CHAPTER 22

Audio Processing
Звуковая Обработка

Audio processing covers many diverse fields, all involved in presenting sound to human listeners. Three areas are prominent: (1) *high fidelity music reproduction*, such as in audio compact discs, (2) *voice telecommunications*, another name for telephone networks, and (3) *synthetic speech*, where computers generate and recognize human voice patterns. While these applications have different goals and problems, they are linked by a common umpire: the human ear. Digital Signal Processing has produced revolutionary changes in these and other areas of audio processing.

Звуковая обработка охватывает много иных областей, вовлеченных в представление звука слушателям. Три области видны: (1) *Высокоточное воспроизведение музыки*, типа на звуковых компакт-дисках, (2) *голосовая линия связи*, другое название телефонные сети, и (3) *синтетической речи*, где компьютеры генерируют и распознают человеческие голоса. В то время как эти приложения имеют различные цели и проблемы, они связаны общим судьей: человеческое ухо. Цифровая Обработка сигналов произвела революционные изменения в этих и других областях звуковой обработки.

Human Hearing
Человеческое Звуковосприятие

The human ear is an exceedingly complex organ. To make matters even more difficult, the information from *two* ears is combined in a perplexing neural network, the human brain. Keep in mind that the following is only a brief overview; there are many subtle effects and poorly understood phenomena related to human hearing.

Человеческое ухо - чрезвычайно комплексный блок. Чтобы делать вопросы даже более трудными, информация от *двух* ушей объединена в озадачивающей нервной системе, человеческом мозге. Имейте в виду, что следующее - только краткий обзор; имеется много тонких эффектов и плохо понятых явлений, связанных с человеческим звуковосприятием.

Figure 22-1 illustrates the major structures and processes that comprise the human ear. The *outer ear* is composed of two parts, the visible flap of skin and cartilage attached to the side of the head, and the *ear canal*, a tube about 0.5 cm in diameter extending about 3 cm into the head. These structures direct environmental sounds to the sensitive *middle and inner ear* organs located safely inside of the skull bones. Stretched across the end of the ear canal is a thin sheet of tissue called the *tympanic membrane* or *ear drum*. Sound waves striking the tympanic membrane cause it to vibrate. The middle ear is a set of small bones that transfer this vibration to the *cochlea* (inner ear) where it is converted to neural impulses. The cochlea is a liquid filled tube roughly 2 mm in diameter and 3 cm in length. Although shown straight in Fig. 22-1, the cochlea is curled up and looks like a small snail shell. In fact, *cochlea* is derived from the Greek word for *snail*.

Рисунок 22-1 иллюстрирует главные структуры и процессы, которые включает человеческое ухо. *Наружное ухо* составлено из двух частей, наружной оболочки из кожи и хряща(ушной раковины), приложенной сбоку к голове, и *ушному каналу*, трубка приблизительно 0.5 см в диаметре, простирающаяся приблизительно на 3 см в голову. Эти структуры направляют звуки окружающей среды к чувствительным органам *среднего и внутреннего уха*, расположенным безопасно внутри костей черепа. Протянут поперек конца ушного канала тонкая полость ткани называемой *барабанной перепонкой* или *ушным барабаном*. Звуковые волны, надавливая на барабанную перепонку, заставляют ее вибрировать. Среднее ухо - набор маленьких костей, которые передают эту вибрацию *улитке* (внутреннее ухо) где эта механическая вибрация преобразовывается в невральные импульсы. Улитка - заполненная жидкостью трубка грубо 2 мм в диаметре и 3 см в длине. Хотя на рис. 22-1 улитка показана прямой, на самом деле она искривлена и напоминает оболочку маленькой *улитки*. Фактически, улитка получена от греческого слова - *спираль*.

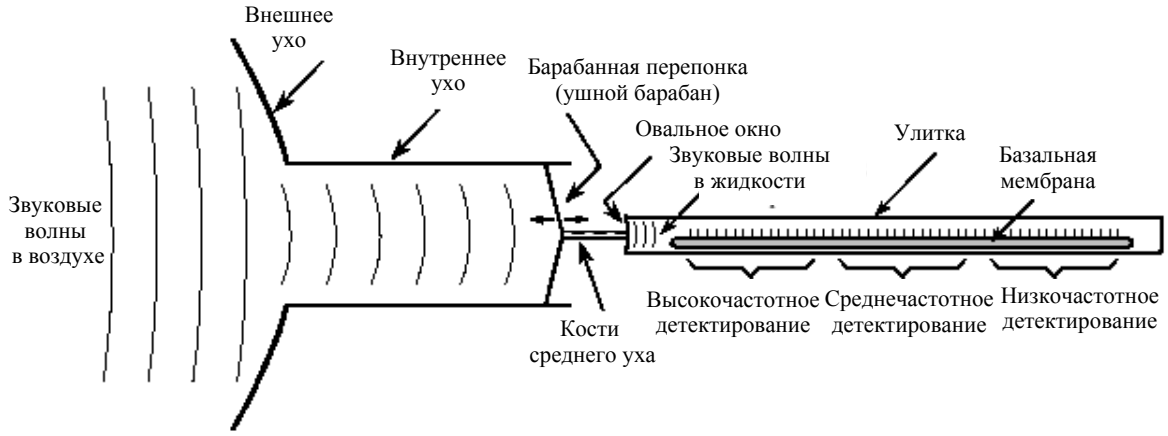

FIGURE 22-1
Functional diagram of the human ear. The outer ear collects sound waves from the environment and channels them to the tympanic membrane (ear drum), a thin sheet of tissue that vibrates in synchronization with the air waveform. The middle ear bones (hammer, anvil and stirrup) transmit these vibrations to the oval window, a flexible membrane in the fluid filled cochlea. Contained within the cochlea is the basilar membrane, the supporting structure for about 12,000 nerve cells that form the cochlear nerve. Due to the varying stiffness of the basilar membrane, each nerve cell only responses to a narrow range of audio frequencies, making the ear a frequency spectrum analyzer.
РИСУНОК 22-1. Функциональная диаграмма человеческого уха.
Наружное ухо собирает звуковые волны от среды и проводит их через канал к барабанной перепонке (ушному барабану), тонкая полость ткани, которая вибрирует синхронно с формой воздушной волны. Кости среднего уха (молоток, наковальня и скоба) передают эти колебания к овальному окну, гибкой мембране в улитку заполненную жидкостью. Содержащаяся в пределах улитки базальная мембрана, структура поддерживающая приблизительно 12000 нервных клеток, которые формируют кохлеарный нерв. Из-за изменяющейся жесткости базальной мембраны, каждая нервная клетка отвечает только на узкий диапазон звуковых частот, делая ухо анализатором спектра частот.

The difference between the loudest and faintest sounds that humans can hear is about 120 dB, a range of one-million in amplitude. Listeners can detect a *change* in loudness when the signal is altered by about 1 dB (a 12% change in amplitude). In other words, there are only about 120 levels of loudness that can be perceived from the faintest whisper to the loudest thunder. The sensitivity of the ear is amazing; when listening to very weak sounds, the ear drum vibrates less than the diameter of a single molecule!

Разность между самыми громкими и самыми слабыми звуками, которые люди могут слышать - приблизительно 120 dB, диапазон одно-миллионных в амплитуде. Слушатели могут обнаруживать изменение в громкости, когда сигнал изменен примерно до 1 dB (изменение 12 % в амплитуде). Другими словами, имеются только приблизительно 120 уровней громкости, которая может быть воспринята от самого слабого шепота до самого громкого

грома. Чувствительность уха удивительна; при прослушивании очень слабых звуков, барабанная перепонка вибрируют меньше чем диаметр отдельной молекулы!

The perception of loudness relates roughly to the sound power to an exponent of 1/3. For example, if you increase the sound power by a factor of *ten*, listeners will report that the loudness has increased by a factor of about *two* ($10^{1/3} \approx 2$). This is a major problem for eliminating undesirable environmental sounds, for instance, the beefed-up stereo in the next door apartment. Suppose you diligently cover 99% of your wall with a perfect soundproof material, missing only 1% of the surface area due to doors, corners, vents, etc. Even though the sound power has been reduced to only 1% of its former value, the perceived loudness has only dropped to about $0.01^{1/3} \approx 0.2$, or 20%

Восприятие громкости имеет грубое отношение с мощностью звука по экспоненте 1/3. Для примера, если Вы усиливаете звуковую мощь в *десять* раз, слушатели скажут, что громкость звука увеличилась приблизительно в *два* раза ($10^{1/3} \approx 2$). Это - важная проблема для устранения нежелательного окружающего шума, например, усиленного стерео за дверью квартиры. Предположите, что Вы старательно закрываете 99 % ваших стен совершенным звуконепроницаемым материалом, пропуская только 1 % площади поверхности из-за дверей, углов, вентиляции, и т.д. Даже при том, что звуковая мощь была сокращена до к 1 % ее прежнего значения, воспринятая громкость понизилась только на $0.01^{1/3} \approx 0.2$, или 20%.

The range of human hearing is generally considered to be 20 Hz to 20 kHz, but it is far more sensitive to sounds between 1 kHz and 4 kHz. For example, listeners can detect sounds as low as 0 dB SPL at 3 kHz, but require 40 dB SPL at 100 hertz (an amplitude increase of 100). Listeners can tell that two tones are different if their frequencies differ by more than about 0.3% at 3 kHz. This increases to 3% at 100 hertz. For comparison, adjacent keys on a piano differ by about 6% in frequency.

Диапазон человеческого слуха, как рассматривается, находится между 20 Гц и 20 кГц, но гораздо более чувствительно к звукам между 1 кГц и 4 кГц. Например, слушатели могут обнаруживать звуки столь же низко как 0 dB SPL в 3 кГц, но требовать 40 dB SPL в 100 герц (амплитудное увеличение 100). Слушатели могут сообщить, что два тона являются отличными, если их частоты отличаются приблизительно более чем на 0.3 % в 3 кГц. При 100 герцах это увеличивается до 3 %. Для сравнения, смежные клавиши фортепьяно, отличаются по частоте примерно на 6 %.

	Вт/см2	Децибелы SPL	Пример звука
	10^{-2}	140 dB	Боль
	10^{-3}	130 dB	Дискомфорт
	10^{-4}	120 dB	
громче ↑	10^{-5}	110 dB	Джек ударник и рок концерт
	10^{-6}	100 dB	
	10^{-7}	90 dB	OSHA ограничитель промышленного шума
	10^{-8}	80 dB	
	10^{-9}	70 dB	
	10^{-10}	60 dB	Нормальный разговор
	10^{-11}	50 dB	
	10^{-12}	40 dB	Едва слышимый в 100 гЦ
	10^{-13}	30 dB	
Тише ↓	10^{-14}	20 dB	Едва слышимый в 10 кГц
	10^{-15}	10 dB	
	10^{-16}	0 dB	Едва слышимый в 3 кГц
	10^{-17}	-10 dB	
	10^{-18}	-20 dB	

TABLE 22-1
Units of sound intensity. Sound intensity is expressed as power per unit area (such as watts/cm 2), or more commonly on a logarithmic scale called *decibels SPL*. As this table shows, human hearing is the most sensitive between 1 kHz and 4 kHz.
ТАБЛИЦА 22-1. Единицы интенсивности звука.
Интенсивность звука выражена как мощность на единицу площади (типа Ватт/см^2), или более обычно в логарифмическом масштабе называемом *децибелами SPL*. Как показано в таблице, человеческий слух наиболее чувствителен между 1 кГц и 4 кГц.

The primary advantage of having *two* ears is the ability to identify the *direction* of the sound. Human listeners can detect the difference between two sound sources that are placed as little as three degrees apart, about the width of a person at 10 meters. This directional information is obtained in two separate ways. First, frequencies above about 1 kHz are strongly *shadowed* by the head. In other words, the ear nearest the sound receives a stronger signal than the ear on the opposite side of the head. The second clue to directionality is that the ear on the far side of the head hears the sound slightly *later* than the near ear, due to its greater distance from the source. Based on a typical head size (about 22 cm) and the speed of sound (about 340 meters per second), an angular discrimination of three degrees requires a timing precision of about 30 microseconds. Since this timing requires the volley principle, this clue to directionality is predominately used for sounds less than about 1 kHz.

Первичное преимущество наличия двух ушей - способность идентифицировать направление звука. Человек может обнаруживать разность между двумя источниками звука, которые помещены друг от друга в трех метрах при расстоянии от них до человека - 10 метров. Эта направленная информация получена двумя отдельными способами. Во первых, частоты выше приблизительно 1 кГц строго *заслонены* головой. Другими словами, ухо расположенное ближе к источнику звука, получает более сильный сигнал, чем ухо на противоположной стороне головы. Второй ключ к направленности - то, что ухо на дальней стороне головы слышит звук слегка позже, чем около уха, из-за его большего расстояния от источника. Основываясь на типичном размере головы (приблизительно 22 см) и скорости звука (приблизительно 340 метров в секунду), угловое выделение(распознование) трех градусов требуют прецизионности синхронизации приблизительно 30 микросекунд. Так как эта синхронизация требует принципа залпа(града; потока), это ключ к использованию направленности преимущественно для звуков меньше чем приблизительно 1 кГц.

Both these sources of directional information are greatly aided by the ability to turn the head and observe the change in the signals. An interesting sensation occurs when a listener is presented with exactly the same sounds to both ears, such as listening to monaural sound through headphones. The brain concludes that the sound is coming from the center of the listener's head!

Обоим этим источникам направленной информации очень помогает способность повернуть голову и наблюдать изменение в сигналах. Интересное ощущение происходит, когда слушателю представлены точно те же самые звуки к обоим ушам, типа моноурального(монофоничесгого) слушания звука через наушники. Мозг заключает, что звук исходит из центра головы слушателя!

While human hearing can determine the *direction* a sound is from, it does poorly in identifying the *distance* to the sound source. This is because there are few clues available in a sound wave that can provide this information. Human hearing weakly perceives that high frequency sounds are nearby, while low frequency sounds are distant. This is because sound waves dissipate their higher frequencies as they propagate long distances. Echo content is another weak clue to dis-

tance, providing a perception of the room size. For example, sounds in a large auditorium will contain echoes at about 100 millisecond intervals, while 10 milliseconds is typical for a small office. Some species have solved this ranging problem by using *active sonar*. For example, bats and dolphins produce clicks and squeaks that reflect from nearby objects. By measuring the interval between transmission and echo, these animals can locate objects with about 1 cm resolution. Experiments have shown that some humans, particularly the blind, can also use active echo localization to a small extent.

