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, полученную 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 («временного окна») \( h(nT) \) (фиг. 12.6). Используя свертку, равенство (12.3) можно записать иначе:

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

Левую часть равенства (12.3) можно представить в виде

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

где

- \( a(\omega, nT) \) — амплитудная характеристика спектра речевого сигнала в момент времени \( nT \);
- \( b(\omega, nT) \) — фазовая характеристика спектра речевого сигнала в момент времени \( nT \).
где $a(ω, nT)$ и $b(ω, nT)$ — действительная и мнимая части кратковременного фурье-преобразования, равные

$$a(ω, nT) = \sum_{r=-\infty}^{n} x(rT) h(nT - rT) \cos ωrT, \quad (12.6a)$$

$$b(ω, nT) = \sum_{r=-\infty}^{n} x(rT) h(nT - rT) \sin ωrT. \quad (12.6b)$$

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

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

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

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 $0 \leq n \leq M - 1$ и что центральные частоты анализа $ω_k$ равны

$$ω_k = \frac{2π}{N} k, \quad k = 0, 1, \ldots, N - 1. \quad (12.7)$$

Тогда (12.3) можно переписать следующим образом:

$$X(ω_k, nT) = \sum_{r=-M+1}^{N-M+1} x(rT) h(nT - rT) e^{-jω_k r T} = \sum_{m=0}^{[M/N]+1} \sum_{r=-M+1}^{n-MN} x(rT) h(nT - rT) e^{-jω_k r T}, \quad (12.8a)$$

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

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

Подстановка $ω_k$ из (12.7) дает

$$X(ω_k, nT) = e^{-j2πNmn} \sum_{l=0}^{N-1} [ \sum_{m=0}^{[M/N]+1} x(nT - lT - mNT) h (lT + mNT) e^{j2πNml}]. \quad (12.10)$$

Здесь $e^{j2πn}$ заменено единицей. Формулу (12.10) можно переписать в виде

$$X(ω_k, nT) = e^{-j2πNmn} \sum_{l=0}^{N-1} [ \sum_{m=0}^{[M/N]+1} g(l, n) e^{j2πNml}]h. \quad (12.11)$$

Фиг. 12.8. Идеальный фильтр низких частот для кратковременного спектрального анализа.
где

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

Соотношение (12.11) показывает, что \( X(\omega_h, nT) \) можно получить, перемножив последовательности \( e^{-j(2\pi N)h} \) и ДПФ последовательности \( g(l, n) \). На фиг. 12.9 иллюстрируется процесс поочередного 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.

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

12.4. Система анализа — синтез 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 сигнала 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

\[ h_k(nT) = h(nT) \cos(\omega_h nT), \] (12.13)

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

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

\[ = \text{Re} \{ e^{j\omega_h nT} X(\omega_h, nT) \}. \] (12.15)

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

\[ y_k(nT) = a(\omega_h, nT) \cos(\omega_h nT) + b(\omega_h, nT) \sin(\omega_h nT). \] (12.16)

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

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

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

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

получим

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

t. е. импульсная характеристика гребенки фильтров равна произведению импульсной характеристики ФНЧ-прототипа и функции, зависящей только от числа фильтров $M$ и их центральных частот $\omega_k$.

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

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

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

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

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

Если $\Delta \omega = 2\pi/N \Delta T$, 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 пик, 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 $\Delta \omega = 2\pi/N \Delta T$, то ясно, что этот случай соответствует измерению кратковременного преобразования Фурье на частотах, расположенных равномерно в диапазоне $0 < \omega < nT$. Если в гребенку фильтров ввести также канал с центром на нулевой частоте, то можно показать, что

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

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 $\tilde{h}(nT)$ 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 обеспечиваемой нулевого частотного разрешения. Если же применить более узкополосные фильтры, то длительность импульсной характеристики гребенки фильтров пропорционально
Глава 12

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. узкое 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 перио-

ду \( N \), 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 временного 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 характеристика 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, размеченных через 100 Гц. Частота дискретизации равна 10 кГц. Пунктирной линией на фиг. 12.14, а слева представлена импульсная характеристика ФНЧ \( h(nT) \) (фильтра Бесселя ш. стого 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 \( t = 10 \text{ мс} \), которое заметное эхо при \( t = 20 \text{ мс} \) (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 производным \( d(nT) \) и \( h(nT) \), указывает 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 приходят идти на ухудшение частотного разрешения. Другой подход основан на том, что если \( d(nT) \) может быть сдвищено вправо относительно \( h(nT) \) (т. е. относительно пунктирной кривой), то амплитуда основного импульса увеличивается, 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 большего 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.
Цифровая обработка речевых сигналов

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

$$a(\omega_k, nT) = \sum_{r=-\infty}^{n} h(nT - rT) \cos \{\omega_k (rT - n_\alpha T)\}, \quad (12.24)$$

$$b(\omega_k, nT) = \sum_{r=-\infty}^{n} h(nT - rT) \sin \{\omega_k (rT - n_\alpha T)\}, \quad (12.25)$$

где $n_\alpha = n_\beta$ — выбранная задержка в числе отсчетов, и использовать для синтеза речи соотношение (12.16), то фактическая импульсная характеристика $k$-го канала будет иметь вид

$$h_k(nT) = h(nT) \cos \{\omega_k (nT - n_\alpha T)\}, \quad (12.26)$$

а суммарная импульсная характеристика будет равна

$$\tilde{h}(nT) = h(nT) \cos \{nT - n_\alpha T\}. \quad (12.27)$$

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

$$y_k(nT) = a(\omega_k, nT) \cos \{\omega_k (nT - n_\alpha T)\} +$$

$$+ b(\omega_k, nT) \sin \{\omega_k (nT - n_\alpha T)\}, \quad (12.28)$$

где $n_\alpha = n_\beta$. Третья возможность состоит в использовании для анализа формул (12.24) и (12.25), а для синтеза — равенства (12.28), если $n_\alpha + n_\alpha = n_\alpha$. 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 — синтез

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, сжатие 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 полиномиальной частотой (т. е. через $T_1$, с) и квантовать по меньшему числу уровней. Эти функции выполняют дискретизатор и квантователь в блоке сжатия спектра.
Эксперимент по моделированию системы анализа — синтеза подробно описан Шаferом 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 семеразрядном 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 используются 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 квазиpériодическим. (Таким образом, в вокодере непосредственно используется модель образования речи с независимым 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 традиционном полосном вокодере (фиг. 12.17) исходный речевой сигнал z(t) 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 yk (n) в той или иной степени представляют огибающую спектра речи. Параметры, характеризующие источник возбуждения, получаются с помощью обнаружения тон — шум, определяющегося, является ли звук звонным (голосовые связки ибируют) или глухим. В первом случае выделитель основного тона определяет основную частоту вибрации связок F0.

Шестнадцать каналных 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