3.15. Наиболее часто импульс в голосовой щели формируется в виде

$$g(n) = \begin{cases} n\,a^n, & n \geqslant 0; \\ 0, & n < 0. \end{cases}$$

а) Найти z-преобразование $g(n)$.
б) Изобразить преобразование Фурье $G(e^{i\omega})$ как функцию ω.
в) Показать, как можно выбрать a, чтобы выполнялось соотношение: $20\log_{10}|G(e^{i0})| - 20\log_{10}|G(e^{i\pi})| = 60$ дБ.

4

Методы обработки речевых сигналов во временной области

4.0. Введение

В гл. 2 и 3 были изложены наиболее эффективные методы цифровой обработки, а также основные свойства речевых сигналов. Рассмотрим теперь применение методов цифровой обработки речевых сигналов. Основной целью обработки речевых сигналов является получение наиболее удобного и компактного представления содержащейся в них информации. Точность представления определяется той информацией, которую необходимо сохранить или выделить. Например, цифровая обработка может применяться для выяснения, является ли данное колебание речевым сигналом. Сходная, но несколько более сложная задача состоит в том, чтобы классифицировать колебания на вокализованную речь, невокализованную речь и паузу (шум). В этих случаях целесообразно использовать такие характеристики сигнала, в которых признаки классификации представлены с максимальной точностью. В других задачах (например, при цифровой передаче) может потребоваться точное восстановление речевого сигнала по его сокращенному представлению. В этой главе рассматриваются методы обработки речевого колебания *во временной области*. В гл. 6—8, напротив, излагаются методы обработки спектрального представления сигнала в частотной области[1].

Примерами временных характеристик речевого сигнала могут служить среднее число переходов через нулевой уровень, энергия сигнала и его корреляционная функция. Эти характеристики часто используются, так как их измерение не требует сложных устройств, а располагая их значениями, можно получить представление о некоторых особенностях сигнала.

В начале главы обсуждаются принципы обработки во временной области; далее приводится несколько примеров такой обработки. В заключение рассматриваются такие задачи, как разделение сигнала на вокализованные и невокализованные сегменты, выделение основного тона, измерение функции кратковременной мощности. Существует множество других задач, которые можно было бы здесь привести. Однако наша цель состоит не в составлении исчерпывающего обзора задач, а в иллюстрации эффективности методов обработки во временной области.

4.1. Текущая обработка речевых сигналов

На рис. 4.1 показана последовательность отсчетов (с частотой 8000 отсч./с), представляющая типичный речевой сигнал. Из рисунка видно, что свойства речевого сигнала изменяются во времени, например характер возбуждения на вокализованных и невокализованных участках, пиковая амплитуда, период основного тона на вокализованных сегментах. Тот факт, что эти изменения видны на осциллограмме речевого сигнала, означает, что методы его обработки во временной области должны обеспечивать хорошее описание таких текущих характеристик сигнала, как мощность, характер возбуждения, основной тон, и, возможно, даже таких параметров голосового тракта, как форматные частоты.

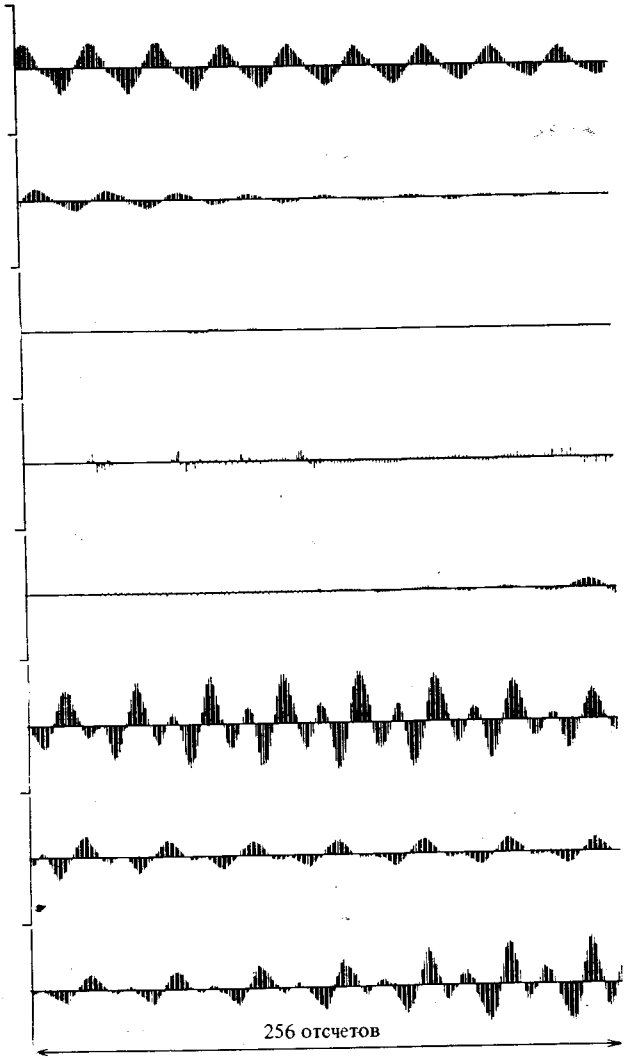

Рис. 4.1. Отсчеты типичного речевого сигнала (частота дискретизации 8 кГц)

[1] Во всех случаях предполагается, что сигнал ограничен по частоте и дискретизирован, по крайней мере, с частотой Найквиста. Предполагается также, что отсчеты квантованы с пренебрежимо малой ошибкой (см. гл. 5, где обсуждаются эффекты квантования).

В основе большинства методов обработки речи лежит предположение о том, что свойства речевого сигнала с течением времени медленно изменяются. Это предположение приводит к методам кратковременного анализа, в которых сегменты речевого сигнала выделяются и обрабатываются так, как если бы они были короткими участками отдельных звуков с отличающимися свойствами. Процедура повторяется так часто, как это требуется. *Сегменты*, которые иногда называют *интервалами* (кадрами) *анализа*, обычно пересекаются. Результатом обработки на каждом интервале является число или совокупность чисел. Следовательно, подобная обработка приводит к новой, зависящей от времени последовательности, которая может служить характеристикой речевого сигнала.

Большинство методов кратковременного анализа, излагаемых в главе, в том числе и кратковременный Фурье-анализ (см. гл. 6), могут быть описаны выражением

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]\, w(n-m). \tag{4.1}$$

Речевой сигнал (возможно, после ограничения частотного диапазона в линейном фильтре) подвергается преобразованию $T[\cdot]$, линейному или нелинейному, которое может зависеть от некоторого управляющего параметра или их совокупности. Результирующая последовательность умножается затем на последовательность значений временного окна (весовой функции), расположенную во времени в соответствии с индексом n. Результаты затем суммируются по всем ненулевым значениям. Обычно, хотя и не всегда, последовательность значений временного окна имеет конечную протяженность. Значение Q_n представляет собой, таким образом, «взвешенное» среднее значение последовательности $T[x(m)]$.

Простым примером, иллюстрирующим изложенное, может служить измерение кратковременной энергии сигнала. Полная энергия сигнала в дискретном времени определяется как

$$E = \sum_{m=-\infty}^{\infty} x^2(m). \tag{4.2}$$

Вычисление этой величины не имеет особого смысла при обработке речевых сигналов, поскольку она не содержит информации о свойствах сигнала, изменяющихся во времени. Кратковременная энергия определяется выражением

$$E_n = \sum_{m=n-N+1}^{n} x^2(m). \tag{4.3}$$

Таким образом, кратковременная энергия в момент n есть просто сумма квадратов N отсчетов от $n-N+1$ до n. Из (4.1) видно, что в (4.3) $T[\cdot]$ есть просто операция возведения в квадрат, а

$$w(n) = \begin{cases} 1, & 0 \leqslant n \leqslant N-1, \\ 0, & \textit{в противном случае}. \end{cases} \tag{4.4}$$

Вычисление кратковременной энергии иллюстрирует рис. 4.2. Отметим, что окно «скользит» вдоль последовательности квадратов значений сигнала, в общем случае вдоль последовательности $T[x(m)]$, ограничивая длительность интервала, используемого в вычислениях.

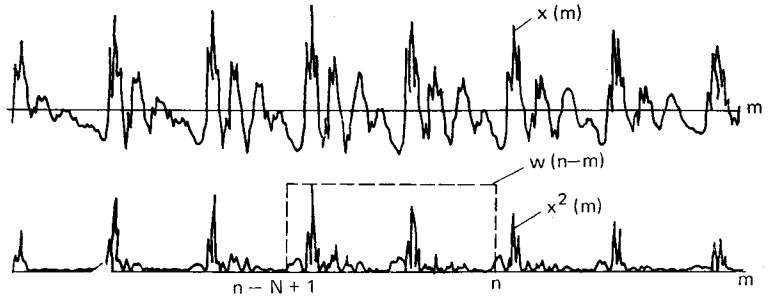

Рис. 4.2. Иллюстрация вычисления функции кратковременной энергии

Более подробно вычисление кратковременной энергии будет обсуждаться в следующем параграфе. Здесь же уместно отметить одно важное свойство преобразования (4.1). Выражение (4.1) описывает дискретную свертку окна $w(n)$ с последовательностью $T[x(n)]$. Таким образом, последовательность Q_n[1] может быть интерпретирована как выходной сигнал линейной инвариантой к сдвигу системы с импульсной характеристикой $h(n) = w(n)$. Такая система изображена на рис. 4.3. Важность этого подхода станет яснее в процессе изучения материала этой главы и гл. 6.

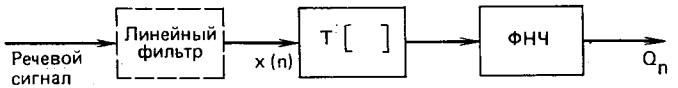

Рис. 4.3. Общее представление принципа кратковременного анализа

4.2. Кратковременная энергия и кратковременное среднее значение сигнала [2]

Как отмечалось выше, амплитуда речевого сигнала существенно изменяется во времени. В частности, амплитуда невокализованных сегментов речевого сигнала значительно меньше амплитуды вокализованных сегментов. Подобные изменения амплитуды хорошо описываются с помощью функции кратковременной энергии

[1] Нижний индекс используется для кратковременных характеристик. Сейчас это не должно вызывать больших затруднений, а в дальнейшем позволит получить простые и ясные обозначения.

[2] Далее везде, где это не вызывает недоразумений, слово «кратковременная» будет опускаться. (*Прим. ред.*)

сигнала. В общем случае определить функцию энергии можно как

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)\, w(n-m)]^2. \qquad (4.5)$$

Это выражение может быть переписано в виде

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)\, h(n-m), \qquad (4.6)$$

где

$$h(n) = w^2(n). \qquad (4.7)$$

Уравнение (4.6) можно интерпретировать в соответствии с рис. 4.4*а*. Сигнал $x^2(n)$ в этом случае фильтруется с помощью линейной системы с импульсной характеристикой $h(n)$.

Выбор импульсной характеристики $h(n)$ или окна составляет основу описания сигнала с помощью функции энергии. Чтобы понять, как влияет выбор окна на функцию кратковременной энергии сигнала, предположим, что $h(n)$ в (4.6) является достаточно длительной и имеет постоянную амплитуду; значение E_n будет при этом изменяться во времени незначительно. Такое окно эквивалентно фильтру нижних частот с узкой полосой пропускания. Полоса фильтра нижних частот не должна быть столь узкой, чтобы выходной сигнал оказался постоянным, иначе говоря, полосу следует выбрать так, чтобы функция энергии отражала изменения амплитуды речевого сигнала. Описанная ситуация выражает противоречие, которое нередко возникает при изучении кратковременных характеристик речевых сигналов. Суть его состоит в том, что для описания быстрых изменений амплитуды желательно иметь узкое окно (короткую импульсную характеристику), однако слишком малая ширина окна может привести к недостаточному усреднению и, следовательно, к недостаточному сглаживанию функции энергии.

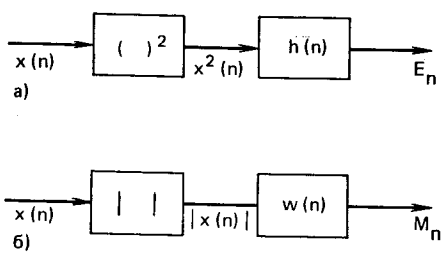

Рис. 4.4. Структурная схема для функции: *а*) кратковременной энергии; *б*) кратковременного среднего значения

Влияние вида окна на вычисление изменяющейся во времени энергии сигнала можно проиллюстрировать на примере использования двух наиболее распространенных окон: прямоугольного

$$h(n) = \begin{cases} 1, & 0 \leqslant n \leqslant N-1, \\ 0, & \textit{в противном случае.} \end{cases} \qquad (4.8)$$

и окна Хемминга

$$h(n) = \begin{cases} 0{,}54 - 0{,}46 \cos(2\pi n/(N-1)), & 0 \leqslant n \leqslant N-1, \\ 0, & \textit{в противном случае.} \end{cases} \qquad (4.9)$$

Прямоугольное окно, как это видно из (4.3), соответствует случаю, когда всем отсчетам на интервале от $(n-N+1)$ до n приписывается одинаковый вес. Частотная характеристика прямоугольного окна (с импульсной характеристикой (4.8)), как легко показать (см. задачу 4.1), равна

$$H(e^{i\Omega T}) = \frac{\sin(\Omega N T/2)}{\sin(\Omega T/2)}\, e^{-i\Omega T (N-1)/2}. \qquad (4.10)$$

Для окна с шириной 51 отсчет ($N=51$) логарифм амплитудно-частотной характеристики представлен на рис. 4.5*а*. Отметим, что

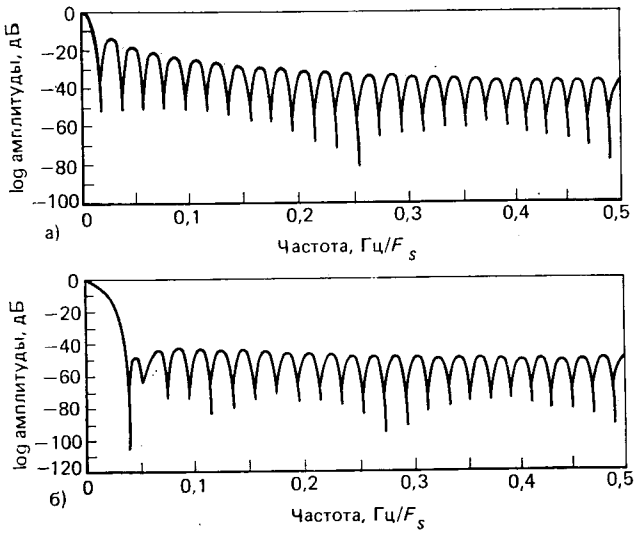

Рис. 4.5. Преобразование Фурье для:
а) прямоугольного окна; *б*) окна Хемминга

первое нулевое значение амплитудно-частотной характеристики (4.10) соответствует частоте

$$F = F_s/N, \qquad (4.11)$$

где $F_s = 1/T$ — частота дискретизации. Это номинальная частота среза фильтра нижних частот, соответствующего прямоугольному окну. Амплитудно-частотная характеристика окна Хемминга при $N=51$ показана на рис. 4.5*б*. Полоса пропускания фильтра с окном Хемминга при одинаковой ширине примерно вдвое превосходит полосу фильтра с прямоугольным окном. Очевидно также, что окно Хемминга обеспечивает большее затухание вне полосы пропускания по сравнению с прямоугольным окном. Затухание, вносимое вне полосы, несущественно зависит от ширины каждого из

окон. Это означает, что увеличение ширины приведет просто к сужению полосы[1]. Если N мало (порядка периода основного тона или менее), то E_n будет изменяться очень быстро, в соответствии с тонкой структурой речевого колебания. Если N велико (порядка нескольких периодов основного тона), то E_n будет изменяться медленно и не будет адекватно описывать изменяющиеся особенности речевого сигнала. Это, к сожалению, означает, что не существует единственного значения N, которое в полной мере удовлетворяло бы перечисленным требованиям, так как период основного тона изменяется от 10 отсчетов (при частоте дискретизации 10 кГц) для высоких женских и детских голосов до 250 отсчетов для очень низких мужских голосов. На практике N выбирают равным 100—200 отсчетов при частоте дискретизации 10 кГц (т. е. длительность порядка 10—20 мс).

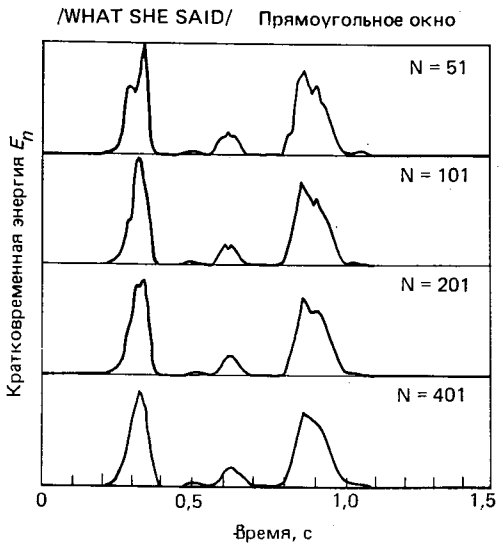

Рис. 4.6. Функции кратковременной энергии для прямоугольных окон различной длительности

На рис. 4.6 и 4.7 показано влияние длительности окна (для прямоугольного окна и окна Хемминга) на энергию, вычисленную по фразе «what she said», произнесенной женским голосом. Легко видеть, что при увеличении N траектория энергии становится более гладкой при использовании любого временного окна.