В то время как человеческий слух может определять *направление* звука от источника, он плох в идентификации *расстояния* до звукового источника. Это - то, потому что имеются немного ключей(улик; признаков), доступных в звуковой волне, которые могли бы обеспечивать эту информацию. Человеческий слух слабо различает, что звуки высокой частоты близлежащие, в то время как низкочастотные звуки отдаленны. Это - то, потому что звуковые волны рассеивают их более высокие частоты, когда они распространяются на длинные расстояния. Содержание эхо - другой слабый ключ к расстоянию, обеспечивая восприятие размером комнаты. Например, звуки в большой аудитории будут содержать, эхо(отражение) интервал приблизительно 100 миллисекунд, в то время как 10 миллисекунд типичны для маленького офиса. Некоторую разновидность решения этой проблемы ранжировки, использует *активный эхолокатор*(активная гидроакустическая станция). Например, летучие мыши и дельфины производят щелчки, и писк, которые отражаются от близлежащих объектов. Измеряя интервал между передачей и отражением, эти животные могут обнаруживать объекты с разрешающей способностью приблизительно 1 см. Эксперименты показали, что некоторые люди, особенно слепые, в маленькой степени, также могут использовать активную локализацию отражения(эхо).

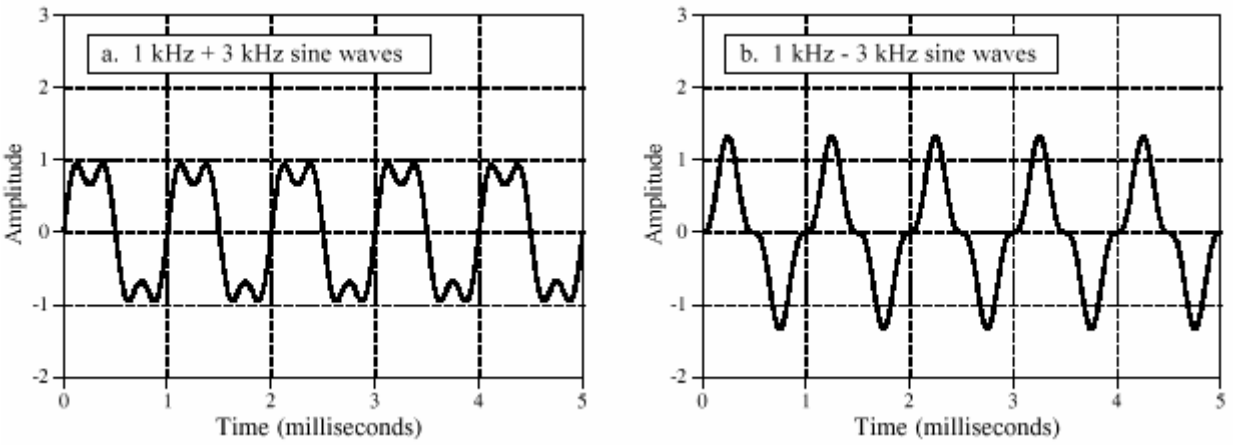

FIGURE 22-2. Phase detection of the human ear.
The human ear is very insensitive to the relative phase of the component sinusoids. For example, these two waveforms would sound identical, because the *amplitudes* of their components are the same, even though their relative *phases* are different.
РИСУНОК 22-2. Фазовое обнаружение человеческого уха.
Человеческое ухо очень нечувствительно относительно фазы составляющих синусоид. Например, эти две формы волны звучали бы идентично, потому что *амплитуды* их компонентов - одинаковы, даже притом, что относительно их *фаз* они различны.

Тембр

The perception of a continuous sound, such as a note from a musical instrument, is often divided into three parts: **loudness**, **pitch**, and **timbre** (pronounced "timber"). *Loudness* is a measure of

sound wave intensity, as previously described. *Pitch* is the frequency of the fundamental component in the sound, that is, the frequency with which the waveform repeats itself. While there are subtle effects in both these perceptions, they are a straightforward match with easily characterized physical quantities.

Восприятие непрерывного звука, типа примечания от музыкального прибора, часто разделяется на три части: **громкость**, **тон**, и **тембр** (явная "древесина"). *Громкость* - мера интенсивности звуковой волны, как предварительно описано. *Тон* - частота фундаментального(основного) компонента в звуке, то есть частота, с который формы волны повторяется самостоятельно. В то время как имеются слабовыраженные эффекты в обоих этих восприятиях, они - прямое соответствие с легко характеризованными физическими количествами.

Timbre is more complicated, being determined by the *harmonic content* of the signal. Figure 22-2 illustrates two waveforms, each formed by adding a 1 kHz sine wave with an amplitude of *one*, to a 3 kHz sine wave with an amplitude of *one-half*. The difference between the two waveforms is that the one shown in (b) has the higher frequency *inverted* before the addition. Put another way, the third harmonic (3 kHz) is phase shifted by 180 degrees compared to the first harmonic (1 kHz). In spite of the very different time domain waveforms, these two signals sound *identical*. This is because hearing is based on the *amplitude* of the frequencies, and is very insensitive to their *phase*. The *shape* of the time domain waveform is only indirectly related to hearing, and usually not considered in audio systems.

Тембр больше усложнен, определяемым *содержанием гармоник* в сигнале. Рисунок 22-2 иллюстрирует две формы волны, каждая сформирована, прибавляя волну синуса 1 кГц с амплитудой 1, к волне синуса 3 кГц с *половиной амплитуды*. Разность между этими двумя формами волны - то, что одна показанная в (b) имеет более высокую частоту *перевернутую(инвертированную* перед добавлением. Излагая другой путь, фаза третьей гармоники (3 кГц) сдвинута на 180 градусов, по сравнению с основной гармоникой (1 кГц). Несмотря на самые различные формы волны домена времени, эти два звуковых сигнала идентичны. Это - то, потому что слух основан на *амплитуде* частот, и почти нечувствительно к их *фазе*. *Форма* формы волны домена времени только косвенно связана со слушанием, и обычно не рассматривается в звуковых системах.

The ear's insensitivity to phase can be understood by examining how sound propagates through the environment. Suppose you are listening to a person speaking across a small room. Much of the sound reaching your ears is reflected from the walls, ceiling and floor. Since sound propagation depends on frequency (such as: attenuation, reflection, and resonance), different frequencies will reach your ear through different paths. This means that the relative phase of each frequency will change as you move about the room. Since the ear disregards these phase variations, you perceive the voice as *unchanging* as you move position. From a physics standpoint, the phase of an audio signal becomes randomized as it propagates through a complex environment. Put another way, the ear is insensitive to phase because it contains little useful information.

Нечувствительность уха к фазе может быть понята, исследуя, как звук распространяется через среду. Предположим, что Вы слушаете человека, говорящего поперек маленького участка комнаты. Многое из звука, достигающего ваших ушей отражено от стенок, потолка и пола. Так как распространение звука зависит от частоты (типа: ослабление, отражение, и резонанс), различные частоты достигнут вашего уха различными путями. Это означает, что относительная фаза каждой частоты изменится, поскольку Вы перемещаетесь по комнате. Так как ухо игнорирует эти фазовые вариации, Вы чувствуете голос как *неизменяемый*, по мере того как Вы перемещаете позицию. С точки зрения физики, фаза звуко-

вого сигнала становится рандомизированной(произвольной?), так как распространяется через комплексную среду. Излагая другой путь, ухо нечувствительно к фазе, потому что это содержит немного полезной информации.

However, it cannot be said that the ear is completely deaf to the phase. This is because a phase change can rearrange the *time sequence* of an audio signal. An example is the chirp system (Chapter 11) that changes an impulse into a much longer duration signal. Although they differ only in their phase, the ear can distinguish between the two sounds because of their difference in duration. For the most part, this is just a curiosity, not something that happens in the normal listening environment.

Однако, нельзя сказать, что ухо полностью глухое к фазе. Это - то, потому что изменение фазы может перестраивать *временную последовательность* (*последовательность времени)* звукового сигнала. Пример - система щебета (глава 11) который изменяет импульс в сигнал намного более длинной продолжительности. Хотя они отличаются только по их фазе, ухо может различать между двумя звуками из-за их разности в продолжительности. Главным образом, это - только любопытство, не кое-что, что случается в нормальной среде прослушивания.

Suppose that we ask a violinist to play a note, say, the *A* below middle *C*. When the waveform is displayed on an oscilloscope, it appear much as the sawtooth shown in Fig. 22-3a. This is a result of the sticky rosin applied to the fibers of the violinist's bow. As the bow is drawn across the string, the waveform is formed as the string sticks to the bow, is pulled back, and eventually breaks free. This cycle repeats itself over and over resulting in the sawtooth waveform.

Предположим, что мы просим, чтобы скрипач запустил примечание(ноту; сопровождение), скажем, *A* ниже середины *C* (от ноты *До* до ноты *Ми*? См. рис. 22-4). Когда форма волны отображена на осциллографе, она кажется сильно остроконечной как показано на рис. 22-3a. Это - результат липкой смолы(канифоли), приложенной к (стекло)волокнам наклонов скрипача. Поскольку смычок выведется поперек струн, форма волны сформирована, поскольку струна придерживается смычка(наклона), отступая, и в конечном счете вырывается на свободу. Эти повторения цикла(периода), самостоятельно(сами себя) много раз приводят к остроконечной форме волны.

Figure 22-3b shows how this sound is perceived by the ear, a frequency of 220 hertz, plus harmonics at 440, 660, 880 hertz, etc. If this note were played on another instrument, the waveform would *look* different; however, the ear would still hear a frequency of 220 hertz plus the harmonics. Since the two instruments produce the same fundamental frequency for this note, they sound similar, and are said to have identical *pitch*. Since the relative amplitude of the *harmonics* is different, they will not sound identical, and will be said to have different *timbre*.

Рисунок 22-3b показывает, как этот звук воспринят ухом, частота 220 герц, плюс гармоники в 440, 660, 880 герц, и т.д. Если бы это примечание(ноту, сопровождение) запускали на другом приборе, форма волны *выглядела* бы различной; однако, ухо все еще слышало бы частоту 220 герц плюс *гармоники* Так как эти два инструмента производят ту же самую фундаментальную частоту для этого примечания(ноты; сопровождения), они звучат подобными, и как считают, имеют идентичный *шаг(тон*). Так как относительная амплитуда *гармоник* различна, они не будут звучать идентично, и будут, как считают, иметь различный *тембр*.

It is often said that timbre is determined by the shape of the waveform. This is true, but slightly misleading. The perception of timbre results from the ear detecting harmonics. While harmonic

content is determined by the shape of the waveform, the insensitivity of the ear to phase makes the relationship very one-sided. That is, a particular waveform will have only one timbre, while a particular timbre has an infinite number of possible waveforms.

Часто говорится, что тембр определен формой формы волны. Это истинно, но слегка вводит в заблуждение. Восприятие тембра следует из уха, обнаруживающего гармоники(флажолеты?). В то время как содержание гармоник определено формой формы волны, нечувствительность уха к фазе делает отношения, очень односторонним. То есть специфическая форма волны будет иметь только один тембр, в то время как специфический тембр имеет бесконечное число возможных форм волны. (флажолеты – возможно музыкальное название гармоник тембровой окраски; акомпанимента; сопровождения?)

The ear is very accustomed to hearing a fundamental plus harmonics. If a listener is presented with the combination of a 1 kHz and 3 kHz sine wave, they will report that it sounds natural and pleasant. If sine waves of 1 kHz and 3.1 kHz are used, it will sound objectionable.

Ухо очень приученно к слушанию фундаментальных положительных гармоник(флажолетов). Если слушателю представлена комбинация синусоидальной волны 1 кГц и 3 кГц, он сообщат, что это звучит естественно и приятно. Если используются синусоидальные волны 1 кГц и 3.1 кГц, это будет звучать нежелательным.

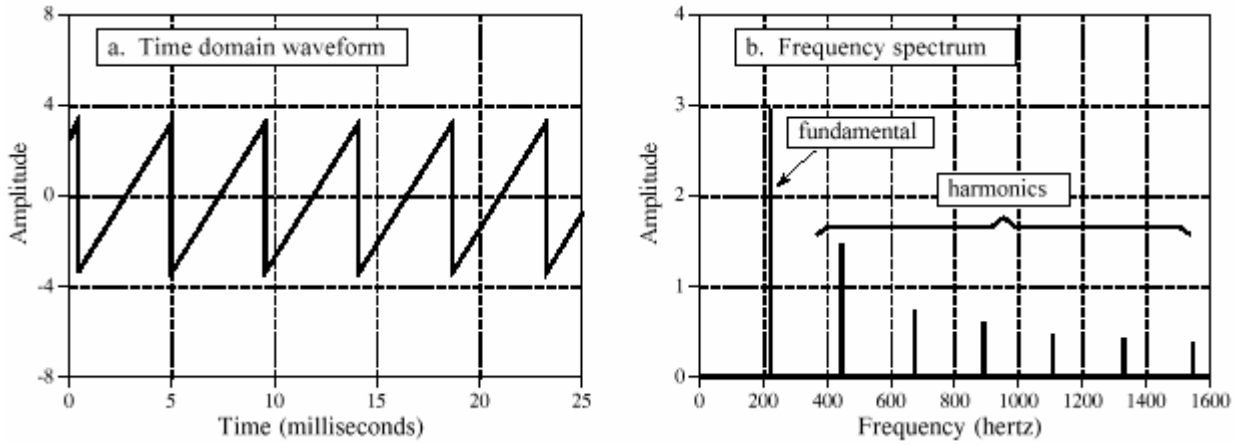

FIGURE 22-3. Violin waveform.
A bowed violin produces a sawtooth waveform, as illustrated in (a). The sound *heard by the ear* is shown in (b), the fundamental frequency plus harmonics.
РИСУНОК 22-3. Скрипичная форма волны.
Смычковая скрипка производит остроконечную форму волны, как иллюстрировано в (a). Звук, который слышит ухо показывается в (b), основной частоты плюс гармоники(флажолеты).

