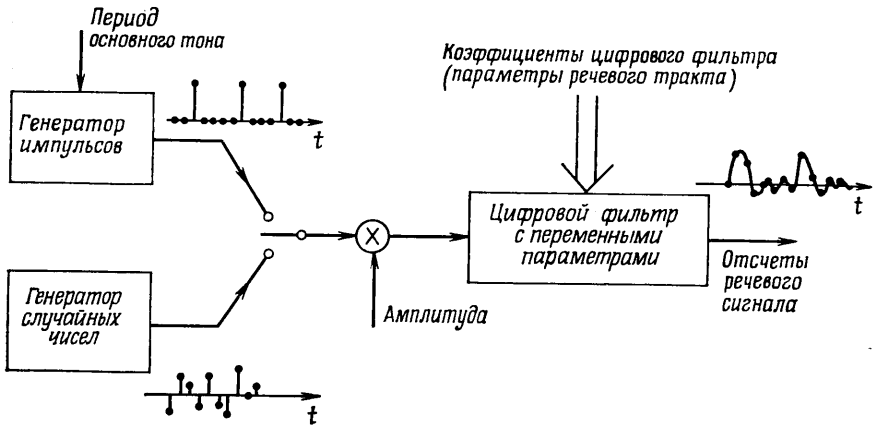


Фиг. 12.5. Цифровая модель образования речи (по Шаферу).

В большинстве случаев это предположение вполне допустимо. Однако в некоторых случаях (например, для глухих взрывных звуков, таких, как *p* в слове pot) оно неверно, и основная модель образования речи становится непригодной. В большей части данной главы будем считать, что предположение о независимости источника и тракта справедливо. В этом случае можно построить простую цифровую модель образования речи (фиг. 12.5). Источниками возбуждения служат генератор импульсов с внешней синхронизацией с периодом основного тона, а также генератор случайных чисел. Генератор импульсов через каждые N_0 отсчетов вырабатывает импульс, соответствующий очередной порции воздуха. Интервал между импульсами называется периодом основного тона. Он равен величине, обратной частоте следования порций воздуха или частоте колебания голосовых связок. Выходная последовательность генератора случайных чисел имитирует и квазислучайный турбулентный поток, и спад давления при образовании глухих звуков.

Каждый из источников (или оба) может быть соединен со входом линейного цифрового фильтра с переменными параметрами, моделирующего голосовой тракт. При этом коэффициенты фильтра отражают свойства голосового тракта в зависимости от времени при непрерывной речи. В среднем через каждые 10 мс коэффициенты фильтра изменяются, отражая тем самым изменение состояния голосового тракта.

Регулировка усиления, введенная между источниками и фильтром, позволяет управлять громкостью выходного сигнала. Последовательность на выходе фильтра эквивалентна речевому сигналу, дискретизованному с соответствующей частотой.

Для управления такой моделью необходимо знать зависимость соответствующих параметров (частоты основного тона, положения переключателя, громкости и коэффициентов фильтра) от времени. Основной задачей почти всех систем анализа речи является оценка параметров модели по реальной речи. Задача большинства систем синтеза речи состоит в том, чтобы, используя эти параметры, полученные некоторым способом, образовать искусственный речевой сигнал, неотличимый на слух от настоящей речи. В системах анализа—синтеза эти две задачи решаются совместно с общей целью увеличения *эффективности* (т. е. понижения частоты дискретизации в системе синтеза до величины, меньшей, чем при обычном представлении речевых сигналов) и *гибкости* (т. е. возможности изменять речь некоторым желаемым образом путем управления параметрами модели). В последующих разделах этой главы обсуждаются различные аспекты нескольких систем, разработанных с учетом этих соображений.

12.3. Кратковременный спектральный анализ

Преобразование Фурье последовательности $x(nT)$, $-\infty < n < \infty$, определяется как

$$X(e^{j\omega T}) = \sum_{n=-\infty}^{\infty} x(nT)\, e^{-j\omega nT}. \tag{12.2}$$

Как было показано в гл. 6, для нестационарных сигналов типа речевых сигналов преобразование Фурье не имеет смысла, так как спектр речи изменяется во времени. Более полезной характеристикой распределения энергии речевого сигнала является преобразование Фурье на коротком интервале, определяемое как

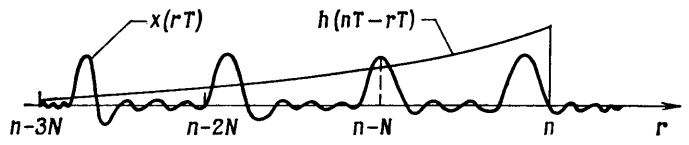
$$X(\omega, nT) = \sum_{r=-\infty}^{n} x(rT)\, h(nT - rT)\, e^{-j\omega rT}. \tag{12.3}$$

Равенство (12.3) можно рассматривать как фурье-преобразование речевого сигнала на бесконечном интервале, если выделить вблизи момента времени nT участок конечной длины с помощью весовой функции («временно́го окна») вида $h(nT)$ (фиг. 12.6). Используя свертку, равенство (12.3) можно записать иначе:

$$X(\omega, nT) = [x(nT)\, e^{-j\omega nT}] * h(nT). \tag{12.4}$$

Левую часть равенства (12.3) можно представить в виде

$$X(\omega, nT) = a(\omega, nT) - jb(\omega, nT), \tag{12.5}$$



Фиг. 12.6. Представление кратковременного спектрального анализа.

