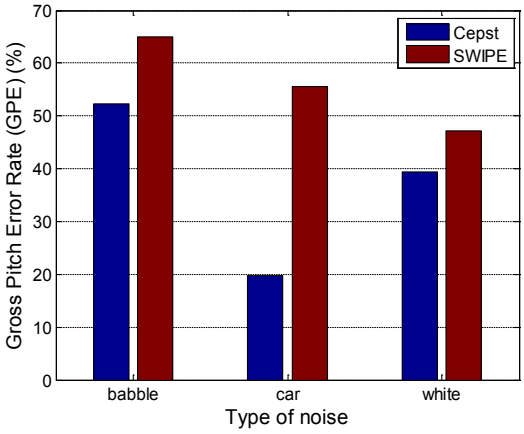

Рис. 5. GPE частотных методов при различных типах шума (Keele)

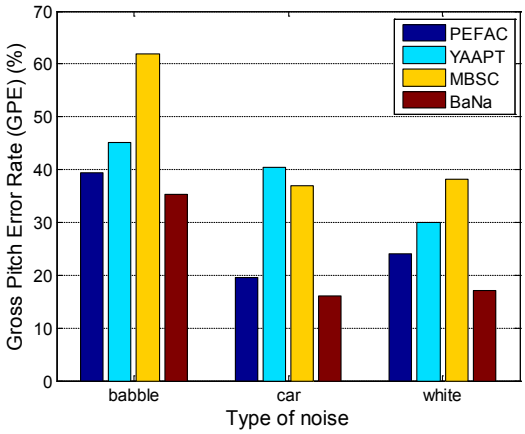

Рис. 6. GPE гибридных методов при различных типах шума (Keele)

Заключение

Оценивание периода (или частоты) основного тона является одной из наиболее важных задач в обработке речи. Различные алгоритмы определения частоты основного тона используются в вокодерах, системах распознавания и верификации дикторов, в устройствах, предназначенных для слабослышащих, в системах распознавания и синтеза речи и т.д. Известные на сегодняшний день эффективные алгоритмы определения основного тона речевых сигналов сильно подвержены влиянию шума.

Работа алгоритмов YIN, YAAPT, BaNa, PEFAC, AUTOC, AMDF, SWIPE', кепстрального метода и MBSC была оценена в условиях шумов с различными спектральными характеристиками. BaNa оказался наиболее робастным к шумам в диапазоне от –5 до 15 дБ. Алгоритмы YIN и PEFAC показали высокую производительность при различных типах шума. Для любой системы, которая опирается на точную оценку основного тона речевого сигнала, это является существенным результатом.

В последние годы в дополнение к классическим методам, упомянутым выше, имеет также место использование методов на основе обучения, например, нейронных

сетей, с целью выявления периода основного тона [19], однако производительность большинства этих методов связана со временем предобработки (например, обнаружение пиков).

Таким образом, разработка методов оценки периода основного тона, которые робастны к аддитивному шуму при различных SNR, остаётся важной областью исследований с большим количеством возможностей.

Мы понимаем, что наши исследования являются условными. Эта статья не может заменить тщательного изучения каждого отдельного метода. Однако мы надеемся, что результаты нашего аналитического обзора и проверки эффективности различных алгоритмов определения периода основного тона могут быть полезны для других исследователей.

Благодарности

Данная работа выполнена при финансовой поддержке Фонда Развития Науки при Президенте Азербайджанской Республики – Грант № EİF-RİTN-MQM-2/İKT-2-2013-7(13)-29/18/1.

Литература

1. *Hess W.J.* Pitch determination of speech signals. — Berlin: Springer-Verlag, 1983.
2. *Rabiner L., Cheng M.J., Rosenberg A.E., McGonegal C.A.* A comparative performance study of several pitch detection algorithms // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1976. — № 5. — P. 399–417.
3. *Veprek P., Scordilis M.S.* Analysis, enhancement and evaluation of five pitch determination techniques // Speech Communication. — 2002. — Vol. 37. — P. 249–270.
4. *De Cheveigne A., Kawahara H. Yin.* A fundamental frequency estimator for speech and music // J. Acoust. Soc. Am. — 2002. — Vol. 111. — № 4. — P. 1917–1930.
5. *Kasi K., Zahorian S.A.* Yet another algorithm for pitch tracking // Proc. of the Int-l Conf. on Acoustics, Speech and Signal Processing. — 2002. — P. 361–364.
6. *Tan L.N., Alwan A.* Multi-Band Summary Correlogram-based Pitch Detection for Noisy Speech // Speech Communication. — 2013. — Vol. 55. — № 78. — P. 841–856.
7. *Ba H., Yang N.* BaNa: a hybrid approach for noise resilient pitch detection // IEEE Statistical Signal Processing Workshop. — 2012. — P. 369–372.
8. *Carey M.J., Parris E.S., Lloyd-Thomas H. and Bennet S.* Robust prosodic features for speaker identification // Proc. of ICSLP. — 1996. — P. 1800–1803.
9. *Xie Y.L., Zhou X., Yao Z.Q., Chen J.X. and Liu M.H.* University of science and technology of China SSIP laboratory NIST SRE 2005 system // NIST Speaker Recognition Evaluation (SRE) Workshop. — 2005.
10. *Adami A.G., Mihaescu R., Reynolds D.A. and Godfrey J.J.* Modeling prosodic dynamics for speaker recognition // Proc. of ICASSP. — 2003. — P. 788–791.
11. *Ross M.J., Shaffer H.L., Cohen A., Freudberg R. and Manley H.J.* Average magnitude difference function pitch extractor // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1974. — Vol. ASSP-22. — № 5. — P. 353–362.
12. *Sondhi M.M.* New methods of pitch extraction // IEEE Trans. Audio Electroacoust. — 1968. — Vol. AU-16. — P. 262–266.