This is the basis of the standard musical scale, as illustrated by the piano keyboard in Fig. 22-4. Striking the farthest left key on the piano produces a fundamental frequency of 27.5 hertz, plus harmonics at 55, 110, 220, 440, 880 hertz, etc. (there are also harmonics between these frequencies, but they aren't important for this discussion). These harmonics correspond to the fundamental frequency produced by other keys on the keyboard. Specifically, every *seventh* white key is a harmonic of the far left key. That is, the eighth key from the left has a fundamental frequency of 55 hertz, the 15th key has a fundamental frequency of 110 hertz, etc. Being harmonics of each other, these keys sound similar when played, and are harmonious when played in unison. For this reason, they are *all* called the note, *A*. In this same manner, the white key immediate right of each *A* is called a *B*, and *they* are all harmonics of each other. This pattern repeats for the seven notes: *A*, *B*, *C*, *D*, *E*, *F*, and *G*.

Это - основание стандартной музыкальной шкалы, как иллюстрировано фортепианной клавиатурой в рис. 22-4. Нажатие самой дальней левой клавиши фортепиано, производит основную частоту 27.5 герц, положительные гармоники(флажолеты) в 55, 110, 220, 440, 880 герц, и т.д. (имеются также гармоники(флажолеты) между этими частотами, но они не важны для этого обсуждения). Эти гармоники(флажолеты) соответствуют основной частоте, произведенной другими клавишами на клавиатуре. Определенно, каждая седьмая белая клавиша - гармоника(флажолет) дальней левой клавиши. То есть восьмая клавиша слева имеет основную частоту 55 герц, 15-ая клавиша имеет основную частоту 110 герц, и т.д. Будучи гармониками(флажолетами) друг друга, эти клавиши издают подобный звук когда нажаты, и гармоничны когда нажаты совместно. По этой причине, они все называются нотой, *A*. Таким же образом, белая клавиша непосредственно справа от каждой называется *B*, и они – *все* (клавиши *B*) гармоники(флажолеты) друг друга. Этот образец повторяется для семи нот: *A*, *B*, *C*, *D*, *E*, *F*, и *G*.

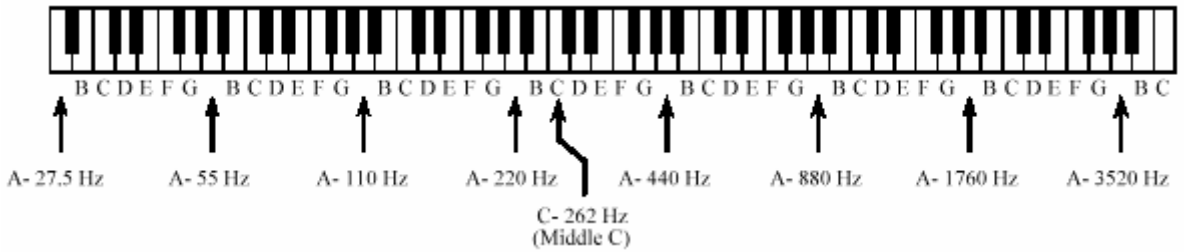

FIGURE 22-4. The Piano keyboard.
The keyboard of the piano is a *logarithmic* frequency scale, with the fundamental frequency doubling after every seven white keys. These white keys are the notes: *A, B, C, D, E, F and G*.
РИСУНОК 22-4. Фортепианная клавиатура.
Клавиатура фортепиано - *логарифмический* частотный масштаб(шкала), с основной частотой, удваивающейся после каждых семи белых клавиш. Эти белые клавиши - ноты: *A, B, C, D, E, F и G*.

The term **octave** means a *factor of two in frequency*. On the piano, one octave comprises eight white keys, accounting for the name (*octo* is Latin for *eight*). In other words, the piano's frequency doubles after every seven white keys, and the entire keyboard spans a little over seven octaves. The range of human hearing is generally quoted as 20 hertz to 20 kHz, corresponding to about ½ octave to the left, and two octaves to the right of the piano keyboard. Since octaves are based on doubling the frequency every fixed number of keys, they are a *logarithmic* representation of frequency. This is important because audio information is generally distributed in this same way. For example, as much audio information is carried in the octave between 50 hertz and 100 hertz, as in the octave between 10 kHz and 20 kHz. Even though the piano only covers about 20% of the frequencies that humans can hear (4 kHz out of 20 kHz), it can produce more than 70% of the audio information that humans can perceive (7 out of 10 octaves). Likewise, the highest frequency a human can detect drops from about 20 kHz to 10 kHz over the course of an adult's lifetime. However, this is only a loss of about 10% of the hearing ability (one octave out of ten). As shown next, this logarithmic distribution of information directly affects the required *sampling rate* of audio signals.

Термин **октава** означает *коэффициент(фактор) два в частоте*. На фортепиано, одна октава включает восемь белых клавиш, объясняя название *октава* (*octo* латинское *восемь*). Другими словами, частота фортепиано удваивается после каждых семи белых клавиш, и полная клавиатура охватывает немного более чем семь октав. Диапазон человеческого слуха вообще цитируется как 20 герц до 20 кГц, соответствующих ½ октавы слева, и двух октав справа на фортепианной клавиатуре. Так как октавы основаны на удвоении частоты каждая установленное число клавиш, они - *логарифмическое* представление частоты. Это важно, потому что звуковая информация вообще распределяется этим же самым путем.

Например, так много звуковой информации несут в октаве между 50 герц и 100 герц, как в октаве между 10 кГц и 20 кГц. Даже при том, что фортепиано покрывает(охватывает) только приблизительно 20 % частот, которые люди могут слышать (4 кГц - 20 кГц), это может производить больше чем 70 % звуковой информации, которую люди могут воспринимать (7 из 10 октав). Аналогично, самая высокая частота человека может обнаруживать перепады приблизительно от 20 кГц до 10 кГц в течение продолжительности жизни взрослого. Однако, это - потеря только приблизительно 10 % способности слуха (одна октава из десяти). Как показано дальше, это логарифмическое распределение информации непосредственно воздействует на требуемую *частоту выборки* аудиосигналов.

Sound Quality vs. Data Rate
Качество Звука против Скорости передачи данных

When designing a digital audio system there are two questions that need to be asked: (1) how good does it need to sound? and (2) what data rate can be tolerated? The answer to these questions usually results in one of three categories. First, **high fidelity music,** where sound quality is of the greatest importance, and almost any data rate will be acceptable. Second, **telephone communication**, requiring natural sounding speech *and* a low data rate to reduce the system cost. Third, **compressed speech**, where reducing the data rate is very important and some unnaturalness in the sound quality can be tolerated. This includes military communication, cellular telephones, and digitally stored speech for voice mail and multimedia.

При проектировании цифровой звуковой системы имеются два вопроса, которые должны задать: (1), как хорошо требуется звучать? и (2), какая скорость передачи данных может допускаться? Ответ на эти вопросы обычно приводит к одной из трех категорий. Во первых, **высокая точность музыки**, где звуковое качество имеет самую большую важность, и почти любая скорость передачи данных, будет приемлема. Во вторых, **телефонная связь**, требует натуральный(естественный) звук речи *и* низкую скорость передачи данных чтобы привести стоимость системы. Третье, **сжатая речь**, где сокращение скорости передачи данных очень важно и некоторая неестественность качестве звука, может допускаться. Это включает военную связь, ячеистые телефоны, и хранении речи в цифровой форме для звуковой почты и мультимедиа.

Table 22-2 shows the tradeoff between sound quality and data rate for these three categories. High fidelity music systems sample fast enough (44.1 kHz), and with enough precision (16 bits), that they can capture virtually all of the sounds that humans are capable of hearing. This magnificent sound quality comes at the price of a high data rate, 44.1 kHz x 16 bits = 706k bits/sec. This is pure brute force.

Таблица 22-2 показывает сделку между качеством звука и скоростью передачи данных для этих трех категорий. Музыкальные системы высокой точности производят выборку достаточно быстро (44.1 кГц), и с достаточной прецизионностью (16 битов), так что они могут фиксировать фактически все звуки, которые люди способны слышать. Это великолепное звуковое качество ценой высокой скорости передачи данных, 44.1 kHz x 16 bits = 706k bits/sec. Это - чистое решение "в лоб"(силовое).

Whereas music requires a bandwidth of 20 kHz, natural sounding speech only requires about 3.2 kHz. Even though the frequency range has been reduced to only 16% (3.2 kHz out of 20 kHz), the signal still contains 80% of the original sound information (8 out of 10 octaves). Telecommunication systems typically operate with a sampling rate of about 8 kHz, allowing natural

sounding speech, but greatly reduced music quality. You are probably already familiar with this difference in sound quality: FM radio stations broadcast with a bandwidth of almost 20 kHz, while AM radio stations are limited to about 3.2 kHz. Voices sound normal on the AM stations, but the music is weak and unsatisfying.

Принимая во внимание, что музыка требует ширины полосы частот 20 кГц, звук натуральной речи требует только приблизительно 3.2 кГц. Даже при том, что частотный диапазон был сокращен к только 16 % (3.2 кГц из 20 кГц), сигнал все еще содержит 80 % первоначальной информации звука (8 из 10 октав). Телекоммуникационые системы типично оперируют с частотой выборки приблизительно 8 кГц, позволяя звук натуральной речи, но очень приведенное музыкальное качество. Вы вероятно уже знакомы с этой разностью в качестве звука: радиостанции ЧМ передают по радио с шириной полосы частот почти 20 кГц, в то время как – радиостанции АМ, ограничены приблизительно 3.2 кГц. Голоса звучат, нормально на – АМ станциях, но музыка слаба и неудовлетворительна.

Voice-only systems also reduce the precision from 16 bits to 12 bits per sample, with little noticeable change in the sound quality. This can be reduced to only 8 bits per sample if the quantization step size is made unequal. This is a widespread procedure called **companding**, and will be discussed later in this chapter. An 8 kHz sampling rate, with an ADC precision of 8 bits per sample, results in a data rate of 64 kbits/sec. This is the *brute force* data rate for natural sounding speech. Notice that speech requires less than 10% of the data rate of high fidelity music.

Системы только - голос - также приводят прецизионность от 16 битов к 12 битов на выборку, с небольшим значимым изменением в качестве звука. Это может быть сокращено всего к 8 битам на выборку, если размер шага квантования сделан неравным. Это - широко распространенная процедура называемая **компандированием**, будет обсуждена позже в этой главе. Частота выборки 8 кГц, с прецизионностью АЦП 8 битов на выборку, приводит к скорости передачи данных битов 64 kbits/sec. Это - скорость передачи данных *решения "в лоб"* для натурального звучания речи. Обратите внимание, что речь требует меньше чем 10 % скорости передачи данных музыки высокой точности.

Sound Quality Required	Bandwidth	Sampling rate	Number of bits	Data rate (bits/sec)	Comments
High fidelity music (compact disc)	5 Hz to 20 kHz	44.1 kHz	16 bit	706k	Satisfies even the most picky audiophile. Better than human hearing.
Telephone quality speech	200 Hz to 3.2 kHz	8 kHz	12 bit	96k	Good speech quality, but very poor for music.
(with companding)	200 Hz to 3.2 kHz	8 kHz	8 bit	64k	Nonlinear ADC reduces the data rate by 50%. A very common technique.
Speech encoded by Linear Predictive Coding	200 Hz to 3.2 kHz	8 kHz	12 bit	4k	DSP speech compression technique. Very low data rates, poor voice quality.

TABLE 22-2
Audio data rate vs. sound quality. The sound quality of a digitized audio signal depends on its *data rate*, the product of its sampling rate and number of bits per sample. This can be broken into three categories, high fidelity music (706 kbits/sec), telephone quality speech (64 kbits/sec), and compressed speech (4 kbits/sec).

ТАБЛИЦА 22-2

Скорость передачи звуковых данных против качества звука. Звуковое качество цифрового аудиосигнала зависит от его *скорости передачи данных*, продукта из его частоты выборки и числа битов на выборку. Это может быть разбито на три категории, музыка высокой точности (706 kbits/sec), речь качества телефона (64 kbits/sec), и сжатая речь (4 kbits/sec).

The data rate of 64k bits/sec represents the straightforward application of sampling and quantization theory to audio signals. Techniques for lowering the data rate further are based on *compressing* the data stream by removing the inherent redundancies in speech signals. Data compression is the topic of Chapter 27. One of the most efficient ways of compressing an audio signal is **Linear Predictive Coding (LPC)**, of which there are several variations and subgroups. Depending on the speech quality required, LPC can reduce the data rate to as little as 2-6 kbits/sec. We will revisit LPC later in this chapter with *speech synthesis*.

Скорость передачи данных 64k bits/sec представляет прямое приложение осуществления выборки и теории квантования к аудиосигналам. Методы для понижения скорости передачи данных далее основаны на сжатии потока данных, удаляя избыток свойственный речевым сигналам. Сжатие Данных – тема главы 27. Один из наиболее эффективных путей сжатия аудиосигнала – **Линейное прогнозирующее Кодирование**(Кодирование с Линейным Предсказанием; Линейное Предиктивное Кодирование; ЛПК), в котором имеется несколько вариаций и подгрупп. В зависимости от требуемого речевого качества, LPC может приводить скорость передачи данных как малые(незначительные) 2-6 2-6 kbits/sec. Мы повторно посетим LPC позже в этой главе с *синтезом речи*.

High Fidelity Audio
Высокая Звуковая Точность

Audiophiles demand the utmost sound quality, and all other factors are treated as secondary. If you had to describe the mindset in one word, it would be: *overkill*. Rather than just matching the abilities of the human ear, these systems are designed to *exceed* the limits of hearing. It's the only way to be sure that the reproduced music is pristine. Digital audio was brought to the world by the **compact laser disc**, or **CD**. This was a revolution in music; the sound quality of the CD system far exceeds older systems, such as records and tapes. DSP has been at the forefront of this technology.

Любители звукотехники требуют предельного звуковое качества, и все другие факторы обработаны как вторичные. Если бы Вы были должны описать *mindset* (*квалифицированный набор?*) одним словом, это было бы: массовое убийство. Скорее чем только соответствие способностям человеческого уха, эти системы предназначены, чтобы *превзойти* пределы слуха. Это - единственный способ убедиться, что воспроизведенная музыка первозданная. Цифровой звук был принесен к миру **компактным лазерным диском**, или **CD**. Это был переворот в музыке; звуковое качество системы CD далеко превышает старшие системы, типа грампластинок и магнитных лент лент. ЦОС был в центре деятельности этой технологии.