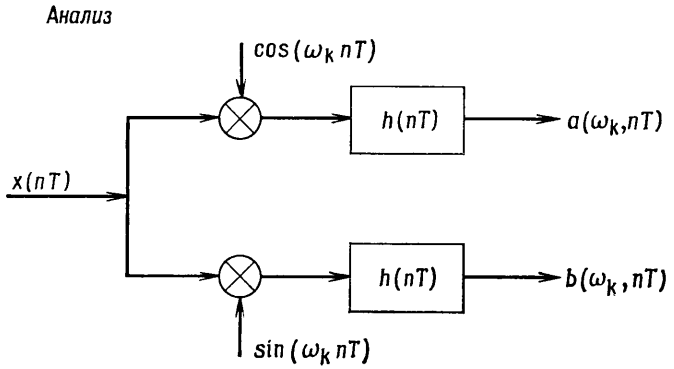
где $a(\omega, nT)$ и $b(\omega, nT)$ — действительная и мнимая части кратковременного фурье-преобразования, равные

$$a(\omega, nT) = \sum_{r=-\infty}^{n} x(rT)\, h(nT - rT) \cos \omega r T, \qquad (12.6а)$$

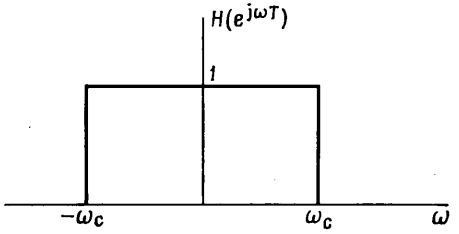
$$b(\omega, nT) = \sum_{r=-\infty}^{n} x(rT)\, h(nT - rT) \sin \omega r T. \qquad (12.6б)$$

Из этих формул вытекает простой способ измерения кратковременных преобразований, который иллюстрируется на фиг. 12.7. Обычно $H(e^{j\omega T})$, преобразование Фурье от $h(nT)$, выбирают таким образом, чтобы аппроксимировать идеальный фильтр нижних частот с частотой среза ω_c, показанный на фиг. 12.8. Тогда $X(\omega, nT)$ соответствует энергии речевого колебания на частоте ω в момент времени nT. Точнее, энергия измеряется в полосе частот от $\omega - \omega_c$ до $\omega + \omega_c$.

В большинстве систем для спектрального анализа речи кратковременное преобразование желательно измерять на N частотах, которые обычно располагаются в полосе $0 \leqslant \omega T \leqslant 2\pi$ равномерно. С этой целью описанные выше измерения проводятся для каждой из N частот. Если $h(nT)$ является импульсной характеристикой



Фиг. 12.7. Простой метод анализа речевого сигнала, основанный на кратковременном спектральном анализе.



Фиг. 12.8. Идеальный фильтр нижних частот для кратковременного спектрального анализа.

КИХ-фильтра, а частоты распределены равномерно, одновременные измерения могут быть выполнены весьма эффективно с применением алгоритма БПФ. Чтобы показать это, положим, что $h(nT)$ отлично от нуля при $0 \leqslant n \leqslant M - 1$ и что центральные частоты анализа ω_k равны

$$\omega_k = \frac{2\pi}{NT} k, \qquad k = 0, 1, \ldots, N-1. \qquad (12.7)$$

Тогда (12.3) можно переписать следующим образом:

$$X(\omega_k, nT) = \sum_{r=n-M+1}^{n} x(rT)\, h(nT - rT)\, e^{-j\omega_k rT} = \qquad (12.8а)$$

$$= \sum_{m=0}^{[M/N]+1} \sum_{r=n-(m+1)N+1}^{n-mN} x(rT)\, h(nT - rT)\, e^{-j\omega_k rT}, \qquad (12.8б)$$

где $[M/N]$ означает целую часть от M/N. Положив $l = n - mN - r$, получим

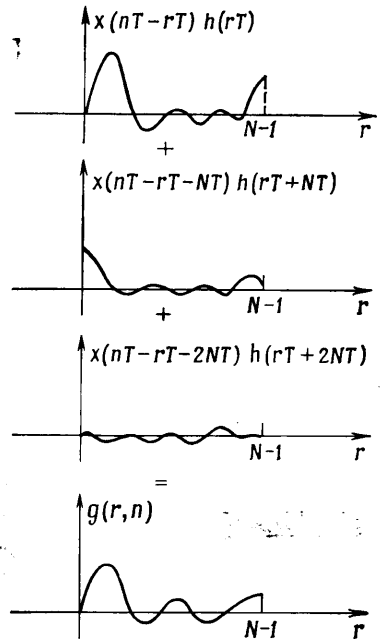
$$X(\omega_k, nT) = \sum_{m=0}^{[M/N]+1} \sum_{l=0}^{N-1} x(nT - rT - mNT)\, h(lT + mNT)\, e^{j\omega_k (l-n+mN)T}. \qquad (12.9)$$