Основное назначение E_n состоит в том, что эта величина позволяет отличить вокализованные речевые сегменты от невокализованных. Как видно из рис. 4.6 и 4.7, значения E_n для невокализованных сегментов значительно меньше, чем для вокализованных. Функция кратковременной энергии может быть использована для приближенного определения момента перехода от вокализованного сегмента к невокализованному и наоборот, а в случае высококачественного речевого сигнала (с большим отношением сигнала к шуму) функцию энергии можно использовать и для отделения речи от пауз.

Одним из недостатков функции кратковременной энергии, определяемой выражением (4.6), является ее чувствительность к большим уровням сигнала (поскольку в (4.6) каждый отсчет возводится в квадрат). Вследствие этого значительно искажается соотношение между значениями последовательности $x(n)$. Простым способом устранения этого недостатка является переход к определению функции среднего значения в виде

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|\, w(n-m), \qquad (4.12)$$

где вместо суммы квадратов вычисляется взвешенная сумма абсолютных значений. На рис. 4.4*б* показано, как соотношение (4.12) может быть представлено посредством линейной фильтрации последовательности $|x(n)|$. Исключение операции возведения в квадрат упрощает арифметические вычисления.

На рис. 4.8 и 4.9 показаны траектории среднего значения, соответствующие рис. 4.6 и 4.7. Различия заметны практически лишь на невокализованных сегментах. При вычислении среднего значения по (4.12) динамический диапазон (отношение максимального значения к минимальному) определяется примерно как квадратный корень из динамического диапазона при обычном вычислении энергии. Таким образом, в данном случае различия в уровнях между вокализованной и невокализованной речью выражены не столь ярко, как при использовании функций энергии.

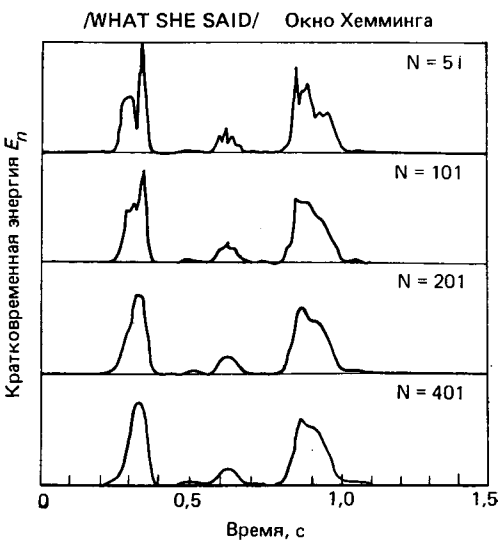

Рис. 4.7. Функции кратковременной энергии для окон Хемминга различной длительности

Поскольку полоса частот при определении как функции энергии, так и среднего значения приближенно совпадает с полосой пропускания используемого фильтра нижних частот, то нет необходимости дискретизировать эти функции столь же часто, как исходный речевой сигнал. Например, для окна длительностью 20 мс достаточна частота дискретизации около 100 Гц. Это означает, что значительная часть информации теряется при использовании подобных кратковременных представлений. Очевидно также, что информация, относящаяся к динамике амплитуд речевого сигнала, сохраняется в весьма удобной форме.

Завершая рассмотрение свойств функций энергии и среднего значения, следует отметить, что используемое окно не обязательно должно быть прямоугольным, или окном Хемминга, или какой-ли-

[1] Здесь нет необходимости в подробном изложении свойств временных окон. Оно содержится в гл. 6.

бо функцией, обычно применяемой в качестве окна при спектральном анализе и при цифровой фильтрации сигналов. Необходимо лишь, чтобы применяемый фильтр обеспечивал адекватное сглаживание. Таким образом, можно использовать фильтр нижних частот, синтезированный любым стандартным способом [1, 2]. Кроме того, фильтр может быть как КИХ-, так и БИХ-фильтром. Импульсная характеристика должна быть всегда положительной, поскольку это гарантирует, что энергия и среднее значение окажутся больше нуля. Фильтры с КИХ (с прямоугольной импульсной характеристикой и характеристикой окна Хемминга) обладают тем преимуществом, что сигнал на их выходе может быть рассчитан сразу для пониженной частоты дискретизации путем сдвига окна более чем на один отсчет входного сигнала. Например, если речевой сигнал дискретизирован с частотой 10 кГц и применяемое окно имеет длительность 20 мс (200 отсчетов), то функция энергии может быть определена при частоте дискретизации около 100 Гц, т. е. 1 раз на каждые 100 отсчетов входного сигнала.

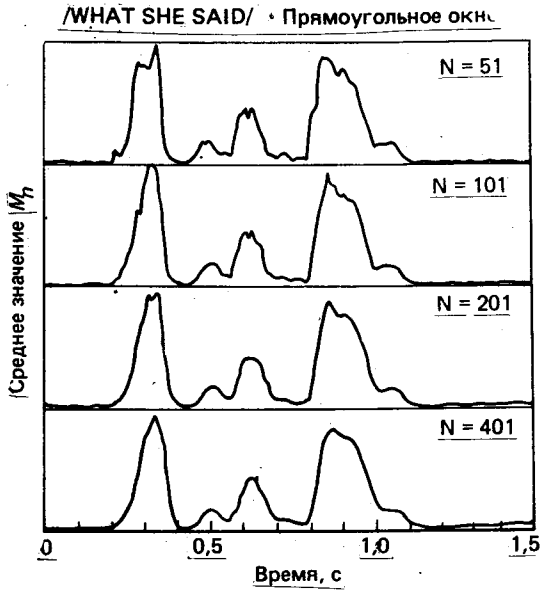

Рис. 4.8. Функции среднего значения для прямоугольных окон различной длительности

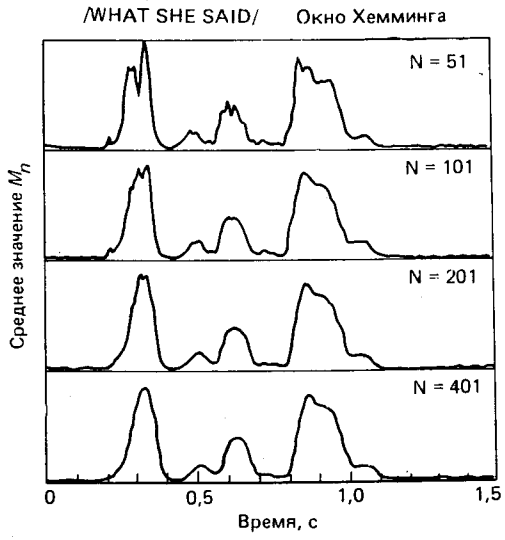

Рис. 4.9. Функции среднего значения для окна Хемминга различной длительности

Совершенно не обязательно использовать окна конечной длительности. Хотя на первый взгляд это кажется необычным, для фильтрации можно использовать фильтр с бесконечно протяженной импульсной характеристикой, если ее z-преобразование представляет собой

рациональную функцию. Примером может служить окно следующего вида:

$$h(n)=\begin{cases} a^n, & n \geqslant 0, \\ 0, & n<0. \end{cases} \tag{4.13}$$

Значение $0<a<1$ позволяет выбирать эффективную длительность окна. Соответствующее z-преобразование окна имеет вид

$$H(z)=1/(1-a\,z^{-1}), \quad |z|>|a|, \tag{4.14}$$

откуда легко видеть, что передаточная функция $H(e^{-i\Omega T})$, как и требуется, сосредоточена в области нижних частот. Подобный фильтр описывается простейшим разностным уравнением, т. е. функция энергии должна удовлетворяться рекуррентному соотношению:

$$E_n=a\,E_{n-1}+x^2(n), \tag{4.15}$$

а среднее значение — рекуррентному соотношению

$$M_n=a\,M_{n-1}+|x(n)|. \tag{4.16}$$

Использование (4.15) и (4.16) приводит к тому, что функции энергии и среднего значения надо вычислять для каждого отсчета входного сигнала, даже если требуется значительно меньшая частота дискретизации. Однако иногда полученные рекуррентные уравнения весьма полезны, например, при кодировании формы речевого сигнала, как это описано в гл. 5. Но если частота дискретизации достаточно снижена, то нерекуррентные методы требуют меньшего объема вычислений (см. задачу 4.4). Другой вопрос, который представляет интерес, относится к определению задержки, связанной с обработкой в фильтре нижних частот. Временные окна (4.8) и (4.9) определены таким образом, что они соответствуют фильтрам, в которых не нарушается принцип причинности (они реализуемы). В силу симметрии импульсной характеристики фильтры имеют абсолютно линейную фазо-частотную характеристику и вносят задержку на $(N-1)/2$ отсчетов. Поэтому исходная функция энергии может быть уточнена с учетом вносимой задержки. При рекуррентной обработке фазо-частотная характеристика нелинейна, поэтому задержку нельзя скомпенсировать полностью.

4.3. Кратковременная функция среднего числа переходов через нуль

При обработке сигналов в дискретном времени считают, что если два последовательных отсчета имеют различные знаки, то произошел переход через нуль. Частота появления нулей в сигнале может служить простейшей характеристикой его спектральных свойств. Это наиболее справедливо для узкополосных сигналов. Например, синусоидальный сигнал с частотой F_0, подвергнутый дискретизации с частотой F_s, имеет F_s/F_0 отсчетов за период. Каждый

период содержит два перехода через нуль, таким образом, среднее число нулевых переходов за большой интервал времени.

$$z = 2F_0/F_s. \tag{4.17}$$

Среднее число нулевых переходов можно принять в качестве подходящей оценки частоты синусоидального колебания.

Речевой сигнал является широкополосным и, следовательно, интерпретация среднего числа переходов через нуль менее очевидна. Однако можно получить грубые оценки спектральных свойств сигнала, основанные на использовании функции среднего числа переходов через нуль для речевого сигнала; рассмотрим способ вычисления этой величины. Определим среднее число переходов через нуль:

$$Z_n = \sum_{m=-\infty}^{\infty} |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]|\, w(n-m), \tag{4.18}$$

где

$$\operatorname{sgn}[x(n)] = \begin{cases} 1, & x(n) \geqslant 0, \\ -1, & x(n) < 0 \end{cases} \tag{4.19}$$

и

$$w(n) = \begin{cases} 1/2N, & 0 \leqslant n \leqslant N-1, \\ 0, & \textit{в противном случае.} \end{cases} \tag{4.20}$$

Операции, входящие в (4.18), представлены в виде структурной схемы на рис. 4.10. Такое представление показывает, что функция среднего числа переходов через нуль имеет те же общие свойства,

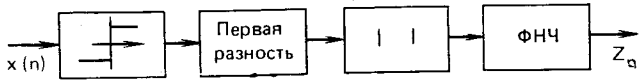

Рис. 4.10. Структурная схема вычисления кратковременной функции среднего нулевых пересечений

что и функции энергии и среднего значения. Может показаться, однако, что вычисления по (4.18) и в соответствии с рис. 4.10 являются более сложными, чем это есть на самом деле. Все, что в действительности требуется, это проверить пары отсчетов с целью определения нулевых пересечений, а затем вычислить среднее по всем N последовательным отсчетам (деление на N, конечно, не обязательно). Как и ранее, может быть вычислено взвешенное среднее и при использовании симметричных окон конечной длительности задержка может быть скомпенсирована точно. Могут быть получены и рекуррентные уравнения, сходные с (4.15) и (4.16) (см. задачу 4.5).

Рассмотрим теперь применение функции среднего числа переходов через нуль для обработки речевых сигналов. Модель речеобразования предполагает, что энергия вокализованных сегментов речевого сигнала концентрируется на частотах ниже 3 кГц, что обусловлено убывающим спектром сигнала возбуждения, тогда как для невокализованных сегментов бóльшая часть энергии лежит в области высоких частот. Поскольку высокие частоты приводят к большому числу переходов через нуль, а низкие — к малому, то существует жесткая связь между числом нулевых пересечений и распределением энергии по частотам. Разумно предположить, что большому числу нулевых пересечений соответствуют невокализованные сегменты, а малому числу — вокализованные сегменты речи. Это, однако, очень расплывчатое утверждение, поскольку мы не определили, что означает «много» или «мало», и количественно определить эти понятия в действительности трудно. На рис. 4.11 представлены гистограммы среднего числа нулевых пересечений (усреднение за 10 мс) как для вокализованных, так и для невокализованных сегментов речевого сигнала. Отметим, что гауссовская кривая хорошо согласуется с приведенными гистограммами. Среднее число пересечений составляет 49 для вокализованных и 14 для невокализованных сегментов длительностью 10 мс. Поскольку оба распределения перекрываются, нельзя вынести однозначное решение о принадлежности сегмента к вокализованным или невокализованным отрезкам только по среднему числу переходов через нуль. Тем не менее, подобное представление весьма полезно при осуществлении такой классификации.

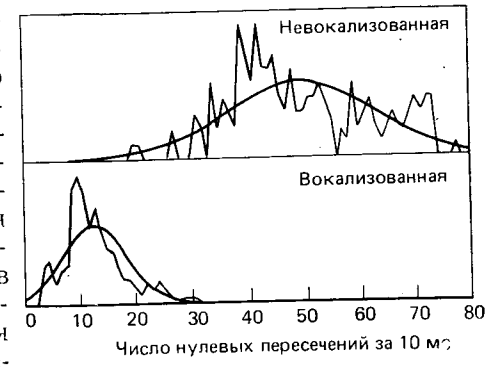

Рис. 4.11. Распределение нулевых пересечений для вокализованной и невокализованной речи

Некоторые результаты измерения среднего числа переходов через нуль представлены на рис. 4.12. В приведенных примерах длительность окна составляла 15 мс (150 отсчетов при частоте дискретизации 10 кГц). Результат вычислялся 100 раз в секунду (окно перемещалось с шагом в 100 отсчетов). Отметим, что так же, как и в случае функций энергии и среднего, функцию среднего числа переходов через нуль можно дискретизировать с очень низкой частотой. Хотя среднее число переходов через нуль изменяется значительно, вокализованные и невокализованные сегменты на рис. 4.12 просматриваются очень четко.

При использовании описания сигнала средним числом переходов через нуль следует иметь в виду ряд практических соображений. Хотя в основу алгоритма вычисления нулевых переходов положено сравнение знаков соседних отсчетов, тем не менее при дискретизации сигнала следует предпринимать специальные меры.

Очевидно, что число нулевых переходов зависит от уровня шума при аналого-цифровом преобразовании, интенсивности фона переменного тока и других шумов, которые могут присутствовать в цифровой системе. Таким образом, с целью уменьшения влияния

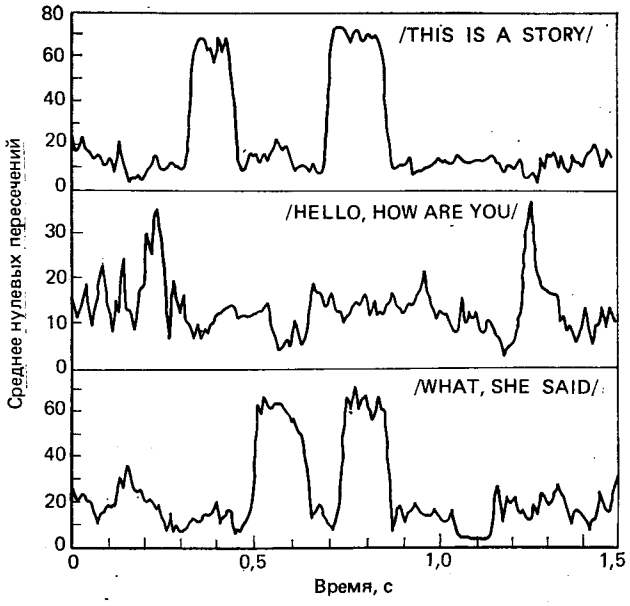

Рис. 4.12. Функция среднего нулевых пересечений для трех различных фраз

этих факторов следует проявлять особую осторожность при аналоговой обработке сигнала, предшествующей дискретизации. Например, часто оказывается более целесообразным использовать полосовой фильтр вместо фильтра нижних частот для уменьшения эффекта наложения при аналого-цифровом преобразовании и устранения фона переменного тока из сигнала. Кроме того, при измерении числа переходов через нуль следует учитывать соотношение между периодом дискретизации и интервалом усреднения N. Период дискретизации определяет точность выделения нулевых пересечений по времени (и по частоте), т. е. чтобы добиться высокой точности, нужна большая частота дискретизации. Вместе с тем от каждого отсчета требуется информация объемом лишь 1 бит (информация только о знаке сигнала).

Вследствие практической ограниченности этого метода было предложено множество сходных представлений сигнала. В каждом из них содержатся дополнительные особенности, направленные на снижение чувствительности оценок к шуму, но все они имеют и свои собственные ограничения. Наиболее заметным среди них является представление сигнала, исследованное Бейкером

[3]. Представление основано на интервалах времени между положительными переходами через нуль (снизу вверх). Бейкер применил это описание для фонетической классификации звуков речи [3].

Другое применение анализа переходов через нуль состоит в получении промежуточного представления речевого сигнала в частотной области. Метод включает фильтрацию речевого сигнала в нескольких смежных частотных диапазонах. Затем по сигналам на выходе фильтров измеряют кратковременную энергию и среднее число переходов через нуль. Совместное использование этих характеристик дает грубое описание спектральных свойств сигнала. Этот подход, предложенный Рэдди и исследованный Вайсенсом [4] и Эрманом [5], положен в основу систем распознавания речи.

4.4. Разделение речи и пауз на основе функций кратковременной энергии и среднего числа переходов через нуль

Задача определения моментов начала и окончания фразы при наличии шума является одной из важных задач в области обработки речи. В частности, при автоматическом распознавании слов важно точно определить моменты начала и окончания слова. Методы обнаружения моментов начала и окончания фразы можно использовать для уменьшения числа арифметических операций, если обрабатывать только те сегменты, в которых имеется речевой сигнал, например, в системах, работающих не в реальном масштабе времени.