13. *Gold B., Rabiner L.R.* Parallel processing techniques for estimating pitch periods of speech in the time domain // J. Acoust. Soc. Amer. — 1969. — Vol. 46. — P. 442–448.
14. *Miller N.J.* Pitch detection by data reduction // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1975. — Vol. ASSP-23. — P. 72–79.
15. *Rabiner L.R.* On the use of autocorrelation analysis for pitch detection // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1977. — Vol. ASSP-25. — № 1. — P. 24–33.
16. *Un C.K., Yang S.C.* A pitch extraction algorithm based on LPC inverse filtering and AMDF // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1977. — Vol. ASSP-25. — P. 526–572.
17. *Moorer J.A.* The optimum comb method of pitch period analysis of continuous digitized speech // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1974. — Vol. 22. — № 3. — P. 330–338.
18. *Medan Y., Yair E. and Chazan D.* Super resolution pitch determination of speech signals // IEEE Trans. on Signal Proc. — 1991. — Vol. 39. — № 1. — P. 40–48.
19. *Barnard E., Cole A.R., Vea M. and Alleva F.* Pitch detection with a neural-net classifier // IEEE Trans. on Signal Proc. — 1991. — Vol. 39. — № 2. — P. 298–307.
20. *Noll A.M.* Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and maximum likelihood estimate // Proc. of the Symposium on Computer Proc. in Communications. — 1969. — P. 779–797.
21. *Seneff S.* Real-time harmonic pitch detector // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1978. — Vol. ASSP-26. — P. 358–365.
22. *Sreenivas T.V., Rao P.V.S.* Pitch extraction from corrupted harmonics of the power spectrum // J. Acoust. Soc. Amer. — 1979. — Vol. 65. — P. 223–228.
23. *Lahat M., Niederjohn R.J. and Krubsack D.A.* A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1987. — Vol. ASSP-35. — P. 741–750.
24. *Serra X.* Musical Sound Modeling with Sinusoids plus Noise // Musical Signal Processing. — 1997.
25. *Noll A.M.* Cepstrum pitch determination // Journal of the Acoustical Society of America. — 1967. — Vol. 41 — № 2. — P. 293–309.
26. *Markel J.* The SIFT algorithm for fundamental frequency estimation // IEEE Trans. Audio Electroacoust. — 1972. — Vol. AU-20. — P. 367–377.
27. *Schroeder M.R.* Period histogram and product spectrum: new methods for fundamental-frequency measurement // Journal of the Acoustical Society of America. — 1968. — P. 829–834.
28. *Atal B.S., Rabiner L.R.* A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition // IEEE Trans. on Acoustics, Speech and Signal Processing. — 1976. — Vol. ASSP-24. — P. 201–212.
29. *Camacho A.* SWIPE: a sawtooth waveform inspired pitch estimator for speech and music: PhD thesis. — Univ. Florida. — 2007.
30. *Gonzalez S., Brookes M.* A pitch estimation filter robust to high levels of noise (PEFAC) // Proc. of European Signal Processing Conf. (EUSIPCO). — 2011. — P. 451–455.
31. *Rabiner L., Schafer R.* Digital Processing of Speech Signals. — NJ: Prentice-Hall, 1978.
32. *Zahorian S.A., Hu H.* A spectral/temporal method for robust fundamental frequency tracking // Journal of the Acoust. Soc. of America. — 2008. — Vol. 123. — № 6. — P. 4559–4571.
33. *Talkin D.* Robust algorithm for pitch tracking // Speech Coding and Synthesis. — 1995. — P. 497–518.
34. *Loughlin P., Tacer B.* On the amplitude- and frequency-modulation decomposition of signals // Journal of the Acoust. Soc. of America. — 1996. — Vol. 100. — № 3. — P. 1594–1601.
35. *Bagshaw P.C., Hiller S.M. and Jack M.A.* Enhanced pitch tracking and the processing of F0 contours for computer and intonation teaching // Proc. European Conf. on Speech Comm. (Eurospeech). — 1993. — P. 1003–1006.
36. *Bagshaw P.C.* Automatic prosodic analysis for computer aided pronunciation teaching, doctoral dissertation. — Edinburgh: University of Edinburgh, 1994.

37. *Plante F., Meyer G. and Ainsworth W.A.* A pitch extraction reference database // Proc. European Conf. on Speech Comm. (Eurospeech). — 1995. — P. 837–840.
38. *Chu W., Alwan A.* Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend // Proc. of ICASSP. — 2009. — P. 3969–3972.
39. *Varga A., Steeneken H.J.* Assessment for automatic speech recognition: II. Noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems // Speech Communication. — 1993. — Vol. 12. — № 3. — P. 247–251.
40. *Drugman T., Alwan A.* Joint robust voicing detection and pitch estimation based on residual harmonics // Proc. of Interspeech. — 2011. — P. 1973–1976.

Сведения об авторах:

Имамвердиев Ядигар Насиб оглы,
кандидат технических наук, руководитель отдела информационной безопасности Института Информационных Технологий Национальной Академии Наук Азербайджана. Автор более 50 научных публикаций. Область научных интересов: цифровая обработка речевых сигналов, идентификация диктора по голосу, информационная безопасность, прикладная криптография. E-mail: yadigar@lan.ab.az

Сухостат Людмила Валентиновна,
научный сотрудник Института Информационных Технологий Национальной Академии Наук Азербайджана. Область научных интересов: идентификация диктора по голосу, цифровая обработка речевых сигналов. E-mail: lsuhostat@hotmail.com