Подстановка ω_k из (12.7) дает

$$X(\omega_k, nT) = e^{-j(2\pi/N)kn} \sum_{l=0}^{N-1} \left[\sum_{m=0}^{[M/N]+1} x(nT - lT - mNT)\, h(lT + mNT)\, e^{j(2\pi/N)kl} \right]. \qquad (12.10)$$

Здесь $e^{j2\pi m}$ заменено единицей. Формулу (12.10) можно переписать в виде

$$X(\omega_k, nT) = e^{-j(2\pi/N)kn} \underbrace{\sum_{l=0}^{N-1} g(l, n)\, e^{j(2\pi/N)lk}}_{\text{ДПФ}}, \qquad (12.11)$$



Фиг. 12.9. Формирование $g\ (r,\ n)$ из $x\ (nT)$ и $h\ (nT)$.

где

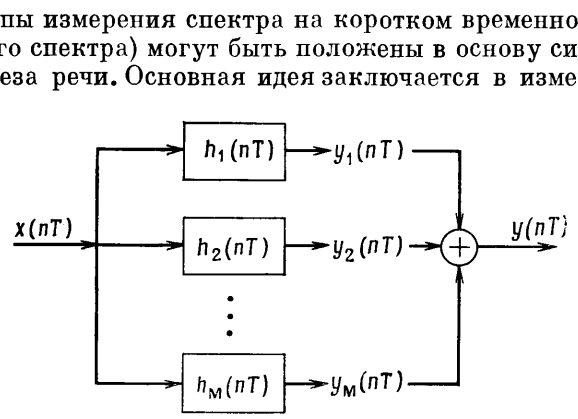
$$g\ (l,\ n) = \sum_{m=0}^{[M/N]+1} x\ (nT - lT - mNT)\ h\ (lT + mNT). \quad (12.12)$$

Соотношение (12.11) показывает, что $X\ (\omega_k,\ nT)$ можно получить, перемножив последовательность $e^{-j\,(2\pi/N)\,kn}$ и ДПФ последовательности $g\ (l,\ n)$. На фиг. 12.9 иллюстрируется процесс почленного получения последовательности $g\ (r,\ n)$ из исходных последовательностей $x\ (rT)$ и $h\ (rT)$.

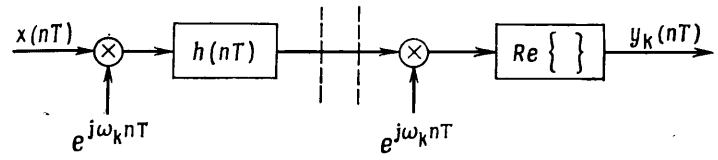
Таким образом, кратковременный фурье-анализ речевых сигналов может быть выполнен либо непосредственно с использованием гребенки цифровых фильтров, либо косвенно с применением БПФ.

12.4. Система анализа — синтеза речи, основанная на кратковременном спектральном анализе

Принципы измерения спектра на коротком временном интервале (текущего спектра) могут быть положены в основу системы анализа—синтеза речи. Основная идея заключается в измерении сигналов на выходах гребенки из M полосовых фильтров и восстановлении речи по этим M сигналам. В упрощенной схеме такой системы (фиг. 12.10) входным речевым сигналом является $x\ (nT)$, а синтезированным колебанием $y\ (nT)$. M полосовых фильтров имеют импульсные характеристики $h_k\ (nT)$, $k = 1, 2, \ldots, M$. Последовательности на выходах полосовых фильтров обозначены через $y_k\ (nT)$, $k = 1, 2, \ldots, M$. Если рассматривать только импульсные характеристики полосовых фильтров вида



Фиг. 12.10. Схема системы анализа — синтеза, основанной на кратковременном спектральном анализе.