Проблема отделения речи от окружающего шума очень сложна, за исключением случаев очень большого отношения сигнал/шум, т. е. в случае высококачественных записей, выполненных в заглушенной камере или звуконепроницаемой комнате. В этих случаях энергия даже наиболее слабых звуков речи (фрикативных согласных) превышает энергию шума и, таким образом, достаточно лишь измерить энергию сигнала. Но подобные условия записи, как правило, не встречаются в реальных ситуациях.

Рассматриваемый ниже алгоритм основан на измерении двух простых характеристик — энергии и числа переходов через нуль. На примере простых ситуаций иллюстрируются трудности, возникающие при обнаружении моментов начала и окончания фразы. На рис. 4.13 представлено колебание (начало слова eight), в котором шум, как это видно из рисунка, легко отделяется от речевого сигнала. В этом случае значительное различие энергий сигнала и шума достаточно для определения момента начала фразы. На рис. 4.14 изображен другой случай (начало слова six), в котором также легко определить начало речевого сигнала. Здесь спектральный состав речи существенно отличается от спектрального состава окружающего шума, что видно по резкому увеличению числа переходов через нуль в сигнале. Следует отметить, что в данном слу-

чае энергия речевого сигнала в начале фразы сравнима с энергией шума.

На рис. 4.15 представлен случай, в котором чрезвычайно трудно выделить начало речевого сигнала. На данном рисунке изображено колебание в начале слова (four). Поскольку это слово начинается со слабого (с малой энергией) фрикативного согласного, очень трудно определить момент его начала. Хотя точка *B* могла бы служить началом слова, в действительности оно начинается в точке *A*. В общем случае очень трудно определить начало или концы слов, в которых: 1) слабые фрикативные согласные (|*f*|, |*th*|, |*h*|) в начале или конце; 2) слабые глухие взрывные звуки (|*p*|, |*t*|, |*k*|) в начале или в конце; 3) носовые звуки в конце; 4) вокализованные фрикативные звуки, которые переходят в невокализованные в конце слова; 5) протяженные гласные звуки в конце слова.

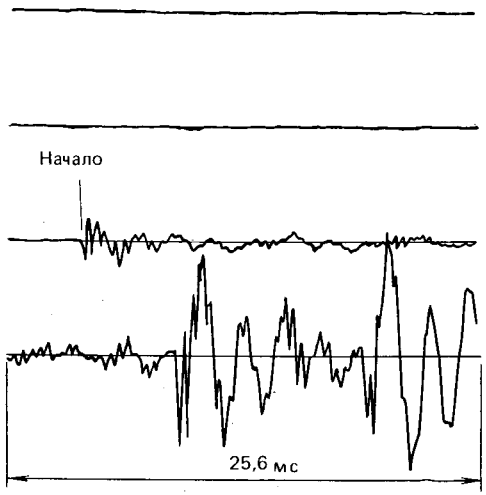

Рис. 4.13. Временна́я диаграмма начала слова */eight/* [6]

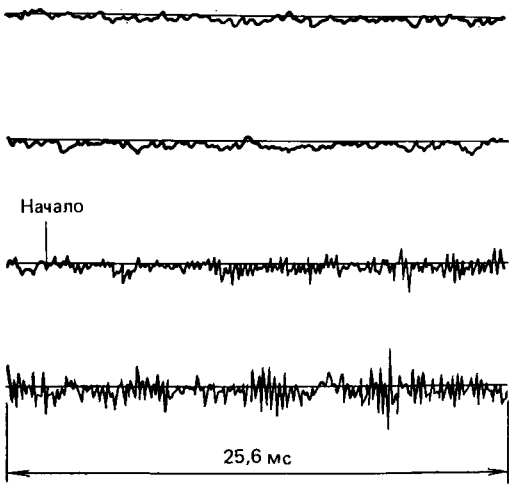

Рис. 4.14. Временна́я диаграмма начала слова */six/* [6]

Несмотря на сложность поставленной выше задачи, энергия и переходы через нуль могут быть совместно использованы в качестве основы для разработки алгоритмов выделения момента начала и конца фразы. Один из таких алгоритмов изучен Рабинером и Самбуром [6] при построении системы распознавания изолированных слов. В этой системе диктор произносит слово в течение интервала времени, установленного для регистрации сигнала. Сигнал далее на всем интервале дискретизируется и заносится в память для последующей обработки. Цель алгоритма состоит в определении начала и конца слова с тем, чтобы при распознавании исключить сегменты, содержащие только шум.

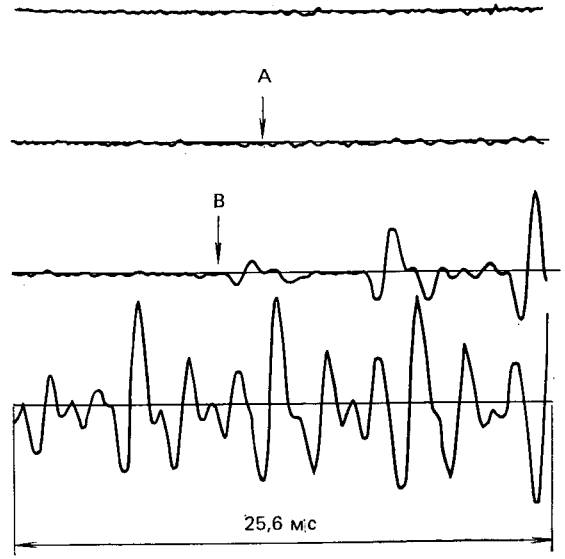

Рис. 4.15. Временна́я диаграмма начала слова */four/* [6]

Алгоритм можно пояснить с помощью рис. 4.16. В качестве основных параметров используются число переходов через нуль в течение 10 мс (4.18) и функция среднего значения (4.12), вычисленные с использованием окна длительностью 10 мс. Обе функции вычисляются на всем интервале с частотой 100 Гц. Предполагается,

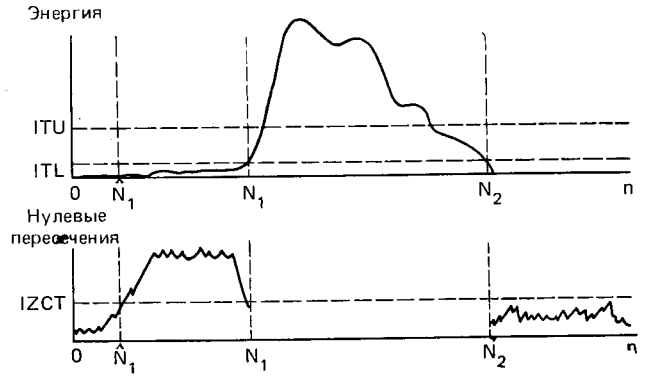

Рис. 4.16. Типичный пример измерений среднего значения и нулевых пересечений для слов с фрикативным звуком в начале [6]

что первые 100 мс не содержат речевого сигнала. По этому участку вычисляется среднее значение и дисперсия каждой из величин (4.12), (4.18) для определения статистических характеристик шума. Затем, с учетом этих характеристик и максимального среднего значения на интервале вычисляются пороги для среднего числа нулевых переходов и энергии сигнала [6]. Определяется фрагмент колебания, на котором траектория среднего значения превышает верхний порог (*ITU* на рис. 4.16). Предполагается, что начало и конец слова лежат вне этого фрагмента. Затем, двигаясь в обратном направлении по оси времени от момента, где M_n впервые превысила порог *ITU*, определяют момент, в котором M_n впервые оказалась меньше нижнего порога *ITL* (точка N_1). Этот момент выбирается в качестве предполагаемого начала. Сходным образом определяется и предполагаемое окончание слова N_2.

Данный двухпороговый алгоритм гарантирует, что провалы в траектории среднего значения не приведут к ложному выделению моментов начала и конца слова. На этом этапе главное — получить данные о том, что начало и конец слова расположены вне интервала от N_1 до N_2. Следующий шаг состоит в перемещении влево от N_1 (вправо от N_2) и сравнении числа переходов через нуль с порогом (*IZCT* на рис. 4.16), вычисленным по начальному участку. Это перемещение не должно превышать 25 интервалов слева от N_1 (справа от N_2). Если число переходов через нуль превышает порог 3 или более раз, начало слова переносится туда, где кривая числа нулевых пересечений впервые превысила порог. В противном случае N_1 считается началом слова. Аналогично поступают и с N_2. На рис. 4.17 показан пример работы алгоритма на типичных изолированных словах. На рисунке представлены восемь функций среднего значения для восьми различных слов двух различных дикторов. Некоторые слова записаны в машинном зале, а другие представляют собой магнитную запись в звуконепроницаемой комнате.

На каждом рисунке помечены начало и конец слова, как они были определены с помощью алгоритма. Например, на рис. 4.17*а* (слово |*nine*|) контроль среднего значения оказался достаточным для определения границ слова. А на примере 4.17*б* (слово |*replace*|) для определения конца слова использована функция числа нулевых пересечений, так как здесь расположен фрикативный звук |*s*|. Несмотря на то что звук |*s*| в конце слова имеет большое среднее значение, конец слова не может быть точно выделен по этому значению из-за высокого порога. В этом случае момент окончания слова уточнен по кривой числа нулевых пересечений. На рис. 4.17*в* конечное |*t*| в слове |*delete*| легко выделяется из-за значительного числа нулевых пересечений на участке длительностью 70 мс, расположенном после смычки и соответствующем взрывному согласному |*t*|. Таким образом, хотя среднее значение и число переходов через нуль малы на интервале 50 мс, соответствующем смычке, алгоритм позволяет правильно определить конец слова за счет большой интенсивности взрывного звука. В то же

время, если интенсивность взрывного звука будет мала, то конец слова будет отождествлен с началом смычки.

На рис. 4.17*г* представлен пример, в котором среднее значение шума было значительным в двух местах до начала слова |*substract*|, однако алгоритм устранил эти моменты из рассмотрения ввиду малого числа переходов через нуль. В этом примере относительно слабый взрывной звук (|*t*|) был правильно идентифициро-

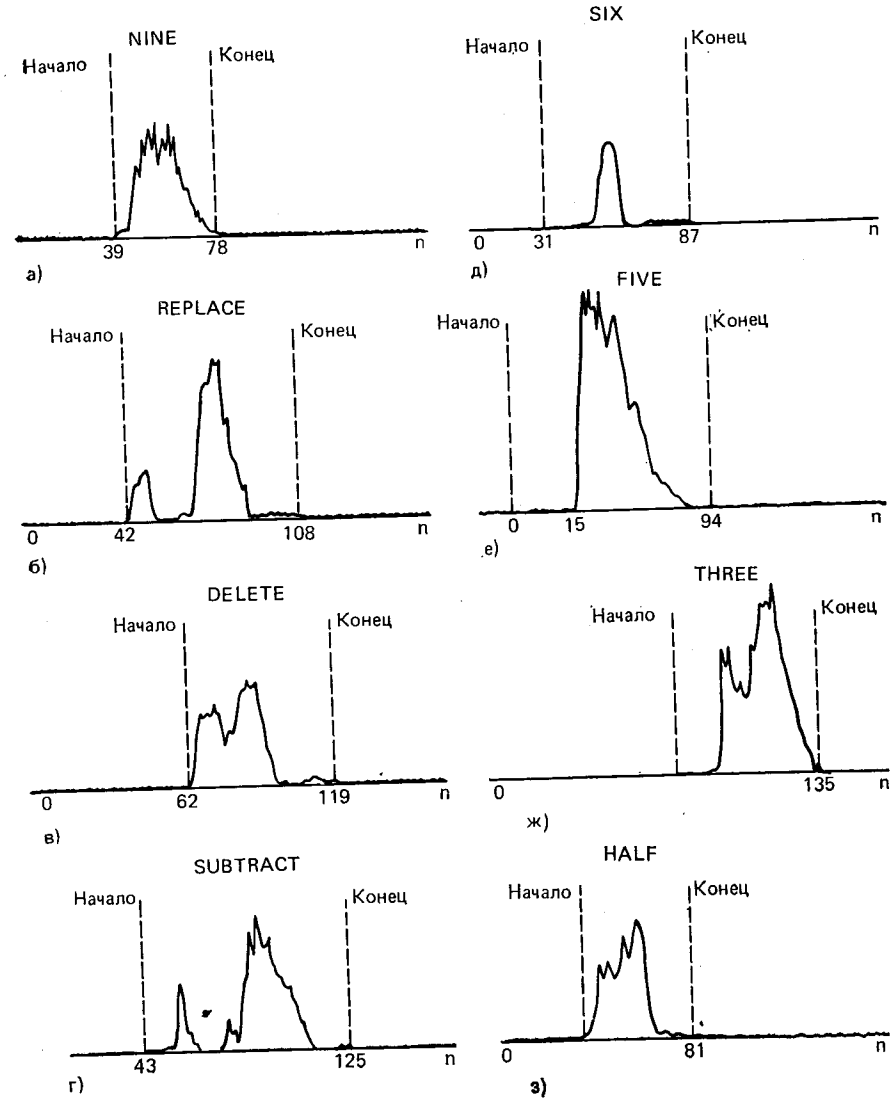

Рис. 4.17. Последовательности среднего значения, иллюстрирующие работу алгоритма выделения конца и начала слова на разных сигналах [6]

ван, как конец слова. На рис. 4.17*д* — показаны слова с фрикативными согласными в начале либо в конце. Во всех случаях алгоритм позволил установить граничные моменты слов так, что значительная часть невокализованного сегмента оказывалась в пределах этих границ.

Это применение функций числа переходов через нулевой уровень и среднего значения показывает, насколько полезны такие характеристики при решении практических задач. Большая практическая ценность рассмотренных способов обусловлена их простотой. Подобные примеры обработки будут встречаться и в последующих параграфах этой главы.

4.5. Оценивание периода основного тона на основе параллельной обработки

Оценивание периода (или частоты) основного тона является одной из наиболее важных задач в обработке речи. Выделители основного тона используются в вокодерах [8], системах распознавания и верификации дикторов [9, 10], в устройствах, предназначенных для глухих [11]. Поскольку задача очень важна, предложен ряд способов ее решения [12—19]. Все они обладают ограничениями и можно с уверенностью сказать, что в настоящее время отсутствует метод выделения основного тона, обеспечивающий удовлетворительные результаты для различных дикторов, в разных областях применения и условиях эксплуатации.

В этом параграфе рассмотрим только один метод выделения основного тона, предложенный Голдом и затем усовершенствованный Голдом и Рабинером [14]. Причины выбора именно этого метода выделения основного тона состоят в следующем: метод с успехом применялся в ряде приложений, основан исключительно на обработке во временной области, требует малых затрат времени при моделировании на универсальной ЭВМ и просто реализуем в спецвычислителе, а также хорошо иллюстрирует принцип параллельной обработки.

Основные положения этого метода:

1. По речевому сигналу формируется несколько импульсных последовательностей, которые сохраняют периодичность входного сигнала и не содержат других его особенностей, бесполезных с точки зрения выделения основного тона.

2. Обработка предполагает использование набора простых выделителей основного тона для каждой последовательности.

3. Оценки основного тона каждой последовательности подвергаются логической обработке для получения результирующей оценки периода основного тона речевого сигнала.

На рис. 4.18 изображена схема, предложенная Голдом и Рабинером [14]. Речевой сигнал дискретизируется с частотой 10 кГц, что позволяет оценить период с точностью $T = 10^{-4}$ с. Далее сигнал сглаживается в фильтре нижних частот с частотой среза около 900 Гц. Можно применить и полосовой фильтр с полосой от 100 до

900 Гц для устранения фона переменного тока питания (фильтр может быть выполнен в виде аналогового устройства до дискретизации или в виде цифрового — после дискретизации).

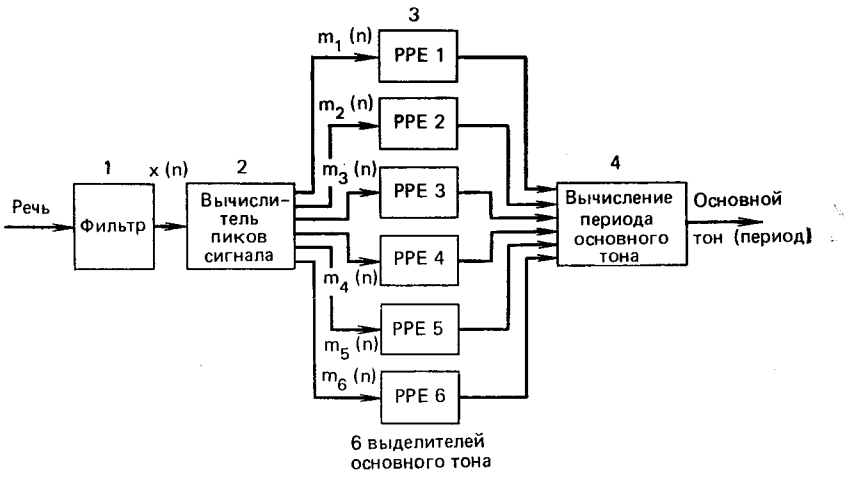

Рис. 4.18. Структурная схема выделения основного тона параллельной обработкой сигнала во временной области

Вслед за фильтрацией определяются локальные максимумы и минимумы в сигнале и по их амплитуде и положению из отфильтрованного сигнала формируется несколько (на рис. 4.18 — шесть) импульсных последовательностей. Каждая импульсная последовательность состоит из положительных импульсов, возникающих в месте расположения максимума или минимума сигнала. Эти шесть последовательностей в [14] имеют следующий вид.

1. $m_1(n)$: импульс, равный по амплитуде максимальному значению сигнала и формирующийся в месте расположения максимума.

2. $m_2(n)$: импульс, равный по амплитуде разности между максимумом и предшествующим минимумом и формирующийся в точке каждого максимума.