Figure 22-5 illustrates the surface of a compact laser disc, such as viewed through a high power microscope. The main surface is shiny (reflective of light), with the digital information stored as a series of dark pits burned on the surface with a laser. The information is arranged in a single track that spirals from the outside to the inside, the same as a phonograph record. The rotation of the CD is changed from about 210 to 480 rpm as the information is read from the outside to the inside of the spiral, making the scanning velocity a constant 1.2 meters per second. (In comparison, phonograph records spin at a fixed rate, such as 33, 45 or 78 rpm). During playback, an optical sensor detects if the surface is reflective or nonreflective, generating the corresponding binary information.

Рисунок 22-5 иллюстрирует поверхность компактного лазерного диска, типа просмотренного через мощный микроскоп. Основная поверхность блестящая (отражает свет), с цифровой информацией, сохраненной как ряд темных ям, выжженных на поверхности лазером. Информация размещается в единственной дорожке, которая расположена спиралью от внешней до внутренней части диска, так же, как делает запись фонограф. Вращение КОМПАКТ-ДИСКА изменяется приблизительно от 210 до 480 оборотов в минуту, по мере того как информация читается от внешней до внутренней части спирали, делая скорость просмотра постоянной 1.2 метра в секунду. (Для сравнении, фонограф делает запись вращения с фиксированной скоростью, типа 33, 45 или 78 оборота в минуту). В течение воспроизведения, оптический датчик обнаруживает, рефлексивна или нерефлексивна поверхность, генерируя соответствующие двоичные данные.

As shown by the geometry in Fig. 22-5, the CD stores about 1 bit per (μm)2 , corresponding to 1 million bits per (mm)2 , and 15 billion bits per disk. This is about the same feature size used in integrated circuit manufacturing, and for a good reason. One of the properties of light is that it cannot be focused to smaller than about one-half wavelength, or 0.3 μm. Since both integrated circuits and laser disks are created by optical means, the fuzziness of light below 0.3 μm limits how small of features can be used.

Как показано геометрией в рис. 22-5, КОМПАКТ-ДИСК сохраняет приблизительно 1 двоичный разряд(бит) на$(\mu m)^2$, соответствуя 1 миллиону битов на mm^2, и 15 миллиардам битов на диск. Это - относительно тот же размер размера особенности(пространства), используемого в интегральных схемах промышленностью, и серьезное основание(соображение). Одно из свойств индикатора - то, что это не может быть сосредоточено к меньшему, чем относительно половины длины волны, или 0.3 μm. Так как обе и интегральных схемы и лазерные диски созданы оптическими средствами, нечеткость индикатора ниже предела 0.3 μm, как маленькая из особенностей может использоваться.

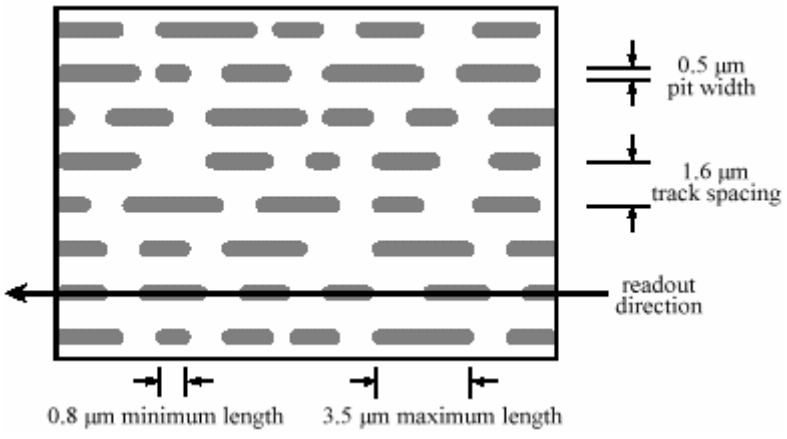

FIGURE 22-5. Compact disc surface. Micron size pits are burned into the surface of the CD to represent ones and zeros. This results in a data density of 1 bit per μm 2 , or one million bits per mm 2 . The pit depth is 0.16 μm.

РИСУНОК 22-5. Поверхность компакт-диска. Ямы размером 1 микрон выжжены на поверхности КОМПАКТ-ДИСКА, чтобы представить и 1 и нули. Это приводит к плотности записи данных 1 bit на μm^2, или один миллион bits на mm^2. Глубина ямки - 0.16 μm.

Figure 22-6 shows a block diagram of a typical compact disc playback system. The raw data rate is 4.3 million bits per second, corresponding to 1 bit each 0.28 μm of track length. However, this is in conflict with the specified geometry of the CD; each pit must be no shorter than 0.8 μm, and no longer than 3.5 μm. In other words, each binary *one* must be part of a group of 3 to 13 *ones*. This has the advantage of reducing the error rate due to the optical pickup, but how do you force the binary data to comply with this strange bunching?

На рисунке 22-6 показана блок-схема типичной системы воспроизведения компакт-диска. Скорость передачи исходных данных - 4.3 миллиона битов в секунду, соответствуя 1 дво-

ичному разряду(биту) на каждые 0.28 μm длины дорожки(фонограммы). Однако, это находится в конфликте с указанной геометрией КОМПАКТ-ДИСКА; каждая яма должна быть не короче чем 0.8 μm, и больше чем 3.5 μm. Другими словами, каждая двоичная *единица* должна быть частью группы из от 3 до 13 *единиц*. Это имеет преимущество сокращения уровень ошибки из-за оптического захвата(звукоснимателя), но как Вы вынуждаете двоичные данные выполнять это странное группирование?

(

The answer is an encoding scheme called **eight-to-fourteen modulation (EFM)**. Instead of directly storing a byte of data on the disc, the 8 bits are passed through a look-up table that pops out 14 bits. These 14 bits have the desired bunching characteristics, and are stored on the laser disc. Upon playback, the binary values read from the disc are passed through the inverse of the EFM look-up table, resulting in each 14 bit group being turned back into the correct 8 bits.

Ответ - схема кодирования называемая **модуляцией "восемь к четырнадцати" (EFM)**. Вместо непосредственно сохранения байта данных на диске, 8 битов пропускают через таблицу перекодировки, которая выталкивает 14 битов. Эти 14 битов имеют желательные характеристики группирования, и сохранены на лазерном диске. При воспроизведении, значения двоичных кодов читаются с диска, проходят через инверсию(декодировку)?) EFM таблицы перекодировки(конвертер частоты выборки?), приводя к группам по 14 двоичных разрядов(бит) каждая, которую поворачивают обратно в правильные(корректные) 8 битов.

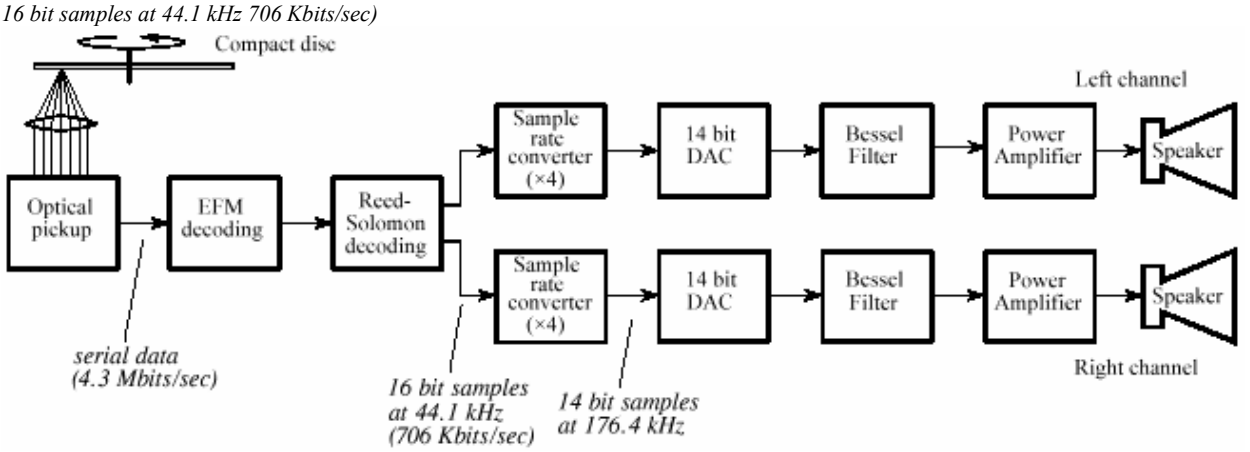

FIGURE 22-6
Compact disc playback block diagram. The digital information is retrieved from the disc with an optical sensor, corrected for EFM and Reed-Solomon encoding, and converted to stereo analog signals.

РИСУНОК 22-6
Блок-схема воспроизведения компакт-диска. Цифровая информация восстановлена(отыскана) с диска с оптическим датчиком, исправлена EFM(модуляцией 8 к 14) и кодированием Рида Соломона, и преобразована(конвертирована) в аналоговые стереосигналы.

In addition to **EFM**, the data are encoded in a format called **two-level Reed-Solomon coding**. This involves combining the left and right stereo channels along with data for error detection and correction. Digital errors detected during playback are either: *corrected* by using the redundant data in the encoding scheme, *concealed* by interpolating between adjacent samples, or *muted* by setting the sample value to zero. These encoding schemes result in the data rate being *tripled*, i.e., 1.4 Mbits/sec for the stereo audio signals versus 4.3 Mbits/sec stored on the disc.

В дополнение к EFM, данные закодированы в формате, называемом **кодом с двумя уровнями Рида Соломона**. Это включает в себя(подразумевает) объединение левых и правых каналов стерео наряду с данными для обнаружения ошибок и исправления. Цифровые

ошибки, обнаруженные в течение воспроизведения - также: *исправлены*, используя избыточные данные в схеме кодирования, *скрытой* интерполируя между смежными выборками, или *приглушенной*, устанавливая значение выборки к нулю. Эти схемы кодирования приводят к утраиваемой скорости передачи данных, то есть, 1.4 Mbits/sec для аудиосигналов стерео против 4.3 Mbits/sec сохранения на диске.

After decoding and error correction, the audio signals are represented as 16 bit samples at a 44.1 kHz sampling rate. In the simplest system, these signals could be run through a 16 bit DAC, followed by a low-pass analog filter. However, this would require high performance analog electronics to pass frequencies below 20 kHz, while rejecting all frequencies above 22.05 kHz, ½ of the sampling rate. A more common method is to use a **multirate** technique, that is, convert the digital data to a higher sampling rate before the DAC. A factor of four is commonly used, converting from 44.1 kHz to 176.4 kHz. This is called **interpolation**, and can be explained as a two step process (although it may not actually be carried out this way). First, three samples with a value of zero are placed between the original samples, producing the higher sampling rate. In the frequency domain, this has the effect of duplicating the 0 to 22.05 kHz spectrum three times, at 22.05 to 44.1 kHz, 41 to 66.15 kHz, and 66.15 to 88.2 kHz. In the second step, an efficient *digital* filter is used to remove the newly added frequencies.

После декодирования и исправления ошибок, аудио-сигналы представлены как выборки 16 двоичных разрядов(бит) с частотой выборки 44.1 кГц. В самой простой системе, эти сигналы могли бы пробегать через ЦАП 16 двоичных разрядов(бит), сопровождаемый аналоговым фильтром низкой частоты. Однако, это требовало бы, чтобы высокоэффективная аналоговая электроника передала частоты ниже 20 кГц, при отклонении всех частот более чем 22.05 кГц, ½ частоты выборки. Более общий(обычный) метод состоит в том, чтобы использовать методику мультичастоты, то есть преобразуйте цифровые данные к более высокой частоте выборки перед ЦАП. Коэффициент(фактор) четыре обычно используется, преобразовывающий от 44.1 кГц до 176.4 кГц. Это называется **интерполяцией**, и можно объясняться как два шага процесса (хотя это не может фактически быть выполнено этим способом). Во первых, три выборки со значением нуля помещены между первоначальными выборками, производя более высокую частоту выборки. В частотном домене, это имеет эффект дублирования спектра от 0 до 22.05 кГц три раза, от 22.05 до 44.1 кГц, от 41 до 66.15 кГц, и от 66.15 до 88.2 кГц. Во втором шаге, эффективный цифровой фильтр используется, чтобы удалить недавно добавленные частоты.

The sample rate increase makes the sampling interval smaller, resulting in a smoother signal being generated by the DAC. The signal still contains frequencies between 20 Hz and 20 kHz; however, the Nyquist frequency has been increased by a factor of four. This means that the analog filter only needs to pass frequencies below 20 kHz, while blocking frequencies above 88.2 kHz. This is usually done with a three pole Bessel filter. Why use a *Bessel* filter if the ear is insensitive to phase? Overkill, remember?

Увеличение частоты выборки делает выборочный интервал меньшим, приводя к более гладкому сигналу, сгенерированному ЦАП. Сигнал все еще содержит частоты между 20 Гц и 20 кГц; однако, частота Найквиста была увеличена на коэффициент(фактор) четыре. Это означает, что аналоговый фильтр должен передать только частоты ниже 20 кГц, при блокировании частот более чем 88.2 кГц. Это обычно делается с 3 полюсным фильтром Бесселя(Бесселери). Почему используется фильтр Бесселя, если ухо нечувствительно к фазе? Массовое убийство(полное поражение?), помните?

Since there are four times as many samples, the number of bits per sample can be reduced from 16 bits to 14 bits, without degrading the sound quality. The correction needed to compensate for the zeroth order hold of the DAC can be part of either the analog or digital filter.

Так как имеются четыре раза, так много выборок, число битов на выборку может быть сокращено от 16 битов до 14 битов, без того, чтобы ухудшить качество звука. Исправление $\sin(x)/x$, необходимое для компенсации хранения нулевого порядка ЦАП может быть частью или аналогового или цифрового фильтра.

Audio systems with more than one channel are said to be in **stereo** (from the Greek word for *solid*, or *three-dimensional*). Multiple channels send sound to the listener from different directions, providing a more accurate reproduction of the original music. Music played through a monaural (one channel) system often sounds artificial and bland. In comparison, a good stereo reproduction makes the listener feel as if the musicians are only a few feet away. Since the 1960s, high fidelity music has used two channels (left and right), while motion pictures have used four channels (left, right, center, and surround). In early stereo recordings (say, the Beatles or the Mamas And The Papas), individual singers can often be heard in only one channel or the other. This rapidly progressed into a more sophisticated **mix-down**, where the sound from many microphones in the recording studio is combined into the two channels. Mix-down is an art, aimed at providing the listener with the perception of *being there*.

