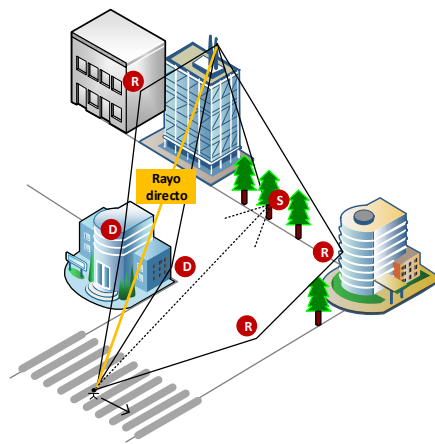


Módulo 1

Comunicación de Datos

Capítulos

- 0. El Modelo de Comunicaciones
- 1. Teoría de la Información
- 2. Teoría de Señal
- 3. Transmisión Digital
- 4. Medios de Transmisión
- 5. Limitaciones de la Transmisión Digital



SEBASTIÀ GALMÉS OBRADOR



Universitat
de les Illes Balears

Escola
Politécnica
Superior

CAPÍTULO 0. EL MODELO DE COMUNICACIONES

El objetivo de toda red de computadores es posibilitar el intercambio de mensajes entre computadores geográficamente dispersos (entendiendo “computador” en la acepción más general del término: ordenador de sobremesa, ordenador portátil, estación de trabajo, tableta, teléfono móvil, teléfono IP, ..., en general, cualquier dispositivo digital). Un **mensaje** se caracteriza cualitativamente por su formato y cuantitativamente por la cantidad de información que contiene. Dicho de otro modo, el mensaje es un contenedor de información en un cierto formato. El **formato** puede ser muy diverso: texto, imagen, audio, vídeo, o también una combinación de estos – mensaje multimedia. El aspecto cuantitativo del mensaje nos lleva en primer lugar al concepto de **información**, un concepto nada sencillo que ha sido objeto de estudio desde disciplinas muy diversas. En efecto, la semántica de este término es lo suficientemente compleja como para tener implicaciones en campos como la filosofía (teoría del conocimiento) o la psicología (pensamiento y lenguaje); sin embargo, estos aspectos quedan fuera de lugar cuando se trata de abordar la información en un contexto como el de las redes de computadores. En este caso se trabaja con un concepto más concreto y práctico de la información, hasta el punto de que ésta resulta medible y, por tanto, utilizable como descriptor cuantitativo de los mensajes. Profundizaremos en este aspecto en el Capítulo 1; por el momento, cabe señalar que, cualesquiera que sean el formato del mensaje y la cantidad de información que contiene, el objetivo de toda red de computadores es reproducir en el destinatario correcto y en un tiempo determinado una réplica aceptable o perfecta del mensaje originado por la fuente. Estos requisitos, fáciles de formular, copan la complejidad que conlleva diseñar una red de computadores.

Tabla 0.1. Tipos de medios de transmisión que pueden formar parte del sistema de transmisión, así como las señales y magnitudes asociadas.

Medio de transmisión	Tipo de señal	Magnitud
Metálico (cobre)	Eléctrica	Voltaje (o corriente)
Espacio libre (aire, vacío)	Electromagnética	Intensidad de campo electromagnético
	Óptica (VLC)	Intensidad óptica (intensidad de luz visible)
Fibra óptica	Óptica	Intensidad óptica
Agua (de mar)	Acústica (pref.)	Intensidad acústica
Cuerpo humano (p.e.)	Química (pref.)	Concentración molecular (p.e)

Para enviar un mensaje de un punto a otro se requiere una señal. En telecomunicaciones, una **señal** es la variación temporal de una cierta magnitud física, o incluso química, en función del mensaje que se pretende enviar. El envío del mensaje tiene lugar con la propagación de la señal a través de un **medio de transmisión**. Precisamente, la naturaleza o tipología de la señal viene determinada en gran medida por el tipo de medio de transmisión utilizado, de manera que, naturaleza o tipo de señal y medio de transmisión son dos conceptos estrechamente relacionados. Esto se pone de manifiesto en la **Tabla 0.1**, donde además se especifican las magnitudes asociadas a cada tipo de señal. En dicha tabla se constatan varias tecnologías actualmente en progreso:

- La tecnología VLC (*Visible Light Communications*), basada en el uso de señales ópticas a través del espacio libre.
- Las comunicaciones submarinas, en cuyo caso las señales acústicas representan la mejor alternativa, dada la baja atenuación que experimentan estas señales al propagarse a través del medio marino.
- Las comunicaciones moleculares. Puesto que sus principales aplicaciones están en el campo de la salud, se ha especificado el cuerpo humano como medio de transmisión, aunque puede concretarse más (sangre, tejidos, órganos, huesos, etc.). En dicho campo, una alternativa la siguen constituyendo las señales electromagnéticas, pero su uso es más discutible, ya que generan calor.

El **modelo de comunicaciones** es el modelo que describe el proceso de envío de un mensaje de una **entidad** a otra, en un sentido amplio de este último término: persona o dispositivo. Por tanto, el modelo de comunicaciones va más allá de las redes de computadores: por ejemplo, el escenario de dos personas conversando a través de la red de telefonía fija encaja en dicho modelo. La **Figura 0.1** muestra los bloques funcionales básicos que componen el modelo de comunicaciones. Son estos:

- **Fuente**. Es la entidad que genera el mensaje.
- **Destinatario**. Es la entidad que recibe el mensaje.

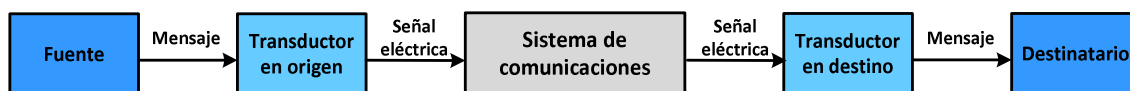


Figura 0.1. Modelo de comunicaciones.

- **Transductores en origen y destino**. En principio, un **transductor** es un dispositivo capaz de cambiar la naturaleza o tipología de una señal. No obstante, el hecho de que, al menos hasta el momento, los dispositivos que integran los sistemas de comunicaciones se basan en circuitos eléctricos o electrónicos, justifica la redefinición del concepto de transductor tomando como referencia la señal eléctrica. Así pues, podemos admitir que un transductor es un dispositivo capaz de cambiar la naturaleza de una señal, ya sea de eléctrica a no eléctrica o viceversa. Así, cuando la fuente genera una señal (portadora inicial del mensaje) que no es eléctrica, el transductor en origen se encarga de transformarla en una señal eléctrica; y al revés, el transductor en destino se encarga de convertir la señal eléctrica entregada por el sistema de comunicaciones en una señal de naturaleza arbitraria, según sea el destinatario. Por ejemplo, en el caso de una comunicación de voz, la fuente sería la persona que habla, el destinatario la persona que escucha, y los transductores en origen y destino serían, respectivamente, el micrófono y el altavoz del teléfono. El micrófono convierte la variación de presión del aire (señal acústica) generada al hablar, en una señal eléctrica, y el altavoz realiza la conversión opuesta. La **Tabla 0.2** recoge los transductores más utilizados en comunicaciones.
- **Sistema de comunicaciones**. El núcleo del sistema de comunicaciones es el llamado **sistema de transmisión**, que puede variar desde un simple **enlace punto a punto** (tramo de medio de transmisión que interconecta directamente los dos extremos) hasta la más compleja red de

computadores o la propia Internet. De manera resumida, el sistema de transmisión es la infraestructura que cubre la separación geográfica entre fuente y destinatario. Justo en sus extremos o internamente, el sistema de transmisión puede contener transductores adicionales, cada vez que cambia el tipo de medio de transmisión. Además, el sistema de comunicaciones está formado por otros dos elementos. El **transmisor** se encarga de preparar la señal que recibe para que pueda ser enviada a través del sistema de transmisión; el **receptor**, por su parte, modifica la señal recibida del sistema de transmisión para compatibilizarla con las especificidades del equipo que sigue a continuación. Transmisor y receptor son, por tanto, dispositivos de adaptación de señales eléctricas, es decir, cambian tan solo el formato de las mismas, pero no introducen ninguna modificación en su naturaleza. Por tanto, no deben confundirse con los transductores. Finalmente, hay que tener en cuenta que toda señal que se transmite experimenta una serie de efectos no deseados o **perturbaciones** (atenuación, distorsión, ruido e interferencias), que pueden llegar a provocar que el mensaje recibido difiera del transmitido. Aunque estas perturbaciones pueden aparecer en cualquier punto entre la fuente y el destinatario, son mayoritariamente introducidas por el sistema de transmisión. Por ello, a efectos de modelado, tal como se indica en la **Figura 0.2**, las perturbaciones se representan conjuntamente como una hipotética señal no deseada que se suma o superpone a la señal transmitida en dicho sistema.

Tabla 0.2. Transductores más habituales en el modelo de comunicaciones.

Transductor	Tipo de señal de entrada	Tipo de señal de salida
Micrófono	Acústica	Eléctrica
Altavoz	Eléctrica	Acústica
Cámara	Óptica	Eléctrica
Antena	Eléctrica/Electromagnética	Electromagnética/Eléctrica
Lámpara fluorescente, LED, diodo láser	Eléctrica	Óptica
Fotodetector, fotodiodo	Óptica	Eléctrica
Hidrófono	Acústica	Eléctrica
Transmisor acústico	Eléctrica	Acústica

No todos los elementos del modelo de comunicaciones tienen por qué estar presentes en un escenario concreto. Los únicos elementos fijos, presentes en cualquier escenario, son la fuente, el destinatario y el sistema de transmisión.

Toda comunicación a través del sistema de transmisión consiste en una cadena de conexiones entre dos dispositivos, finales y/o intermedios. Por tanto, los fundamentos físicos y matemáticos que rigen la transmisión de información entre dos dispositivos conectados directamente constituyen la base de todo el proceso. Del estudio de estos fundamentos se encarga la **Comunicación de Datos**. Por tanto, el modelo de comunicaciones que se presupone en este ámbito es el mostrado en la **Figura 0.3**, donde el sistema de transmisión queda reducido a un enlace punto a punto, pero el concepto de entidad (fuente, destinatario) queda ampliado en lo que a dispositivos se refiere, ya que tales dispositivos pueden ser finales (computador, estación de trabajo, teléfono móvil, tableta, etc.), pero también intermedios (conmutador, enrutador, punto de acceso, etc.). Por otro lado, el estudio de

las formas más complejas del sistema de transmisión pertenece ya al ámbito de las **Redes de Computadores**.

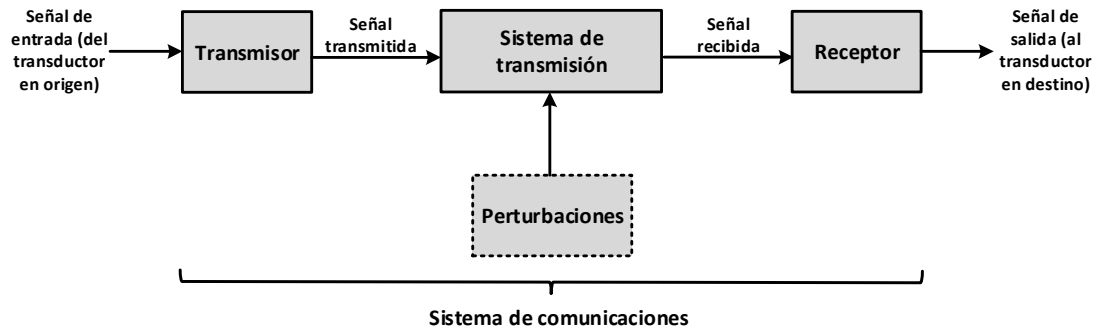


Figura 0.2. Elementos del sistema de comunicaciones.

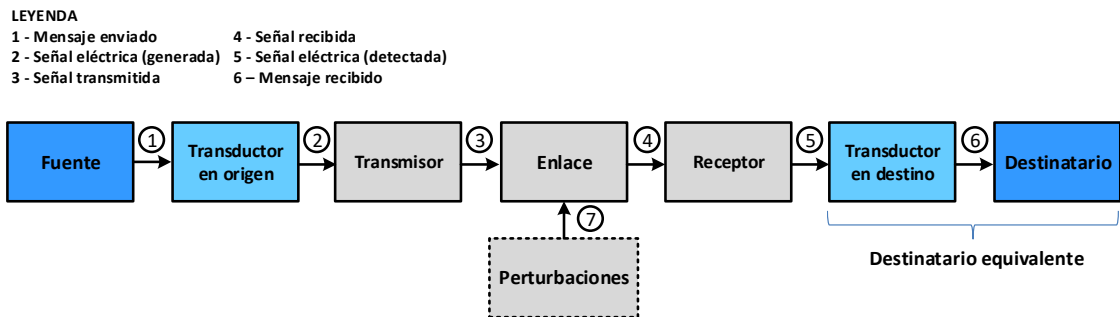


Figura 0.3. Modelo de comunicaciones en Comunicación de Datos.

CAPÍTULO 1. TEORÍA DE LA INFORMACIÓN

Una red de computadores es un conjunto de dos o más computadores interconectados con la finalidad de compartir recursos e intercambiar información. Precisamente, el objetivo general de este primer capítulo es acercarnos a la **Teoría de la Información**, resolviendo algunas cuestiones básicas como:

- ¿Cómo se mide la cantidad de información?
- ¿Cómo se representa la información generada por un dispositivo conectado a una red?
- ¿Cómo podemos caracterizar matemáticamente el proceso de transmisión de la información a través de un canal, que por naturaleza es imperfecto?

El capítulo culmina con la formulación del Teorema Fundamental de Shannon para la capacidad de un canal de transmisión. Este teorema establece el máximo teórico para la velocidad con que se puede transmitir la información a través de un canal imperfecto, es decir, un canal que puede introducir diferencias entre la información recibida y la transmitida.