3. $m_3(n)$: импульс, равный по амплитуде разности между текущим максимумом и предшествующим максимумом и возникающий в точке каждого максимума (если эта разность отрицательна, то импульс обращается в нуль).

4. $m_4(n)$: импульс, равный по амплитуде минимальному отрицательному значению, взятому со знаком «минус», и возникающий в точке каждого минимума.

5. $m_5(n)$: импульс, равный сумме значений сигнала в точке минимума, взятого со знаком «минус», и сигнала в точке предшествующего максимума; формируется в точке каждого минимума.

6. $m_6(n)$: импульс, равный сумме минимального значения сигнала, взятого со знаком «минус», и предшествующего минимально-

го значения (если эта разность отрицательная, то импульс обращается в нуль).

На рис. 4.19 и 4.20 показаны два примера — синусоидальный сигнал и сумма синусоидального сигнала и его второй гармоники вместе с импульсными последовательностями, определенными выше. Очевидно, что импульсные последовательности имеют тот же период, что и исходный сигнал, хотя $m_5(n)$ на рис. 4.20 ближе к периодической последовательности с периодом, равным половине основного периода. Целью формирования импульсных последовательностей является упрощение текущего оценивания периода основного тона. Работа простейшего устройства оценивания иллюстрируется рис. 4.21. Каждая импульсная последовательность обрабатывается нелинейной системой с переменными параметрами (названной в [13] выделителем основного тона с экспоненциальной границей раздела). Когда на входе появляется импульс достаточно большой амплитуды, на выходе цепи устанавливается постоянный сигнал, равный этой амплитуде. Этот сигнал поддерживается неизменным в течение фиксированного интервала времени $\tau(n)$. В конце интервала выходной сигнал начинает уменьшаться по экспоненте. Когда входной импульс превысит уровень экспоненциально затухающего сигнала на выходе, процесс повторяется. Скорость затухания и длительность интервала зависят от последней оценки основного тона [14]. В результате такого своеобразного сглаживания получается квазипериодическая импульсная последовательность, показанная на рис. 4.21. Длительность каждого импульса представляет собой оценку периода основного тона, который обновляется с частотой 100 Гц.

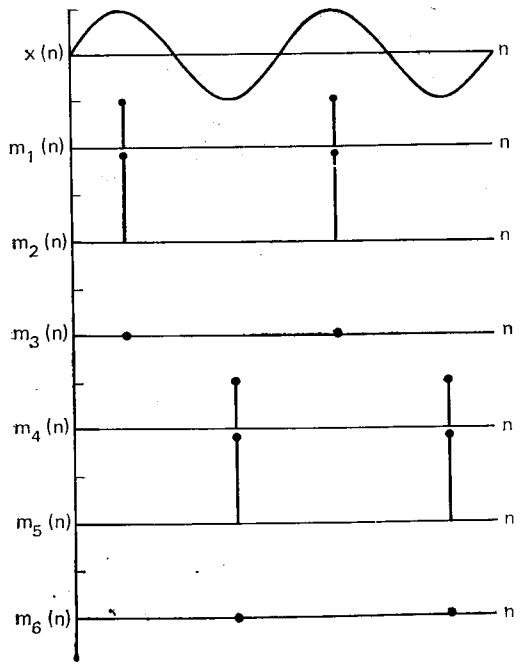

Рис. 4.19. Входной сигнал (синусоида) и соответствующая последовательность импульсов, сформированная по пикам и провалам.

Этот метод применяется в каждом из шести каналов, и таким образом получается шесть оценок периода основного тона. Полученные текущие оценки рассматриваются совместно с двумя последними оценками в каждом из шести каналов. Все оценки затем сравниваются и за оценку периода основного тона принимается та,

которая чаще всего встречается в данном множестве. Метод дает очень хорошие результаты при оценивании периода основного тона на вокализованных сегментах речевого сигнала. Для невокализованных сегментов возникает значительный разброс в оценках периода в каждом из шести каналов. Если такой разброс обнаруживается, то речь классифицируется как невокализованная. Весь процесс повторяется периодически для получения периода основного тона как функции времени и разделения всего сигнала на вокализованные и невокализованные участки.

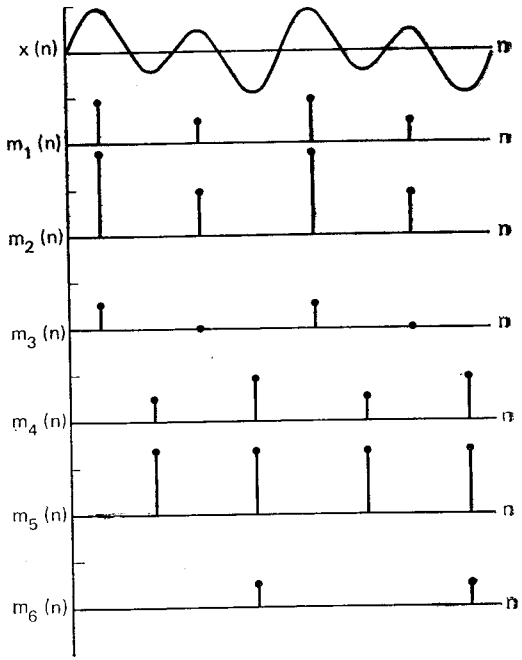

Рис. 4.20. Входной сигнал (ослабленная основная гармоника в сумме со второй гармоникой) и соответствующая последовательность импульсов, сформированная по пикам и провалам

Хотя описанный способ может показаться чрезмерно хитроумным, такая схема выделения периода основного тона может быть рекомендована как для целей технической реализации, так и для моделирования в универсальной ЭВМ. Действительно, при использовании современных ЭВМ метод позволяет обрабатывать сигнал почти в реальном масштабе времени (с коэффициентом трансформации времени, равным 2).

Работу схемы выделения периода основного тона иллюстрирует рис. 4.22, где приведен результат обработки синтезированного речевого сигнала. Преимущество использования синтетической ре-

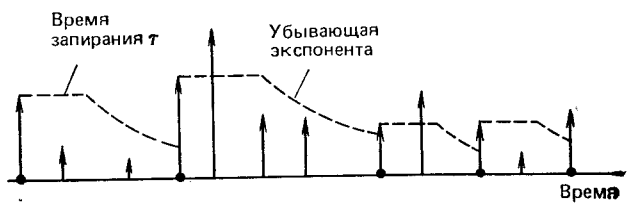

Рис. 4.21. Иллюстрация работы отдельного устройства оценивания периода основного тона во временной области [14]

чи состоит в том, что истинный период основного тона известен точно (поскольку он задается при синтезе). Это позволяет установить точность алгоритма. Недостаток синтетической речи заключается в том, что она формируется с помощью простой модели и по-

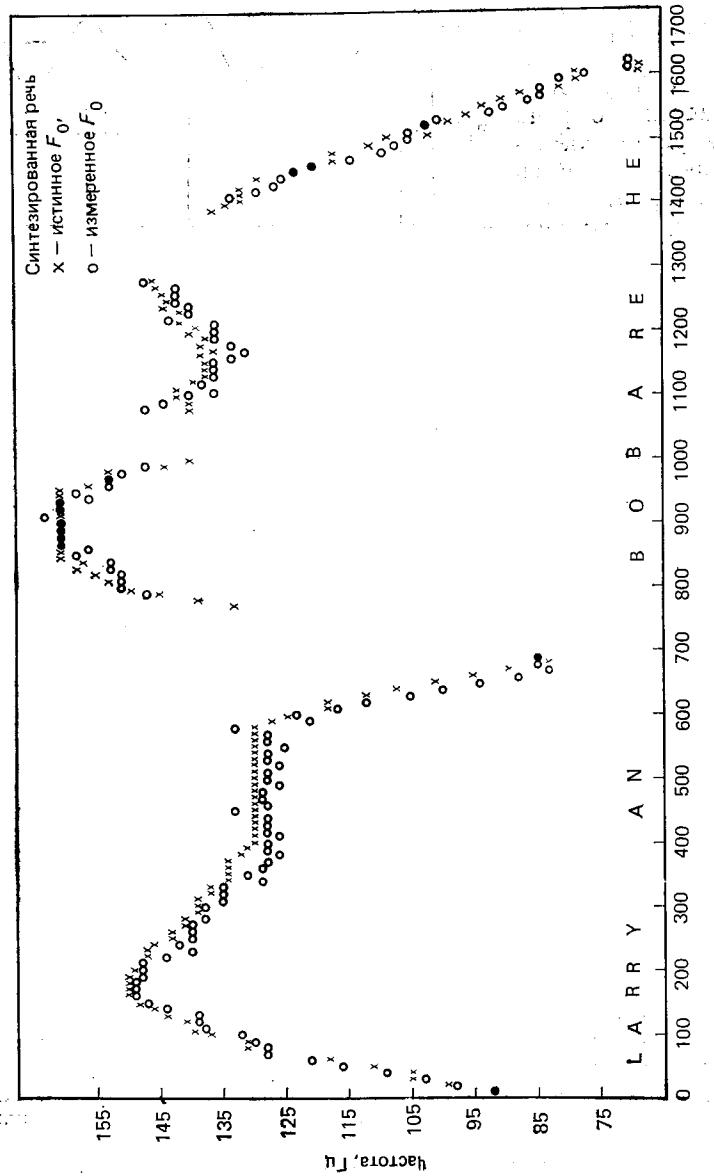

Рис. 4.22. Истинное и измеренное значение частоты основного тона для синтезированной фразы [14]

этому не может отражать всех свойств натурального речевого сигнала. Во всяком случае, эксперимент на синтезированном сигнале показал, что метод позволяет следить за периодом основного тона с точностью до двух интервалов дискретизации. Более того, было обнаружено, что в начале вокализованных сегментов (т. е. первые 10—30 мс) речь часто классифицируется как невокализованная. Этот результат обусловлен тем, что для устойчивого выделения требуется примерно три периода. Таким образом, возникает задержка приблизительно на два периода основного тона. В проведенном недавно сравнительном анализе алгоритмов выделения основного тона на реальном речевом сигнале и для различных условий оценивания этот метод дал хорошие результаты по сравнению с другими известными алгоритмами [12].

В заключение подчеркнем несколько основных положений метода. Во-первых, речевой сигнал обрабатывается с целью получения совокупности импульсных последовательностей, сохраняющих только периодичность сигнала (или фиксирующих ее отсутствие). Из-за такого упрощения структуры речевого сигнала для получения хорошей оценки основного тона оказывается возможным использовать простейшее устройство оценивания. И, во-вторых, несколько оценок основного тона рассматриваются в совокупности для повышения качества выделения. Таким образом, простота обработки сигнала достигнута ценой увеличения сложности логической части алгоритма. Поскольку логические операции осуществляются значительно реже (100 раз в секунду), чем дискретизация сигнала, скорость обработки оказывается высокой. Аналогичный подход был использован Барнвеллом [15] при разработке выделителя основного тона с помощью четырех простых схем выделения по нулевым пересечениям с последующей совместной обработкой решений для получения надежной оценки.

4.6. Кратковременная автокорреляционная функция

Автокорреляционная функция детерминированного сигнала в дискретном времени определяется выражением

$$\varphi(k) = \sum_{m=-\infty}^{\infty} x(m)\, x(m+k). \tag{4.21}$$

Если сигнал случайный или периодический, то $\varphi(k)$ целесообразно определить по-другому:

$$\varphi(k) = \lim_{N\to\infty} \frac{1}{(2N+1)} \sum_{m=-N}^{N} x(m)\, x(m+k). \tag{4.22}$$

Во всех случаях представление сигнала с помощью автокорреляционной функции позволяет отразить определенные свойства сиг-

нала. Например, если сигнал имеет период в P отсчетов, то легко показать, что

$$\varphi(k)=\varphi(k+P), \tag{4.23}$$

т. е. автокорреляционная функция периодического сигнала тоже периодическая. Автокорреляционная функция обладает и рядом других важных свойств: 1) она является четной функцией; 2) достигает максимального значения при $k=0$, т. е. $|\varphi(k)|\leqslant\varphi(0)$ для всех k; 3) величина $\varphi(0)$ равна полной энергии для детерминированного сигнала и средней мощности для случайного или периодического сигнала.

Таким образом, по автокорреляционной функции можно определить энергию сигнала и, кроме того, его периодические свойства. Если рассмотреть (4.23) и свойства 1 и 2, то можно отметить, что для периодического сигнала автокорреляционная функция достигает максимального значения в точках $0, \pm P, \pm 2P, \ldots$, т. е. при любом временно́м расположении сигнала его период можно оценить путем определения местоположения первого максимума автокорреляционной функции. Это свойство автокорреляционной функции позволяет использовать ее для оценки периодичности в любом сигнале, в том числе и речевом. Более того (см. гл. 8), автокорреляционная функция содержит значительно больше информации о тонкой временно́й структуре сигнала. Таким образом, весьма важно рассмотреть, как следует изменить приведенное определение для описания сигнала с помощью кратковременной автокорреляционной функции.

Используя развитый выше подход к определению кратковременных характеристик, определим корреляционную функцию в виде

$$R_n(k)=\sum_{m=-\infty}^{\infty} x(m)\,w(n-m)\,x(m+k)\,w(n-k-m). \tag{4.24}$$

Это уравнение можно интерпретировать следующим образом: сначала выделяется сегмент речевого сигнала с помощью функции временного окна, затем к взвешенному таким образом речевому сигналу применяется преобразование (4.21). Легко установить, что

$$R_n(-k)=R_n(k). \tag{4.25}$$

Используя это соотношение, можно выразить $R_n(k)$ в виде (4.10). Сначала заметим, что

$$R_n(k)=R_n(-k)=\sum_{m=-\infty}^{\infty} x(m)\,x(m-k)\,[w(n-m)\,w(n+k-m)]. \tag{4.26}$$

Если обозначить

$$h_k(n)=w(n)\,w(n+k), \tag{4.27}$$

то (4.26) можно переписать в виде

$$R_n(k)=\sum_{m=-\infty}^{\infty} x(m)\,x(m-k)\,h_k(n-m). \tag{4.28}$$

Таким образом, значение автокорреляционной функции при задержке k и в момент n получается путем фильтрации последовательности $x(n)x(n-k)$ в фильтре с импульсной характеристикой $h_k(n)$. Это изображено на рис. 4.23.

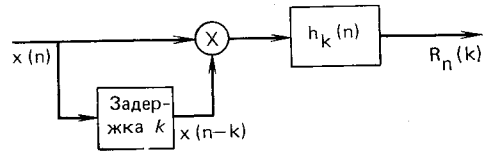

Рис. 4.23. Структурная схема вычисления кратковременной автокорреляции

Вычисление автокорреляционной функции обычно осуществляют, переписав соотношение (4.24) в виде

$$R_n(k)=\sum_{m=-\infty}^{\infty}[x(n+m)\,w'(m)]\,[x(n+m+k)\,w'(k+m)], \tag{4.29}$$

где $w'(n)=w(-n)$. Уравнение (4.29) показывает, что начало отсчета времени во входной последовательности действительно сдвигается к n-му отсчету, после чего она умножается на временно́е окно w' для выделения короткого сегмента речи. Если окно имеет конечную длительность, как в (4.8) и (4.9), то результирующая последовательность также будет иметь конечную длительность и (4.29) запишется в виде

$$R_n(k)=\sum_{m=0}^{N-1-k}[x(n+m)\,w'(m)]\,[x(n+m+k)\,w'(k+m)]. \tag{4.30}$$

Отметим, что при использовании в качестве w' прямоугольного окна или окна Хемминга уравнение (4.30) соответствует нереализуемому фильтру в (4.28). Для окон конечной длительности это, однако, не приводит к затруднениям, поскольку всегда, даже при обработке в реальном масштабе времени, может быть введена необходимая задержка.

Для вычисления автокорреляционной функции при задержке k в соответствии с (4.30) требуется N умножений для вычисления $x(n+m)w'(m)$ и $(N-k)$ умножений и сложений для получения суммы задержанных произведений. Таким образом, для вычисления совокупности значений корреляционной функции, как это нужно при выделении периодичности, требуется выполнять большой объем вычислений. Его можно сократить путем использования специальных свойств уравнения (4.30). Несколько таких примеров описано в приложении.

В отличие от (4.30) соотношение (4.28) можно использовать и при вычислении совокупности значений корреляционной функции,

так как при правильном выборе окна $R_n(k)$ можно вычислять рекуррентно (см. задачу 4.7).

На рис. 4.24 представлены три примера автокорреляционных функций, вычисленных по речевому сигналу, дискретизированному

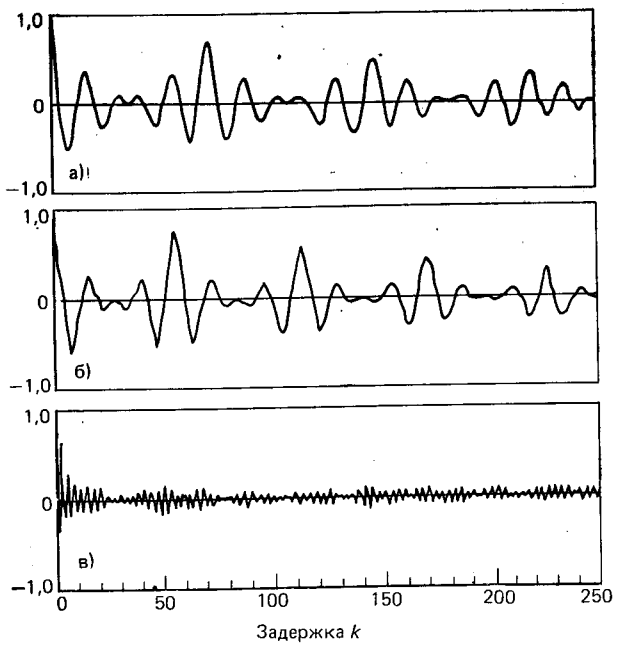

Рис. 4.24. Автокорреляционная функция вокализованной речи (*а*, *б*) и невокализованной речи (*в*), полученная с прямоугольным окном длительностью $N=401$

с частотой 10 кГц с использованием (4.30) при $N=401$. Как видно из рисунка, автокорреляционные функции вычислены для задержек $0 \leqslant k \leqslant 250$. Первые два примера соответствуют сегментам вокализованной, а последний — невокализованной речи[1]. Для первого сегмента пики корреляции возникают при задержках, кратных 72, что указывает на наличие периода, равного 7,2 мс, или частоты основного тона, равной примерно 140 Гц. Заметим, что даже очень короткий сегмент речи отличается от сегмента строго периодического сигнала. В течение интервала длительностью 401 отсчет и «период» сигнала и его форма изменяются. Это одна из причин, по которой пики уменьшаются по амплитуде с ростом задержки. Для другого вокализованного сегмента (взятого в другом месте фразы) видна сходная периодичность, только теперь пики возникают при задержках, кратных 58, что показывает наличие основного тона с периодом 5,8 мс. Наконец, в автокорреляционной

[1] Здесь и на последующих рисунках автокорреляционная функция нормирована таким образом, что $R_n(0)=1$.