Фиг. 12.11. Обработка, выполняемая в k-м канале.

$$h_k\ (nT) = h\ (nT) \cos\ (\omega_k nT), \quad (12.13)$$

где $h\ (nT)$ — импульсная характеристика фильтра нижних частот (т. е. импульсная характеристика полосовых фильтров равна промодулированной импульсной характеристике фильтра нижних частот), то последовательности на выходах полосовых фильтров будут равны

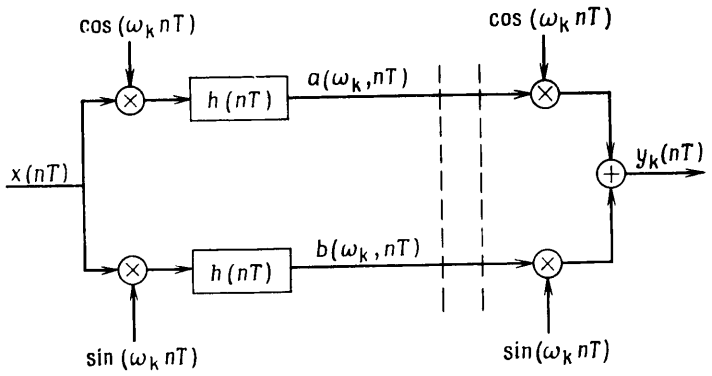
$$y_k\ (nT) = \sum_{r=-\infty}^{n} x\ (rT)\ h\ (nT - rT) \cos\ [\omega_k\ (nT - rT)] = \quad (12.14)$$

$$= \mathrm{Re}\ [e^{j\omega_k nT}\ X\ (\omega_k,\ nT)]. \quad (12.15)$$

Здесь $X\ (\omega_k,\ nT)$ определены формулой (12.3). Таким образом, каждый канал этой системы может быть построен по схеме, приведенной на фиг. 12.11. Поскольку в $X\ (\omega_k,\ nT)$ можно выделить действительную и мнимую части [см. (12.5)], то равенство (12.15) приводится к виду

$$y_k\ (nT) = a\ (\omega_k,\ nT) \cos\ (\omega_k nT) + b\ (\omega_k,\ nT) \sin\ (\omega_k nT). \quad (12.16)$$

Соответствующая схема представлена на фиг. 12.12. Пунктирными линиями отмечены точки передачи и приема, если система предназначена для сжатия полосы речевого сигнала. Сплошные линии, заключенные между пунктирными, обозначают канал связи (считается, что он не вносит ошибок). Чтобы добиться сужения полосы, передаваемые параметры $a\ (\omega_k,\ nT)$ и $b\ (\omega_k,\ nT)$ следует квантовать и дискретизовать с меньшей частотой, чем при передаче речевых сигналов. Дальнейшее обсуждение этого вопроса содержится в разд. 12.6.



Фиг. 12.12. Обработка, выполняемая в k-м канале при использовании действительных чисел.

12.5. Особенности анализа речи

Качество представления речи рассматриваемой системой зависит от того, насколько полно гребенка из M фильтров представляет спектр речевого сигнала. Простой способ оценки качества состоит в определении импульсной характеристики всей системы и анализе ее преобразования Фурье. Если обозначить импульсную характеристику гребенки фильтров через $\tilde{h}(nT)$, то

$$\tilde{h}(nT) = \sum_{k=1}^{M} h_k(nT) = h(nT) \sum_{k=1}^{M} \cos(\omega_k nT). \qquad (12.17)$$

Обозначив сумму косинусов в (12.17) через $d(nT)$, т. е.

$$d(nT) = \sum_{k=1}^{M} \cos(\omega_k nT), \qquad (12.18)$$

получим

$$\tilde{h}(nT = h(nT)\, d(nT), \qquad (12.19)$$

т. е. импульсная характеристика гребенки фильтров равна *произведению* импульсной характеристики ФНЧ-прототипа и функции, зависящей только от числа фильтров M и их центральных частот ω_k.

Чтобы понять насколько хорошо $\tilde{h}(nT)$ аппроксимирует единичный импульс (возможно, с некоторой задержкой), можно проанализировать либо саму характеристику $\tilde{h}(nT)$, либо ее преобразование Фурье. В частном случае равномерного расположения фильтров гребенки по частоте, когда

$$\omega_k = \Delta\omega k \qquad (12.20)$$

($\Delta\omega$ постоянно), $d(nT)$ можно найти, вычислив сумму (12.18)

$$d(nT) = \frac{1}{2}\left[\sum_{k=-M}^{M} e^{jk\Delta\omega nT} - 1\right], \qquad (12.21)$$