ÍNDICE

1.1.	MEDIDA DE LA INFORMACIÓN	6
1.2.	ENTROPÍA	9
1.3.	CODIFICACIÓN DE FUENTE	11
1.4.	TAXONOMÍA DE CÓDIGOS FUENTE. MÉTODOS Y FORMATOS DE COMPRESIÓN	19
1.5.	ALGORITMOS DE HUFFMAN Y LEMPEL-ZIV	22
1.6.	TRANSMISIÓN DE LA INFORMACIÓN	30

1.1. MEDIDA DE LA INFORMACIÓN

Recordemos que uno de los atributos que caracterizan el mensaje es la cantidad de información que contiene. El análisis completo de este atributo nos llevaría, por tanto, al concepto de información y a su cuantificación. Sin embargo, las tecnologías de la información y las comunicaciones ponen su énfasis directamente en la medida de la información, dejando el concepto para otros campos, como la filosofía y la psicología. Por consiguiente, nuestro interés no está en las implicaciones metafísicas o intelectuales de un mensaje, sino en el hecho de que este mensaje fue seleccionado entre un conjunto de mensajes posibles. Dos ejemplos pueden arrojar un poco de luz sobre esta idea.

En el primer ejemplo, una persona hace girar una bola del mundo y la frena marcando un punto cualquiera, de forma completamente aleatoria. Se supone que todos los puntos de la superficie terrestre son seleccionables con igual probabilidad. A continuación, envía un mensaje a otra persona acerca del punto elegido de acuerdo con tres posibles alfabetos, tal como se indica en la **Tabla 1.1**. El alfabeto 1 tan solo corresponde a indicar si la latitud del punto elegido es Norte o Sur, y si su longitud es Este u Oeste. Por lo tanto, dicho alfabeto contiene un total de 4 mensajes, resultantes de las 4 combinaciones posibles de los dos valores de latitud y los dos de longitud. Es evidente que el receptor del mensaje adquiere “relativamente poca” información sobre el punto elegido. El alfabeto 2, que resulta de subdividir cada opción del alfabeto 1 en 3 partes iguales, contiene un total de 36 mensajes posibles para especificar longitud y latitud. Por tanto, con el uso de este alfabeto, el mensaje enviado contiene “más” información sobre el punto elegido que en el caso anterior. Finalmente, si quien envía el mensaje utiliza el alfabeto 3, que contiene 900 mensajes posibles resultantes de la división en 5 partes iguales de cada opción del alfabeto 2, estará proporcionando “mucho más” información sobre el punto elegido que en los casos anteriores. Podríamos seguir así, ampliando sucesivamente el alfabeto, de modo que un solo mensaje aportara cada vez “más información”.

En el segundo ejemplo planteamos un alfabeto sencillo que contiene pocos mensajes, si bien no todos tienen la misma probabilidad. Imaginemos que preguntamos por el resultado de una partida de ajedrez en la que juega el campeón del mundo frente a un principiante, y podemos recibir uno de estos cuatro posibles mensajes:

“Ha ganado el campeón del mundo” (*mensaje 1*)

“La partida ha acabado en tablas” (*mensaje 2*)

“Ha ganado el principiante después de una larga partida” (*mensaje 3*)

“Ha ganado el principiante al poco de empezar la partida” (*mensaje 4*)

Es obvio que conforme vamos descendiendo en esta secuencia, el mensaje recibido contiene “más información”, pues corresponde a un evento menos esperado.

Tabla 1.1. Alfabetos manejados en el envío de las coordenadas geográficas de un punto sobre la superficie terrestre elegido al azar (N: Norte; S: Sur; E: Este; O: Oeste).

Coordenada	Alfabeto 1	Alfabeto 2	Alfabeto 3
Latitud	[0,90]° N [0,90]° S	[0,30]° N [0,30]° S	[0,6]° N / [0,6]° S
			[6,12]° N / [6,12]° S
			[12,18]° N / [12,18]° S
			[18,24]° N / [18,24]° S
			[24,30]° N / [24,30]° S
		[30,60]° N [30,60]° S	[30,36]° N / [30,36]° S
			[36,42]° N / [36,42]° S
			[42,48]° N / [42,48]° S
			[48,54]° N / [48,54]° S
			[54,60]° N / [54,60]° S
		[60,90]° N [60,90]° S	[60,66]° N / [60,66]° S
			[66,72]° N / [66,72]° S
			[72,78]° N / [72,78]° S
			[78,84]° N / [78,84]° S
			[84,90]° N / [84,90]° S
Longitud	[0,180]° E [0,180]° O	[0,60]° E [0,60]° O	[0,12]° E / [0,12]° O
			[12,24]° E / [12,24]° O
			[24,36]° E / [24,36]° O
			[36,48]° E / [36,48]° O
			[48,60]° E / [48,60]° O
		[60,120]° E [60,120]° O	[60,72]° E / [60,72]° O
			[72,84]° E / [72,84]° O
			[84,96]° E / [84,96]° O
			[96,108]° E / [96,108]° O
			[108,120]° E / [108,120]° O
		[120,180]° E [120,180]° O	[120,132]° E / [120,132]° O
			[132,144]° E / [132,144]° O
			[144,156]° E / [144,156]° O
			[156,168]° E / [156,168]° O
			[168,180]° E / [168,180]° O

Los dos ejemplos ponen de manifiesto que un mensaje aporta tanta más información cuanto más **incertidumbre** resuelve. Dicho de otra manera, la información contenida en un mensaje es tanto mayor cuanto menor es la probabilidad del evento representado por dicho mensaje, o, lo que es lo mismo, cuanto menor es la probabilidad de que la fuente seleccione dicho mensaje para ser enviado al destinatario. Este razonamiento llevó al ingeniero y matemático **Claude Shannon** a introducir una medida de la información que fuera inversamente proporcional a la probabilidad. Concretamente, estableció que, si x_i representa un mensaje arbitrario cuya probabilidad es $p(x_i) = p_i$, la información $I(x_i) = I_i$ contenida en dicho mensaje viene dada por la siguiente expresión logarítmica, en la que b representa la base:

$$I_i = -\log_b p_i = \log_b \frac{1}{p_i} \quad (1.1)$$

Obsérvese que esta definición matemática implica que, efectivamente, si $p_i < p_j$, entonces $I_i > I_j$, puesto que la función logaritmo es monótona creciente para cualquier base.

Podemos comprobar la coherencia de esta definición con los dos ejemplos anteriores. En el primero de ellos, utilizábamos alfabetos de tamaño cada vez mayor, siempre formados por mensajes con igual probabilidad. Entonces, al aplicar (1.1) a un mensaje cualquiera de uno cualquiera de estos alfabetos, de tamaño genérico S , la información contenida en dicho mensaje resulta ser la siguiente:

$$I_i = \log_b \frac{1}{p_i} = \log_b \frac{1}{1/S} = \log_b S \quad (1.2)$$

Este resultado indica que, en el supuesto de mensajes con igual probabilidad, la información aportada por cualquiera de ellos es directamente proporcional al tamaño del alfabeto, ya que la función logaritmo es monótona creciente. Esta conclusión es lógica si tenemos en cuenta que cuanto mayor es el número de mensajes equiprobables que conforman un alfabeto, mayor es la incertidumbre del receptor en cuanto a cuál fue el mensaje elegido por la fuente. O, dicho de otra manera, cuanto mayor es el número de mensajes equiprobables que conforman un alfabeto, mayor es el grado de concreción, es decir, de información, aportada por cada uno.

En el segundo ejemplo, referido a un único alfabeto formado por mensajes con probabilidades desiguales, la ecuación (1.1) implica algo que ya esperábamos:

$$I(\text{mensaje 1}) < I(\text{mensaje 2}) < I(\text{mensaje 3}) < I(\text{mensaje 4})$$

Además, la medida logarítmica de la información propuesta por Shannon goza de ciertas propiedades matemáticas deseables:

- (1) $I_i \geq 0$ para cualquier $0 \leq p_i \leq 1$.
- (2) $I_i \rightarrow 0$ cuando $p_i \rightarrow 1$.
- (3) Si x_i y x_j son dos mensajes independientes enviados por una fuente, la información aportada por ambos mensajes es $I_{ij} = \log_b \frac{1}{p(x_i, x_j)} = \log_b \frac{1}{p(x_i) \cdot p(x_j)} = \log_b \frac{1}{p_i p_j} = \log_b \frac{1}{p_i} + \log_b \frac{1}{p_j} = I_i + I_j$. Es decir, la información total aportada por dos mensajes independientes es igual a la suma de las informaciones contenidas en cada uno.

El trabajo fundamental de Shannon se recoge en su artículo *A Mathematical Theory of Communication*, publicado en 1948 y reconocido como el punto de partida de una nueva disciplina, híbrida entre la ingeniería y las matemáticas, llamada **Teoría de la Información**. En dicho trabajo, Shannon no sólo aborda la definición y medida de la información, sino también el análisis de la capacidad de un canal de comunicaciones, entendida en términos de cuánta información puede transmitir por unidad de tiempo, y también sienta las bases de la codificación de fuente y la codificación de canal (las estudiaremos más adelante), ambas enfocadas a aprovechar al máximo la capacidad del canal.

No sería justo ensalzar la obra de Shannon sin mencionar importantes aportaciones previas de otros autores, como las de **Harry Nyquist** y **Ralph Hartley** sobre transmisión de la información, o las herramientas de teoría de probabilidad desarrolladas hasta el momento por **Norbert Wiener**. Y también la aportación posterior de **Warren Weaver**, quien en 1949 reeditó el artículo de Shannon en forma de libro, esta vez titulado *The Mathematical Theory of Communication*. Weaver era un científico e ingeniero cuya contribución al libro consistió en resaltar las importantes implicaciones, la generalidad y el potencial de la teoría desarrollada por Shannon. De hecho, el modelo de comunicaciones que aparece en dicho libro, del que básicamente el modelo de la **Figura 0.1** es una reproducción (en la que hemos incluido los transductores), pasó a conocerse también como modelo de Shannon-Weaver¹. Y en la carátula del libro aparecen también ambos autores... eso sí, el nombre de Shannon en letras más grandes.

Para terminar este apartado, conviene considerar un último punto: si bien ya disponemos de una fórmula que nos mide la cantidad de información contenida en un mensaje, ¿en qué unidades se expresa esta medida? Como se constata en el artículo de Shannon, esto depende de la elección de la base b para el logaritmo de la ecuación (1.1). Lo habitual en Teoría de la Información es elegir $b = 2$, en cuyo caso la unidad de medida es el **bit**. En dicha base, la cantidad de información contenida en un mensaje se concreta de la siguiente manera:

$$I_i = -\log_2 p_i = \log_2 \frac{1}{p_i} \text{ (bits)} \quad (1.3)$$

Esta expresión permite definir **1 bit** como la cantidad de información que resulta de la elección de una entre dos alternativas igualmente probables ($I = \log_2 \frac{1}{1/2} = 1$).

Aunque en el artículo de Shannon se adopta el bit como unidad de medida de la información, en él aparece una referencia inicial a otras dos posibles unidades: las unidades decimales, correspondientes a $b = 10$ (logaritmo decimal), y las unidades naturales (*nats*, contracción introducida posteriormente), correspondientes a $b = e$ (logaritmo natural).

1.2. ENTROPÍA

En adelante vamos a suponer que el alfabeto con el que trabaja la fuente está formado por S mensajes elementales llamados **símbolos**, de manera que cualquier otro mensaje se construye encadenando tales símbolos. Así pues, sea $X = \{x_1, x_2, \dots, x_S\}$ dicho alfabeto. Cada símbolo x_i es un mensaje elemental que se genera con probabilidad p_i y transmite una información $I_i = -\log_2 p_i$, cumpliéndose lógicamente que $\sum_{i=1}^S p_i = 1$. Estas características definen una **fente discreta**.

¹ La denominación “modelo de Shannon-Weaver” es más popular en el ámbito de las ciencias sociales, en las que el modelo se utiliza para caracterizar la comunicación humana. En los ámbitos científico-tecnológicos, como la ingeniería, las matemáticas, la física y la biología, es más habitual hablar de “modelo de Shannon” o “teoría de Shannon”.

Además, supondremos que dicha fuente es **estacionaria**, lo cual significa que las probabilidades p_i no varían con el tiempo.

Sentadas estas premisas, podemos caracterizar la cantidad de información transmitida por un símbolo como una variable aleatoria discreta de valores I_1, I_2, \dots, I_S , cuyas probabilidades son, precisamente, p_1, p_2, \dots, p_S . Por lo tanto, la información media transmitida por un símbolo no es más que la media estadística de dicha variable aleatoria, representada por $H(X)$ y calculada de la siguiente manera:

$$H(X) = \sum_{i=1}^S p_i \cdot I_i = \sum_{i=1}^S p_i \cdot \log_2 \frac{1}{p_i} \quad (1.4)$$

A esta magnitud se la conoce como **entropía de la fuente X** (es decir, de la fuente cuyo alfabeto es X). Shannon adoptó este término y su notación por la similitud de la expresión (1.4) con la correspondiente a la entropía termodinámica, un concepto de Mecánica Estadística.

El valor mínimo de la entropía dada por la ecuación (1.4) es 0, correspondiendo al caso en que $p_j = 1$ para algún j , y $p_i = 0 \forall i \neq j$, es decir, la fuente emite siempre el mismo símbolo y, por tanto, no aporta ninguna información (el símbolo no resuelve ninguna incertidumbre). Por otro lado, se puede demostrar que la máxima entropía tiene lugar cuando todos los símbolos tienen la misma probabilidad, es decir, $p_i = 1/S \forall i$, en cuyo caso $H(X) = \log_2 S$. Así pues, podemos afirmar lo siguiente:

$$0 \leq H(X) \leq \log_2 S \quad (1.5)$$

Consideremos una **fente binaria**, es decir, una fuente discreta cuyo alfabeto contiene solamente dos símbolos ($S = 2$), que podemos identificar con los dígitos binarios 0 y 1. Para el caso general en que las probabilidades de ambos símbolos son distintas, es decir, p y $1 - p$, podemos comprobar que la entropía $H(X)$ de dicha fuente se concreta en una función que lógicamente sólo depende de p , llamada $\Omega(p)$, cuya expresión es la siguiente:

$$H(X) = \Omega(p) = p \cdot \log_2 \frac{1}{p} + (1 - p) \cdot \log_2 \frac{1}{1-p} \quad (1.6)$$

A esta **entropía binaria** $\Omega(p)$ se la conoce también como “función de herradura”, a juzgar por la forma de su gráfica mostrada en la **Figura 1.4**. Como podemos observar en dicha figura, el valor máximo de la entropía binaria es 1 bit, correspondiente al caso en que los dos símbolos tienen la misma probabilidad ($\Omega(p)|_{max} = \log_2 2 = 1$). Para el resto de situaciones, dicha entropía es siempre inferior a 1, alcanzando el valor 0 cuando uno de los símbolos presenta probabilidad 1. Estos resultados concuerdan perfectamente con la expresión (1.5). Además, permiten poner de manifiesto la diferencia entre el bit como unidad de información y el dígito binario como símbolo, aunque etimológicamente el término bit proviene de la contracción de *binary digit* (dígito binario), que fue acuñada por el químico y matemático **John W. Tuckey**. Si las probabilidades de ambos dígitos son distintas, la información transmitida por uno de ellos es inferior a 1 bit, mientras que la información transmitida por el otro es superior a 1 bit. Y la información promedio transmitida por

dígito binario, calculada mediante la expresión (1.6), resulta ser inferior a 1 bit. Por lo tanto, el hecho de que una fuente transmita una cantidad de dígitos binarios por unidad de tiempo, no significa necesariamente que transmita esa cantidad de información en la misma unidad de tiempo. Para recalcar esta diferencia y evitar ambigüedades, se acuñó el término *binit* como contracción alternativa de *dígito binario*, distinguiéndolo del término *bit* como unidad de información, aunque, ciertamente, el nuevo término no tuvo mucho éxito.