Звуковые системы больше чем с одним каналом, как считают, находятся в стерео (от греческого слова *solid* (*телесный; объемный?*), или *three-dimensional* (*трехмерный*)). Множественные каналы посылают звук слушателю от различных направлений, обеспечивая более точное воспроизводство первоначальной музыки. Музыка, проигранная через монофоническую систему (один канал) часто звучит искусственной(поддельной) и мягкой(декоративной). Для сравнения, хорошее воспроизводство стерео создает восприятие слушателя, как будто музыканты - на расстоянии только несколько футов. Начиная с 1960-ых, музыка высокой точности использовала два канала (левый и правый), в то время как кинофильмы использовали четыре канала (левый, правый, средний, и фоновый(surround - окружающий)). В ранних стерео записях (скажем, Beatles или Мамы и Папы), индивидуальных певцов можно часто слышать только в одном или в другом канале. Это быстро прогрессировало в более сложное(изощренное) объединение всех треков(сведение фонограмм) в одну, где звук от многих микрофонов в студии записи объединен в два канала. Объединение всех треков в один(сведение фонограмм) - искусство, нацеленное на обеспечение слушателя восприятием *being there(присутствия там?)*.

The four channel sound used in motion pictures is called **Dolby Stereo**, with the home version called **Dolby Surround Pro Logic**. ("Dolby" and "Pro Logic" are trademarks of Dolby Laboratories Licensing Corp.). The four channels are encoded into the standard left and right channels, allowing regular two-channel stereo systems to reproduce the music. A Dolby decoder is used during playback to recreate the four channels of sound. The left and right channels, from speakers placed on each side of the movie or television screen, is similar to that of a regular two-channel stereo system. The speaker for the center channel is usually placed directly above or below the screen. Its purpose is to reproduce speech and other visually connected sounds, keeping them firmly centered on the screen, regardless of the seating position of the viewer/listener. The surround speakers are placed to the left and right of the listener, and may involve as many as twenty speakers in a large auditorium. The surround channel only contains midrange frequencies (say, 100 Hz to 7 kHz), and is *delayed* by 15 to 30 milliseconds. This delay makes the listener perceive that speech is coming from the screen, and not the sides. That is, the listener hears the speech coming from the front, followed by a delayed version of the speech coming from the

sides. The listener's mind interprets the delayed signal as a reflection from the walls, and ignores it.

Четыре звуковых канала, используемые в кинофильмах называется **Dolby Стерео**, с домашней версией, по имени **Dolby Surround Pro Logic**. ("Dolby" и " Pro Logic " являются торговыми марками Dolby Laboratories Licensing Corp). Четыре канала закодированы в стандартно, в левом и правом каналах, позволяя регулярным(обычным; симметричным?) двухканальным стереосистемам воспроизвести музыку. Декодер Долби используется в течение воспроизведения, чтобы воссоздать четыре канала звука. Динамики левого и правого каналов, размещены по обе стороны от кино или телевизионного экрана, являются подобными таковому регулярной(обычной; симметричной?) двухканальной стереосистемы. Динамик для среднего канала обычно помещается непосредственно выше или ниже экрана. Его цель состоит в том, чтобы воспроизвести речь и другие визуально связанные звуки, сохраняя их твердо центрированными на экране, независимо от позиции видео/ слушателя. Окружающие динамики помещены слева и справа от слушателя, и могут включать в себя целых двадцать динамиков для большой аудитории. Канал окружения содержит только частоты среднего диапазона (скажем, 100 Гц - 7 кГц), и *запаздывает* на 15 - 30 миллисекунд. Эта задержка заставит слушателя чувствовать, что речь исходит от экрана, а не стороны. То есть слушатель слышит речь, исходящую спереди, сопровождаемой отсроченной версией речи(эхом), исходящей с боков. Мнение(Мысль; интеллект; психика?) слушателя интерпретирует отсроченный сигнал как отражение от стенок, и игнорирует это.

Companding
Компандирование

The data rate is important in telecommunication because it is directly proportional to the *cost* of transmitting the signal. Saving bits is the same as saving money. **Companding** is a common technique for reducing the data rate of audio signals by making the quantization levels *unequal*. As previously mentioned, the loudest sound that can be tolerated (120 dB SPL) is about one-million times the amplitude of the weakest sound that can be detected (0 dB SPL). However, the ear cannot distinguish between sounds that are closer than about 1 dB (12% in amplitude) apart. In other words, there are only about 120 different loudness levels that can be detected, spaced logarithmically over an amplitude range of one-million.

Скорость передачи данных важна в дальней связи, потому что это - непосредственно пропорционально стоимости передачи сигнала. Сохранение битов - тот же самое как сохранение денег. Компандирование - обычная методика для сокращения скорости передачи данных аудиосигналов, делая неравными уровни квантований. Как предварительно упомянуто, самый громкий звук, который может допускаться (120 dB SPL) – в одну-миллионную раза, по отношению к амплитуде самого слабого звука, который может быть услышан (0 dB SPL). Однако, ухо не может разделить звуки, которые являются ближе чем приблизительно 1 dB (12 % в амплитуде). Другими словами, имеются только приблизительно 120 различных уровней громкости, которые могут быть обнаружены, расположенные через логарифмические интервалы по амплитудному диапазону один миллион.

This is important for digitizing audio signals. If the quantization levels are equally spaced, 12 bits must be used to obtain telephone quality speech. However, only 8 bits are required if the quantization levels are made *unequal*, matching the characteristics of human hearing. This is

quite intuitive: if the signal is small, the levels need to be very close together; if the signal is large, a larger spacing can be used.

Это важно для отцифровывания аудиосигналов. Если уровни квантований одинаково располагаются, 12 битов должны использоваться, чтобы получить телефонную качественную речь. Однако, только 8 битов требуются, если уровни квантований сделаны неравными, соответствуя характеристикам человеческого слуха. Это весьма интуитивно: если сигнал маленький, уровни должны быть очень близко друг к другу; если сигнал большой, больший интервал может использоваться.

Companding can be carried out in three ways: (1) run the analog signal through a nonlinear circuit before reaching a linear 8 bit ADC, (2) use an 8 bit ADC that internally has unequally spaced steps, or (3) use a linear 12 bit ADC followed by a digital look-up table (12 bits in, 8 bits out). Each of these three options requires the same nonlinearity, just in a different place: an analog circuit, an ADC, or a digital circuit.

Компандирование может быть выполнено тремя способами: (1) прогоняют аналоговый сигнал через нелинейную схему перед достижением линейного 8 разрядного АЦП, (2) используют 8 разрядный АЦП, который внутренне неравноценно расположил шаги, или (3) используют линейный 12 разрядный АЦП, сопровождаемых цифровой таблицей перекодировки (12 битов в 8 битов и обратно). Каждый из этих трех параметров требует той же самой нелинейности, только в различном месте: аналоговая схема, АЦП, или цифровая схема.

Two nearly identical standards are used for companding curves: **μ255 law** (also called **mu law**), used in North America, and **"A" law**, used in Europe. Both use a logarithmic nonlinearity, since this is what converts the spacing detectable by the human ear into a linear spacing. In equation form, the curves used in μ255 law and "A" law are given by:

Два почти идентичных стандарта используются для кривых компандирования: **закон μ255** (также называемый закон мю), используемый в Северной Америке, и **закон "A"**, используемый в Европе. Оба используют логарифмическую нелинейность, так как это - то, что преобразовывает интервал, обнаруживаемый человеческим ухом в линейный интервал. В форме уравнения, кривые, используемые в законеμ255 и законе "A" даются:

EQUATION 22-1
Mu law companding. This equation provides the nonlinearity for μ255 law companding. The constant, μ, has a value of 255, accounting for the name of this standard.

$$y = \frac{\ln(1+\mu x)}{\ln(1+\mu)} \qquad \text{for } 0 \le x \le 1$$

УРАВНЕНИЕ 22-1
компандирование законом мю. Это уравнение обеспечивает нелинейность для закона компандирования μ255. Константа, μ, имеет значение 255, объясняя название этого стандарта.

EQUATION 22-2. "A" law companding. The constant, *A*, has a value of 87.6.
УРАВНЕНИЕ 22-2. Закон "A" компандирования. Константа, A, имеет значение 87.6.

$$y = \frac{1+\ln(Ax)}{1+\ln(A)} \qquad \text{for } 1/A \le x \le 1$$

$$y = \frac{Ax}{1+\ln(A)} \qquad \text{for } 0 \le x \le 1/A$$

Figure 22-7 graphs these equations for the input variable, *x*, being between –1 and +1, resulting in the output variable also assuming values between -1 and +1. Equations 22-1 and 22-2 only handle positive input values; portions of the curves for negative input values are found from

symmetry. As shown in (a), the curves for μ255 law and "A" law are nearly identical. The only significant difference is near the origin, shown in (b), where μ255 law is a smooth curve, and "A" law switches to a straight line.

На рисунке 22-7 показаны графики этих уравнений для входной переменной, *x*, находящийся между -1 и +1, приводя к переменной выхода, также принимающей значения между -1 и +1. Уравнения 22-1 и 22-2 управляют только положительными значениями ввода; части кривых для отрицательных входных значений найдены от симметрии. Как показано в (a), кривые для закона μ255 и закона "A" закона почти идентичны. Единственная существенная разность - около начала координат, показанного в (b), где гладкая кривая закона μ255 и закона "A" - переходит в прямую линию.

Producing a stable nonlinearity is a difficult task for analog electronics. One method is to use the logarithmic relationship between current and voltage across a *pn* diode junction, and then add circuitry to correct for the ghastly temperature drift. Most companding circuits take another strategy: approximate the nonlinearity with a group of straight lines. A typical scheme is to approximate the logarithmic curve with a group of 16 straight segments, called **cords**. The first bit of the 8 bit output indicates if the input is positive or negative. The next three bits identify which of the 8 positive or 8 negative cords is used. The last four bits break each cord into 16 equally spaced increments. As with most integrated circuits, companding chips have sophisticated and proprietary internal designs. Rather than worrying about what goes on inside of the chip, pay the most attention to the pinout and the specification sheet.

Создание устойчивой нелинейности - трудная задача для аналоговой электроники. Один метод состоит в том, чтобы использовать логарифмические отношения между током и напряжением через pn перехода диода, и затем добавить схему, чтобы устранить ужасный температурный дрейф. Большинство цепей компандирования принимают другую стратегию: аппроксимируют нелинейность с группой прямых линии. Типичная схема состоит в том, чтобы аппроксимировать логарифмическую кривую с группой 16 прямолинейных сегментов, называемых **cords(шнурами)**. Первый двоичный разряд(бит) из выхода 8 двоичных разрядов(бит) указывает, является ли ввод положительным или отрицательным. Следующие три бита выделяют, который из 8 положительных или 8 отрицательных шнуров используется. Последние четыре бита разбивают каждый шнур на 16 одинаково раздельных приращений. Как с большинством интегральных схем, чипы компандирования имеют сложные(изощренные; тонкие) и частные внутренние проекты. Скорее чем беспокойство о том что происходит внутри чипа, уделите больше внимания цоколевке(разводке выводов) и листу спецификации.

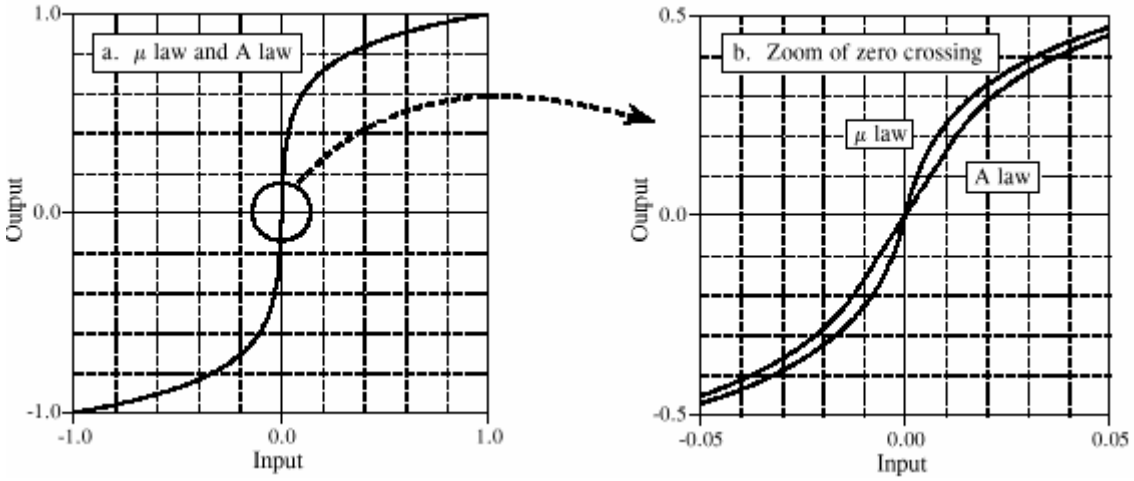

FIGURE 22-7

Companding curves. The μ255 law and "A" law companding curves are nearly identical, differing only near the origin. Companding increases the amplitude when the signal is small, and decreases it when it is large.
РИСУНОК 22-7
Кривые Компандирования. Кривые компандирования закона μ255 и закона "A" почти идентичны, отличаясь только около начала координат. Компандирование увеличивает амплитуду, когда сигнал маленький, и уменьшает его, когда он большой.

Speech Synthesis and Recognition
Синтез и Распознавание Речи

Computer generation and recognition of speech are formidable problems; many approaches have been tried, with only mild success. This is an active area of DSP research, and will undoubtedly remain so for many years to come. You will be very disappointed if you are expecting this section to describe how to build speech synthesis and recognition circuits. Only a brief introduction to the typical approaches can be presented here. Before starting, it should be pointed out that most commercial products that produce human sounding speech do not *synthesize* it, but merely play back a digitally recorded segment from a human speaker. This approach has great sound quality, but it is limited to the prerecorded words and phrases.

Компьютерное порождение и распознавание речи - огромные проблемы; много подходов были испытаны, только с умеренным успехом. Это - активная область исследования ЦОС, и несомненно останется такой на много лет вперед. Вы будете очень разочарованы, если Вы ожидаете, что этот раздел описывает, как формировать речевой синтез и цепи распознавания. Только краткое введение в типичные подходы может быть представлено здесь. Перед стартом, это должно быть указано, что большинство коммерческих изделий, которые производят человеческую звучащую речь, не *синтезирует* это, но просто воспроизводят в цифровой форме записанный сегмент от динамика человека. Этот подход имеет высокое звуковое качество, но ограничен записанными заранее словами и фразами.