функции для невокализованной речи отсутствует ярко выраженная периодичность, что говорит о непериодическом характере сигнала в данном случае. Видно, что автокорреляционная функция невокализованной речи представляет собой шумоподобное колебание, напоминающее речевой сигнал, по которому оно вычислено.

На рис. 4.25 приведены примеры применения временно́го окна Хемминга. Сравнивая эти результаты с соответствующими результатами на рис. 4.24, отметим, что прямоугольное окно значительно сильнее выявляет периодичность в сигнале, чем окно Хемминга. Этот результат не покажется неожиданным, если учесть, что окно Хемминга вносит затухание на концах обрабатываемого сегмента речи.

Примеры на рис. 4.24 и 4.25 рассчитаны для $N=401$. Важным вопросом является выбор N для надежного обнаружения периодичности. Здесь мы вновь сталкиваемся с противоречивыми требо-

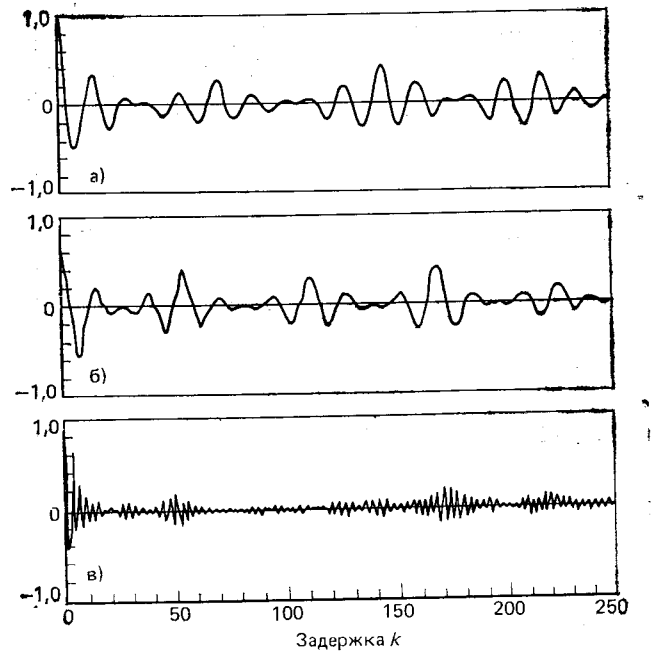

Рис. 4.25. Автокорреляционная функция вокализованной речи (*а*, *б*) и невокализованной речи (*в*), полученная с окном Хемминга длительностью $N=401$

ваниями. Вследствие изменения свойств речевого сигнала N следовало бы выбирать как можно меньше. С другой стороны очевидно, что для измерения периодичности по автокорреляционной функции окно должно иметь длительность, равную по крайней мере двум периодам основного тона. Фактически из-за конечной длительности взвешенного речевого сигнала, используемого при

вычислении $R_n(k)$, по мере увеличения k используется все меньше и меньше данных [см. верхний предел в сумме (4.30)]. Это приводит к уменьшению амплитуды пиков автокорреляционной функции при увеличении k, что видно в случае периодической импульсной последовательности (см. задачу 4.8) и легко может быть продемонстрировано на речевом сигнале. На рис. 4.26 показано влияние прямоугольного окна различной длительности на вид корреляционной функции. Пунктиром изображена функция

$$R(k) = 1 - k/N, \quad |k| < N, \tag{4.31}$$

которая представляет собой автокорреляционную функцию прямоугольного временно́го окна. Очевидно, что эта пунктирная линия является границей значений амплитуды автокорреляционной функции сигнала. В задаче 4.8 показано, что для периодической последовательности максимумы будут лежать точно на этой линии. В

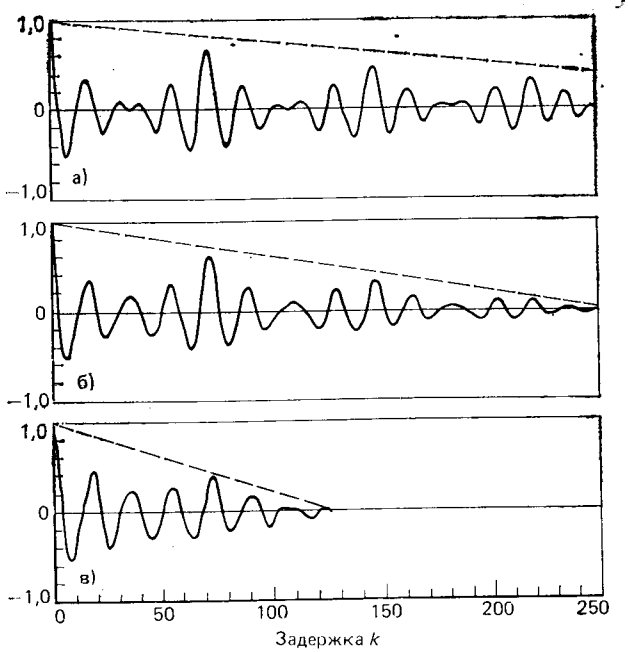

Рис. 4.26. Автокорреляционная функция вокализованной речи, полученная с прямоугольным окном при $N=401$ (*а*), 251 (*б*) и 125 (*в*)

рассматриваемом примере максимумы более отдалены от нее при $N=401$, чем в двух других случаях. Это происходит из-за того, что форма сигнала и период основного тона больше изменяются на интервале, равном 401 отчету, чем на более коротких интервалах. В этом состоят причины общего затухания амплитуды корреляционной функции.

Рисунок 4.26*в* соответствует окну длительностью 72 отсчета. Поскольку период основного тона в этом примере того же порядка, окно не охватывает даже двух полных периодов. Это случай, которого следует избегать. Но сделать это очень трудно из-за широкого диапазона значений периода основного тона. Один способ состоит в выборе столь длительного окна, чтобы оно охватывало даже наибольший период основного тона. Но это, очевидно, приведет к нежелательному усреднению большого числа периодов, когда период основного тона мал. Другой подход состоит в изменении длительности окна в соответствии с ожидаемым периодом основного тона. Еще один подход, позволяющий использовать короткие окна, состоит в модификации определения автокорреляционной функции.

Модифицированная кратковременная корреляционная функция определяется выражением

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(m)\, w_1(n-m)\, x(m+k)\, w_2(n-m-k). \tag{4.32}$$

Это выражение можно переписать в виде

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(n+m)\, \hat{w}_1(m)\, x(n+m+k)\, \hat{w}_2(m+k), \tag{4.33}$$

где

$$\hat{w}_1(m) = w_1(-m), \tag{4.34а}$$

$$\hat{w}_2(m) = w_2(-m). \tag{4.34б}$$

Для того чтобы исключить затухание, обусловленное переменным верхним пределом в (4.30), выберем окно $\hat{w}_2$ таким, чтобы оно включало отсчеты вне ненулевого интервала окна w_1, т. е. определим временны́е окна в виде

$$\hat{w}_1(m) = \begin{cases} 1, & 0 \leqslant m \leqslant N-1 \\ 0, & \text{в противном случае} \end{cases} \tag{4.35а}$$

и

$$\hat{w}_2(m) = \begin{cases} 1, & 0 \leqslant m \leqslant N-1+K \\ 0, & \text{в противном случае} \end{cases} \tag{4.35б}$$

где K — наибольшая требуемая задержка. Таким образом, уравнение (4.33) можно переписать в виде

$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)\, x(n+m+k) \quad 0 \leqslant k \leqslant K, \tag{4.36}$$

т. е. усреднение осуществляется по всем N отсчетам, включая отсчеты вне интервала от n до $n+N-1$. Различия в исходных данных для (4.30) и (4.36) изображены на рис. 4.27. На рис. 4.27*а* изображен исходный речевой сигнал, а на рис. 4.27*б* N отсчетов, выделенных с помощью прямоугольного временно́го окна. В слу-

чае прямоугольного окна этот сегмент будет использован в обоих сомножителях в (4.30) и будет сомножителем $x(n+m)w_1(m)$ в (4.36). На рис. 4.27*в* изображен другой сомножитель в (4.36), включающий K дополнительных отсчетов.

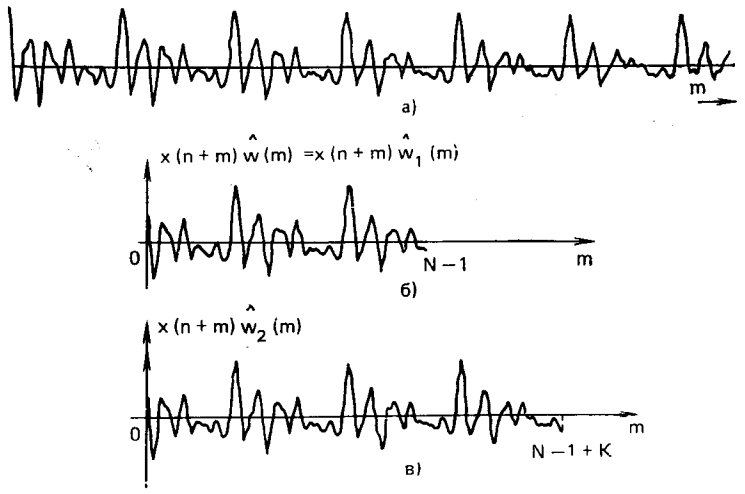

Рис. 4.27. Отсчеты, применяемые для вычисления кратковременной автокорреляционной функции

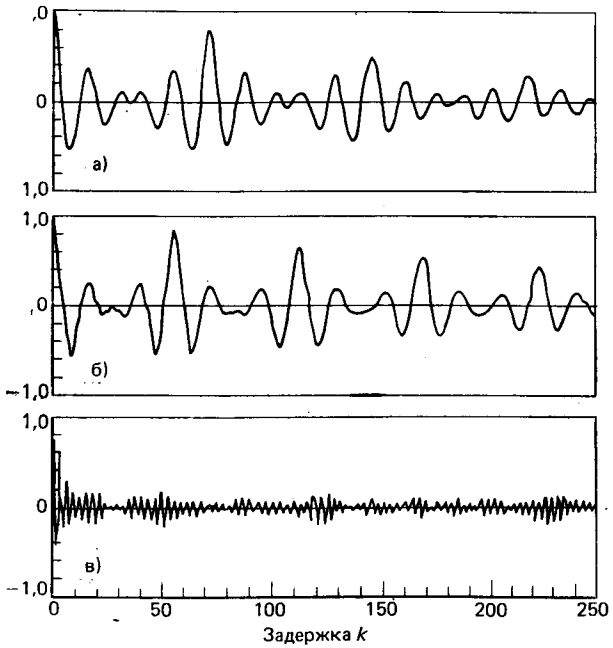

Рис. 4.28. Модифицированная автокорреляционная функция вокализованной речи для сегментов рис. 4.24 при $N=401$

Выражение (4.36) далее называется *модифицированной* кратковременной корреляционной функцией. Однако, строго говоря, это *взаимная* корреляционная функция двух сегментов речи $x(n+m)\hat{w}_1(m)$ и $x(n+m)\hat{w}_2(m)$. Таким образом, $\hat{R}_n(k)$ имеет свойства взаимной корреляционной функции, а не автокорреляционной, например, $\hat{R}_n(-k)\neq\hat{R}_n(k)$. Тем не менее $\hat{R}_n(k)$ имеет пики при задержках, кратных периоду сигнала, и эти пики не затухают с ростом k. На рис. 4.28 представлены модифицированные автокорреляционные функции, соответствующие примерам рис. 4.24. Поскольку при $N=401$ эффект изменения формы сигнала преобладает над краевым эффектом на рис. 4.24, то оба рисунка выглядят почти одинаково. Сравнение рис. 4.29 и 4.26 показывает, что различия более заметны для малых значений N. Очевидно, что пики

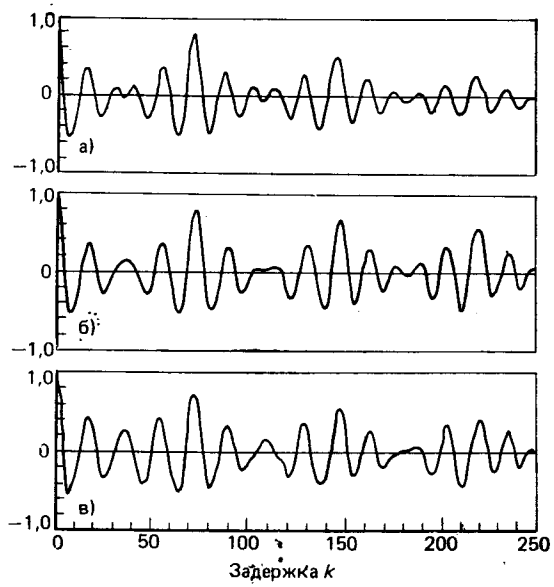

Рис. 4.29. Модифицированная автокорреляционная функция вокализованной речи при $N=401$ (*а*), 251 (*б*) и 125 (*в*) (см. рис. 4.26)

на рис. 4.29 меньше, чем при $k=0$, только вследствие изменения периодичности на интервале от n до $n+N-1+K$, который используется в уравнении (4.36). В задаче 4.8 показано, что для строго периодической последовательности все пики будут иметь одинаковые амплитуды.

4.7. Кратковременная функция среднего значения разности

Как отмечалось выше, вычисление автокорреляционной функции требует выполнения большого числа арифметических операций, даже при использовании упрощений, изложенных в приложе-

нии. Метод, исключающий необходимость умножений, основан на том, что для строго периодической функции с периодом P последовательность

$$d(n)=x(n)-x(n-k) \tag{4.37}$$

будет равна нулю при $k=0\pm P,\ \pm 2P, \ldots$ Для сегментов вокализованного речевого сигнала естественно ожидать, что последовательность $d(n)$ будет близка к нулю (но не равна ему) при k, кратном периоду основного тона. Кратковременное среднее значение величины $d(n)$ как функция k будет мало, если k близко к периоду основного тона. Кратковременная функция среднего значения разности (КФСР) определяется как

$$\gamma_n(k)=\sum_{m=-\infty}^{\infty}|x(n+m)\,w_1(m)-x(n+m-k)\,w_2(m-k)|. \tag{4.38}$$

Очевидно, что если $x(n)$ близка к периодической функции на интервале, выделенном с помощью временно́го окна, то $\gamma_n(k)$ будет иметь глубокие провалы при $k=P,\ 2P\ldots$ Отметим, что целесообразно применять прямоугольные окна. Если оба окна имеют одинаковую длительность, то получается функция, сходная с автокорреляционной функцией (4.30). Если длительность $w_2(n)$ превышает длительность $w_1(n)$, то ситуация аналогична вычислению модифицированной автокорреляции (4.36). Можно показать [16], что

$$\gamma_n(k)\approx\sqrt{2}\,\beta(k)\,[\hat{R}_n(0)-\hat{R}_n(k)]^{1/2}, \tag{4.39}$$

причем $\beta(k)$ в (4.39) изменяется в пределах от 0,6 до 1,0 на различных сегментах речи, но слабо зависит от k.

На рис. 4.30 представлена КФСР для речевых сегментов рис. 4.24 и 4.28 при окне той же протяженностью. Легко видеть, что КФСР действительно имеет вид (4.39); функция содержит глубокие провалы в точках, кратных периоду основного тона, для вокализованной речи, и не имеет заметных провалов для невокализованной речи.

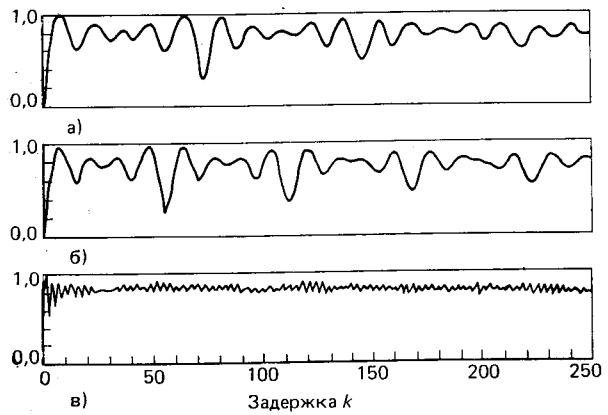

Рис. 4.30. Функция КФСР (нормированная) для сегментов речи рис. 4.24, 4.28

Для получения КФСР требуется выполнить операции сложения, вычитания и вычисления модуля, в то время как для вычисления автокорреляционной функции требуется выполнить операции сложения и умножения. При использовании системы счисления с плавающей запятой, где на операцию сложения и умножения требуется приблизительно одно и то же время, оба метода при одинаковых окнах сравнимы по быстродействию. Однако с точки зрения технической реализации, когда применяется система счисления с фиксированной запятой, КФСР имеет некоторые преимущества. В этом случае на выполнение операции умножения требуется значительно больше времени, чем операции сложения. Кроме того, для вычисления суммы произведений требуется либо масштабирование, либо удвоенная точность вычислений. С этой точки зрения именно КФСР использовалась в цифровых системах обработки речи, функционирующих в реальном масштабе времени.