1.3. CODIFICACIÓN DE FUENTE

La **codificación de fuente** es el proceso de representar cada símbolo o cadena de símbolos mediante una **palabra código**, es decir, una combinación de dígitos binarios. Esto último no debe sorprender, pues la lógica binaria es la que manejan los ordenadores y los dispositivos digitales en general. Aunque la asignación de palabras código puede parecer una tarea relativamente sencilla, en realidad sólo es así cuando los símbolos transmitidos por la fuente son mutuamente independientes, es decir, se trata de una **fente discreta sin memoria** (en inglés DMS – *Discrete Memoryless Source*). Sólo en este caso se puede alcanzar una codificación de fuente apropiada (en términos de eficiencia, concepto que vemos más adelante) asociando una palabra código a cada símbolo. Es la llamada **codificación escalar**, en contraposición a la **codificación vectorial** o **codificación por bloques**, consistente en asignar palabras código a agrupaciones de símbolos. Cuando se trata de codificar una **fente discreta con memoria**, la codificación vectorial permite alcanzar eficiencias sustancialmente mayores que la codificación escalar. En esta sección nos centraremos en la codificación escalar de fuentes discretas sin memoria.

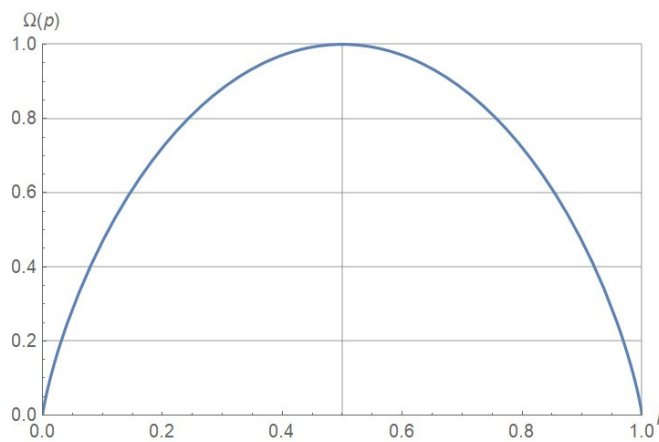


Figura 1.4. Entropía binaria. Indica la cantidad media de información contenida en un dígito binario.

Supongamos una fuente discreta sin memoria que emite símbolos de un alfabeto X de tamaño S , siendo $H(X)$ su entropía. La codificación de fuente más simple consiste en representar cada símbolo mediante una palabra código formada por $\lceil \log_2 S \rceil$ dígitos binarios. No obstante, ésta no es necesariamente la opción más eficiente: la información media contenida en cada dígito binario se puede expresar en este caso como la ratio $H(X)/\lceil \log_2 S \rceil$, pero, según la ecuación (1.5), $H(X) \leq$

$\log_2 S$, lo que significa que dicha ratio puede ser menor, incluso significativamente menor, que la unidad. Esto equivale a afirmar que el dígito binario se “aprovecha” poco; sólo en el caso excepcional de que todos los símbolos tengan la misma probabilidad, la ratio anterior alcanza el valor máximo $\log_2 S / \lceil \log_2 S \rceil$ (1, si S es potencia de 2) y el dígito binario transporta una información muy cercana a 1 bit.

Cuando el cociente $H(X) / \lceil \log_2 S \rceil$ es bajo, no hay forma de aumentarlo, pues tanto la entropía como el tamaño del alfabeto son parámetros intrínsecos de la fuente y no admiten modificación. No obstante, recordemos que ese cociente mide la información media por dígito binario cuando la codificación de fuente se basa en utilizar palabras del mismo tamaño, es decir, cuando trabajamos con **codificación de longitud fija**. Pero una alternativa sería trabajar con **codificación de longitud variable**, consistente en asignar palabras código más cortas a los símbolos más probables, con el fin de reducir el tamaño medio de las palabras código con respecto al caso de codificación fija. Si \bar{L} representa el tamaño medio de las palabras código o **longitud media del código**, la información media por dígito binario cuando el tamaño de las palabras código puede variar, viene dada por el cociente $H(X) / \bar{L}$. A diferencia de lo que ocurría antes, este cociente es susceptible de aumentar a costa de disminuir el denominador. Esto es lo que persigue, esencialmente, la codificación de fuente.

La idea de variar el tamaño de la palabra código en función de su probabilidad ya fue aplicada a mediados del siglo XIX en la elaboración del famoso **código Morse** para telegrafía, denominado así en honor a su principal autor, el inventor y pintor americano **Samuel Morse**. En dicho código se asignan combinaciones cortas de puntos y rayas (en lugar de ceros y unos) a los caracteres más frecuentes en los textos en lengua inglesa, y combinaciones más largas a los menos frecuentes². Así pues, el código Morse puede considerarse un precursor de la codificación de fuente. No obstante, fue Shannon quien formalizó matemáticamente esta técnica.

Más específicamente, sea X' el conjunto de palabras código x'_1, x'_2, \dots, x'_S asignadas a los símbolos del alfabeto X por un código $C_{XX'}$. Suponemos además que este código es **unívocamente descifrable**, lo que significa que no hay ambigüedad posible en la decodificación de los símbolos a partir de una secuencia cualquiera de dígitos binarios (de lo contrario, el código destruiría parte de la información generada por la fuente, al no permitir que dicha información fuera completamente descifrada³). Además, sea L_i la longitud de la palabra código x'_i , con $i = 1 \dots S$. La longitud media del código introducida anteriormente, \bar{L} , obedece a la siguiente expresión (recuérdese que p_1, p_2, \dots, p_S son las probabilidades de los símbolos del alfabeto X):

$$\bar{L} = \sum_{i=1}^S p_i \cdot L_i \quad (1.7)$$

² Estrictamente, el código Morse no es binario, pues además del punto y la raya hace uso del espaciado entre caracteres y el espaciado entre palabras (silencios equivalentes a una raya y siete puntos respectivamente).

³ De momento estamos interesados en códigos unívocamente descifrables, si bien más adelante comprobaremos que algunas aplicaciones y servicios toleran un cierto grado de pérdida de información.

y enunció el teorema de codificación de fuente, según el cual, cualquiera que sea la entropía de la fuente, es posible construir un código tal que la información media por dígito binario se acerque tanto como se quiera a 1 bit.

Entonces, el **teorema de codificación de fuente** o **primer teorema de Shannon** establece que, para una fuente discreta sin memoria, el valor de \bar{L} correspondiente a un código unívocamente descifrable⁴ está acotado inferior y superiormente de la siguiente manera, donde ε es una cantidad positiva que puede hacerse tan pequeña como se quiera (lo que equivale a decir que siempre podremos encontrar un código mejor que uno dado):

$$H(X) \leq \bar{L} < H(X) + \varepsilon \quad (1.8)$$

Cuanto más se acerque el tamaño medio de las palabras código al valor de la entropía, tanto mejor. Así pues, en el mejor de los casos, $\bar{L} = H(X)$. En general, para cuantificar el grado de cercanía entre \bar{L} y $H(X)$ se utiliza el siguiente cociente (η):

$$\eta = \frac{H(X)}{\bar{L}} \quad (1.9)$$

Éste es, precisamente, el cociente introducido anteriormente para cuantificar la información media por dígito binario. De la primera desigualdad de la ecuación (1.8) se desprende que su valor es siempre menor o igual que 1, y por supuesto es siempre mayor o igual que 0, pues tanto numerador como denominador son magnitudes claramente no negativas. Además, según hemos visto, interesa que dicho cociente se acerque lo máximo posible a 1. Por tanto, la expresión (1.9) tiene todas las características de una medida de eficiencia, y de hecho se conoce como **eficiencia del código fuente**. No obstante, aunque las medidas de eficiencia no suelen tener unidades, la definición dada por (1.9) implica que, estrictamente, la eficiencia de la codificación de fuente quedaría expresada en bits de información por dígito binario (es decir, en bits/binit). De todos modos, dicha eficiencia suele manejarse en la práctica como si fuera adimensional.

En el caso particular de un código de longitud fija, se cumple $\bar{L} = \lceil \log_2 S \rceil$, y la eficiencia dada por la ecuación (1.9) se concreta de la siguiente manera:

$$\eta = \frac{H(X)}{\lceil \log_2 S \rceil} \quad (1.10)$$

También habíamos postulado esta expresión al caracterizar la información media por dígito binario en codificación de longitud fija.

Según el enunciado del teorema de codificación de fuente, $H(X) \leq \bar{L}$ es una condición necesaria para que un código sea unívocamente descifrable. Significa esto que, si un determinado código no

⁴ En su formulación completa, el primer teorema de codificación de fuente de Shannon abarca también códigos no unívocamente descifrables.

verifica esta condición, podemos descartarlo como unívocamente descifrable, pero si la verifica, no podemos asegurar nada, pues la condición es necesaria, pero no suficiente.

La demostración del teorema de codificación de fuente se apoya en un resultado intermedio, formulado a través del llamado **teorema de Kraft-McMillan**, pues se atribuyó conjuntamente a los americanos **Leon G. Kraft**, ingeniero, y **Brockway McMillan**, científico y funcionario estatal. El teorema establece que la condición necesaria para que un código sea unívocamente descifrable viene dada por la siguiente expresión:

$$K = \sum_{i=1}^S 2^{-L_i} \leq 1 \quad (1.11)$$

A esta inecuación se la conoce como **desigualdad de Kraft-McMillan**. En ella, K es el valor del sumatorio, el cual es inversamente proporcional a las longitudes de las palabras código. Si dicha condición no se cumple para un código dado, podemos descartarlo como unívocamente descifrable, pero si se cumple, no podemos afirmar nada. Por tanto, las comprobaciones $H(X) \leq \bar{L}$ y $K \leq 1$ son equivalentes a la hora de analizar el carácter descifrable del código.

La formulación del teorema de codificación de fuente es coherente con dos argumentos ya introducidos, tal como se expone a continuación (se utiliza CUD como acrónimo de código unívocamente descifrable):

$$\begin{aligned} CUD &\Rightarrow K \leq 1 \text{ (teorema de Kraft-McMillan)} \\ K \leq 1 &\Rightarrow \bar{L} \geq H(X) \text{ (demostración del teorema de codificación de fuente)} \\ \text{Por tanto, } CUD &\Rightarrow \bar{L} \geq H(X) \text{ (teorema de codificación de fuente)} \end{aligned}$$

No debe sorprender que las condiciones comprobaciones $H(X) \leq \bar{L}$ y $K \leq 1$ sean sólo necesarias, pero no suficientes. Por ejemplo, un código podría satisfacer cualquiera de ellas siendo un **código singular**, es decir, un código que contuviera dos o más palabras código exactamente iguales, por lo que claramente no sería unívocamente descifrable. La explicación es que ambas condiciones se basan en las longitudes de las palabras código solamente, pero no en el contenido de las mismas.

El teorema de Kraft-McMillan tiene una segunda parte, la cual establece que, dado un conjunto de valores naturales que satisfagan la desigualdad (1.11), siempre es posible encontrar un código unívocamente descifrable cuyas longitudes de palabra código coincidan con tales valores. Esta segunda parte del teorema, juntamente con la primera, permiten afirmar que la condición necesaria y suficiente para que exista un código unívocamente descifrable es que se satisfaga la desigualdad (1.11).

Para ilustrar los conceptos anteriores, consideremos la **Tabla 1.2**, donde se proponen diversos códigos para una fuente que trabaja con un alfabeto de 4 símbolos con las probabilidades indicadas. La tabla incluye los principales parámetros tanto del alfabeto como de los códigos propuestos. Los valores de tales parámetros merecen algunos comentarios:

Tabla 1.2. Ejemplos de código fuente para un alfabeto de 4 símbolos con probabilidades desiguales. Se indican los parámetros principales, tanto del alfabeto como de los códigos.

Símbolos	Probabilidades	Código I	Código II	Código III	Código IV	Código V
x_1	1/2	00	0	0	1	0
x_2	1/4	01	01	10	00	1
x_3	1/8	10	011	110	01	10
x_4	1/8	11	0111	111	10	11
Parámetros del alfabeto						
S	4					
$H(X)$	1.75					
Parámetros de los códigos						
K		1	0.9375	1	1.25	1.5
\bar{L}		2	1.875	1.75	1.5	1.25
η		0.875	0.9333	1	1.1667	1.4

- En primer lugar, podemos constatar que, efectivamente, en el caso de probabilidades desiguales, la selección de un código de longitud fija (código I) no es la opción más eficiente.
- Observamos también que el tamaño medio de las palabras código disminuye progresivamente al pasar del código I al código III, a la vez que la eficiencia aumenta. En particular, con el código III se alcanza la mínima longitud media de palabra código posible preservando su carácter descifrable, es decir, 1.75 (valor de la entropía), y por tanto la máxima eficiencia posible.
- Los códigos IV y V parecen mejorar el código III al alcanzar una longitud media de palabra código inferior a 1.75 (¡y una eficiencia superior a 1!), pero no son unívocamente descifrables. En ellos, la longitud media del código es inferior a la entropía de la fuente, y la desigualdad de Kraft-McMillan no se cumple. Por ejemplo, la secuencia binaria 1001001 puede ser decodificada de múltiples maneras según el código IV ($x_1x_2x_4x_3$, $x_1x_2x_1x_2x_1$, $x_4x_3x_2x_1$, etc.) o según el código V ($x_2x_1x_1x_2x_1x_1x_2$, $x_3x_1x_3x_1x_2$, $x_3x_1x_2x_1x_1x_2$, etc.). Definitivamente, estos dos últimos códigos alcanzan longitudes medias inferiores a 1.75 a costa de vulnerar la univocidad del descifrado e incurrir en la pérdida de información entre fuente y destinatario.