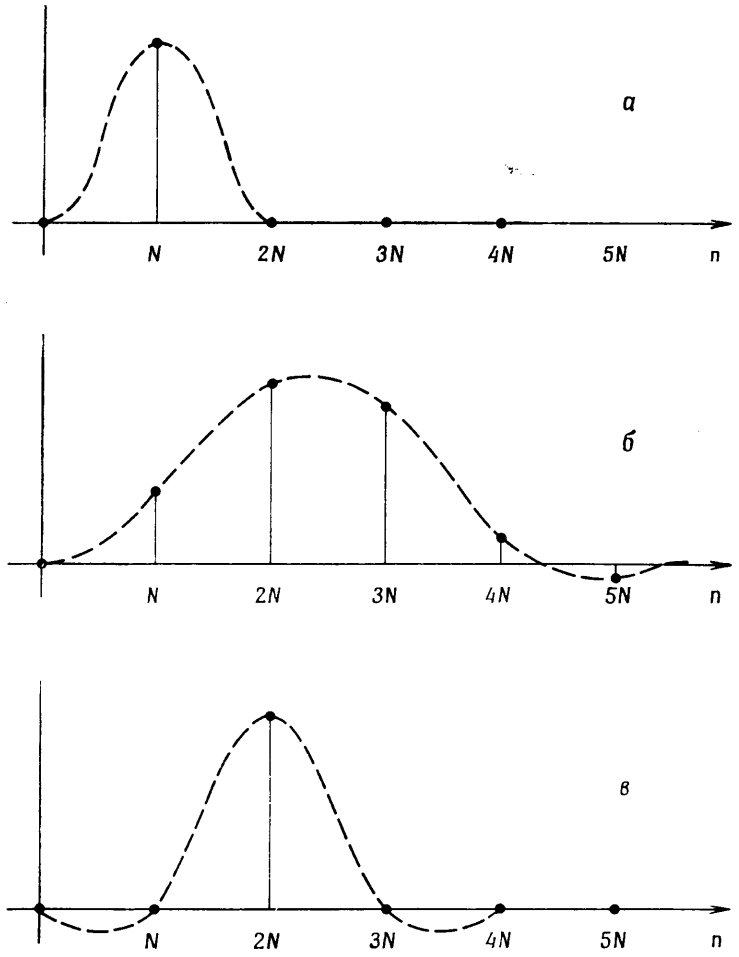
$$d(nT) = \frac{1}{2}\left[\frac{\sin\left[\left(M+\frac{1}{2}\right)\Delta\omega nT\right]}{\sin\left[(\Delta\omega/2)\, nT\right]} - 1\right]. \qquad (12.22)$$

Если $\Delta\omega = 2\pi/NT$, где N — целое, то последовательность $d(nT)$ периодична с периодом в N отсчетов. Если же отношение $2\pi/\Delta\omega T$ не равно целому числу, то последовательность $d(nT)$ непериодична, но имеет пики, следующие через NT секунд.

Особенно интересен случай, когда N — целое и нечетное (аналогичные результаты можно получить для четных N), а $M = (N-1)/2$. Поскольку $\Delta\omega = 2\pi/NT$, то ясно, что этот случай соответствует измерению кратковременного преобразования Фурье на частотах, расположенных равномерно в диапазоне $0 < \omega < \pi/T$. Если в гребенку фильтров ввести также канал с центром на нулевой частоте, то можно показать, что

$$d(nT) = \frac{\sin(\pi n)}{\sin(\pi n/N)} = \begin{cases} N, & n = 0, \pm N, \pm 2N, \ldots, \\ 0 & \text{при других } n. \end{cases} \qquad (12.23)$$