4.8. Оценивание периода основного тона по автокорреляционной функции

Как показано в § 4.6, кратковременная автокорреляционная функция является удобной характеристикой сигнала, по которой можно производить текущее оценивание периода основного тона. В данном параграфе рассматривается несколько особенностей применения автокорреляционных выделителей основного тона.

Одно из основных ограничений применения автокорреляционной функции состоит в том, что она «содержит» излишне много сведений о сигнале. (В гл. 8 показано, что по 10—12 значениям автокорреляционной функции можно достаточно точно оценить передаточную функцию голосового тракта.) В результате (см. рис. 4.26) автокорреляционная функция имеет много побочных пиков. Большинство этих пиков обусловлено откликом голосового тракта, состоящего из затухающих колебаний, которые определяют форму речевого колебания на каждом периоде основного тона. На рис. 4.26*а* и *б* пик, соответствующий периоду основного тона, имеет наибольшую амплитуду, однако на рис. 4.26*в* пик в точке $k=15$ фактически больше, чем в точке $k=72$. Так получилось из-за малой протяженности окна по сравнению с периодом основного тона. Быстрое изменение форматных частот также может привести к подобному эффекту. В том случае, если побочные пики автокорреляционной функции превышают пик основного тона, простая процедура выделения наибольшего пика приведет к ошибкам.

Для устранения этих затруднений целесообразно так обработать речевой сигнал, чтобы подчеркнуть его периодичность и устранить несущественные в данном случае особенности его тонкой структуры. Этот подход, введенный в § 4.5, допускает использова-

ние очень простого выделителя основного тона. Методы, которые используют этот вид обработки сигнала, иногда называют методами «выравнивания спектра», поскольку они основаны на устранении влияния передаточной функции речевого тракта. При этом каждая гармоника обработанного сигнала имеет одинаковую амплитуду. Предложен ряд методов спектрального сглаживания [17], однако в данном случае наиболее удобным оказался метод центрального ограничения.

В методе, предложенном Сондхи [17], центральное ограничение речи осуществляется посредством нелинейного преобразования:

$$y(n) = C[x(n)], \tag{4.40}$$

где функция $C[\cdot]$ изображена на рис. 4.31. Рисунок 4.32 иллюстрирует обработку речевого сигнала с использованием центрального ограничения. Сегмент речевого сигнала, используемый при

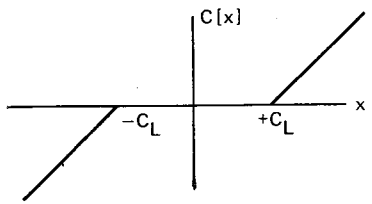

Рис. 4.31. Центральное ограничение

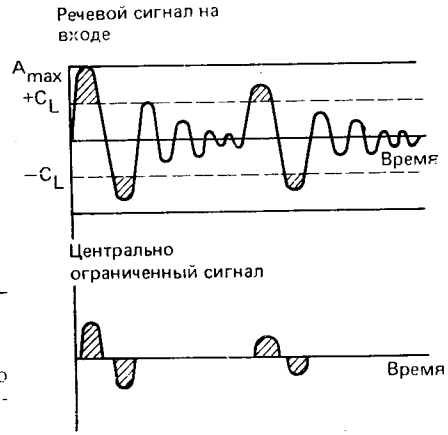

Рис. 4.32. Пример обработки речевого сигнала с помощью центрального ограничения [17]

вычислении автокорреляционной функции, показан сверху на рис. 4.32. Для этого сегмента определяется максимальная амплитуда A_{max} и уровень ограничения C_L устанавливается как некоторый процент от A_{max} (в [17] — 30%). Как видно из рис. 4.31, для отсчетов, превышающих C_L, выходной сигнал центрального ограничителя равен разности входного сигнала и уровня ограничения. Для отсчетов, меньших уровня ограничения, сигнал на выходе равен нулю. Нижняя кривая на рис. 4.32 — сигнал на выходе ограничителя. В отличие от метода, изложенного в § 4.5, где пики сигнала отображались импульсами, в данном случае преобразованный сигнал состоит из части входного сигнала, превысившей порог ограничения.

Рисунок 4.33 [18] иллюстрирует процесс вычисления автокорреляционной функции с помощью центрального ограничения. На рис. 4.33*а* показан вокализованный сегмент сигнала длиной 300 отсчетов ($F_s = 10$ кГц). На автокорреляционной функции, показан-

ной справа, имеется четкий пик, соответствующий основному тону. Однако имеется и много побочных пиков, обусловленных затухающими колебаниями в голосовом тракте. На рис. 4.33*в* изображен сигнал после центрального ограничения, где уровень ограничения установлен так, как это показано на рис. 4.33*в* (в данном случае

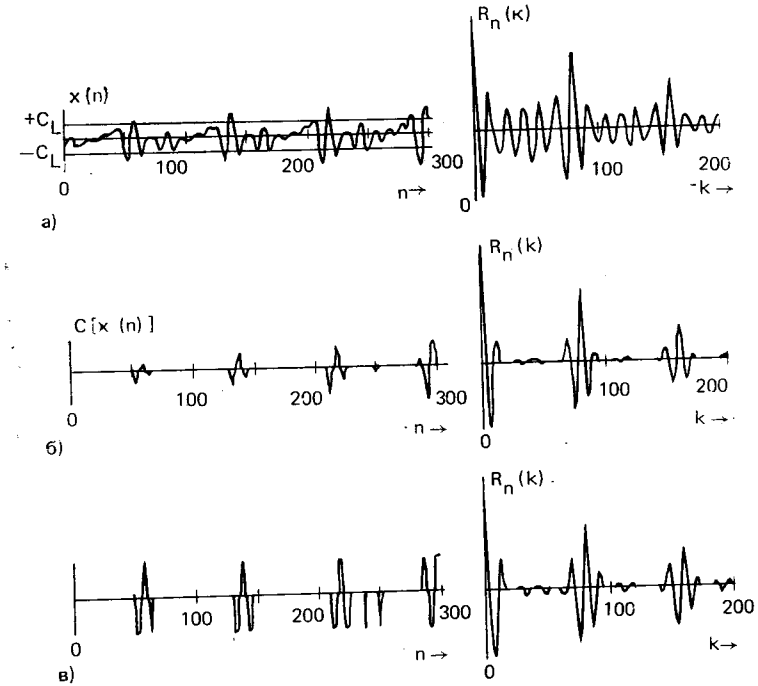

Рис. 4.33. Временна́я диаграмма речевого сигнала и корреляционная функция: *а*) без ограничения; *б*) с центральным ограничением; *в*) с трехуровневым центральным ограничением (Функции корреляции нормированы) [18]

он составляет 68% от максимальной амплитуды на первых 100 отсчетах). Отметим, что все, что остается при этом от речевого сигнала, представляет собой несколько импульсов, расположенных там же, где и исходные импульсы основного тона. Автокорреляционная функция имеет теперь значительно меньше побочных пиков, что уменьшает вероятность ошибки.

Рассмотрим эффект ограничения уровня. При высоком уровне ограничения количество отсчетов, превышающих его, будет небольшим. Это приведет к малому количеству посторонних пиков в автокорреляционной функции, что иллюстрируется рис. 4.34, на котором изображена автокорреляционная функция сегмента речи рис. 4.26*а* для уменьшающихся уровней ограничения. Очевидно, что по мере уменьшения уровня ограничения через ограничитель проходит больше пиков, что приводит к усложнению формы автокорреляционной функции (случай, когда уровень ограничения ра-

вен нулю, соответствует рис. 4.26*а*). Из данного рисунка можно сделать вывод, что наиболее надежное оценивание периода основного тона достигается при наиболее высоком уровне ограничения. Использование слишком высокого уровня может, однако, также привести к трудностям. Может случиться так, что амплитуда речевого сигнала будет заметно изменяться на протяжении сегмента.

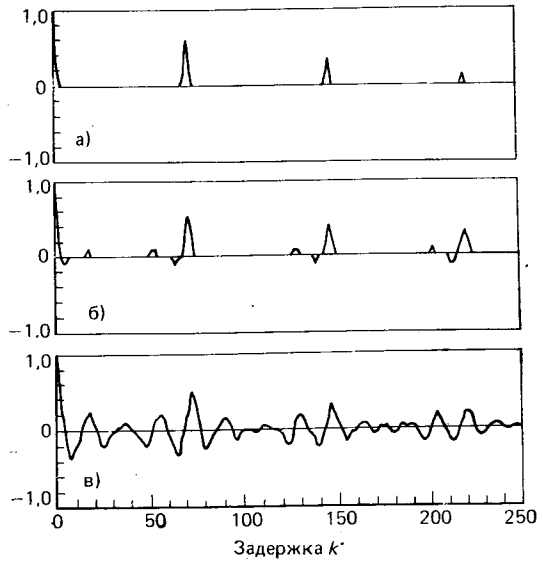

Рис. 4.34. Автокорреляционная функция на выходе центрального ограничителя при $N=401$; *а*) $C_L=80\%$ от максимума; *б*) 64%; *в*) 48% (Сегмент речи соответствует рис. 4.26*а*)

Тогда, если уровень ограничения выставлен по наибольшей амплитуде сигнала, бо́льшая часть колебания окажется ниже уровня и будет потеряна. Поэтому Сондхи выбрал величину уровня ограничения как 30% от максимального значения. В методах, позволяющих использовать более высокие уровни (от 60 до 80%), определяется максимальное значение сигнала на первой и последней трети обрабатываемого сегмента и затем в качестве уровня ограничения выбирается меньшее из этих значений. Этот способ иллюстрируется рис. 4.33*б*.

Решение задачи устранения побочных максимумов корреляционной функции может быть существенно облегчено применением центрального ограничения перед вычислением автокорреляционной функции. Однако другая трудность при автокорреляционном спи-сании сигнала (которая не устраняется даже при центральном ограничении) состоит в большом объеме вычислений. Простая модификация функции центрального ограничения приводит к значительному упрощению автокорреляционной функции, практически без потерь точности оценивания основного тона. Модифицирован-

ная функция представлена на рис. 4.35. Выходной сигнал ограничителя, как это следует из рис. 4.35, равен $+1$, если $x(n)>C_L$, и равен -1, если $x(n)<-C_L$. Во всех других случаях он равен нулю. Устройство, описываемое данной функцией, будем далее называть трехуровневым центральным ограничителем. На рис. 4.33*в* по-

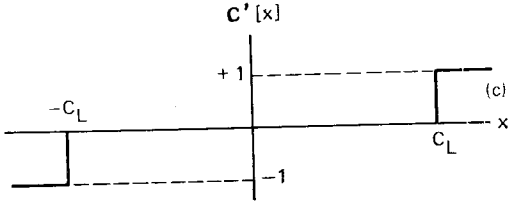

Рис. 4.35. Трехуровневое центральное ограничение

казан сигнал на выходе трехуровневого центрального ограничителя в случае, когда на вход поступает сигнал рис. 4.33*а*. Несмотря на то что при этом все пики, превысившие уровень ограничения, имеют одну и ту же амплитуду, корреляционная функция сигнала практически не отличается от изображенной на рис. 4.33*б*. Таким образом, и в этом случае подавляется большинство побочных пиков, что позволяет точно оценить период основного тона.

Вычисление автокорреляционной функции сигнала на выходе трехуровневого ограничителя отличается простотой. Если обозначить выходной сигнал через $y(n)$, то слагаемое $y(n+m)y(n+m+k)$ в автокорреляционной функции

$$R_n(k)=\sum_{m=0}^{N-k-1} y(n+m)\,y(n+m+k) \tag{4.41}$$

может принимать только три различных значения:

$$y(n+m)\,y(n+m+k)=\begin{cases}0, & y(n+m)=0 \text{ или } y(n+m+k)=0\\ +1, & y(n+m)=y(n+m+k),\\ -1, & y(n+m)\neq y(n+m+k).\end{cases} \tag{4.42}$$

Таким образом, при технической реализации алгоритма вычислений требуется выполнить лишь логические операции и накопление значений автокорреляционной функции в реверсивном счетчике для каждого k.

Обсуждая дальнейшие детали технической реализации, отметим, что целесообразно использовать модифицированную автокорреляционную функцию, определенную выражением (4.36), как в случае трехуровневого ограничения, так и при другом центральном ограничении, поскольку в этом случае пики автокорреляционной функции не уменьшаются при увеличении задержки. При вычислении КФСР по (4.38) сигнал также может быть подвергнут одному из видов ограничения. В действительности существует множество комбинаций, которые можно использовать в разных ситуациях [18].

Алгоритмов оценивания тона по кратковременной корреляционной функции уже предложено много и несомненно, что будет пред-

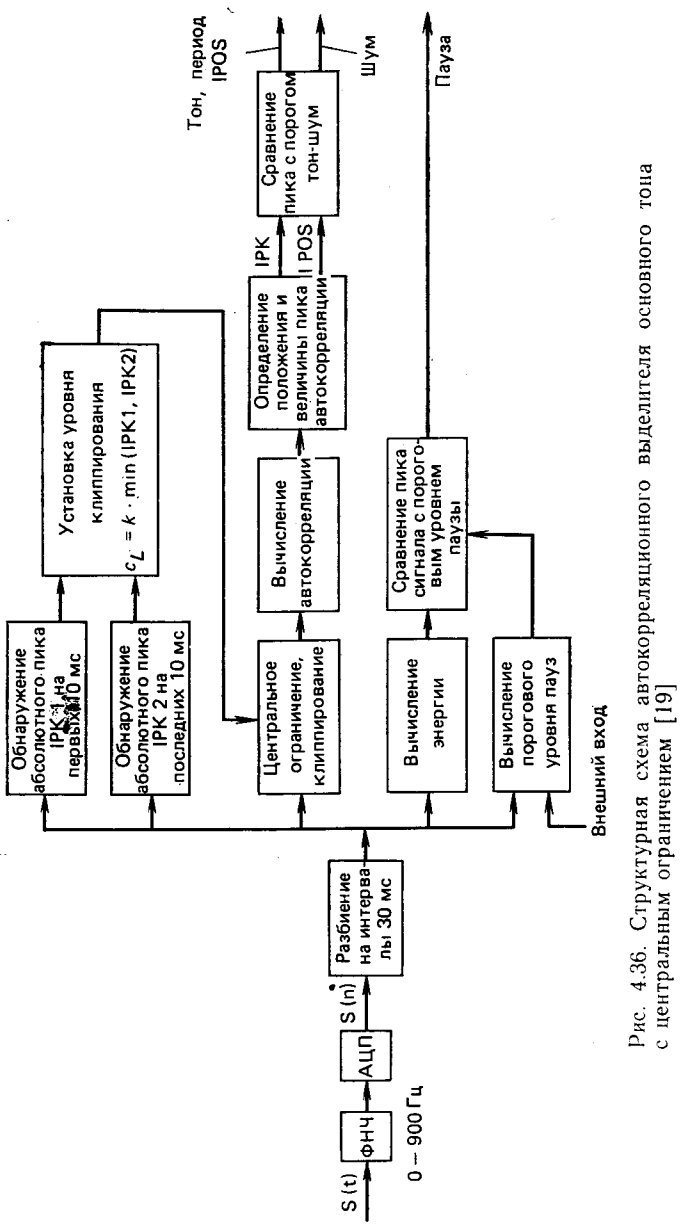

Рис. 4.36. Структурная схема автокорреляционного выделителя основного тона с центральным ограничением [19]

ложено еще больше. Мы завершим этот параграф рассмотрением одного способа, который был реализован в цифровой форме [19]. Подробный алгоритм представлен на рис. 4.36, а его краткое описание приводится ниже.

1. Речевой сигнал пропускается через фильтр нижних частот с частотой среза 900 Гц и дискретизируется с частотой 10 кГц.

2. Выделяются сегменты длительностью 30 мс (300 отсчетов) через каждые 10 мс. Таким образом, соседние сегменты имеют перекрытие 20 мс.

3. Вычисляется среднее значение (4.12) с прямоугольным окном протяженностью 100 отсчетов. Максимальное значение сигнала на каждом сегменте сравнивается с порогом, вычисленным по сегменту шумового фона длительностью 50 мс. Если пиковый уровень сигнала превосходит порог, то принимается решение о наличии речевого сигнала на сегменте и продолжается его обработка, в противном случае сегмент классифицируется как пауза и дальнейшая обработка не производится.

4. Уровень ограничения составляет фиксированный процент (68%) минимального значения двух максимумов сигнала, рассчитанных по первым и последним 100 отсчетам.

5. Речевой сигнал преобразуется в трехуровневом ограничителе и далее вычисляется автокорреляционная функция в области предполагаемого значения периода основного тона.