Estos ejemplos nos han permitido comprobar que la codificación de fuente es esencialmente un proceso de **compresión** de la información, o, mejor dicho, de compresión de la representación de la información. Cuanto mayor sea el nivel de compresión alcanzado, es decir, cuanto mayor sea la eficiencia de la codificación de fuente, menor es el número medio de dígitos binarios necesarios para representar la misma cantidad de información. Por ello, no debe extrañar que la codificación de fuente se utilice ampliamente para ahorrar memoria en cualquier tipo de dispositivo de almacenamiento (disco duro, CD, USB, etc.), y para reducir el tiempo de transmisión en los enlaces de comunicaciones.

Otra forma de ver la codificación de fuente es como un proceso de substitución de la fuente discreta original X por una fuente binaria equivalente con la misma entropía por dígito binario. En efecto, desde la perspectiva del destinatario, la fuente original codificada se comporta como una fuente binaria, pues tanto la una como la otra emiten, en definitiva, secuencias de ceros y unos. Entonces, para que sean equivalentes, sólo falta imponer que la entropía por dígito binario de la fuente binaria coincida con la entropía por dígito binario de la fuente original, expresiones (1.6) y (1.9) respectivamente, es decir:

$$\Omega(p) = \frac{H(X)}{L} = \eta \quad (1.12)$$

Resumiendo, la eficiencia de la codificación de fuente no es más que la entropía de la fuente binaria equivalente (afirmación cierta incluso si la fuente original es una fuente discreta con memoria). Por tanto, desde el punto de vista de la cantidad de información que el destinatario recibe por unidad de tiempo, las dos fuentes son indistinguibles (otra cosa es que el destinatario realmente trabaja con la fuente original, y por ello requiere conocer el código fuente para decodificar correctamente la secuencia de ceros y unos recibidos).

Además de que un código fuente sea unívocamente descifrable y eficiente, otra característica deseable es que sea **decodificable instantáneamente**. Esto significa que cada símbolo es decodificable justo cuando se acaba de recibir su palabra código, sin necesidad de prolongar la recepción de la secuencia de dígitos binarios e incurrir en el retraso que esto conlleva. Dicho de otra manera, la decodificación símbolo a símbolo es posible. Para que un código sea decodificable instantáneamente debe ocurrir que ninguna palabra código sea prefijo de otra. Un código que satisface esta condición es un **código prefijo**. Así pues, los términos “código decodificable instantáneamente” y “código prefijo” se pueden usar indistintamente. Por ejemplo, los códigos I y III de la **Tabla 1.2** son prefijo, y por tanto decodificables instantáneamente; en cambio, el código II, en el que cada palabra código es prefijo de la siguiente, claramente no lo es. En este código II, el dígito 0 marca el comienzo de cada palabra código, característica que lo hace unívocamente descifrable, pero no es posible deducir el símbolo enviado hasta que no se recibe el siguiente dígito 0.

Una forma gráfica y sencilla de constatar si un código es prefijo, además de alguna otra propiedad, es mediante su representación en forma de árbol. En la construcción del **diagrama de árbol** de un código, cada rama representa un dígito binario que se añade, y cada nodo en el que desemboca una rama puede estar ocupado por la palabra código que resulta de encadenar todos los dígitos añadidos, o puede estar vacío si, con los dígitos añadidos, todavía no corresponde a ninguna palabra código. En el proceso de construcción se adopta un convenio para las ramas, consistente en que las ramas hacia la izquierda corresponden a uno de los dos dígitos binarios, y las ramas hacia la derecha corresponden al otro dígito. Así, si se aplica el convenio de añadir el dígito 1 con las ramas hacia la derecha y el 0 con las ramas hacia la izquierda, el árbol resultante para el código I es el mostrado en la **Figura 1.5**. Aplicando el mismo convenio, los árboles correspondientes a los códigos II y III se muestran respectivamente en las **Figuras 1.6** y **1.7**. En el caso de la **Figura 1.6**, se han contrastado

las ramas que conducen a nodos que contienen palabras código frente a muchas ramificaciones y nodos no utilizados. La representación de la **Figura 1.7** ya solo incluye las ramas y nodos vacíos que conducen a nodos ocupados por palabras código.

Observando detenidamente los diagramas de árbol de los códigos I, II y III, no resulta difícil comprender que un código prefijo es aquel cuyas palabras código se ubican en nodos hoja. Es lo que ocurre con los códigos I y III, a diferencia del código II.

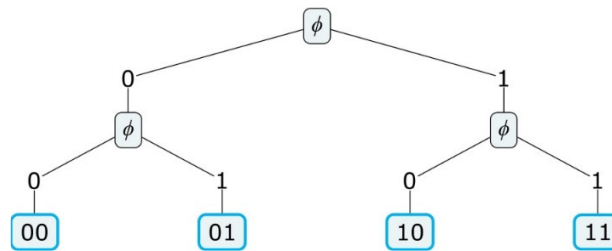


Figura 1.5. Representación en árbol del código I.

Uno de los resultados adicionales del teorema de Kraft-McMillan es que, dado un código unívocamente descifrable, siempre existe un código prefijo con exactamente la misma distribución de longitudes de las palabras código. Por ejemplo, la distribución de longitudes del código prefijo mostrado en la **Figura 1.8** es (1, 2, 3, 4), exactamente la misma que la del código II. Así pues, el **teorema de Kraft-McMillan** puede reformularse en términos de códigos prefijo en lugar de códigos unívocamente descifrables, es decir, la condición necesaria para que un código sea prefijo es que cumpla la desigualdad dada por (1.11). Análogamente, la segunda parte del teorema queda también reformulada, estableciendo que, dado un conjunto de valores naturales que satisfagan la desigualdad (1.11), siempre es posible encontrar un código prefijo cuyas longitudes de palabra código coincidan con tales valores.

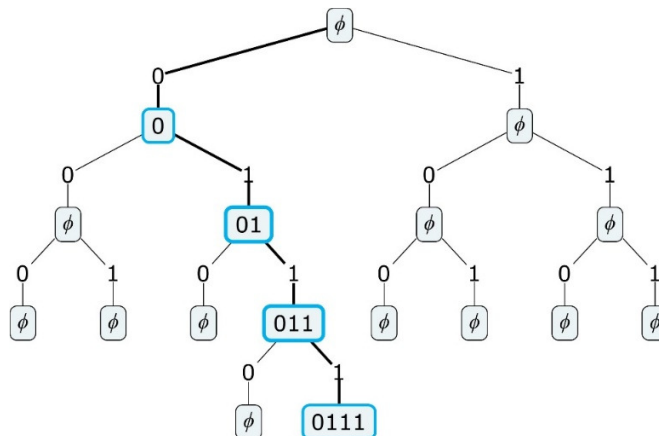


Figura 1.6. Diagrama de árbol del código II, resaltado en el seno de un árbol mayor que contiene muchas ramas y nodos sin utilizar.

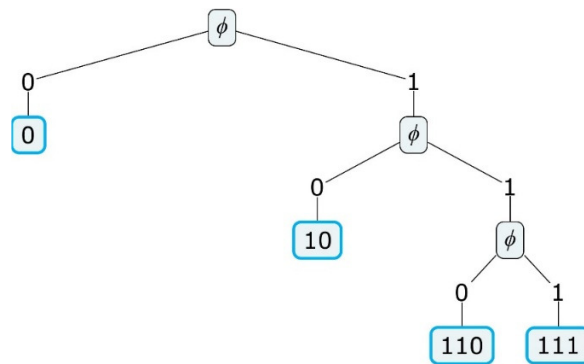


Figura 1.7. Diagrama de árbol del código III.

Si bien los códigos I, III y el representado en la **Figura 1.8** son prefijo, solamente en los dos primeros se cumple $K = 1$ (**Tabla 1.2**), mientras que para el tercero se satisface $K < 1$, pues este código es equivalente al código II en términos de distribución de longitudes de palabras código. Esta diferencia se puede justificar también a través de sus diagramas de árbol: se puede demostrar que, si el árbol resultante presenta tantas ramas hacia la derecha como hacia la izquierda, entonces $K = 1$; de lo contrario, $K < 1$. De todos modos, aunque es cierto que un valor alto de K (sin superar la unidad) indica que las palabras código son de corta longitud, la importancia de dicho valor es relativa, pues en su cálculo no se tiene en cuenta la probabilidad de cada longitud de palabra código. En este sentido, resultan mucho más significativas la longitud media de las palabras código (\bar{L}), en cuya expresión se pondera cada longitud por su probabilidad, y, en último término, la eficiencia (η). No debe extrañar, por ello, que la progresión de K entre los códigos I y III de la **Tabla 1.2** no esté alineada con las progresiones de \bar{L} y η , que son monótonas (una decreciente y la otra creciente) y permiten valorar claramente la calidad de los códigos.

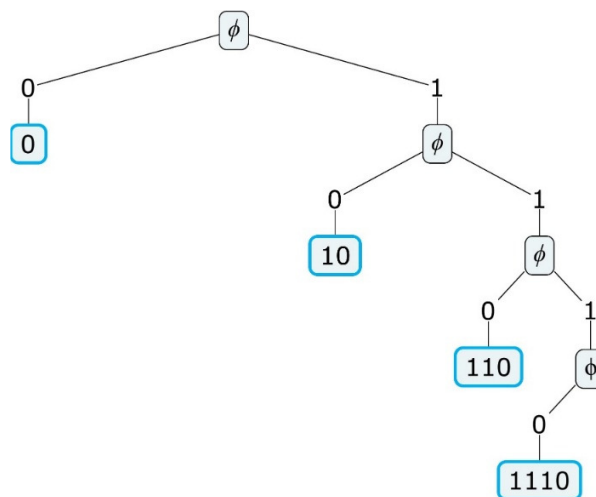


Figura 1.8. Código prefijo con la misma distribución de longitudes que el código II.

Podemos formular ya los requisitos de diseño definitivos que debe satisfacer un código fuente:

- Que sea prefijo. Puesto que todo código prefijo es unívocamente descifrable, y además conlleva el menor retardo de decodificación posible, los métodos de compresión más utilizados para fuentes sin memoria se centran en este tipo de códigos.
- Que sea lo más eficiente posible.

Estas dos condiciones nos permiten redefinir un **código óptimo** como aquel código prefijo cuya longitud media de palabra código es igual a su entropía (eficiencia 100%). De nuevo, una solución sub-óptima puede consistir en un código prefijo cuya eficiencia sea algo inferior al 100%.

1.4. TAXONOMÍA DE CÓDIGOS FUENTE. MÉTODOS Y FORMATOS DE COMPRESIÓN

Ya hemos visto que la codificación de fuente es esencialmente un proceso de compresión de la (representación de la) información o, si se prefiere, de eliminación de redundancia (al reducir la cantidad media de dígitos binarios por símbolo, es como si elimináramos dígitos sobrantes, es decir, redundantes). En el caso de la codificación escalar, típicamente utilizada para fuentes discretas sin memoria, podemos afirmar que hay compresión siempre y cuando el tamaño medio de las palabras del código seleccionado satisfaga la condición $\bar{L} < \lceil \log_2 S \rceil$ (se toma como referencia la longitud de las palabras código en codificación de longitud fija). Así pues, la **tasa de compresión** r en codificación escalar puede definirse de la siguiente manera:

$$r = \frac{\lceil \log_2 S \rceil}{\bar{L}} \quad (1.13)$$

Interesa que $r \geq 1$. El caso $r = 1$ corresponde al uso de codificación de longitud fija ($\bar{L} = \lceil \log_2 S \rceil$), lo que significa que no hay compresión. Esto puede ocurrir cuando, a pesar de que el alfabeto admite compresión, ésta no se aplica, o simplemente cuando la compresión no es aplicable porque todos los símbolos presentan la misma probabilidad. En este último caso, la fuente discreta sin memoria carece de redundancia, por lo que ésta no puede explotarse para reducir el tamaño medio de las palabras código. De hecho, puede demostrarse matemáticamente que, en el caso de símbolos igualmente probables, el uso de un código de longitud variable daría lugar a $r \leq 1$, correspondiendo el caso $r < 1$ a una expansión.

La **Figura 1.9** muestra, en forma de diagrama de Venn, la clasificación de los códigos fuente o de compresión de datos⁵ escalares, coherente con los conceptos y resultados de la sección anterior.

Donde la compresión de datos puede resultar más efectiva y proporcionar las más altas tasas de compresión es en la codificación vectorial de las fuentes discretas con memoria, es decir, fuentes en las que la probabilidad de emitir un símbolo está condicionada por la secuencia de símbolos anteriores y, en consecuencia, presentan un elevado grado de redundancia. Sin embargo, la tasa de

⁵ Cuando se trata de reducir el tamaño de un fichero para que ocupe menos memoria en un computador, se suele hablar de **compresión de datos**, mientras que en comunicación de datos se utiliza más el término **codificación de fuente**.

compresión dada por (1.13) no es aplicable en codificación vectorial (¿qué codificación fija podríamos tomar como referencia?), por lo que en este caso se redefine la **tasa de compresión** mediante la siguiente expresión, donde b_u es el número de dígitos binarios requeridos para representar una determinada cantidad de información antes de comprimir, y b_c el número de dígitos binarios que resultan después de la compresión:

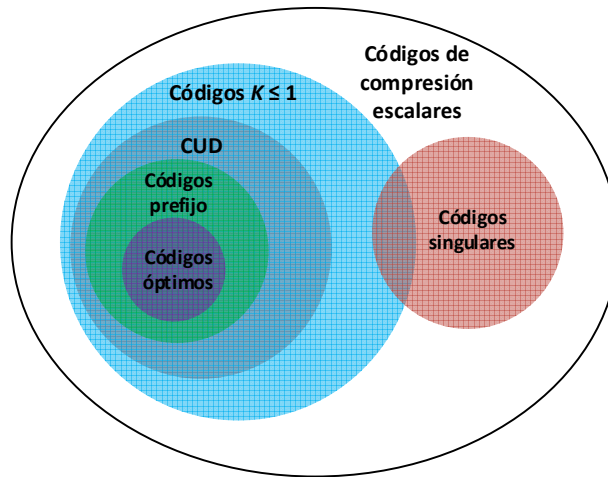


Figura 1.9. Taxonomía de los códigos de compresión escalares, típicamente utilizados para fuentes discretas sin memoria (CUD: código unívocamente descifrable).

$$r = \frac{b_u}{b_c} \quad (1.14)$$

De nuevo interesa que $r \geq 1$. La expresión (1.14) también es aplicable en codificación escalar, produciendo en este caso el mismo resultado que la expresión (1.13). Por lo tanto, la definición de la tasa de compresión dada por (1.14) puede considerarse general.