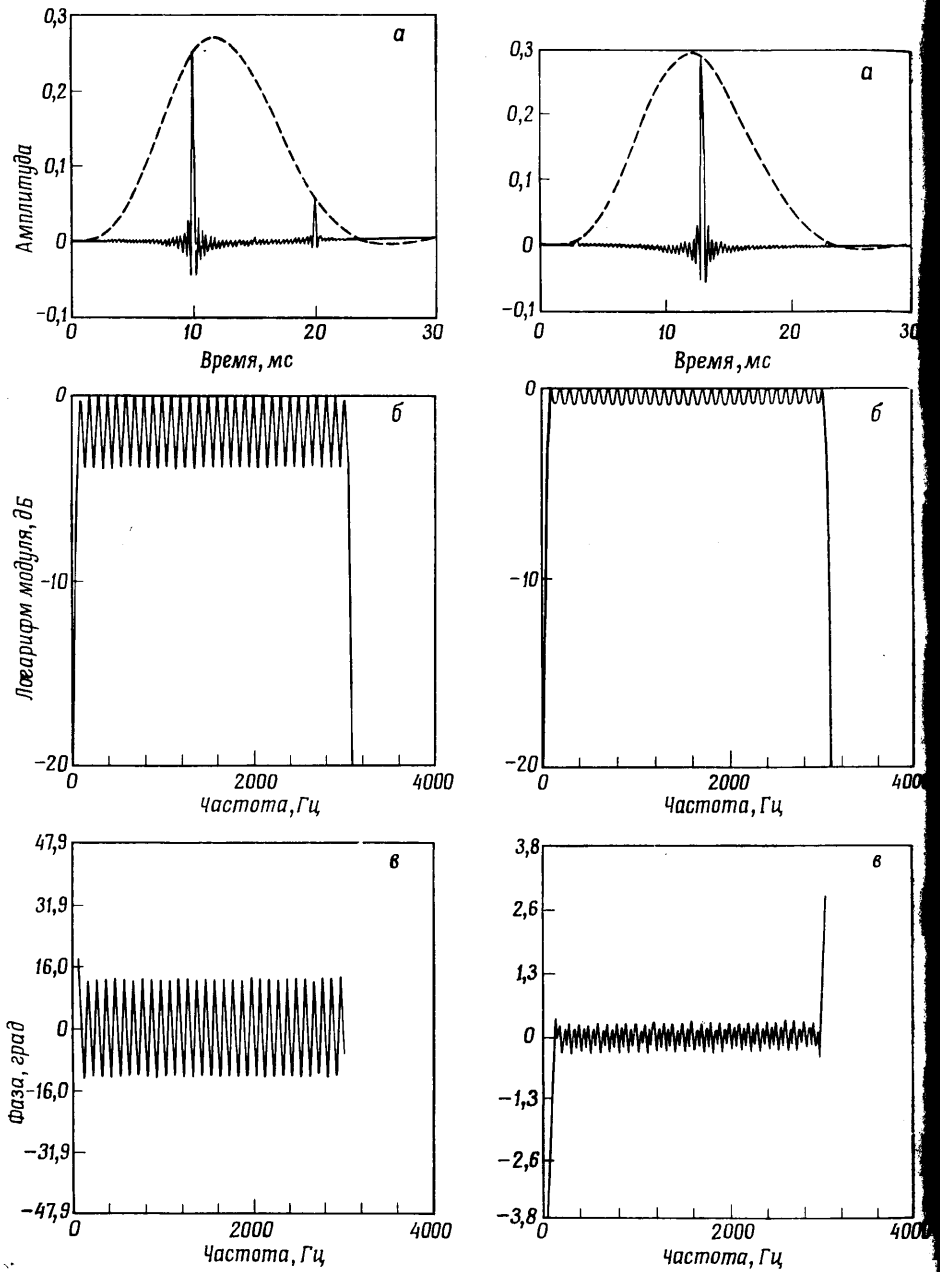
Итак, в рассматриваемых условиях $d(nT)$ представляет собой периодическую последовательность импульсов с периодом NT, обратно пропорциональным разнесению каналов по частоте. Поскольку $\tilde{h}(nT) = h(nT)\, d(nT)$, то ясно, что импульсная характеристика всей гребенки фильтров также является последовательностью импульсов. А так как идеальная импульсная характеристика — это одиночный задержанный импульс, то импульсную характеристику ФНЧ-прототипа $h(nT)$ следует выбирать так, чтобы в последовательности $d(nT)$ остался только один импульс. Зафиксируем T и N, при этом разнесение частот фильтров $\Delta\omega$ будет фиксировано. Тогда если выбрать импульсную характеристику фильтра-прототипа очень короткой, длиной менее $2N$, то суммарная импульсная характеристика будет такой, как показано на фиг. 12.13, *а*. Здесь же пунктиром изображена импульсная характеристика ФНЧ-прототипа, служащая временным окном для входного сигнала. Она совмещена с последовательностью импульсов, представляющей импульсную характеристику гребенки фильтров. В данном случае эта последовательность состоит лишь из одного импульса. Однако подобные короткие импульсные характеристики $h(nT)$ соответствуют довольно широкой полосе ФНЧ, не обеспечивающей нужного частотного разрешения. Если же применять более узкополосные фильтры, то длительность импульсной характеристики гребенки фильтров пропорционально



Фиг. 12.13. Компромисс между разрешением по времени и разрешением по частоте.

увеличится (фиг. 12.13, *б*), причем она будет состоять из нескольких импульсов, и в синтезированном речевом сигнале появится реверберация. Таким образом, условие хорошего частотного разрешения (т. е. узкие полосы фильтров) вступает в противоречие с условием отсутствия реверберации. Существует, однако, способ, позволяющий (по крайней мере теоретически) в точности согласовать выходной сигнал со входным. Он иллюстрируется на фиг. 12.13, *в*. Здесь используется более широкополосный фильтр, но принято, что значения $h\,(nT)$ в точках, кратных периоду N, должны равняться нулю. В этом случае суммарная импульсная характеристика состоит из единственного импульса, задержанного на $2N$. Таким образом, выходной сигнал представляет собой задержанную масштабированную копию входного сигнала. Форму соответствующего временно́го окна можно рассчитать. Итак, с помощью кратковременного фурье-преобразования теоретически можно представить речевой сигнал без искажений.