Nearly all techniques for speech synthesis and recognition are based on the model of human speech production shown in Fig. 22-8. Most human speech sounds can be classified as either **voiced** or **fricative**. Voiced sounds occur when air is forced from the lungs, through the vocal cords, and out of the mouth and/or nose. The vocal cords are two thin flaps of tissue stretched across the air flow, just behind the Adam's apple. In response to varying muscle tension, the vocal cords vibrate at frequencies between 50 and 1000 Hz, resulting in periodic puffs of air being injected into the throat. Vowels are an example of voiced sounds. In Fig. 22-8, voiced sounds are represented by the pulse train generator, with the pitch (i.e., the fundamental frequency of the waveform) being an adjustable parameter.

Почти все методы для речевого синтеза и распознавания основаны на модели речи человека, показанной на рис. 22-8. Наиболее человеческие речевые звуки могут быть классифицированы как или **вокал** или **фрикатив**. Вокальные звуки происходят, когда воздух выходит от легких, через голосовые связки изо рта и-или носа. Голосовые связки - две тонких откидных створки ткани, протянутой поперек тока воздуха, только позади Кадыка. В ответ на изменяющуюся ригидность мышц, голосовые связки вибрируют в частотах между 50 и 1000 Гц, приводя к периодическим затяжкам воздуха, вводимого в горло. Гласные - пример высказанных звуков. В рис. 22-8, высказанные звуки представлены генератором импульсной последовательность, с тоном (то есть, основная частота формы волны) являющийся корректируемым параметром.

In comparison, *fricative* sounds originate as random noise, not from vibration of the vocal cords. This occurs when the air flow is nearly blocked by the tongue, lips, and/or teeth, resulting in air

turbulence near the constriction. Fricative sounds include: *s*, *f*, *sh*, *z*, *v*, and *th*. In the model of Fig. 22-8, fricatives are represented by a *noise generator*.

Для сравнения, *фрикативные* звуки происходят как случайный шум, не от вибрации голосовых связок. Это происходит, когда ток воздуха почти блокирован языком, губами, и-или зубами, приводя к воздушной турбулентности около сжатия. фрикативные звуки включит: *s*, *f*, *sh*, *z*, *v*, и *th*. В модели рис. 22-8, фрикативы представлены *генератором шума*.

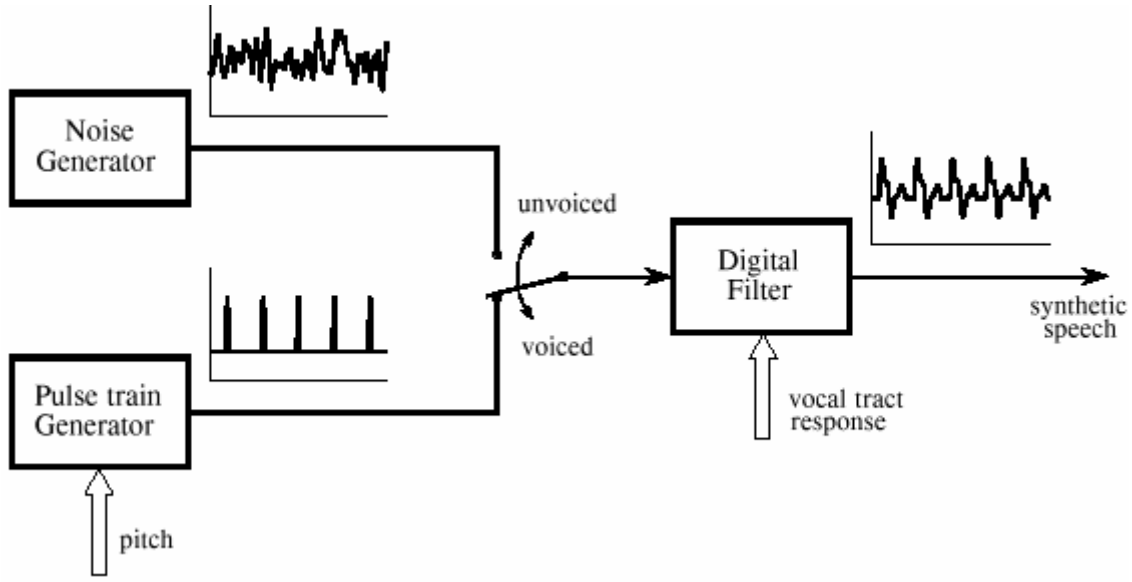

FIGURE 22-8
Human speech model. Over a short segment of time, about 2 to 40 milliseconds, speech can be modeled by three parameters: (1) the selection of either a periodic or a noise excitation, (2) the pitch of the periodic excitation, and (3) the coefficients of a recursive linear filter mimicking the vocal tract response.
РИСУНОК 22-8
Модель человеческой речи. По короткому сегменту времени, приблизительно от 2 до 40 миллисекунд, речь может быть смоделирована тремя параметрами: (1) выбранным или периодическим или шумовым возбуждением, (2) шагом(тоном) периодического возбуждения, и (3) коэффициентами рекурсивного линейного фильтра характеристики вокального тракта.

Both these sound sources are modified by the acoustic cavities formed from the tongue, lips, mouth, throat, and nasal passages. Since sound propagation through these structures is a linear process, it can be represented as a linear filter with an appropriately chosen impulse response. In most cases, a *recursive* filter is used in the model, with the recursion coefficients specifying the filter's characteristics. Because the acoustic cavities have dimensions of several centimeters, the frequency response is primarily a series of resonances in the kilohertz range. In the jargon of audio processing, these resonance peaks are called the **format frequencies**. By changing the relative position of the tongue and lips, the format frequencies can be changed in both frequency and amplitude.

Figure 22-9 shows a common way to display speech signals, the **voice spectrogram**, or **voiceprint**. The audio signal is broken into short segments, say 2 to 40 milliseconds, and the FFT used to find the frequency spectrum of each segment. These spectra are placed side-by-side, and converted into a grayscale image (low amplitude becomes light, and high amplitude becomes dark). This provides a graphical way of observing how the frequency content of speech changes with time. The segment length is chosen as a tradeoff between *frequency resolution* (favored by longer segments) and *time resolution* (favored by shorter segments).

На рисунке 22-9 показан обычный способ отображать речевые сигналы, **спектрограмма голоса**, или **сонограмма**. Аудиосигнал разбит на короткие сегменты, скажем от 2 до 40 миллисекунд, и БПФ использовано чтобы найти спектр частот каждого сегмента. Эти спектры помещены бок о бок, и преобразованы в полутоновое изображение (низкая амплитуда становится световой, и высокая амплитуда, становится темной). Это обеспечивает графический путь наблюдения, как частотное содержание речи изменяется со временем. Длина сегмента выбрана как сделка между частотной разрешающей способностью (оцененной более длинными сегментами) и разрешающей способностью времени (оцененной более короткими сегментами).

As demonstrated by the *a* in *rain*, voiced sounds have a periodic time domain waveform, shown in (a), and a frequency spectrum that is a series of regularly spaced harmonics, shown in (b). In comparison, the *s* in *storm*, shows that fricatives have a noisy time domain signal, as in (c), and a noisy spectrum, displayed in (d). These spectra also show the shaping by the format frequencies for both sounds. Also notice that the time-frequency display of the word *rain* looks similar both times it is spoken.

Как демонстрируется *a* в слове *rain*(слово английское!), высказанные звуки имеют период формы волны домена времени, показанную в (a), и спектр частот, который является рядом регулярно раздельных гармоник(флажолетов), показанный в (b). Для сравнения, *s* в слове *storm* (слово английское!), показывает, что фрикативы имеют шумный сигнал домена времени, как в (c), и шумный спектр, отображенный в (d). Эти спектры также показывают формирование формата частоты для обоих звуков. Также обратите внимание, что дисплей частоты времени слова *rain* выглядит подобным оба раза, когда его произносят.

Over a short period, say 25 milliseconds, a speech signal can be approximated by specifying three parameters: (1) the selection of either a periodic or random noise excitation, (2) the frequency of the periodic wave (if used), and (3) the coefficients of the digital filter used to mimic the vocal tract response. Continuous speech can then be synthesized by continually updating these three parameters about 40 times a second. This approach was responsible for one the early commercial successes of DSP: the *Speak & Spell*, a widely marketed electronic learning aid for children. The sound quality of this type of speech synthesis is poor, sounding very mechanical and not quite human. However, it requires a very low data rate, typically only a few kbits/sec.

В течение короткого периода, скажем 25 миллисекунд, речевой сигнал может быть аппроксимирован, определяя три параметра: (1) выбор или периодического или случайного шумового возбуждения, (2) частота периодической волны (если используется), и (3) коэффициенты цифрового фильтра, используемые чтобы подражать вокальному частотному тракту. Слитная речь может тогда синтезироваться, непрерывно модифицируя эти три параметра приблизительно 40 раз в секунду. Этот подход был одним из важных(достойных) ранних коммерческих успехов ЦОС: *Говорим и Записываем* по буквам, широко выставленная на продажу электронная помощь изучения для детей. Звуковое качество этого типа речевого синтеза бедное(плохое), звуча очень механическим и не совсем человеческим. Однако, это требует очень низкой скорости передачи данных, типично только несколько kbits/sec.

This is also the basis for the **linear predictive coding** (**LPC**) method of speech compression. Digitally recorded human speech is broken into short segments, and each is characterized according to the three parameters of the model. This typically requires about a dozen bytes per segment, or 2 to 6 kbytes/sec. The segment information is transmitted or stored as needed, and then reconstructed with the speech synthesizer.

Это - также основание для метода **линейного прогнозирующего кодирования** (линейного предиктивного кодирования - ЛПК) речевого сжатия. В цифровой форме зарегистрированная человеческая речь разбита на короткие сегменты, и каждый характеризован согласно трем параметрам модели. Это типично требует около дюжины байтов на сегмент, или от 2 до 6 kbytes/sec. Информация сегмента передана или сохранена как необходимо, и тогда восстановлена речевым синтезатором.

Speech recognition algorithms take this a step further by trying to recognize patterns in the extracted parameters. This typically involves comparing the segment information with templates of previously stored sounds, in an attempt to identify the spoken words. The problem is, this method does not work very well. It is useful for some applications, but is far below the capabilities of human listeners. To understand why speech recognition is so difficult for computers, imagine someone unexpectedly speaking the following sentence:

Алгоритмы Распознавания речи берут это шагом далее, пробуя распознать образцы в извлеченных параметрах. Это типично включает в себя(подразумевает) сравнение информации сегмента с шаблонами предварительно сохраненных звуков, в попытке идентифицировать говорящиеся слова. Проблема, этот метод работает не очень хорошо. Это полезно для некоторых приложений, но далеко ниже возможностей человеческих слушателей. Чтобы понимать, почему распознавание речи настолько трудно для компьютеров, вообразите кого ни будь с неожиданным(неправильным) произношением следующего предложения:

Larger run medical buy dogs fortunate almost when.
Большие выполненные медицинские собаки покупки, удачливые почти, когда.

Of course, you will not understand the meaning of this sentence, because it has none. More important, you will probably not even understand all of the individual words that were spoken. This is basic to the way that humans perceive and understand speech. Words are recognized by their sounds, but also by the *context* of the sentence, and the *expectations* of the listener. For example, imagine hearing the two sentences:

Конечно, Вы не поймете значение этого предложения, потому что это не имеет ни одного. Более важно, Вы вероятно, даже не поймете правильно все индивидуальные слова, которые сказаны. Это основное к пути, которым люди чувствуют и понимают речь. Слова распознаются по их звукам, но также и *контекстом* предложения, и ожиданием слушателя. Например, вообразите слышать два предложения:

The child wore a spider ring on Halloween.
Ребенок носил кольцо паукообразной гемангиомы на Halloween(на кануне дня всех святых).

He was an American spy during the war.
Он был Американский шпион в течение войны.

Even if exactly the same sounds were produced to convey the underlined words, listeners *hear* the correct words for the context. From your accumulated knowledge about the world, you know that children don't wear secret agents, and people don't become spooky jewelry during wartime. This usually isn't a conscious act, but an inherent part of human hearing.

Даже если точно те же самые звуки были произведены, чтобы передать подчеркнутые слова, слушатели слышат правильные слова по контексту. От вашего накопленного знания относительно мира, Вы знаете, что дети не носят секретных агентов(шпионов), и люди не носят драгоценностей в течение военного времени. Это обычно не сознательное действие, но свойственная часть человеческого слушания(слухового восприятия).