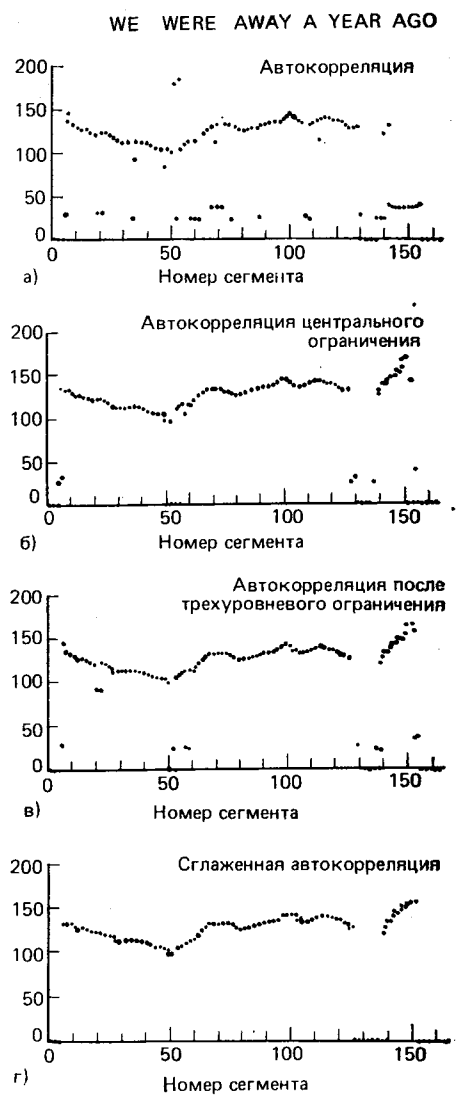

Рис. 4.37. Выходной сигнал автокорреляционного выделителя основного тона: *а*) без ограничения; *б*) с центральным ограничением; *в*) с трехуровневым центральным ограничением (см. рис. 4.35); *г*) с нелинейным сглаживанием сигнала (*в*) [18]

6. Определяется максимальный пик автокорреляционной функции и сравнивается с порогом, который составляет примерно 30% от $R_n(0)$. Если этот пик оказался меньше порога, то сегмент классифицируется как невокализованный, а в противном случае — как вокализованный с периодом основного тона, соответствующим положению этого пика. Выше приведен алгоритм, реализованный в цифровом виде [19]. Однако имеется возможность значительно изменить вычислительный процесс. Например, на шагах 4 и 5 алгоритма можно использовать центральное ограничение и обычную процедуру вычисления автокорреляционной функции или вообще устранить всякое ограничение. Другая возможность заключается в использовании КФСР (и поиске минимумов, а не максимумов) как с каким-либо ограничением, так и без него.

На рис. 4.37 изображена кривая основного тона для трех вариантов алгоритма. На рис. 4.37*а* приведена траектория основного тона для случая вычисления автокорреляционной функции без какого-либо ограничения. Следует отметить наличие значительного разброса оценок вследствие ошибок в оценивании, обусловленных, очевидно, тем, что побочные пики оказываются больше, чем пики в области основного тона. Поскольку нормированный период основного тона в среднем расположен между 100 и 150, уменьшение автокорреляционной функции вызывает значительное затухание пика в области основного тона. Поэтому побочные пики автокорреляционной функции оказываются больше основного. На рис. 4.37*б* и *в* иллюстрируется применение центрального ограничения и трехуровневого центрального ограничения при вычислении автокорреляционной функции. Видно, что большинство ошибок здесь устранено и, более того, между этими двумя результатами нет существенного различия. Небольшое количество ошибок остается в обоих случаях. Эти ошибки можно эффективно устранить путем применения нелинейного сглаживания с помощью метода, излагаемого ниже. Пример показан на рис. 4.37*г*.

4.9. Медианное сглаживание и обработка речи

Часто для сглаживания шумоподобной компоненты в сигнале используются линейные фильтры. Однако для некоторых приложений, вследствие особенностей обрабатываемых данных, линейное сглаживание не является адекватным. Примером может служить траектория основного тона на рис. 4.37*в*, содержащая очевидные ошибки, которые следует устранить, переместив ошибочные точки на основную траекторию. Обычный линейный фильтр не только не устранит этих ошибок, но и внесет искажения на резких переходах от вокализованного сегмента к невокализованному. Последнему соответствует нулевое значение периода основного тона. В таких случаях требуется нелинейный алгоритм обработки, устраняющий большие ошибки и не приводящий к значительным искажениям. Хотя идеального алгоритма такого типа не существует, можно показать, что комбинация алгоритма вычисления медианы и линейного сглаживания (впервые предложенная Тьюки [20]) обладает требуемыми свойствами [21].

Сущность линейного сглаживания состоит в разделении таких сигналов, спектры которых почти не пересекаются. Для нелинейного сглаживания целесообразно ввести разделение с учетом того, является ли сигнал «гладким» или шумоподобным. Таким образом, сигнал можно представить в виде

$$x(n) = S[x(n)] + R[x(n)], \tag{4.43}$$

где $S[x]$ — гладкая компонента, а $R[x]$ — шумоподобная компонента сигнала x. Нелинейностью, позволяющей разделить эти компоненты, является текущая медиана сигнала. Выходным сигналом устройства медианного сглаживания $M_L[x(n)]$ является медиана L отсчетов $x(n), \ldots, x(n-L+1)$. Текущие медианы протяженностью L обладают рядом полезных для сглаживания свойств:

1. $M_L[\alpha x(n)] = \alpha M_L[x(n)]$.
2. Медианы не «смазывают» основных разрывов в сигнале, если сигнал не имеет других разрывов среди $L/2$ отсчетов.
3. Медианы приблизительно повторяют полиномиальный тренд низкого порядка. Следует отметить, что алгоритмы медианного сглаживания, как и другие нелинейные алгоритмы обработки, не обладают принципом суперпозиции, т. е.

$$M_L[\alpha x_1(n) + \beta x_2(n)] \neq \alpha M_L[x_1(n)] + \beta M_L[x_2(n)]. \tag{4.44}$$

Хотя в общем случае медианы сохраняют резкие разрывы в сигнале, применение такой обработки позволяет значительно сгладить нежелательные шумоподобные компоненты сигнала. Хорошие результаты дает совместное использование линейных методов и методов медианного сглаживания. Поскольку текущие медианы обеспечивают некоторое сглаживание сигнала, линейная система может быть низкого порядка. Например, вполне подходит фильтр с импульсной характеристикой

$$h(n) = \begin{cases} 1/4 & n = 0, \\ 1/2 & n = 1, \\ 1/4 & n = 2. \end{cases} \tag{4.45}$$

На рис. 4.38*а* приведена структурная схема системы совместного сглаживания. Сигнал на выходе устройства приблизительно равен

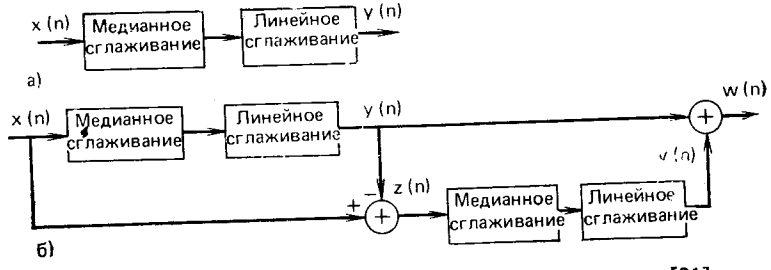

Рис. 4.38. Структурная схема нелинейного сглаживания [21]

$S[x(n)]$. Так как приближение не идеальное, далее применяется повторное сглаживание рис. 4.38*б*. Поскольку

$$y(n) = S[x(n)], \tag{4.46}$$

то

$$z(n) = x(n) - y(n) = R[x(n)]. \tag{4.47}$$

Повторное нелинейное сглаживание $z(n)$ дает корректирующий сигнал, который прибавляется к $y(n)$ для получения $w(n)$, более точно описывающей $S[x(n)]$. Сигнал $w(n)$ удовлетворяет соотношению

$$w(n) = S[x(n)] + S[R[x(n)]]. \tag{4.48}$$

Если $z(n) = R[x(n)]$ точно, т. е. устройство нелинейного сглаживания идеально, то $S[R[x(n)]]$ равно нулю и корректирующее слагаемое будет отсутствовать.

При использовании устройства нелинейного сглаживания с алгоритмом рис. 4.38 следует учитывать задержки в каждой его ветви. Медианное сглаживание вносит задержку на $(L-1)/2$ отсчетов, а линейное — в соответствии с импульсной характеристикой фильтра. Например, устройство медианного сглаживания на пять отсчетов вносит задержку в два отсчета, а окно Хемминга на три отсчета — задержку на один отсчет. На рис. 4.39 изображен реализуемый вариант устройства рис. 4.38*б*.

Наконец, при технической реализации устройства медианного сглаживания (рис. 4.39) надо установить граничные условия, т. е.

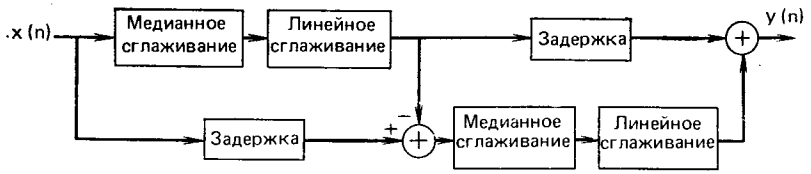

Рис. 4.39. Система нелинейного сглаживания с компенсирующей задержкой [21]

определить, как вычислять текущую медиану в начале и конце сегмента сигнала. Хотя возможны различные подходы к выбору граничных условий, для речевого сигнала целесообразно доопределить сигнал за пределы сегмента, полагая его периодически повторяющимся.

На рис. 4.40 показаны результаты применения различных устройств сглаживания к функции среднего числа нулевых пересечений. Исходная функция (рис. 4.40*а*) имеет шумоподобную компоненту из-за малого времени усреднения. На рис. 4.40*г* показан выходной сигнал устройства медианного сглаживания (последовательное пятиточечное и трехточечное сглаживание). Можно заметить, что сигнал имеет квазипрямоугольную форму, что обусловлено наличием высокочастотных компонент в сглаживаемом сигнале.

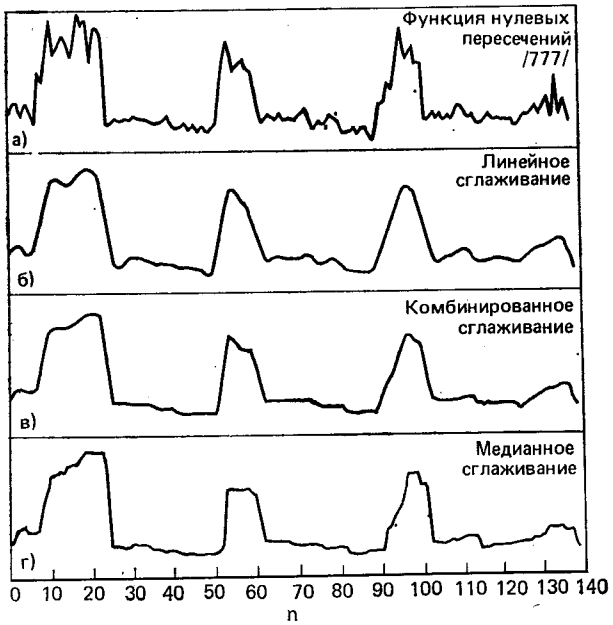

Рис. 4.40. Пример нелинейного сглаживания функции нулевых пересечений [21]

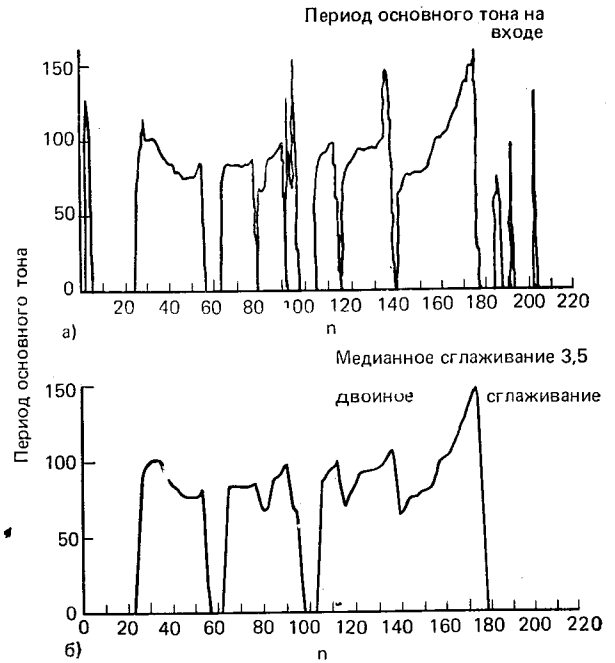

Рис. 4.41. Пример нелинейного сглаживания траектории основного тона [21]

Входной сигнал устроиства линейного сглаживания (КИХ-фильтр нижних частот 19-го порядка) изображен на рис. 4.40*б* и иллюстрирует «смазывание» сигнала при возникновении в нем резких переходов. На рис. 4.40*в* показан сигнал на выходе устройства при совместном использовании медианного сглаживания, описанного выше, и линейного сглаживания с трехточечным временны́м окном типа (4.45). В данном случае оценка хорошо следует за изменениями входного сигнала и бо́льшая часть шума устранена.

На рис. 4.41 показан пример использования нелинейного сглаживания для обработки траектории основного тона, содержащей ряд очевидных ошибок оценивания. Полезным свойством медианного сглаживания является возможность устранения отдельных ошибок в исходном сигнале совместно со сглаживанием траектории. Как это видно из рис. 4.41, применение совместного сглаживания позволяет устранить значительные ошибки при оценивании и адекватно сгладить сигнал, сохранив резкие переходы тон/шум.

4.10. Заключение

В гл. 4 рассмотрены некоторые характеристики речевого сигнала, полученные путем обработки во временно́й области. Мы подробно изучили методы обработки, так как они широко применяются при анализе речи, и надеемся, что подробное изложение будет способствовать их эффективному использованию. В данную главу включены также несколько примеров устройств обработки речевых сигналов на основе совместного использования функций кратковременной энергии, переходов через нуль и автокорреляционной функции речевого сигнала. Это сделано с тем, чтобы показать, как используются основные методы обработки сигналов при построении систем обработки речи.

ПРИЛОЖЕНИЕ. СОКРАЩЕНИЕ ОБЪЕМА ВЫЧИСЛЕНИЙ ПРИ РАСЧЕТЕ АВТОКОРРЕЛЯЦИОННОЙ ФУНКЦИИ

Вычисление K значений автокорреляционной функции по N-точечному окну требует по крайней мере KN операций умножения и сложения. Поскольку для многих практических приложений как K, так и N велики (например, $K=250$ и $N=401$), для уменьшения объема вычислений желательно использовать некоторые свойства автокорреляционной функции. Рассмотрим три метода, позволяющие сократить число арифметических операций при вычислении автокорреляционной функции.

Первый метод предложенный Бланкеншипом [22], основан на том, что при $m\neq 0$ большинство входных отсчетов используется при умножении дважды, т. е. для модифицированной автокорреляционной функции при $k=1$ получаем

$$\begin{aligned}\hat{R}_n(1) = &\sum_{m=0}^{N-1} x(m+n)\,x(m+n+1) = \\ &= x(n)\,x(n+1) + x(n+1)\,x(n+2) + \ldots + \\ &+ x(n+N-1)\,x(n+N) = \\ &= x(n+1)\,[x(n) + x(n+2)] + x(n+3)\,[x(n+2) + \\ &+ x(n+4)] + \ldots\end{aligned} \tag{П.1}$$

Таким образом, когда $k\neq 0$, можно, используя выражение (4.36), сократить число умножений без увеличения числа сложений. Формально автокорреляционная функция может быть записана в виде

$$\hat{R}(k) = B(k) + C(k) \tag{П.2}$$

(для упрощения индекс n опущен), где

$$\underset{\text{четная компонента}}{N = 2qk} \qquad \underset{\text{произвольная компонента}}{+ak} \qquad + b \tag{П.3}$$

Здесь $a=0$ или 1 и b находится в области

$$0 \leqslant b < k. \tag{П.4}$$

В уравнении (П2)

$$B(k) = \sum_{j=0}^{q-1} \sum_{i=1}^{k} x(2jk+i+k)\,[x(2jk+i) + x(2jk+i+2k)], \tag{П.5}$$

и если $a=0$, то

$$C(k) = \sum_{i=1}^{b} x(2qk+i)\,x(2qk+i+k) \tag{П.6}$$

или, если $a=1$, то

$$\begin{aligned}C(k) = &\sum_{i=1}^{b} x(2qk+i+k)\,[x(2qk+i) + x(2qk+i+2k)] + \\ &+ \sum_{i=b+1}^{k} x(2qk+i)\,x(2qk+i+k).\end{aligned} \tag{П.7}$$

Например, рассмотрим $N=60$ с $k=6$, 7 и 8. Величины q, a и b в (П3) равны:

k	q	a	b
6	5	0	0
7	4	0	4
8	3	1	4

Поскольку значения q, a и b получены, то можно использовать уравнение (П2) и (П5)—(П7) для вычисления $\hat{R}(k)$. Легко показать, что число умножений, необходимых для вычисления $\hat{R}(k)$, удовлетворяет неравенству.

$$N_M < \frac{1}{2}(N+k). \tag{П.8}$$

Таким образом, при $k \ll N$ этот метод дает примерно двукратное сокращение числа умножений. Если вычислять небольшое число значений автокорреляционной функции (например, при линейном предсказании методами гл. 8), этот метод очень эффективен. Например, Бланкеншип показал, что если $K=12$ и $N=128$, то при прямом вычислении автокорреляционной функции потребуется 1664 умножения ($N_1(K+1)$), в то время как при использовании модифицированной процедуры требуется лишь 912 умножений, т. е. в 1,825 раз меньше. Если требуется вычислить много значений автокорреляционной функции, как в примерах § 4.6, этот метод дает незначительную экономию времени.