Otro aspecto a considerar para cualquier tipo de fuente, con o sin memoria, es si la compresión introduce pérdida de información, algo que hasta el momento dábamos por supuesto que no ocurría. La **compresión sin pérdidas**, también llamada **compresión reversible**, permite la reconstrucción perfecta de los datos originales a partir de los datos comprimidos. En el caso de la codificación escalar para fuentes discretas sin memoria, sería el tipo de compresión realizada por los códigos unívocamente descifrables de la **Figura 1.9**. La **compresión con pérdidas**, en cambio, permite sólo una reconstrucción aproximada de los datos originales, y es por ello una **compresión irreversible**. Puede parecer extraño que se admitan códigos de compresión con pérdidas, pero si están bien diseñados permiten alcanzar tasas de compresión muy elevadas antes de que la degradación sea perceptible por el usuario final. Naturalmente nos estamos refiriendo a la compresión de mensajes en formato multimedia (audio, imágenes, vídeo o una combinación de ellos), mensajes que finalmente van a ser procesados por los sistemas de percepción humana, visual y/o auditiva, cuyas limitaciones permiten tolerar ciertos grados de distorsión en la información recibida. Por tanto, cuando se trata de manejar datos multimedia, estas limitaciones se aprovechan en los códigos de compresión con pérdidas para reducir considerablemente la ocupación de

memoria en los dispositivos de almacenamiento, los tiempos de descarga desde servidores, o los tiempos de transmisión de los datos generados por las aplicaciones de *streaming* (de vídeo, de audio, telefonía por Internet, etc.). Podríamos resumir diciendo que los códigos de compresión sin pérdidas explotan la redundancia presente en los datos, mientras que los códigos de compresión con pérdidas explotan la redundancia presente en los datos y las limitaciones de la percepción humana.

La **Tabla 1.3** muestra los principales métodos (algoritmos) y formatos de compresión sin pérdida. Los métodos se han agrupado entre aquellos diseñados para fuentes sin memoria, o simplemente que no explotan la memoria de las fuentes, y aquellos que sí la explotan. En cuanto a los formatos de compresión, la lista completa es enorme, mucho más amplia que lo mostrado en la tabla, si bien ésta refleja los de uso más frecuente.

Por su parte, la **Tabla 1.4** muestra los métodos y formatos de compresión con pérdidas más comunes. Puesto que estos métodos se utilizan para datos que van a ser procesados por los sistemas de percepción humana, algunos de ellos se apoyan en sofisticados modelos psicoacústicos o psicovisuales a fin de eliminar la máxima redundancia posible. Estos modelos son muy completos, pues combinan los efectos del órgano receptor con la interpretación realizada por el cerebro, y su uso en la construcción de códigos da lugar a la llamada **codificación perceptual**. Así pues, algunos de los formatos que aparecen en la **Tabla 1.4** incorporan este tipo de codificación. Por otro lado, centrándonos en la última columna de la **Tabla 1.4**, que muestra los diversos formatos de compresión agrupados por tipos de datos, podemos resaltar dos aspectos:

Tabla 1.3. Métodos y formatos de compresión **sin pérdidas** más importantes.

Método de compresión		Formato de compresión
Sin memoria	Algoritmo de Huffman	Propósito general ZIP RAR
	Codificación aritmética	Audio FLAC
	Código de Shannon-Fano	MPEG-4 ALS Imágenes
Con memoria	Algoritmo de Lempel-Ziv	GIF
	RLE (<i>Run Length Encoding</i>)	PNG
	Codificación contextual	TIFF
	BWT (<i>Burrows-Wheeler Transform</i>)	JPEG-LS
	PPM (<i>Prediction by Partial Matching</i>)	JPEG 2000 sin pérdidas
	Codificación predictiva sin pérdidas	Vídeo Motion JPEG 2000 sin pérdidas

- En primer lugar, la distinción entre **voz** y **audio**, siendo esta última categoría más general al englobar voz, música y cualquier señal audible.

- En segundo lugar, la referencia al concepto de **contenedor multimedia**, que no es un formato, sino un repositorio de formatos de compresión para los diferentes tipos de datos que pueden formar parte de un mensaje multimedia. Cualquier fichero con una extensión del tipo contenedor incluye también una cierta cantidad de metadatos.

Globalmente, las **Tablas 1.3 y 1.4** merecen algunos comentarios adicionales:

- Primeramente, constatar que no existe una relación biunívoca entre métodos de compresión y formatos de compresión. En general, los formatos de compresión implementan variaciones y/o combinaciones de los métodos de compresión.
- En segundo lugar, el hecho de que buena parte de los formatos de la **Tabla 1.3** y todos los de la **Tabla 1.4** están optimizados para tipos de datos, aplicaciones o entornos específicos, y por lo tanto cada escenario dicta las mejores opciones. Incluso en algunos casos, el formato recomendado puede no coincidir con ninguno de los mostrados en ambas tablas, y además es posible que dicho formato dé lugar a una distorsión mucho mayor que la que cabría esperar, pero esto no constituye ninguna desventaja si con ello se consigue realzar aquello que pretende percibirse. Es el caso, por ejemplo, de las imágenes médicas (resonancias, ecografías, etc.).

En general, a cada formato de compresión le corresponde una extensión de fichero, mientras que a cada contenedor le corresponden varias. Por ejemplo, `.png` es la extensión de un fichero que contiene una imagen comprimida según el formato PNG, mientras que `.mp4` y `.m4a` son extensiones de ficheros MP4 – la primera representando un flujo multimedia, y la segunda un flujo sólo de audio.

Para terminar esta sección, una breve referencia a las grandes **plataformas multimedia**, que no deben confundirse con los contenedores multimedia. Ejemplos de tales plataformas son Windows Media de Windows, QuickTime de Apple, RealMedia de ReaNetworks o DivX de DivX, por citar algunas de las más conocidas. Consisten en soluciones propietarias que integran bibliotecas y programas de desarrollo y/o reproducción de contenidos digitales multimedia. Por lo general, estas plataformas son compatibles con la mayoría de los formatos de compresión expuestos en las **Tablas 1.3 y 1.4**, pero además suelen incluir formatos propios, con sus extensiones de fichero específicas.

1.5. ALGORITMOS DE HUFFMAN Y LEMPEL-ZIV

En esta sección describiremos dos de los métodos de compresión sin pérdidas más relevantes, uno para fuentes sin memoria (algoritmo de Huffman), y otro para fuentes con memoria (algoritmo de Lempel-Ziv). Empecemos por el primero.

Algoritmo de Huffman

Para el caso de fuentes discretas sin memoria, la mejor codificación de fuente se obtiene cuando se minimiza la longitud media de palabra código manteniendo el código unívocamente descifrable. Además, es deseable que dicho código sea prefijo, aunque ello no influye en la eficiencia de la codificación de fuente. Es un código óptimo en la **Figura 1.9**. En este contexto, el **segundo teorema**

de **codificación de fuente** de Shannon constituye una referencia fundamental, pues establece que, dada una fuente discreta sin memoria con entropía $H(X)$, es posible construir un código prefijo cuya longitud media de palabra, \bar{L} , cumpla la siguiente condición:

$$H(X) \leq \bar{L} < H(X) + 1 \quad (1.15)$$

Tabla 1.4. Métodos y formatos de compresión **con pérdidas** más frecuentes. Leyenda: a: audio; v: vídeo; d: datos; m: multimedia.

Método de compresión		Formato de compresión
Codificación vía transformada ⁶	DCT (<i>Discrete Cosine Transform</i>)	Voz
	KLT (<i>Karhunen-Loève Transform</i>)	CELP
Codificación predictiva con pérdidas	DFT (<i>Discrete-Fourier Transform</i>)	AMR
	Wavelet Compression	Audio
Compresión fractal	DPCM (<i>Differential Pulse Code Modulation</i>)	SILK (orientado a voz)
	DM (<i>Delta Modulation</i>)	LDAC
		MP3 → AAC
		OPUS
		Dolby Digital
		G.719, G.722, G.722.1, G.722.2
		Imágenes
		TIFF con pérdidas
		JPEG
		JPEG 2000
		Vídeo
		H.264/MPEG-4 AVC
		Motion JPEG
		AV1
		Contenedores
		Motion JPEG 2000 (a/v)
		AVI (a/v)
		WebM (a/v)
		FLV/F4V (a/v/d)
		Matroska (m)
		MP4 (m)

El algoritmo de Huffman, desarrollado por **David A. Huffman** en 1952, genera un código que verifica el teorema anterior. Esto no garantiza que la eficiencia sea 1, pues la longitud media de palabra código resultante puede ser ligeramente superior a la entropía de la fuente. No obstante, lo que

⁶ Existen versiones sin pérdidas de algunos métodos basados en transformadas, como DCT sin pérdidas o *lossless wavelet compression*.

puede afirmarse es que el algoritmo de Huffman genera el mejor código prefijo posible, es decir, el código prefijo de mayor eficiencia.

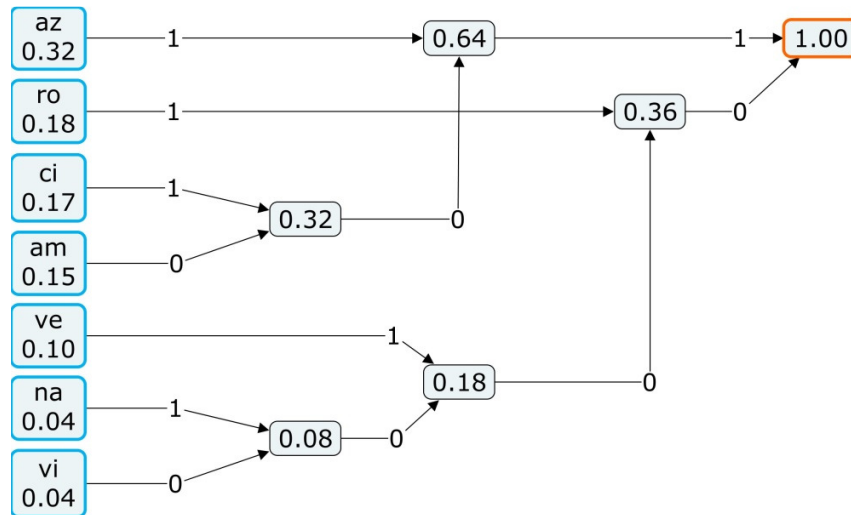


Figura 1.10. Aplicación del algoritmo de Huffman a la fuente de los 7 colores fundamentales.

La mejor forma de comprender el algoritmo es a través de un ejemplo. Supongamos una fuente que emite imágenes de baja resolución en base únicamente a los siete colores fundamentales del espectro visible. Representemos cada uno de estos colores por sus dos primeras letras, de forma que el alfabeto asociado queda descrito como $X = \{ro, na, am, ve, ci, az, vi\}$ ($S = 7$ símbolos). Admitamos que se trata de una fuente discreta sin memoria, cuyos símbolos tienen respectivamente las siguientes probabilidades: $\{0.18, 0.04, 0.15, 0.10, 0.17, 0.32, 0.04\}$. El primer paso para aplicar el algoritmo de Huffman es ordenar los símbolos por probabilidades decrecientes. A partir de aquí, tal como se muestra en la **Figura 1.10**, el proceso de codificación consiste en agrupar sucesivamente los dos símbolos con probabilidades más bajas, donde un símbolo en este contexto puede referirse a un símbolo original o a uno que resulte de agrupaciones anteriores. En cada agregación se asigna un 1 a una de las ramas incidentes en la nueva agrupación, y un 0 a la otra. En la **Figura 1.10**, se ha adoptado el criterio de asignar un 0 a la rama inferior y un 1 a la superior en todos los casos, pero realmente el criterio de asignación es irrelevante, mientras se asignen dígitos distintos a ambas ramas. Es más, dicho criterio puede modificarse durante la ejecución del algoritmo. La codificación final para cada símbolo original se obtiene al recorrer de derecha a izquierda las ramas del árbol resultante. En consecuencia, según las asignaciones de unos y ceros a las ramas, el código resultante varía y, por tanto, puede afirmarse que, dada una fuente discreta sin memoria, existen al menos tantos códigos de Huffman como asignaciones distintas de ceros y unos. Además, puede ocurrir que en un paso del algoritmo haya varias opciones de agrupación, y según el orden en que se escojan, el código resultante también varía. No obstante, todos estos códigos presentan la misma eficiencia y son prefijo. El alfabeto y el código obtenidos para las asignaciones de la **Figura 1.10**, así como los parámetros básicos de ambos, se muestran en la **Tabla 1.5**. Observamos que la eficiencia obtenida es muy alta, casi un 97.7%, y es la máxima que puede alcanzarse en la codificación escalar de la fuente dada.

Tabla 1.5. Resultado del algoritmo de Huffman para la fuente de los 7 colores fundamentales.