Во многих практических системах неудобно выбирать параметры так, чтобы суммарная импульсная хакатеристика изображалась кривой фиг. 12.13, *в*. Однако уравнения анализа и синтеза можно изменить таким образом, чтобы улучшить характеристику системы, даже если оптимальные характеристики при этом не достигаются. Этот подход иллюстрируется на фиг. 12.14 на примере системы анализа—синтеза, содержащей 39 каналов, размещенных через 100 Гц. Частота дискретизации равна 10 кГц. Пунктирной линией на фиг. 12.14, *а* слева представлена импульсная характеристика ФНЧ $h\,(nT)$ (фильтра Бесселя шестого порядка), а сплошной — суммарная импульсная характеристика $\tilde{h}\,(nT)$. Последняя кривая иллюстрирует импульсно-периодический характер последовательности $d\,(nT)$ в случае, когда не все каналы анализа используются при синтезе. Видно, что, кроме основного импульса при $t = 10$ мс, имеется заметное эхо при $t = 20$ мс [период $d\,(nT)$ составляет 10 мс]. На суммарной частотной характеристике эхо проявляется в виде пульсаций модуля и фазы [фиг. 12.14, *б* и *в*, слева], а на слух — как реверберация в синтезированном выходном сигнале. Рассмотренный пример, а также то, что $\tilde{h}\,(nT)$ является произведением $d\,(nT)$ и $h\,(nT)$, указывают на два пути улучшения суммарной характеристики гребенки фильтров. Как уже отмечалось, при заданном разнесении каналов можно расширить полосу ФНЧ, сократив таким образом длительность $h\,(nT)$. Согласно фиг. 12.14, *а* (слева), это приведет к увеличению амплитуды первого импульса и уменьшению второго. Однако в этом случае приходится идти на ухудшение частотного разрешения. Другой подход основан на том, что если $d\,(nT)$ может быть сдвинуто вправо относительно $h\,(nT)$ (т. е. относительно пунктирной кривой), то амплитуда основного импульса увеличится, а эхо станет меньше. В то же время импульс последовательности $d\,(nT)$ в точке $nT = 0$, который был полностью подавлен множителем $h\,(nT)$, при сдвиге вправо будет увеличиваться по амплитуде (фиг. 12.14, *а*, график справа). Поэтому при заданном частотном разрешении существует оптимальная задержка $d\,(nT)$ относительно $h\,(nT)$, при которой $\tilde{h}\,(nT)$ состоит из большого центрального импульса и двух одинаковых малых импульсов слева и справа от него. Можно показать, что при этом условии для заданного частотного разрешения обеспечиваются минимальные пульсации модуля и фазы.



Фиг. 12.14. Импульсные и частотные характеристики двух гребенок фильтров.

Задержку $d(nT)$ относительно $h(nT)$ можно ввести как при анализе, так и при синтезе речи. Если формулы (12.6а) и (12.6б) записать в виде

$$a(\omega_k, nT) = \sum_{r=-\infty}^{n} h(nT - rT)\, x(rT) \cos[\omega_k (rT - n_a T)], \quad (12.24)$$

$$b(\omega_k, nT) = \sum_{r=-\infty}^{n} h(nT - rT)\, x(rT) \sin[\omega_k (rT - n_a T)], \quad (12.25)$$

где $n_a = n_0$ — выбранная задержка в числе отчетов, и использовать для синтеза речи соотношение (12.16), то фактическая импульсная характеристика k-го канала будет иметь вид

$$h_k(nT) = h(nT) \cos[\omega_k (nT - n_0 T)], \quad (12.26)$$

а суммарная импульсная характеристика будет равна

$$\tilde{h}(nT) = h(nT)\, d(nT - n_0 T). \quad (12.27)$$

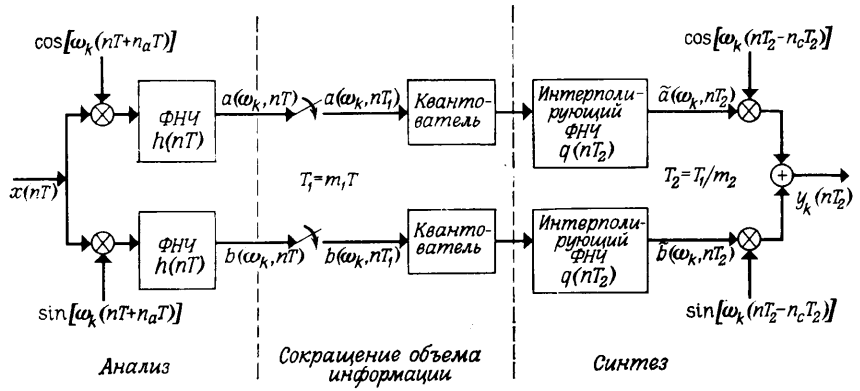
Ту же импульсную характеристику канала можно получить иначе, строя систему анализа на основе формул (12.6а) и (12.6б) и заменяя равенство (12.16) соотношением

$$y_k(nT) = a(\omega_k, nT) \cos[\omega_k (nT - n_c T)] + b(\omega_k, nT) \sin[\omega_k (nT - n_c T)], \quad (12.28)$$

где $n_c = n_0$. Третья возможность состоит в использовании для анализа формул (12.24) и (12.25), а для синтеза — равенства (12.28), если $n_a + n_c = n_0$. Программа проектирования системы обеспечивает такой выбор параметров, что импульсная и частотная характеристики системы соответствуют кривым, приведенным в правой части фиг. 12.14.