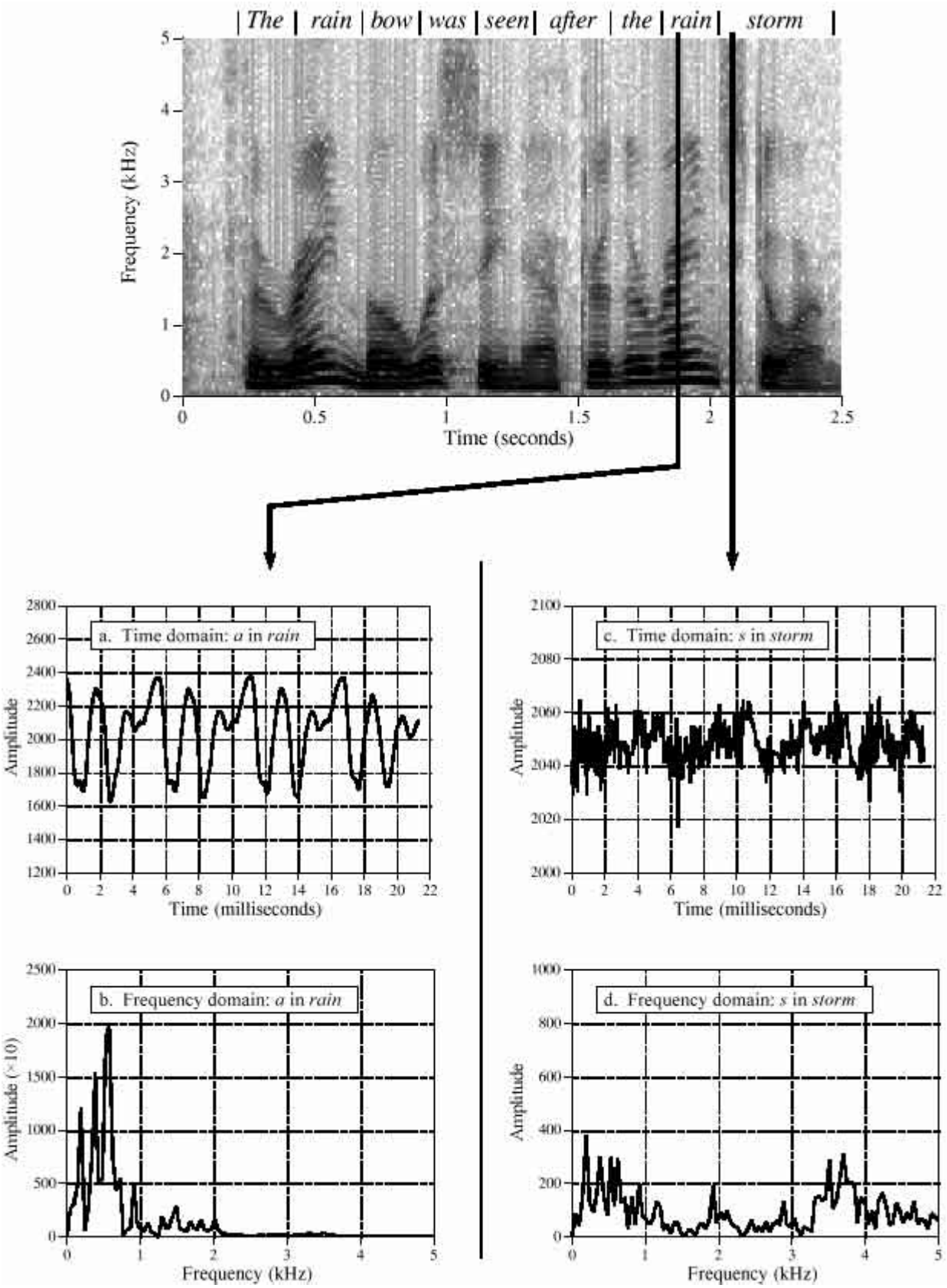

РИСУНОК 22-9
Спектрограмма Голоса. Спектрограмма фразы: "*The rainbow was seen after the rain storm.*" Рисунки (a) и (b) показывают в сигнале домена времени частоту звука *a* произнесенного в слове *rain*. Рисунки (c) и (d) в сигнале домена времени и частоту звука *s* в слове *storm*..

Most speech recognition algorithms rely only on the sound of the individual words, and not on their context. They attempt to *recognize words*, but not to *understand speech*. This places them at a tremendous disadvantage compared to human listeners. Three annoyances are common in speech recognition systems: (1) The recognized speech must have distinct pauses between the words. This eliminates the need for the algorithm to deal with phrases that sound alike, but are composed of different words (i.e., *spider ring* and *spy during*). This is slow and awkward for people accustomed to speaking in an overlapping flow. (2) The vocabulary is often limited to only a few hundred words. This means that the algorithm only has to search a limited set to find the best match. As the vocabulary is made larger, the recognition time and error rate both increase. (3) The algorithm must be *trained* on each speaker. This requires each person using the system to speak each word to be recognized, often needing to be repeated five to ten times. This personalized database greatly increases the accuracy of the word recognition, but it is inconvenient and time consuming.

Большинство алгоритмов распознавания речи основано только на звуке индивидуальных слов, а не на их контексте. Они пытаются признавать слова, но не понимать речь. Это размещает их в огромный недостаток, сравненный с человеческими слушателями. Три раздражения обычны в системах распознавания речи: (1) признанная речь должна иметь отличные паузы между словами. Это устраняет потребность в алгоритме, чтобы иметь дело с фразами, которые звучат подобными, но составлены из различных слов (то есть, *spider ring* и *spy during*). Это медленно и неуклюже для людей, приученных к разговору в накладывающемся потоке. (2) словарь часто ограничивается только несколькими сотнями слов. Это означает, что алгоритм должен искать только ограниченный набор, чтобы найти лучшее соответствие. Поскольку словарь сделан большим, время распознавания и уровень ошибки увеличены. (3) алгоритм должен быть обучен на каждом динамике. Это часто требует от каждого человека, использующего систему, повторить каждое слово, которое должно будет опознано, пять - десять раз. Эта индивидуализированная база данных очень увеличивает точность распознавания слова, но это неудобно и потребляет время.

The prize for developing a successful speech recognition technology is enormous. Speech is the quickest and most efficient way for humans to communicate. Speech recognition has the potential of replacing writing, typing, keyboard entry, and the electronic control provided by switches and knobs. It just needs to work a little better to become accepted by the commercial marketplace. Progress in speech recognition will likely come from the areas of artificial intelligence and neural networks as much as through DSP itself. Don't think of this as a technical *difficulty*; think of it as a technical *opportunity*.

Приз за разработку успешной технологии распознавания речи огромен. Речь - самый быстрый и наиболее эффективный путь для людей, чтобы связаться. Распознавание речи имеет потенциал замены записи, печатания, входа клавиатуры, и электронного управления, обеспеченного выключателями и кнопками. Только требоваться работать немного лучше, чтобы стать принятым коммерческим рынком. Прогресс распознавания речи будет вероятно исходить из областей искусственного интеллекта и невральных сетей (и вообще) чего угодно через ЦОС непосредственно. Не думайте о этом как о технической *трудности*; думайте о этом как о технической *возможности*.

Nonlinear Audio Processing
Нелинейная Звуковая Обработка

Digital filtering can improve audio signals in many ways. For instance, *Wiener filtering* can be used to separate frequencies that are mainly signal, from frequencies that are mainly noise (see Chapter 17). Likewise, *deconvolution* can compensate for an undesired convolution, such as in the restoration of old recordings (also discussed in Chapter 17). These types of linear techniques are the backbone of DSP. Several *nonlinear* techniques are also useful for audio processing. Two will be briefly described here.

Цифровая фильтрация может улучшать аудиосигналы многими способами. Например, *фильтрация Винера* может использоваться, чтобы отделить частоты, которые являются главным образом сигналом, от частот, которые являются главным образом шумом (см. главу 17). Аналогично, *деконволюция* может компенсировать нежелательную свертку, типа в восстановлении старых записей (также обсуждалось в главе 17). Эти типы линейных методов - основа ЦОС. Несколько *нелинейных* методов также полезны для звуковой обработки. Два из них будет кратко описано здесь.

The first nonlinear technique is used for reducing wideband noise in speech signals. This type of noise includes: magnetic tape hiss, electronic noise in analog circuits, wind blowing by microphones, cheering crowds, etc. Linear filtering is of little use, because the frequencies in the noise completely overlap the frequencies in the voice signal, both covering the range from 200 hertz to 3.2 kHz. How can two signals be separated when they overlap in both the time domain *and* the frequency domain?

Первая нелинейная методика используется для сокращения широкополосного шума в речевых сигналах. Этот тип шума включает: посторонний шум магнитной ленты, электронный шум в аналоговых цепях, перемотка, дыхание просачивающееся через микрофон, приветствия толпы(аплодисменты; приветственные выкрики), и т.д. Линейная фильтрация имеет небольшое применение, потому что частоты в шуме полностью накладываются на частоты в сигнале голоса, оба покрывают диапазон от 200 герц до 3.2 кГц. Как могут быть разделены два сигнала когда они накладываются, и в домене времени и в частотном домене?

Here's how it is done. In a short segment of speech, the amplitude of the frequency components are greatly *unequal*. As an example, Fig. 22-10a illustrates the frequency spectrum of a 16 millisecond segment of speech (i.e., 128 samples at an 8 kHz sampling rate). Most of the signal is contained in a few large amplitude frequencies. In contrast, (b) illustrates the spectrum when only random noise is present; it is very irregular, but more uniformly distributed at a low amplitude.

Имеется, как это сделано. В коротком сегменте речи, амплитуды частотных компонентов очень *неравны*. Как пример, рис. 22-10 а иллюстрирует спектр частот из 16 миллисекундных сегментов речи (то есть, 128 выборок с частотой выборки 8 кГц). Большинство сигнала содержится в нескольких больших амплитудных частотах. Напротив, (b) иллюстрирует спектр, когда только случайный шум присутствует; это очень неправильно, но более равномерно распределено в низкой амплитуде.

Now the key concept: if both signal and noise are present, the two can be partially separated by looking at the *amplitude* of each frequency. If the amplitude is large, it is probably mostly signal, and should therefore be retained. If the amplitude is small, it can be attributed to mostly noise, and should therefore be discarded, i.e., set to zero. Mid-size frequency components are adjusted in some smooth manner between the two extremes.

Теперь ключевой концепт: если и сигнал и шум присутствуют, они могут быть частично отделены смотрящий на амплитуду каждой частоты. Если *амплитуда* большая, это - вероятно главным образом сигнал, и поэтому должен быть сохранен. Если амплитуда маленькая, это может быть приписано главным образом шуму, и поэтому должно быть отвергнуто, то есть, устанавливать на нуль. Частотные компоненты среднего размера откорректированы некоторым гладким способом между этими двумя крайностями.

Another way to view this technique is as a *time varying Wiener filter*. As you recall, the frequency response of the Wiener filter passes frequencies that are mostly signal, and rejects frequencies that are mostly noise. This requires a knowledge of the signal and noise spectra *beforehand*, so that the filter's frequency response can be determined. This nonlinear technique uses the same idea, except that the Wiener filter's frequency response is recalculated for each segment, based on the spectrum *of that segment*. In other words, the filter's frequency response changes from segment-to-segment, as determined by the characteristics of the signal itself.

Другой способ рассматривать эту методику как *изменение времени Винеровским фильтром*. Как Вы помните, частотная характеристика фильтра Винера передает частоты, которые являются главным образом сигналом, и отклоняет частоты, которые являются главным образом шумом. Это требует знания сигнала и шумовых спектров *заранее*, так, чтобы частотная характеристика фильтра могла быть определена. Эта нелинейная методика использует ту же самую идею, за исключением того, что частотная характеристика Винеровского фильтра повторно рассчитана для каждого сегмента, основано на спектре *этого сегмента*. Другими словами, частотная характеристика фильтра изменяется от "сегмента к сегменту", как определено в соответствии характеристиками из сигнала непосредственно(автоматически?).

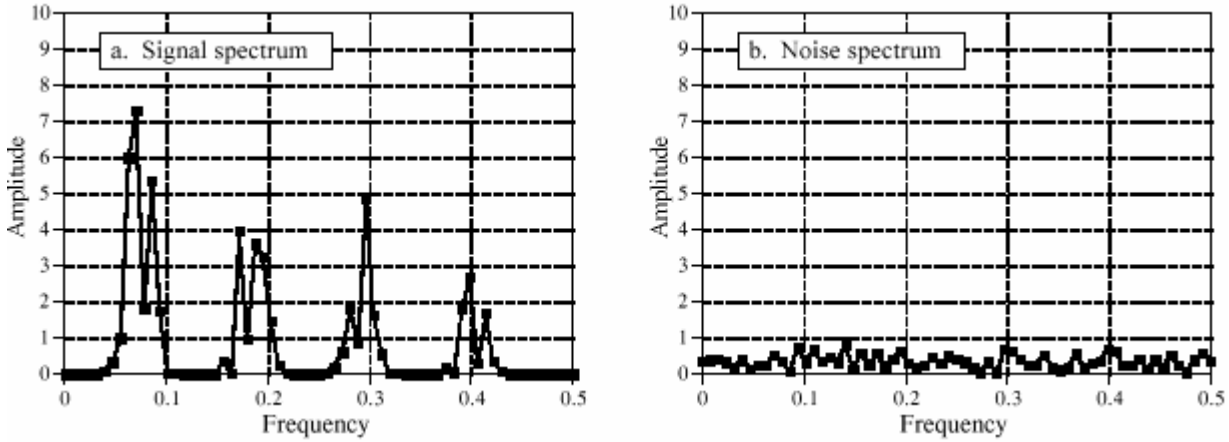

FIGURE 22-10
Spectra of speech and noise. While the frequency spectra of speech and noise generally overlap, there is some separation if the signal segment is made short enough. Figure (a) illustrates the spectrum of a 16 millisecond speech segment, showing that many frequencies carry little speech information, *in this particular segment*. Figure (b) illustrates the spectrum of a random noise source; all the components have a small amplitude. (These graphs are not of real signals, but illustrations to show the noise reduction technique).

РИСУНОК 22-10
Спектры речи и шума. В то время как частотные спектры речи и шума вообще накладываются, имеется некоторое разделение, если сегмент сигнала сделан достаточно коротким. Рисунок (a) иллюстрирует спектр из 16 миллисекундных сегмент речи, показывая, что много частот несут небольшую речевую информацию, в этом специфическом сегменте. Рисунок (b) иллюстрирует спектр случайного источника помех; все компоненты имеют маленькую амплитуду. (Эти графики не имеют реальных сигналов, но иллюстрация, чтобы показать методику шумоподавления).

One of the difficulties in implementing this (and other) nonlinear techniques is that the overlap-add method for filtering long signals is not valid. Since the frequency response changes, the time domain waveform of each segment will no longer align with the neighboring segments. This can be overcome by remembering that audio information is encoded in frequency patterns that change over time, and not in the shape of the time domain waveform. A typical approach is to divide the original time domain signal into *overlapping* segments. After processing, a smooth window is applied to each of the over-lapping segments before they are recombined. This provides a smooth transition of the frequency spectrum from one segment to the next.

Одна из трудностей в осуществлении этого (и других) нелинейного метода - то, что перекрытие-добавленный метод для фильтрации длинных сигналов, не имеет силы. Начиная с изменений частотной характеристики, форма волны домена времени каждого сегмента больше не будет выравниваться с соседними сегментами. Это может быть преодолено, помня, что звуковая информация закодирована в частотных образцах, которые изменяются через какое-то время, а не в форме формы волны домена времени. Типичный подход состоит в том, чтобы делить первоначальный сигнал домена времени на *перекрывающиеся* сегменты. После обработки, гладкое(сглаживающее) окно применяется к каждому из накладывающихся сегментов прежде, чем они рекомбинируются. Это обеспечивает гладкое(плавное) перемещение спектра частот от одного сегмента до следующего.

The second nonlinear technique is called **homomorphic** signal processing. This term literally means: *the same structure*. Addition is not the only way that noise and interference can be combined with a signal of interest; multiplication and convolution are also common means of mixing signals together. If signals are combined in a nonlinear way (i.e., anything other than addition), they cannot be separated by linear filtering. Homomorphic techniques attempt to separate signals combined in a nonlinear way by making the problem *become* linear. That is, the problem is converted to the *same structure* as a linear system.