Símbolos	Probabilidades	Código de Huffman		
<i>ro</i>	0.18	01		
<i>na</i>	0.04	0001		
<i>am</i>	0.15	100		
<i>ve</i>	0.10	001		
<i>ci</i>	0.17	101		
<i>az</i>	0.32	11		
<i>vi</i>	0.04	0000		
Parámetros				
S	$H(X)$	K	\bar{L}	η
7	2.52017	1	2.58	0.976812

Podríamos aumentar ligeramente la eficiencia obtenida (97.7%) si aplicáramos el algoritmo de Huffman a agrupaciones de símbolos desde el primer momento. Para ilustrar esta idea sin que el tamaño del problema se desborde, consideremos una fuente más sencilla, con un alfabeto formado exclusivamente por los tres colores primarios (rojo, verde y azul), es decir, $X = \{ro, ve, az\}$ ($S = 3$ símbolos). Seguimos suponiendo que se trata de una fuente discreta sin memoria, cuyos símbolos presentan, por ejemplo, las siguientes probabilidades: $\{0.37, 0.27, 0.36\}$. La construcción del código de Huffman para esta fuente se muestra en la **Figura 1.11**. A continuación, supongamos que en lugar de codificar los símbolos elementales, codificamos directamente agrupaciones de 2 símbolos. Es decir, pasamos a trabajar con los 9 símbolos que resultan de formar todas las parejas posibles con los 3 colores primarios. Las nuevas probabilidades resultan naturalmente de multiplicar las correspondientes a los símbolos iniciales, puesto que hemos supuesto que la fuente es sin memoria y, por tanto, los símbolos elementales se generan de forma independiente. La **Figura 1.12** muestra la aplicación del algoritmo de Huffman a ese nuevo conjunto de símbolos, y la **Tabla 1.6** recoge las dos codificaciones obtenidas para la fuente de 3 colores. Observamos que con la agrupación de símbolos de dos en dos la eficiencia mejora; y seguiría mejorando si agrupáramos los símbolos de tres en tres, de cuatro en cuatro, y así sucesivamente. Es decir, la eficiencia de la **codificación por bloques** aumenta con el tamaño de los mismos. Este resultado es consecuencia del segundo teorema de codificación de fuente de Shannon. En efecto, dado que la entropía resultante de agrupar los símbolos de una fuente discreta sin memoria en bloques de tamaño B es $B \cdot H(X)$, la aplicación de dicho teorema se concreta en la siguiente acotación – recordemos la ecuación (1.15):

$$B \cdot H(X) \leq \bar{L}_B < B \cdot H(X) + 1 \quad (1.16)$$

Tabla 1.6. Resultados del algoritmo de Huffman para la fuente de los 3 colores primarios, sin agrupación (codificación escalar) y con agrupación de símbolos de 2 en 2 (codificación vectorial).

Símbolos	Probabilidades				Código	Símbolos	Probabilidades	Código
<i>ro</i>	0.37				0	<i>roro</i>	0.1369	110
<i>ve</i>	0.27				10	<i>rove</i>	0.0999	010
<i>az</i>	0.36				11	<i>roaz</i>	0.1332	101
Parámetros codificación escalar						<i>vero</i>	0.0999	001
<i>S</i>	<i>K</i>	<i>H(X)</i>	\bar{L}	η	<i>veve</i>	0.0729	1110	
3	1	1.57137	1.63	0.964028	<i>veaz</i>	0.0972	1111	
Parámetros codificación vectorial						<i>azro</i>	0.1332	100
<i>S</i>	<i>K</i>	<i>H(X)</i>	\bar{L}	η	<i>azve</i>	0.0972	000	
9	1	3.14273	3.1701	0.991367	<i>azaz</i>	0.1296	011	

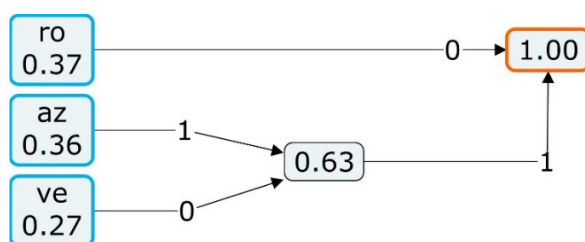


Figura 1.11. Aplicación del algoritmo de Huffman a la fuente de los 3 colores primarios.

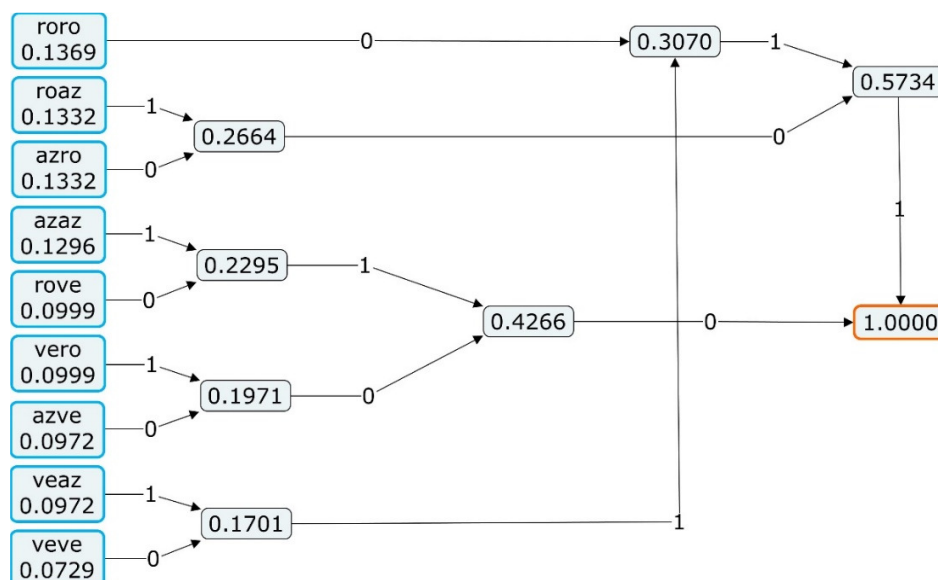


Figura 1.12. Aplicación del algoritmo de Huffman sobre parejas de colores primarios.

En esta expresión, $\overline{L_B}$ representa el número medio de dígitos binarios por bloque de B símbolos (es decir, la longitud media de palabra del código resultante de la agrupación). Si dividimos los tres miembros de (1.16) por B , obtenemos el siguiente resultado:

$$H(X) \leq \frac{\overline{L_B}}{B} < H(X) + \frac{1}{B} \quad (1.17)$$

El cociente $\overline{L_B}/B$ no es más que el número medio de dígitos binarios por símbolo elemental, es decir, lo que hemos venido definiendo como \bar{L} . Por tanto, podemos reescribir la expresión anterior de la siguiente forma:

$$H(X) \leq \bar{L} < H(X) + \frac{1}{B} \quad (1.18)$$

Esto significa que cuanto mayor es el tamaño de los bloques, más se acerca la longitud media de palabra a la entropía de la fuente, lo que no hace más que confirmar el primer teorema de codificación de fuente de Shannon, según el cual siempre es posible encontrar un código cuya longitud media de palabra se acerque tanto como queramos al mínimo dado por la entropía: la expresión (1.18) no es más que una reproducción de la expresión (1.8) en la que hemos reemplazado el término ε por $1/B$.

La codificación de Huffman que acabamos de describir es **estática**, en el sentido de que se basa en el conocimiento de la estadística de la fuente (distribución de probabilidades de los símbolos) para determinar las palabras código. Sin embargo, exceptuando algunos casos concretos, como la compresión de texto en una determinada lengua, dicha estadística es una incógnita. Por ello se han desarrollado las llamadas versiones **cuasi-dinámica** o **cuasi-adaptativa** y **dinámica** o **adaptativa** del algoritmo de Huffman. En el primer caso, se realiza un análisis previo de la fuente a fin de determinar las probabilidades de los símbolos, para a continuación obtener las palabras código según el procedimiento ya descrito (por tanto, se requieren dos pasadas sobre los datos a codificar, lo cual es una desventaja); en el segundo caso no se requiere ningún análisis estadístico preliminar, sino que el árbol de codificación se va construyendo a medida que se transmiten los símbolos. Otras variaciones del algoritmo de Huffman original son los códigos de Huffman (para fuentes con memoria), la codificación de Huffman para bloques de longitud variable (eliminando la restricción de que todos los bloques tengan que estar formados por el mismo número de símbolos), y la codificación de Huffman mediante dígitos multinivel (no binarios). En general, en cualquiera de sus versiones, la codificación de Huffman es una de las más utilizadas, concretamente en formatos como JPEG, TIFF, PNG y MP3, entre otros.

Acabamos de constatar que la codificación por bloques mejora la eficiencia, pero el ejemplo de la **Tabla 1.6** también refleja que esta mejora no es muy sustancial. De hecho, esto ocurre en general con las fuentes sin memoria, para las que la codificación de Huffman escalar suele producir una eficiencia muy alta, con poco margen para ser mejorada mediante codificación por bloques. Donde realmente se comprueban las ventajas de este último tipo de codificación es en su aplicación a fuentes con memoria. Lo vemos a continuación a través del algoritmo de Lempel-Ziv.

Algoritmo de Lempel-Ziv

El algoritmo de Lempel-Ziv pertenece a la categoría de **códigos lexicográficos** o **códigos basados en diccionario**. Supongamos que se trata de comprimir un texto, es decir, una concatenación de caracteres alfanuméricos y signos de puntuación. Una opción sería realizar un primer análisis del texto para identificar cadenas repetidas (se trata de comprimir aprovechando la memoria presente en los datos), con el fin de elaborar un diccionario de equivalencias asignando códigos más breves a las cadenas más frecuentes. A continuación, en una segunda etapa, se convertiría el texto utilizando las equivalencias del diccionario. Si bien esta técnica permitiría alcanzar buenos ratios de compresión, es poco práctica en lo que a consumo de recursos se refiere, pues además del tiempo invertido en la primera etapa (análisis del texto), requiere guardar el diccionario junto con el texto codificado (de otro modo no sería posible la decodificación), lo cual incrementa el tamaño del archivo de salida. Todavía peor, estos inconvenientes se magnificarían si se tratara de comprimir otros tipos de datos, como imágenes, voz o vídeo.

El algoritmo de Lempel-Ziv representa una solución ingeniosa a todos estos inconvenientes, pues permite realizar la compresión en una sola etapa, elaborando el diccionario a medida que se codifica la fuente, y sin requerir que ese diccionario forme parte del archivo de salida. Resulta por ello aplicable a cualquier tipo de información, y de hecho algunas utilidades de compresión como GZIP o PKZIP (en esencia, ZIP) se basan en diferentes implementaciones del algoritmo Lempel-Ziv básico. También forma parte de algunos formatos de compresión gráfica, como GIF o TIFF (en este último caso de forma opcional). De todos modos, a la hora de describir e ilustrar el algoritmo, resulta más sencillo asumir que el documento es un texto, y es lo que vamos a suponer a continuación.

La versión original del algoritmo fue desarrollada en 1977 por **Abraham Lempel** y **Jacob Ziv**, de ahí que comúnmente se le denomina LZ77. Posteriormente se desarrollaron versiones mejoradas, una en 1978 por los propios Lempel y Ziv (LZ78), y otra en 1984 por **Terry Welch** (LZW). Aunque todas las versiones giran en torno a una misma idea, nos centraremos en la descripción de LZ78, que es especialmente sencilla. El proceso básico del algoritmo consiste en diseccionar el texto en frases distintas según el orden de lectura, es decir, de izquierda a derecha en lenguaje occidental, entendiendo por frase cualquier secuencia de caracteres de mínima longitud que todavía no ha aparecido (obsérvese que esta definición de frase no se corresponde con su significado habitual). Por ejemplo, supongamos que se trata de codificar el texto *abracadabra pata de cabra*. La primera frase según la definición que acabamos de dar siempre coincide con la primera letra, en este caso *a*. A continuación, la siguiente frase más corta no aparecida es la letra *b*, y luego la *r*. Siguiendo, la frase más corta no aparecida hasta el momento es *ac*, luego *ad*, y así sucesivamente. La disección completa queda de la siguiente manera, donde se ha representado el espacio en blanco mediante un asterisco para que quede constancia explícita (el espacio en blanco es un carácter más):

a|b|r|ac|ad|ab|ra||p|at|a*|d|e|*c|abr|a*

Tabla 1.7. Codificación de Lempel-Ziv (versión LZ78) del texto *abracadabra pata de cabra*. El asterisco indica espacio en blanco.

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Frase	<i>a</i>	<i>b</i>	<i>r</i>	<i>ac</i>	<i>ad</i>	<i>ab</i>	<i>ra</i>	*	<i>p</i>	<i>at</i>	<i>a</i> *	<i>d</i>	<i>e</i>	* <i>c</i>	<i>abr</i>	<i>a</i>
Codificación	0 <i>a</i>	0 <i>b</i>	0 <i>r</i>	1 <i>c</i>	1 <i>d</i>	1 <i>b</i>	3 <i>a</i>	0 *	0 <i>p</i>	1 <i>t</i>	1 *	0 <i>d</i>	0 <i>e</i>	8 <i>c</i>	6 <i>r</i>	1

En total 16 frases, con la última repetida al no haber más caracteres a continuación. Estas frases constituyen el diccionario para este texto. El proceso de codificación se basa en el hecho de que cada frase resulta de alguna frase anterior a la que se ha añadido un carácter (excepto la última frase y las primeras correspondientes a caracteres individuales). Partiendo de esta observación, la construcción del código pasa por enumerar correlativamente las frases y representar cada una de ellas mediante una referencia a una frase ya aparecida junto con el nuevo carácter. La **Tabla 1.7** ilustra este proceso. Obsérvese que en la codificación se utiliza el valor 0 cuando no hay ninguna referencia anterior (como si el conjunto vacío ocupara la posición 0). En el caso de la última frase, no hace falta añadir ningún carácter a la referencia numérica.