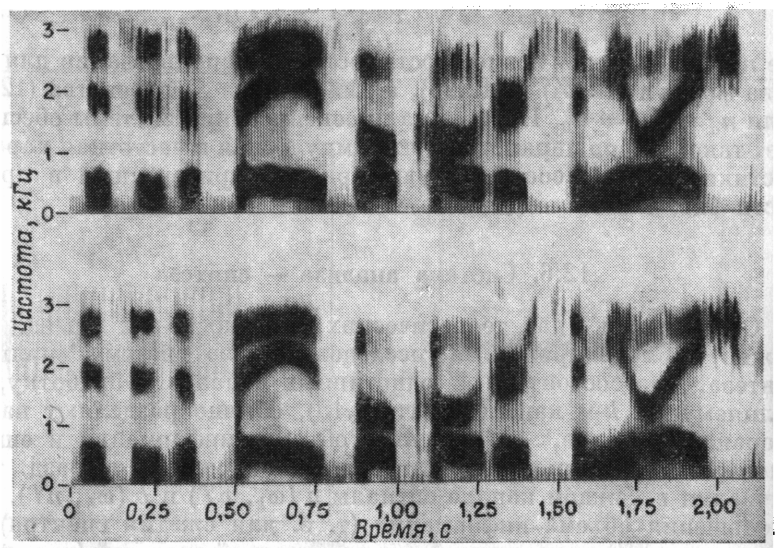
12.6. Система анализа — синтеза

Основываясь на теоретических предпосылках разд. 12.5 можно промоделировать и исследовать всю систему анализа—синтеза. Она состоит из M однотипных каналов. Обработку, выполняемую в k-м канале (фиг. 12.15), обычно разделяют на три операции: анализ, сокращение объема информации и синтез. Блок анализа выполняет алгоритмы, описанные в разд. 12.5, вычисляя в каждом канале сигналы $a(\omega_k, nT)$ и $b(\omega_k, nT)$. Для уменьшения объема информации (т. е. для сжатия спектра) эти сигналы нужно дискретизовать с пониженной частотой (т. е. через T_1 с) и квантовать по меньшему числу уровней. Эти функции выполняют дискретизатор и квантователь в блоке сжатия спектра.



Фиг. 12.15. Полная схема системы анализа — синтеза, включая дискретизацию и квантование.

Подходящие значения T_1 и числа уровней квантования определяются из экспериментов по разборчивости речи. Блок синтеза аналогичен рассмотренному в разд. 12.5, за исключением интерполирующих ФНЧ, предназначенных для интерполяции принятых значений $a(\omega_k, nT)$ и $b(\omega_k, nT)$ к новому периоду дискретизации T_2, отличному от T_1, периода дискретизации при анализе.



Фиг. 12.16. Сравнение спектрограмм естественного и синтезированного высказываний.

Эксперимент по моделированию системы анализа—синтеза подробно описан Шафером и Рабинером. Отметим лишь, что удовлетворительное качество речи получается при скорости передачи информации 15 000 бит/с, что примерно в четыре раза меньше, чем в системе с импульсно-кодовой модуляцией (ИКМ) при частоте дискретизации 8 кГц и семиразрядном логарифмическом кодировании. Если параметры спектра не квантованы по уровню, то спектрограммы исходного и синтезированного речевого сигналов (фиг. 12.16) трудно отличить.

12.7. Полосный вокодер

Полосный вокодер — это система анализа — синтеза речи, основанная на знании механизмов ее образования и восприятия. В частности, в полосном вокодере используется нечувствительность органов слуха к фазе сигнала и воспроизводится только кратковременный энергетический спектр речевого сигнала (это эквивалентно использованию амплитуды кратковременного фурье-преобразования без учета его фазы). Огибающая спектра речи измеряется с помощью гребенки полосовых фильтров, причем предполагается, что ее форма определяется характеристикой фильтра, образованного голосовым трактом. Источник возбуждения считается шумовым или импульсным квазипериодическим. (Таким образом, в вокодере непосредственно используется модель образования речи с независимыми источником возбуждения и голосовым трактом.) Существуют различные методы восстановления речи на основе измеренных параметров. В данном разделе описано несколько схем вокодера и рассмотрены факторы, влияющие на выбор их конструктивных параметров.

В типичном полосном вокодере (фиг. 12.17) исходный речевой сигнал $x(n)$ анализируется гребенкой полосовых фильтров (в данном случае их 16), неравномерно перекрывающих диапазон, существенный для восприятия речи (обычно от 0 до 3 кГц). Особенности проектирования этих фильтров будут рассмотрены ниже. Колебания на выходах полосовых фильтров детектируются и проходят через ФНЧ, выходные сигналы которых $y_k(n)$ в той или иной степени представляют огибающую спектра речи. Параметры, характеризующие источник возбуждения, получаются с помощью обнаружителя тон—шум, определяющего, является ли звук звонким (голосовые связки вибрируют) или глухим. В первом случае выделитель основного тона определяет основную частоту вибрации связок F_0.

Шестнадцать канальных сигналов, сигнал тон—шум и значение высоты основного тона кодируются и передаются по каналу связи к приемнику. Предположим, что передача происходит без ошибок. Тогда задача приемника сводится к восстановлению речи