Вторая нелинейная методика называется **гомоморфной** обработкой сигналов. Этот термин буквально означает: *та же самая структура*. Добавление - не единственный путь, которым шум и интерференция могут быть объединены с сигналом, представляющим интерес; умножение и свертка - также обычные средства смешивания сигналов вместе. Если сигналы объединены нелинейным способом (то есть, чем - нибудь другим чем добавление), они не могут быть отделены линейной фильтрацией. Гомоморфные методы пытаются отделять, сигналы, объединенные нелинейным способом, создавая проблему *стать* линейными. То есть проблема преобразования к той же *самой структуре* как линейная система.

For example, consider an audio signal transmitted via an AM radio wave. As atmospheric conditions change, the received amplitude of the signal increases and decreases, resulting in the loudness of the received audio signal slowly changing over time. This can be modeled as the audio signal, represented by $a[\ ]$, being *multiplied* by a slowly varying signal, $g[\ ]$, that represents the changing gain. This problem is usually handled in an electronic circuit called an *automatic gain control* (AGC), but it can also be corrected with nonlinear DSP.

Например, рассмотрите аудиосигнал, переданный через радиоволну AM. При изменении атмосферных условий, полученная амплитуда сигнала увеличивается и уменьшается, приводя к громкости полученного аудиосигнала, медленно изменяющейся через какое-то время. Это может быть смоделировано как аудиосигнал, представленыый $a[\ ]$, *умноженный* медленно изменяющимся сигналом, $g[\ ]$, который представляет изменение усиления(увеличения). Эта проблема обычно обрабатывается в электронной схеме называемой

автоматической регулировкой усиления (АРУ), но это может также быть исправлено с нелинейным ЦОС.

As shown in Fig. 22-11, the input signal, $a[\ ] \times g[\ ]$, is passed through the logarithm function. From the identity, $\log (x + y) = \log x + \log y$, this results in two signals that are combined by addition, i.e., $\log a[\ ] + \log g[\ ]$. In other words, the *logarithm* is the homomorphic transform that turns the nonlinear problem of *multiplication* into the linear problem of *addition*.

Как показано в рис. 22-11, входной сигнал, $a[\ ] \times g[\ ]$, проходит через функцию логарифма. Из тождества, $\log (x + y) = \log x + \log y$, это приводит к двум сигналам, которые объединены сложением, то есть, $\log a[\ ] + \log g[\ ]$. Другими словами, логарифм - гомоморфная трансформанта, которая поворачивает нелинейную проблему умножения в линейную задачу сложения.

Next, the added signals are separated by a conventional linear filter, that is, some frequencies are passed, while others are rejected. For the AGC, the gain signal, $g[\ ]$, will be composed of very low frequencies, far below the 200 hertz to 3.2 kHz band of the voice signal. The logarithm of these signals will have more complicated spectra, but the idea is the same: a high-pass filter is used to eliminate the varying gain component from the signal.

Затем, сложенные сигналы отделены обычным линейным фильтром, то есть некоторые частоты пропускают, в то время как другие отклонены. Для АРУ, сигнал усиления, $g[\ ]$, будет составлен из очень низких частот, далеко ниже полосы голосового сигнала 200 герц - 3.2 кГц. Логарифмирование этих сигналов больше усложнит спектры, но идея - та же самая: фильтр верхних частот используется, чтобы устранить изменяющийся компонент усиления от сигнала.

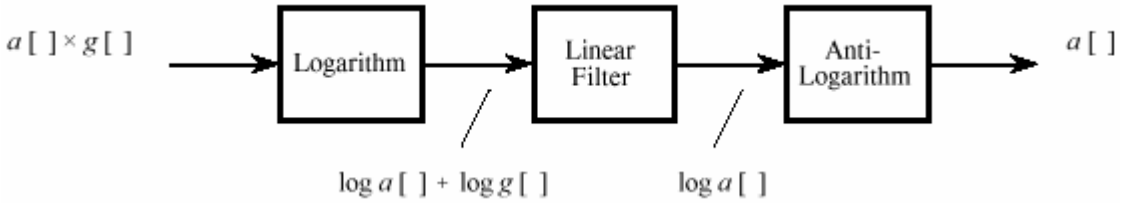

FIGURE 22-11
Homomorphic separation of multiplied signals. Taking the logarithm of the input signal transforms components that are *multiplied* into components that are *added*. These components can then be separated by linear filtering, and the effect of the logarithm undone.

РИСУНОК 22-11
Гомоморфное разделение мультиплицированных(умноженных) сигналов. Взятие логарифма входного сигнала преобразовывает компоненты, которые *умножены* в компоненты, которые *сложены*. Эти компоненты могут тогда быть отделены линейной фильтрацией, и эффектом уничтожения логарифма.

In effect $\log a[\ ] + \log g[\ ]$, is converted into $\log a[\ ]$. In the last step, the logarithm is undone by using the exponential function (the anti-logarithm, or $e^x[\ ]$), producing the desired output signal, $\log a[\ ]$

В действительности $\log a[\ ] + \log g[\ ]$, преобразован в $\log a[\ ]$. В последнем шаге, логарифм уничтожен, используя показательную функцию (антилогарифм, или $e^x[\ ]$), производя желательный сигнал выхода $\log a[\ ]$.

Figure 22-12 shows a homomorphic system for separating signals that have been *convolved*. An application where this has proven useful is in removing echoes from audio signals. That is, the audio signal is convolved with an impulse response consisting of a delta function plus a shifted

and scaled delta function. The homomorphic transform for convolution is composed of two stages, the *Fourier transform*, changing the convolution into a multi-plication, followed by the *logarithm*, turning the multiplication into an addition. As before, the signals are then separated by linear filtering, and the homomorphic transform undone.

Рисунок 22-12 показывает гомоморфную систему для отделения сигналов, которые были *свернуты*. Приложение, где это имеет полезность, находится в устранении, эхо из аудио-сигналов. То есть аудиосигнал свернут с импульсной передаточной функцией, состоящей из дельта функции плюс сдвинутая и масштабируемая дельта функция. Гомоморфная трансформанта(преобразование) для свертки составлено из двух стадий, *Преобразования Фурье*, изменяя свертку в умножение, сопровождаемое *логарифмированием*, поворачивая умножение в сложение. Как прежде, сигналы тогда отделены линейной фильтрацией, и удалением гомоморфного преобразования.

An interesting twist in Fig. 22-12 is that the linear filtering is dealing with frequency domain signals in the same way that time domain signals are usually processed. In other words, the time and frequency domains have been swapped from their normal use. For example, if FFT convolution were used to carry out the linear filtering stage, the "spectra" being multiplied would be in the *time domain*. This role reversal has given birth to a strange jargon. For instance, *cepstrum* (a rearrangment of *spectrum*) is the Fourier transform of the logarithm of the Fourier transform. Likewise, there are *long-pass* and *short-pass* filters, rather than low-pass and high-pass filters. Some authors even use *Quefrency Alanysis* and *liftering*.

Интересный трюк (поворот; характерная черта; отличительная особенность) в рис. 22-12 - то, что линейная фильтрация имеет дело с сигналами частотного домена, таким же образом обычно обрабатываются эти сигналы домена времени. Другими словами, домены времени, и частотные домены поменялись по отношению к их нормальному использованию. Например, если бы свертка БПФ использовалась, чтобы выполнить стадию линейной фильтрации, умноженные "спектры" были бы в домене времени. Эта инверсия роли родила странный жаргон. Например, *cepstrum* - *косинус-преобразование фурье логарифма энергетического спектра* (перестройка (перегруппировка; перекомпоновка; перераспределение) спектра)) - Преобразование Фурье логарифма из преобразования Фурье. Аналогично, имеются и фильтры с *коротким проходом* и фильтры с *длинным проходом*, скорее чем фильтры *низких частот* и фильтры *верхних частот*. Некоторые авторы даже используют *Quefrency Alanysis* and *liftering*.

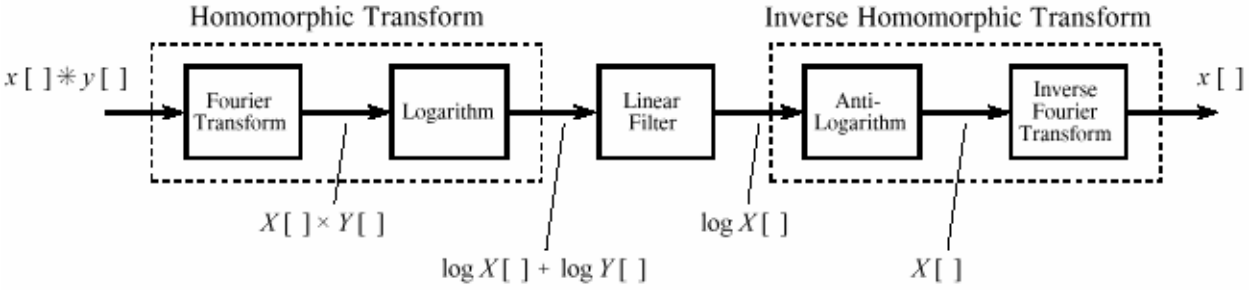

FIGURE 22-12
Homomorphic separation of convolved signals. Components that have been *convolved* are converted into components that are *added* by taking the Fourier transform followed by the logarithm. After linear filtering to separate the added components, the original steps are undone.

РИСУНОК 22-12
Гомоморфное разделение свернутых сигналов. Компоненты, которые были свернуты, преобразованы в компоненты, которые сложены, беря Преобразование Фурье, сопровождаемое логарифмированием. После линейной фильтрации, чтобы отделить добавленные компоненты, первоначальные шаги уничтожены.

Keep in mind that these are simplified descriptions of sophisticated DSP algorithms; homomorphic processing is filled with subtle details. For example, the logarithm must be able to handle both negative and positive values in the input signal, since this is a characteristic of audio signals. This requires the use of the *complex logarithm*, a more advanced concept than the logarithm used in everyday science and engineering. When the linear filtering is restricted to be a *zero phase* filter, the complex log is found by taking the simple logarithm of the absolute value of the signal. After passing through the zero phase filter, the sign of the original signal is reapplied to the filtered signal.

Имейте в виду, что они - упрощенные описания сложных алгоритмов ЦОС; гомоморфная обработка заполнена тонкими подробностями. Например, логарифм должен быть способен обработать, и отрицательные и положительные значения во входном сигнале, так как это - характеристика звуковых сигналов. Это требует использования *комплексного логарифма*, более преждевременная концепция, чем логарифм, используемый в каждодневной науке и разработке. Когда линейная фильтрация ограничена, чтобы быть фильтром *нулевой фазы*, комплексный логарифм найден, беря простой логарифм абсолютного значения сигнала. После прохождения через фильтр нулевой фазы, знак первоначального сигнала повторно обращается к фильтрованному сигналу.

Another problem is *aliasing* that occurs when the logarithm is taken. For example, imagine digitizing a continuous *sine wave*. In accordance with the sampling theorem, two or more samples per cycle is sufficient. Now consider digitizing the logarithm of this continuous sine wave. The sharp corners require many more samples per cycle to capture the waveform, i.e., to prevent aliasing. The required sampling rate can easily be 100 times as great after the log, as before. Further, it doesn't matter if the logarithm is applied to the continuous signal, or to its digital representation; the result is the same. Aliasing will result unless the sampling rate is high enough to capture the sharp corners produced by the nonlinearity. The result is that audio signals may need to be sampled at 100 kHz or more, instead of only the standard 8 kHz.

Другая проблема - *наложение спектров*(псевдочастоты; замещение частот?), которое происходит, когда логарифм принят. Например, вообразите отцифровывать непрерывную *синусоидальную волну*. В соответствии с выборочной теоремой, две или больше выборки в цикл(период) достаточны. Теперь рассмотрите отцифровывание логарифма этой непрерывной синусоидальной волны. Крутые углы требуют намного больше выборок в цикл(период), чтобы фиксировать форму волны, то есть, предотвращать наложение спектров. Требуемая частота выборки может легко быть 100 раз столь же большая после логарифмирования, как прежде. Далее, это не имеет значения, если логарифм применяется к непрерывному сигналу, или к его цифровому представлению; результат - тот же самый. Наложение спектров прекратится, если частота выборки не достаточно высока, чтобы фиксировать крутые углы, произведенные нелинейностью. Результат - то, что аудиосигналы могут нуждаться в производстве выборок с частотой в 100 кГц или больше, вместо стандартых 8 кГц.

Even if these details are handled, there is no guarantee that the linearized signals *can* be separated by the linear filter. This is because the spectra of the linearized signals can overlap, even if the spectra of the original signals do not. For instance, imagine adding two sine waves, one at 1 kHz, and one at 2 kHz. Since these signals do not overlap in the frequency domain, they can be completely separated by linear filtering. Now imagine that these two sine waves are multiplied. Using homomorphic processing, the log is taken of the combined signal, resulting in the log of one sine wave plus the log of the other sine wave. The problem is, the logarithm of a sine wave contains many harmonics. Since the harmonics from the two signals overlap, their complete separation is not possible.

Даже если эти подробности обработаны, не имеется никакой гарантии, что линеаризованные сигналы *могут* быть отделены линейным фильтром. Это - то, потому что спектры линеаризованных сигналов могут накладываться, даже если спектры первоначальных сигналов этого не делают. Например, вообразите сложение двух синусоидальных волн, одна в 1 кГц, и одна в 2 кГц. Так как эти сигналы не накладываются в частотном домене, они могут быть полностью отделены линейной фильтрацией. Теперь вообразите, что эти две синусоидальных волны мультиплицированы(умножены). Используя гомоморфную обработку, логарифмированием принимаемы объединенный сигнала, приводит к логарифму одной синусоидальной волны плюс логарифм другой синусоидальной волны. Проблема, логарифм синусоидальной волны содержит много гармоник(флажолетов). Начиная с гармоник(флажолетов) от перекрытия двух сигналов, их полное разделение невозможно.

In spite of these obstacles, homomorphic processing teaches an important lesson: signals should be processed in a manner *consistent* with how they are formed. Put another way, the first step in any DSP task is to understand how information is represented in the signals being processed.

Несмотря на эти препятствия, гомоморфная обработка преподает важный урок: сигналы должны быть обработаны способом *совместимым* с тем, как они сформированы. Излагая другой путь, первый шаг в любую задачу ЦОС состоит в том, чтобы понять, как представлена информация в обрабатываемых сигналах.