La codificación binaria final se obtiene al expresar tanto los números como los caracteres en binario. Por lo que se refiere a los números, la cantidad de dígitos binarios utilizados es el que potencialmente puede ser necesario en función de la posición que ocupa la frase a codificar, independientemente de que ello implique una representación con ceros a la izquierda. Es decir, al codificar por ejemplo las frases 9 a 16, la referencia a una frase anterior puede requerir 4 dígitos binarios, y por tanto este es el número de dígitos que se utilizarán en todas estas frases cualquiera que sea el valor real referenciado. Así, si la referencia es a la frase 1, como ocurre con las frases 10, 11 y 16, su codificación binaria será 0001. En general, la regla a seguir es que se utilizan k dígitos binarios desde la posición $2^{k-1} + 1$ hasta la posición 2^k , ambas inclusive. Este criterio se aplica para que el decodificador pueda discernir claramente los dígitos binarios que corresponden a la referencia numérica de los que representan el carácter que sigue a continuación. Precisamente, para representar los caracteres se puede utilizar cualquiera de los códigos conocidos (por ejemplo, ASCII) o incluso un código de Huffman si se conoce su estadística. Supongamos que se utiliza el código ASCII de 7 bits para representar los caracteres. Sin compresión, y dado que el texto completo consta de 25 caracteres, el tamaño del archivo generado sería de 175 dígitos, mientras que con la codificación LZ78, el tamaño resultante es de 155 dígitos (se deja al lector como ejercicio). Ciertamente la tasa de compresión no parece muy buena; es más, para textos de longitud similar o textos muy aleatorios, con poco grado de redundancia, la codificación de Lempel-Ziv puede generar incluso archivos más largos que el original sin comprimir. El hecho es que esta codificación muestra sus ventajas sobre textos largos y archivos grandes en general, formados por interminables secuencias de 0s y 1s con alto grado de redundancia, para los que permite alcanzar tasas de compresión excelentes (del orden del 50% o incluso mejores).

El ejemplo que acabamos de ver pone de manifiesto que, en efecto, el algoritmo de Lempel-Ziv no requiere guardar o enviar ningún diccionario para el decodificador: el diccionario se va construyendo a medida que se codifica el archivo (una sola etapa), y forma parte de la codificación misma; el decodificador, por su parte, reconstruye el diccionario al tiempo que va decodificando, siempre sobre la base de frases ya descubiertas. Otra gran ventaja del algoritmo es que, a diferencia de otros como el código de Huffman, no requiere conocer la estadística de la fuente, lo cual es coherente con la gran mayoría de situaciones reales. Esta última característica permite ubicar el algoritmo de Lempel-Ziv en la categoría de los llamados **códigos universales**.

1.6. TRANSMISIÓN DE LA INFORMACIÓN

Hasta aquí hemos sentado las bases de la codificación de fuente, y hemos visto que el objetivo de la misma es eliminar la redundancia presente en los datos. Con esto se consigue ahorrar memoria cuando hay que almacenar estos datos y/o reducir el tiempo de transmisión cuando hay que enviarlos. La transmisión de datos es precisamente el otro gran ámbito de interés de la teoría de la información, en especial por el hecho de que los canales de transmisión pueden producir perturbaciones en los datos enviados. En esta sección nos ocupamos precisamente del efecto de dichas perturbaciones desde la perspectiva de la teoría de la información. Aunque las perturbaciones son muy variadas, por ahora es suficiente con englobarlas bajo el término “ruido”.

Recordemos que el proceso de comunicación consiste esencialmente en la transferencia de información de una fuente a un destinatario remoto, y que la fuente selecciona los símbolos de un cierto alfabeto $X = \{x_1, x_2, \dots, x_S\}$. En el mejor de los casos, el destinatario debería recibir los símbolos enviados por la fuente, pero estos pueden ser alterados por el ruido del canal, de modo que en general el destinatario maneja un conjunto distinto de palabras código, hecho que podemos asimilar a que manejara un alfabeto distinto $Y = \{y_1, y_2, \dots, y_N\}$, con $N \geq S$. Por ejemplo, supongamos que el alfabeto de la fuente consta de $S = 5$ símbolos, codificados en longitud fija de 3 dígitos binarios; dado que el ruido del canal puede alterar cualquiera de los dígitos, el destinatario puede recibir cualquier palabra código que resulte de la combinación de 3 de ellos, como si trabajara con un alfabeto de tamaño $N = 8$. Esta visión del proceso de comunicación como una transformación de alfabetos se ilustra en la **Figura 1.13**, la cual no es más que una abstracción de la **Figura 1.3**, consistente en haber substituido todos los elementos entre fuente y destinatario por un canal equivalente. Éste es un canal discreto, ya que transforma un **conjunto discreto** de símbolos (alfabeto X) en otro conjunto también discreto (alfabeto Y), y además es ruidoso.

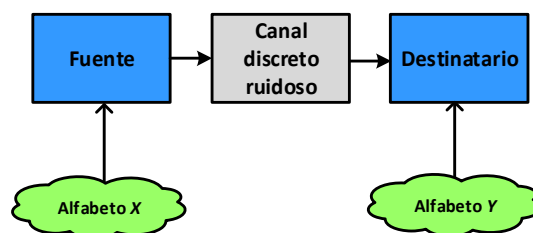


Figura 1.13. Transmisión a través del canal discreto.

Definamos las siguientes probabilidades, donde i y j son índices, el primero variando entre 1 y S y el segundo entre 1 y N :

- $p(x_i)$: probabilidad de que la fuente seleccione el símbolo x_i .
- $p(y_j)$: probabilidad de que el destinatario reciba el símbolo y_j .
- $p(x_i|y_j)$: probabilidad de que se haya transmitido el símbolo x_i dado que se ha recibido el símbolo y_j .
- $p(y_j|x_i)$: probabilidad de que se reciba el símbolo y_j dado que se ha transmitido el símbolo x_i .

Observemos que las dos últimas probabilidades son condicionadas. En adelante supondremos que el canal es **invariante en el tiempo**, es decir, que las probabilidades condicionadas no dependen del tiempo, y **sin memoria**. Formalmente, la ausencia de memoria en el canal se puede expresar de la siguiente manera, al considerar una secuencia de k símbolos transmitidos $(u_1 u_2 u_3 \dots u_l \dots u_k)$, donde $u_l = x_i \forall l = 1 \dots k, \forall i = 1 \dots S$, y los correspondientes k símbolos recibidos $(v_1 v_2 v_3 \dots v_l \dots v_k)$, donde $v_l = y_j \forall l = 1 \dots k, \forall j = 1 \dots N$:

$$p(v_1 v_2 v_3 \dots v_k | u_1 u_2 u_3 \dots u_k) = \prod_{l=1}^k p(v_l | u_l) \quad (1.19)$$

Es decir, la probabilidad de una secuencia de símbolos de salida dada una secuencia de símbolos de entrada se reduce al producto de probabilidades de cada símbolo de salida dado el correspondiente símbolo de entrada. Dicho aún de forma más sencilla, la probabilidad de que se reciba un determinado símbolo sólo depende del símbolo actualmente transmitido, pero no de los símbolos anteriores. Las dos propiedades, invariabilidad en el tiempo y ausencia de memoria, caracterizan el canal como un **canal discreto sin memoria** (DMC – *Discrete Memoryless Channel*)⁷.

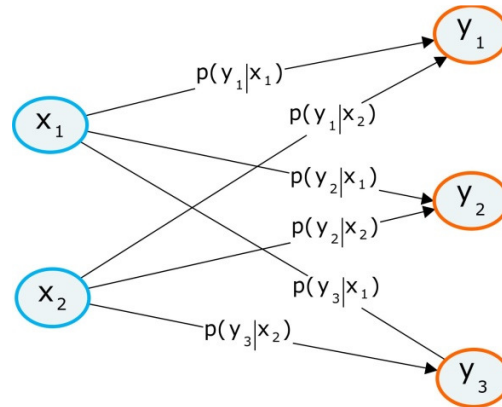


Figura 1.14. Diagrama de probabilidades de transición hacia adelante para un canal con 2 símbolos de entrada y 3 de salida.

En el caso de un canal discreto sin memoria, las probabilidades condicionadas definidas más arriba, $p(x_i|y_j)$ y $p(y_j|x_i)$, adquieren una especial relevancia, pues con cualquier conjunto de ellas queda

⁷ No confundir el canal discreto sin memoria con la fuente discreta sin memoria.

caracterizado el efecto del canal. Concretamente, $p(x_i|y_j)$ son las llamadas **probabilidades de transición hacia atrás** (*backward transition probabilities*), mientras que $p(y_j|x_i)$ son las **probabilidades de transición hacia delante** (*forward transition probabilities*). A modo de ejemplo, la **Figura 1.14** muestra el diagrama de probabilidades de transición hacia delante para un canal con 2 símbolos de entrada y 3 de salida. Si se pretende que el destinatario reciba y_1 cuando la fuente emite x_1 , y que el destinatario reciba y_2 cuando la fuente emite x_2 , todas las transiciones entre símbolos con diferente subíndice representan transmisiones erróneas.

Una forma alternativa de representar las probabilidades de transición hacia delante es mediante la **matriz de probabilidades de transición**, donde $p_{ij} = p(y_j|x_i) \forall i = 1 \dots S, \forall j = 1 \dots N$:

$$\begin{pmatrix} p_{11} & \cdots & p_{1N} \\ \vdots & \ddots & \vdots \\ p_{S1} & \cdots & p_{SN} \end{pmatrix} \quad (1.20)$$

El análisis de la cantidad de información que se transfiere a través de un canal discreto sin memoria empieza con el concepto de información mutua. La **información mutua** entre un símbolo de entrada x_i y un símbolo de salida y_j , denotada por $I(x_i; y_j)$, no es más que la información transferida a través del canal cuando se transmite x_i y se recibe y_j . Recordemos que información en nuestro contexto es un concepto vinculado al grado de incertidumbre; por tanto, la información mutua tal como la acabamos de definir no es más que la reducción de incertidumbre acerca del símbolo x_i por el hecho de recibir el símbolo y_j , es decir:

$$I(x_i; y_j) = \log_2 \frac{1}{p(x_i)} - \log_2 \frac{1}{p(x_i|y_j)} \quad (1.21)$$

En esta expresión, el primer término es la incertidumbre absoluta de que el símbolo seleccionado por la fuente sea x_i , mientras que el segundo término es la incertidumbre de que el símbolo seleccionado por la fuente sea x_i si se ha recibido el símbolo y_j . La diferencia entre ambas incertidumbres es obviamente la información transmitida por el canal cuando la fuente transmite x_i y el destinatario recibe y_j . Por la propiedad del logaritmo de un cociente, podemos reescribir la expresión anterior de la siguiente manera:

$$I(x_i; y_j) = \log_2 \frac{p(x_i|y_j)}{p(x_i)} \quad (1.22)$$

Esta definición es coherente con lo que sucede en algunos casos extremos. Por ejemplo, supongamos que la recepción del símbolo y_j permite asegurar que el símbolo transmitido es x_i , es decir, $p(x_i|y_j) = 1$; entonces se obtiene $I(x_i; y_j) = \log_2 \frac{p(x_i|y_j)}{p(x_i)} = \log_2 \frac{1}{p(x_i)} = I(x_i)$, es decir, la información transmitida por el canal es la información contenida en el símbolo x_i . En el otro extremo, si la recepción del símbolo y_j no resuelve ninguna incertidumbre acerca del símbolo x_i , es decir, $p(x_i|y_j) = p(x_i)$, entonces resulta $I(x_i; y_j) = \log_2 \frac{p(x_i|y_j)}{p(x_i)} = \log_2 \frac{p(x_i)}{p(x_i)} = \log_2 1 = 0$. Es lógico, pues si la transmisión a través del canal no resuelve ninguna incertidumbre, la información

transferida es nula. En general, los escenarios reales se sitúan en algún punto intermedio entre ambos extremos.

El término $I(x_i; y_j)$ dado por la expresión (1.22) corresponde a una pareja concreta de símbolos de entrada y salida. Al ponderar dicho término por la probabilidad de que en efecto se transmita el símbolo x_i y se reciba el símbolo y_j , es decir, $p(x_i, y_j)$, y extenderlo a todas las combinaciones posibles de tales símbolos, obtenemos la **información mutua media** $I(X; Y)$:

$$I(X; Y) = \sum_{i=1}^S \sum_{j=1}^N p(x_i, y_j) \cdot I(x_i; y_j) = \sum_{i=1}^S \sum_{j=1}^N p(x_i, y_j) \cdot \log_2 \frac{p(x_i|y_j)}{p(x_i)} \quad (1.23)$$

Esta magnitud se mide en bits por símbolo, pues representa la cantidad media de información enviada al transmitir un símbolo. Aplicando algunas leyes básicas de la teoría de probabilidades, podemos manipular la expresión (1.23) para obtener el siguiente resultado (se deja al lector como ejercicio), donde $H(X|Y)$ queda definida a través de la expresión (1.25):

$$I(X; Y) = H(X) - H(X|Y) \quad (1.24)$$

$$H(X|Y) = \sum_{i=1}^S \sum_{j=1}^N p(x_i, y_j) \cdot \log_2 \frac{1}{p(x_i|y_j)} \quad (1.25)$$

Sin conocer todavía qué representa el término $H(X|Y)$, podemos afirmar que en el mejor de los casos debería anularse, pues implicaría que $I(X; Y) = H(X)$, es decir, la información media por símbolo transportado por el canal coincidiría con la información media por símbolo seleccionado por la fuente (entropía de la fuente), y por tanto el canal estaría transmitiendo la información generada por la fuente sin pérdida de integridad de la misma, es decir, sin equivocación. Sería el caso de un canal sin ruido. Por ello, al término $H(X|Y)$ se le conoce precisamente como **equivocación**, y representa la información media que se pierde en el canal por efecto del ruido.

El principal objetivo de diseño del sistema que comunica la fuente con el destinatario es maximizar la cantidad de información media transferida por símbolo, es decir, la información media mutua $I(X; Y)$. Se trata, pues, de un problema de optimización. Para caracterizarlo, conviene poner de manifiesto cuáles son las variables de optimización. Se puede comprobar que si conocemos las probabilidades de entrada $p(x_i)$ (características de la fuente) y las probabilidades de transición hacia delante $p(y_j|x_i)$ (características del canal), los otros dos conjuntos de probabilidades, es decir, las probabilidades de salida $p(y_j)$ y las probabilidades de transición hacia atrás $p(x_i|y_j)$, quedan determinados:

$$p(y_j) = \sum_{i=1}^S p(x_i, y_j) = \sum_{i=1}^S p(y_j|x_i) \cdot p(x_i) \quad (1.26a)$$

$$p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)} = \frac{p(y_j|x_i) \cdot p(x_i)}{p(y_j)} \quad (1.26b)$$

Estas expresiones confirman que podemos plantear la maximización de la información mutua media tomando como variables de optimización las probabilidades de entrada y las probabilidades de transición hacia delante (también podríamos escoger las probabilidades de transición hacia atrás y las probabilidades de salida). Además, tanto las probabilidades de entrada, por un lado, como las probabilidades de transición hacia delante, por otro, tienen que satisfacer las restricciones habituales, es decir, que sean valores no negativos y que la suma de todos ellos sea igual a la unidad. Así pues, el problema de optimización queda formulado de la siguiente manera, donde C_s representa el valor máximo de la información mutua media, es decir, lo que se conoce como **capacidad del canal discreto**:

$$C_s = \max_{\substack{p(x_i) \\ p(y_j|x_i)}} I(X;Y) \quad (1.27a)$$

tal que:

$$p(x_i) \geq 0 \quad (1.27b)$$

$$\sum_{i=1}^S p(x_i) = 1 \quad (1.27c)$$

$$p(y_j|x_i) \geq 0 \quad (1.27d)$$

$$\sum_{j=1}^N p(y_j|x_i) = 1 \quad (1.27e)$$

Recordemos que las probabilidades de entrada son características de la fuente, mientras que las probabilidades de transición hacia delante lo son del canal. Es importante observar que, si bien las probabilidades de entrada eran fijas al tratar el problema de la codificación de fuente, el efecto final de la misma, como veremos a continuación, puede verse como un reajuste del alfabeto y de su entropía, como si efectivamente la fuente fuera una variable del problema.