Модификация алгоритма предложена Кендаллом [23] в случае определения автокорреляционной функции в соответствии с (4.30). Обозначая взвешенный речевой сигнал $x(n)w(n)$ через $\hat{x}(n)$ и опуская индекс n, перепишем (4.30):

$$R(k) = \sum_{m=0}^{N-1-k} \hat{x}(m)\,\hat{x}(m+k), \tag{П.9}$$

что можно представить, полагая N четным, в виде

$$R(k)=\sum_{m=0}^{(N-k)/2-1}[\hat{x}(2m)+\hat{x}(2m+k+1)][\hat{x}(2m+1)+\hat{x}(2m+k)]-$$
$$-A(k)-B(k) \qquad (\text{П.10})$$

при четном k,

$$R(k)=\sum_{m=0}^{(N-k-1)/2-1}[\hat{x}(2m)+\hat{x}(2m+k+1)][\hat{x}(2m+1)+\hat{x}(2m+k)]-$$
$$-A(k)-B(k)+\hat{x}(N-1-k)\,\hat{x}(N-1), \qquad (\text{П.11})$$

при нечетном k, где $A(k)$ и $B(k)$ получаются с помощью рекурсивных соотношений

$$A(k)=A(k+2)+\hat{x}(N-2-k)\,\hat{x}(N-1-k) \qquad (\text{П.12})$$

при четном k с начальным условием $A(N)=0$,

$$A(k)=A(k+1) \qquad (\text{П.13})$$

при нечетном k и

$$B(k)=B(k+2)+\hat{x}(k)\,\hat{x}(k+1) \qquad (\text{П.14})$$

при четном k с начальным условием $B(N)=0$ и

$$B(k)=B(k+2)+\hat{x}(k)\,\hat{x}(k+1) \qquad (\text{П.15})$$

при нечетном k с начальным условием $B(N-1)=0$. **Уравнения (П10) и (П11)** показывают, что число умножений, необходимое для вычисления $R(k)$, примерно равно $(N-k-1)/2$, т. е. составляет половину от обычно требуемого числа, однако число сложений увеличилось примерно на 50%. Легко видеть, что число умножений в данном случае сократилось для любых k, а не только для $k \ll N$, как в предыдущем случае.

Третий метод ускорения вычислений автокорреляционной функции заключается в применении быстрого преобразования Фурье (БПФ), в котором автокорреляционная функция определяется как обратное преобразование Фурье спектральной плотности мощности последовательности сигнала (т. е. квадрата модуля спектра сигнала) [1, 2, 24]. При этом методе требуется двукратное применение БПФ и операция возведения в квадрат. Для того чтобы избежать наложения при вычислении автокорреляционной функции, применяется 2 N-точечное дискретное преобразование Фурье (вычисляемое с помощью БПФ), в котором N-точечная последовательность заполняется нулями. Процесс вычисления квадрата модуля спектра требует примерно $2N$ умножений, а для получения $2N$ точек БПФ необходимо осуществить $2N\log_2(2N)$ умножений для вычисления всех N значений автокорреляционной функции. Таким образом, использование БПФ для вычисления автокорреляции предполагает выполнение следующего количества операций умножения:

$$N_F=2\cdot 2N\log_2(2N)+2N. \qquad (\text{П.16})$$

Кендалл [23] показал, что прямое модифицированное вычисление автокорреляционной функции при $N \leqslant 256$ более эффективно, чем применение БПФ с точки зрения количества операций умножения. Если при вычислении эффективности учитывать и операции сложения, то прямой метод более эффективен при $N \leqslant 128$.

Задачи

4.1. Прямоугольное окно определяется выражением

$$w_R(n)=\begin{cases}1, & 0\leqslant n\leqslant N-1;\\ 0, & \textit{в противном случае.}\end{cases}$$

Окно Хемминга определяется выражением

$$w_H(n)=\begin{cases}0{,}54-0{,}46\cos[2\pi n/(N-1)], & 0\leqslant n\leqslant N-1;\\ 0, & \textit{в противном случае.}\end{cases}$$

а) Показать, что преобразование Фурье прямоугольного окна имеет вид

$$w_R(e^{i\omega})=\frac{\sin(\omega N/2)}{\sin(\omega/2)}e^{-i\omega(N-1)/2}.$$

б) Изобразить $W_R(e^{i\omega})$ как функцию ω, опуская множитель $e^{-i\omega(N-1)/2}$ с линейной фазой.

в) Выразить $w_H(n)$ через $w_R(n)$ и таким образом получить выражение $W_H(e^{i\omega})$ через $W_R(e^{i\omega})$.

г) Изобразить отдельные члены в $W_H(e^{i\omega})$ (опуская общий для них множитель с линейной фазой.) Рисунок должен иллюстрировать изменения в частотной характеристике окна Хемминга, приводящие к улучшению подавления высших частот.

4.2. Кратковременная энергия последовательности определяется в виде

$$E_n=\sum_{m=-\infty}^{\infty}[x(m)\,w(n-m)]^2.$$

В случае выбора

$$w(m)=\begin{cases}a^m, & m\geqslant 0;\\ 0, & m<0\end{cases}$$

можно получить рекуррентную формулу для E_n.

а) Вывести рекуррентное разностное уравнение для E_n через E_{n-1} и входной сигнал $x(n)$.

б) Изобразить схему цифрового фильтра, соответствующую полученному уравнению.

в) Показать, каким основным свойством должна обладать последовательность $h(m)=w^2(m)$, чтобы можно было получить рекуррентное уравнение.

4.3. Кратковременная энергия сигнала определяется выражением

$$E_n=\sum_{m=-N}^{N}h(m)\,x^2(n-m).$$

Предположим, что мы хотим вычислять E_n по каждому отсчету входного сигнала.

а) Пусть $h(m)$ имеет вид:

$$h(m)=\begin{cases}a^{|m|}, & |m|\leqslant N;\\ 0, & |m|>N.\end{cases}$$

Получить рекуррентное (разностное) уравнение для E_n.

б) Какой выигрыш по числу умножений достигается применением рекуррентных соотношений вместо непосредственного вычисления E_n?

в) Изобразить схему цифрового фильтра, соответствующего рекуррентной формуле для E_n. (Поскольку $h(m)$ нереализуема, следует предусмотреть соответствующую задержку.)

4.4. Предположим, что среднее значение модуля оценивается для каждых L отсчетов входного сигнала. Один из способов состоит в использовании окна конечной длительности

$$M_n=\sum_{m=n-N+1}^{n}|x(m)|\,w(n-m).$$

В этом случае M_n вычисляется только один раз для каждых входных отсчетов. Другой подход состоит в применении окна, для которого сложно получить ре-

куррентную формулу, например $M_n = aM_{n-1} + |x(n)|$. В этом случае M_n необходимо вычислять по каждому отсчету, даже если требуется получить его только для каждых L отсчетов.

а) Определить, сколько операций умножения и сложения требуется для вычисления среднего значения модуля для каждых L отсчетов с использованием окна конечной длительности.

б) Повторить вычисления п. а) для рекуррентного случая.

в) В каких случаях окно конечной длительности оказывается более эффективным с этой точки зрения.

4.5. Среднее число переходов через нуль определяется уравнениями (4.18) и (4.20): $z_n = \frac{1}{2N} \sum_{m=n-N+1}^{n}{}' |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]|$. Показать, что z_n можно представить в виде

$$z_n = z_{n-1} + \frac{1}{2N} \{|\operatorname{sgn}[x(n)] - \operatorname{sgn}[x(n-1)]| - |\operatorname{sgn}[x(n-N)] - \operatorname{sgn}[x(n-N-1)]|\}.$$

4.6. Чтобы показать, как с помощью параллельной обработки добиться высокой точности выделения основного тона, рассмотрите следующий идеализированный случай. Предположим, что имеется семь выделителей основного тона; вероятность правильного оценивания периода равна p, а вероятность ошибки $1-p$. Эти оценки обрабатываются таким образом, что полная ошибка возникает тогда, когда период ошибочно оценен в четырех или более выделителях основного тона.

а) Получить явное выражение для вероятности ошибки при параллельном оценивании основного тона. (Указание: рассмотреть результат работы каждого выделителя в рамках схемы Бернулли с вероятностью ошибки $1-p$ и вероятностью правильного решения p.)

б) Изобразить графически зависимость полной вероятности ошибки от p.

в) Определить, при каком значении p полная вероятность ошибки менее 0,05?

4.7. В соответствии с (4.24) кратковременная автокорреляционная функция определяется выражением

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)\, w(n-m)\, x(m+k)\, w(n-k-m).$$

а) Показать, что $R_n(k) = R_n(-k)$, т. е. что $R_n(k)$ — четная функция k.

б) Показать, что $R_n(k)$ можно представить в виде

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)\, x(m-k)\, h_k(n-m),$$

где $h_k(n) = w(n)\, w(n+k)$.

в) Пусть

$$w(n) = \begin{cases} a^n, & n \geqslant 0; \\ 0, & n < 0. \end{cases}$$

Определить импульсную характеристику $h_k(n)$ для вычисления k-го значения корреляционной функции.

г) Определить z-преобразование $h_k(n)$ в п. в) и из него получить рекуррентное уравнение для $R_n(k)$. Изобразить схему цифрового фильтра для вычисления $R_n(k)$ как функции n при временно́м окне, приведенном в п.в).

д) Определить то же, что в п.в) и г) для случая

$$w(n) = \begin{cases} na^n, & n \geqslant 0; \\ 0, & n < 0. \end{cases}$$

4.8. Пусть дана периодическая импульсная последовательность

$$x(m) = \sum_{r=-\infty}^{\infty} \delta(m - rP).$$

а) Используя уравнение (4.30) с прямоугольным окном, длина которого N удовлетворяет соотношению $QP < N-1 < (Q+1)P$, где Q — целое, определить и изобразить $R_n(k)$ для $0 \leqslant k \leqslant N-1$.

б) Как изменится результат п.а), если в качестве окна использовать окно Хемминга той же длины?

в) Определить и построить кратковременную модифицированную корреляционную функцию $\hat{R}_n(k)$, задаваемую для тех же значений N.

4.9. Корреляционная функция случайного или периодического сигнала определяется выражением

$$\Phi(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{m=-N}^{N} x(m)\, x(m+k).$$

Кратковременная автокорреляционная функция определяется выражением

$$R_n(k) = \sum_{m=0}^{N-|k|-1} x(n+m)\, w'(m)\, x(n+m+k)\, w'(m+k).$$

Модифицированная кратковременная автокорреляционная функция равна

$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)\, x(n+m+k).$$

Показать, являются ли справедливыми следующие соотношения:

Если $x(n) = x(n+P)$, $-\infty < n < \infty$, то

а)
$$\Phi(k) = \Phi(k+P), \qquad -\infty < k < \infty;$$
$$R_n(k) = R_n(k+P), \qquad -(N-1) \leqslant k \leqslant N-1$$
$$\hat{R}_n(k) + \hat{R}_n(k+P), \qquad -(N-1) \leqslant k \leqslant N-1.$$

б)
$$\Phi(-k) = \Phi(k), \qquad -\infty < k < \infty;$$
$$R_n(-k) = R_n(k). \qquad -(N-1) \leqslant k \leqslant N-1;$$
$$\hat{R}_n(-k) = \hat{R}_n(k), \qquad -(N-1) \leqslant k \leqslant N-1.$$

в)
$$\Phi(k) \leqslant \Phi(0), \qquad -\infty < k < \infty;$$
$$R_n(k) \leqslant R_n(0), \qquad -(N-1) \leqslant k \leqslant N-1;$$
$$\hat{R}_n(k) \leqslant \hat{R}_n(0), \qquad -(N-1) \leqslant k \leqslant N-1.$$

г) $\Phi(0)$ равна мощности сигнала; $R_n(0)$ равна кратковременной энергии сигнала; $\hat{R}_n(0)$ равна кратковременной энергии сигнала.

4.10. Рассмотрим сигнал $x(n) = \cos \omega_0 n$, $-\infty < n < \infty$.

а) Определить корреляционную функцию для $x(n)$ по (4.21).

б) Изобразить $\Phi(k)$ как функцию от k.

г) Рассчитать и построить автокорреляционную функцию сигнала

$$y(n) = \begin{cases} 1, & x(n) \geqslant 0; \\ 0, & x(n) < 0. \end{cases}$$

4.11. Кратковременная функция среднего разности значений (КФСР) сигнала $x(n)$ определяется выражением [см. (4.38)]

$$\gamma_n(k) = \frac{1}{N}\sum_{m=0}^{N-1} |x(n+m) - x(n+m-k)|.$$

а) Используя неравенство [16]

$$\frac{1}{N}\sum_{m=0}^{N-2} |x(m)| \leqslant \left[\frac{1}{N}\sum_{m=0}^{N-1} |x(m)|^2\right]^{1/2},$$

показать, что $\gamma_n(k) \leqslant [2(\hat{R}_n(0) - \hat{R}_n(k))]^{1/2}$. Этот результат приводит к равенству (4.39).

б) Изобразить $\gamma_n(k)$ и величину $[2(\hat{R}_n(0) - \hat{R}_n(k))]^{1/2}$ при $0 \leqslant k \leqslant 200$ для сигнала $x(n) = \cos(\omega_0 n)$ с $N = 200$, $\omega_0 = 200\pi/(10\,000)$.

4.12. Рассмотрим входной сигнал $x(n) = A\cos(\omega_0 n)$ трехуровневого центрального ограничителя вида

$$y(n) = \begin{cases} 1, & x(n) > C_L; \\ 0, & |x(n)| \leqslant C_L; \\ -1, & x(n) < -C_L. \end{cases}$$

а) Изобразить $y(n)$ как функцию n для $C_L = 0{,}5A$, $0{,}75A$ и A.

б) Изобразить автокорреляционные функции для $y(n)$ и значений C_L из п.а).

в) Обсудить влияние взаимного расположения C_L и A. Пусть A изменяется во времени, удовлетворяя неравенству $0 < A(n) \leqslant A_{max}$. Рассмотреть ситуацию, которая может возникнуть при C_L, близком к A_{max}.

5
Цифровое представление речевых сигналов

5.0. Введение

«Если бы я смог заставить поток электричества изменяться по интенсивности точно в соответствии с изменением плотности воздуха во время распространения в нем звука, я бы смог передавать по телеграфу любые звуки, даже звуки речи» — А. Г. Белл [1]. Эта простая идея, имеющая столь важное значение для истории связи, кажется сегодня очевидной. Принцип, изложенный в открытии Белла, положен в основу множества устройств и систем, предназначенных для записи, передачи или обработки речевых сигналов и в которых речевой сигнал отражает колебания плотности звуковых (речевых) волн. Это относится и к цифровым системам, в которых речевой сигнал представлен последовательностью своих мгновенных значений.

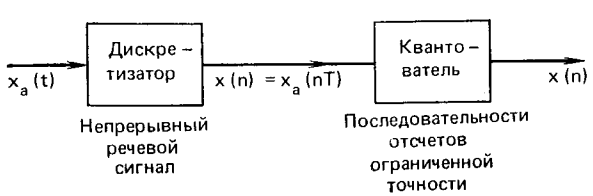

Рис. 5.1. Общая схема цифрового представления

Общая схема цифрового представления речевого сигнала изображена на рис. 5.1. Из рисунка следует, что речевое колебание как непрерывная функция времени подвергается дискретизации, чаще всего периодической, в результате которой образуется последовательность отсчетов $x_a(nT)$. Эти отсчеты могут в общем случае принимать непрерывное множество значений. Поэтому для получения цифрового, т. е. дискретного по амплитуде и по времени, представления необходимо проквантовать каждый отсчет до конечного множества значений.

Мы увидим далее, что рис. 5.1 достаточно полно отражает процесс формирования цифрового представления речевого сигнала. Может быть не во всех случаях можно разделить эту процедуру на два отдельных этапа, но основные операции — дискретизация и квантование — свойственны всем методам, приведенным в данной главе.

В начале главы изложены вопросы дискретизации применительно к речевым сигналам. Далее излагаются методы квантования отсчетов речевого колебания.

5.1. Дискретизация речевых сигналов

Теорема дискретизации уже обсуждалась в гл. 2. Последовательность отсчетов сигнала, как показано в гл. 2, единственным образом описывает аналоговый сигнал, если он ограничен по полосе частот и частота дискретизации по крайней мере вдвое больше наивысшей частоты спектра сигнала. Поскольку нас интересует цифровое представление речевых сигналов, изучим спектральные свойства речи. В соответствии с изложенным в гл. 3 описанием гласных и фрикативных звуков речевой сигнал не ограничен по полосе частот, хотя его спектр быстро спадает в области высоких частот. На рис. 5.2 изображены спектры типичных звуков речи.

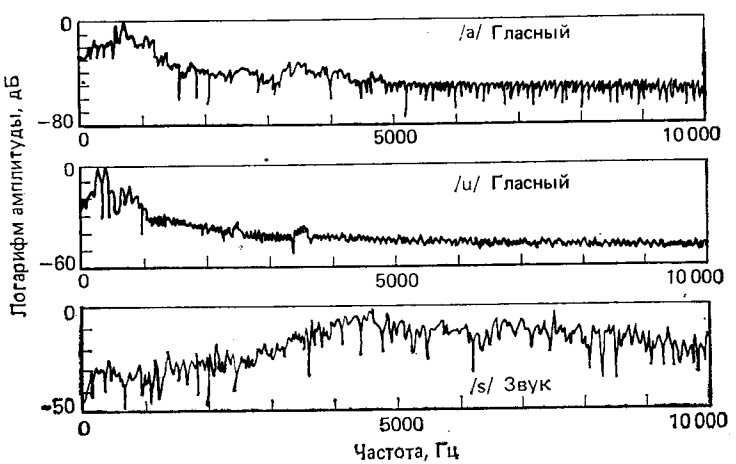

Рис. 5.2. Спектры вокализованных звуков /a/ и /u/ и невокализованного /s/ при частоте дискретизации 20 кГц

Видно, что для вокализованных звуков наивысшая частота, ниже которой максимумы спектра меньше уровня 40 дБ, составляет около 4 кГц. С другой стороны, для невокализованных звуков