Si $R_{s,max}$ es la máxima velocidad de transmisión de símbolos que permite el canal, es decir, la máxima cantidad de símbolos que el canal puede transmitir por unidad de tiempo, entonces la máxima velocidad de transferencia de información de dicho canal vendrá dada por la siguiente expresión:

$$C = R_{s,max} \cdot C_s \quad (1.28)$$

C es la llamada **capacidad del canal de transmisión**. Observemos que hay coherencia entre las unidades de medida de las diversas magnitudes: C_s se mide en bits por símbolo (o sus múltiplos), $R_{s,max}$ en símbolos por segundo (o sus múltiplos) y C en bits por segundo (o sus múltiplos).

La capacidad del canal de transmisión es objeto del **teorema fundamental de Shannon para un canal ruidoso**, el cual establece lo siguiente (más o menos literalmente):

Si la capacidad de un canal es C , y la velocidad con que la fuente transmite información es $R \leq C$, entonces se puede encontrar un sistema de codificación tal que la salida de la fuente se transmita a través del canal con una tasa de errores arbitrariamente pequeña. Al contrario, si $R > C$, entonces no es posible transmitir la información sin errores.

Este enunciado se completa con una expresión matemática para C , que a su vez requiere sendas expresiones para $R_{s,max}$ y C_s . La expresión de C_s es precisamente la solución al problema de optimización planteado en (1.27), y constituye otra de las grandes contribuciones de Shannon. No obstante, tanto dicha expresión como la de $R_{s,max}$ se formulan en términos de ciertos parámetros físicos del canal, cuya comprensión requiere el conocimiento y tratamiento de las señales así como de los medios de transmisión que las propagan. Por lo tanto, volveremos a la capacidad del canal de transmisión en la parte final de este bloque dedicado a la comunicación de datos, cuando hayamos tratado las señales y los medios de transmisión. De todos modos, esto no es obstáculo para indagar ya en el significado de la expresión “sistema de codificación”, que aparece en el enunciado del teorema fundamental de Shannon para un canal ruidoso. Recordemos que la codificación de fuente puede reinterpretarse como la substitución de una fuente discreta cualquiera X por una fuente binaria cuya entropía $\Omega(p)$ satisface la ecuación (1.12), es decir, se cumple $\Omega(p) = \eta$, donde η es la eficiencia. Si el código fuente es lo suficientemente bueno, entonces $\eta \approx 1$ y, por tanto, $\Omega(p) \approx 1$, lo cual significa que los dos dígitos binarios se generan prácticamente con la misma probabilidad, como se desprende de la **Figura 1.4**; por tanto, podemos afirmar que un primer efecto de la codificación de fuente es equiparar las probabilidades de ambos dígitos en la fuente binaria equivalente. Por otro lado, si la fuente original carece de memoria, la fuente binaria equivalente también, mientras que, si la fuente original tiene memoria y, nuevamente, la codificación de fuente es lo suficientemente buena como para eliminar el máximo de redundancia, entonces la fuente binaria equivalente resultante es nuevamente sin memoria. En conclusión, cuando la codificación de fuente es de calidad, podemos afirmar que la fuente binaria equivalente carece de memoria (hemos eliminado toda redundancia) y genera ambos dígitos con la misma probabilidad (hemos maximizado la eficiencia).

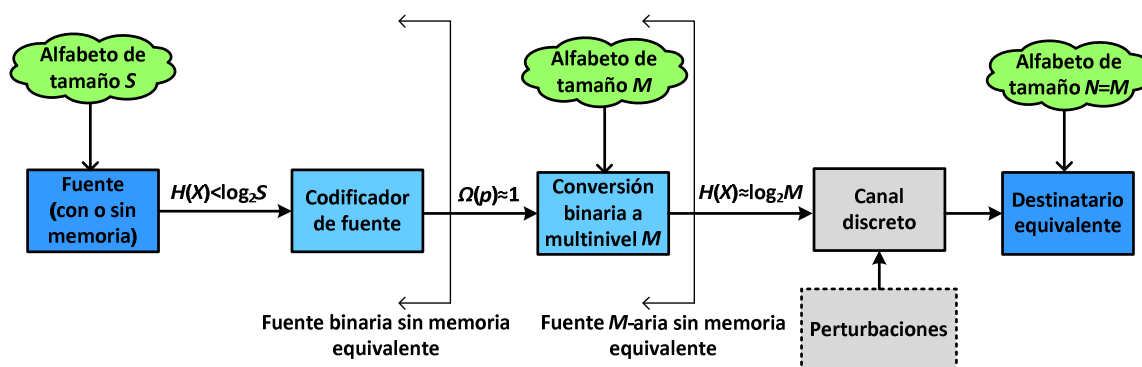


Figura 1.15. Efecto de la codificación de fuente. El conversor binario a multinivel es un primer elemento de adaptación de la fuente al canal (es decir, hace el papel de transmisor).

La **Figura 1.15** muestra la transformación de la fuente original en una fuente binaria sin memoria cuya entropía es (prácticamente) óptima. Puesto que los símbolos que finalmente se transmiten los define el canal, un paso adicional en el proceso de transformación de la fuente es la agrupación de los dígitos binarios a la salida del codificador de fuente en palabras código de longitud fija, el valor de la cual es $\log_2 M$ si M es el número total de símbolos que maneja el canal. Esta transformación viene impuesta por el canal, y no representa ninguna mejora de la fuente, pues su optimización se

alcanzó ya con el codificador de fuente; se trata, simplemente, de una reformulación de dicha fuente en términos de alfabeto. La conversión binaria a multinivel se muestra también en la **Figura 1.15**, donde se pone de manifiesto que a la entrada del canal es como si conectáramos una fuente discreta sin memoria (DMS) definida sobre un alfabeto de tamaño M formado por símbolos igualmente probables. Además, y a diferencia de lo que ocurría con S , el valor de M es siempre una potencia de 2. Se confirma, pues, que la codificación de fuente representa una transformación de la misma en términos de alfabeto y entropía, desde una situación en general por debajo de la óptima, con una entropía que puede ser muy inferior al máximo dado por $\log_2 S$, a una situación óptima con una entropía máxima de valor 1 si nos referimos a la fuente binaria equivalente, o de valor $\log_2 M$ si nos referimos a la fuente M -aria equivalente. En consecuencia, podemos afirmar que la codificación de fuente es la parte de la solución al problema formulado en (1.27) que contribuye a maximizar la información mutua media en términos de las probabilidades de entrada. Por su parte, la llamada **codificación de canal**, que estudiaremos más adelante, es la otra parte de la solución, la que se encarga de maximizar la información mutua media en términos de las probabilidades de transición hacia delante. En definitiva, codificación de fuente y codificación de canal son dos sub-problemas independientes que conjuntamente constituyen el **sistema de codificación** al que se refiere el teorema de Shannon.

La codificación de fuente y la codificación de canal se pueden ver también, respectivamente, como los procesos de maximizar la entropía y minimizar el término de equivocación en la expresión (1.24). Minimizar el término de equivocación equivale a optimizar las probabilidades de transición hacia delante de un canal que sigue viniendo representado por un diagrama similar al de la **Figura 1.14**, si bien ahora con alfabetos de entrada y salida coincidentes (consecuencia de que M es una potencia de 2). Aunque trataremos la codificación de canal más adelante, tan solo añadir que dicha codificación no consiste en modificar el canal físico (medio de transmisión) entre fuente y destinatario, sino en proteger los mensajes de la fuente de forma que el canal equivalente resultante presente unas probabilidades de transición hacia delante optimizadas. Con la codificación de fuente sucedía algo parecido: no modificábamos la fuente original, pero representábamos los símbolos de forma que la fuente resultante equivalente fuera óptima.

Para terminar esta sección, consideremos un caso particular de canal discreto, el llamado **canal binario simétrico**. Dado que la codificación de fuente y la conversión binaria a multinivel equivalen conjuntamente a una fuente M -aria sin memoria con símbolos igualmente probables y palabras código de longitud fija, el efecto del canal puede verse símbolo a símbolo o dígito binario a dígito binario. En este último caso, podemos representarlo mediante el sencillo diagrama de probabilidades de transición hacia delante mostrado en la **Figura 1.16**. Damos por supuesto que los dos dígitos binarios son igualmente probables, coherentemente con la hipótesis de una codificación de fuente óptima⁸. Las probabilidades de transición hacia delante se reducen a una probabilidad de error p , provocada por la presencia de ruido en el canal, y una probabilidad de acierto

⁸ De todos modos, algunos textos adoptan una definición más general del canal binario simétrico, en la que los dos dígitos binarios presentan probabilidades desiguales.

complementaria de valor $1 - p$. Observemos que estas probabilidades se dan tanto cuando se transmite un 1 como cuando se transmite un 0, de ahí el carácter simétrico del canal.

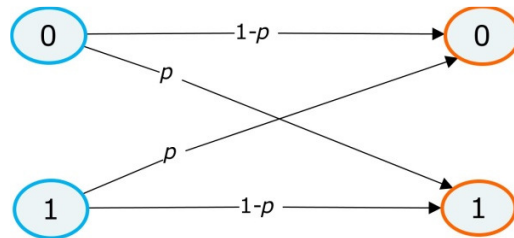


Figura 1.16. Canal binario simétrico.

CAPÍTULO 2. TEORÍA DE SEÑAL

Como ya hemos señalado en el capítulo anterior, la información se transmite mediante señales que se propagan a través de los medios de transmisión. Resulta conveniente, pues, adquirir los conocimientos más básicos de **Teoría de la Señal**, es decir:

- ¿Qué es una señal y cómo podemos caracterizarla?
- Potencia de una señal.
- Espectro de una señal.

Estos conceptos son básicos para comprender y analizar los efectos que experimentan las señales cuando se propagan a través de los medios de transmisión, efectos que estudiaremos más adelante.

ÍNDICE

2.1.	CARACTERIZACIÓN DE LAS SEÑALES	39
2.2.	LA SEÑAL SINUSOIDAL	41
2.3.	POTENCIA DE SEÑAL	45
2.4.	ESPECTRO DE UNA SEÑAL	53

2.1. CARACTERIZACIÓN DE LAS SEÑALES

En el ámbito de las comunicaciones, una **señal** $s(t)$ es cualquier variación temporal de una magnitud eléctrica, electromagnética, óptica, acústica o química⁹ que contenga un mensaje. El envío de este mensaje tiene lugar con la propagación de la señal a través de un **medio de transmisión**. Así pues, se puede ver la señal como el mecanismo de transporte del mensaje de un lugar a otro. Por tanto, una señal transporta información y energía (esto último lo veremos más adelante al tratar la potencia de una señal). Si bien estas afirmaciones se corresponden con la concepción tradicional de *señal*, en la práctica también se utiliza este término para designar variaciones temporales de las magnitudes especificadas más arriba, aunque dichas variaciones no contengan información, sino solamente energía. Iremos viendo algunos casos de estos usos.

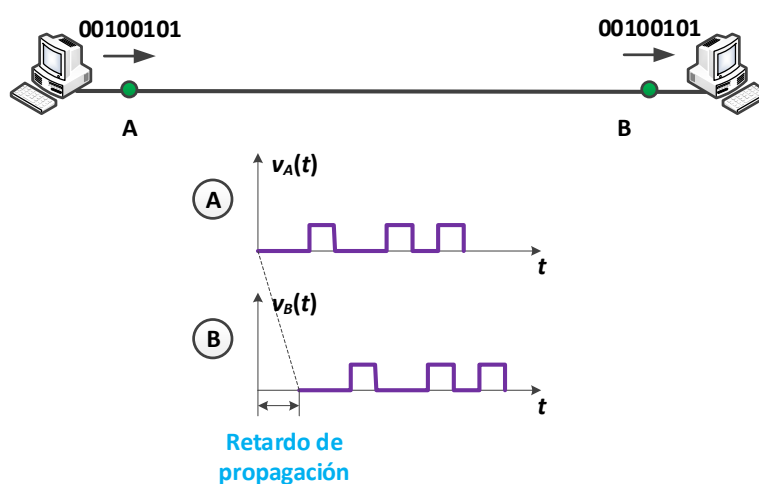


Figura 2.1. Envío de una secuencia de dígitos binarios mediante una señal de voltaje entre dos computadores conectados directamente a través de una línea de cobre.

La **Figura 2.1** muestra un ejemplo sencillo consistente en la transmisión del carácter "%" (mensaje) entre dos ordenadores conectados directamente mediante una línea de cobre (medio de transmisión). En código ASCII (código habitual para representar la información en los ordenadores) el carácter anterior es 25h (formato hexadecimal), es decir, 00100101 (formato binario). En este caso, la magnitud física que varía en función de la información a transmitir es un voltaje, correspondiendo un nivel alto al "1" binario y un nivel bajo (nulo) al "0" binario. Obsérvese que entre el mensaje y la señal ha tenido lugar un paso previo de **codificación**, que consiste en representar el mensaje en formato binario. Éste es un rasgo inequívoco de las comunicaciones digitales, que son las que nos ocupan en redes de computadores. El envío de la información del punto A al punto B de la figura tiene lugar con la propagación de la señal de voltaje a través de la

⁹ Las señales químicas han cobrado importancia en los últimos años con el auge de las comunicaciones moleculares.

línea de cobre. Los dos diagramas temporales mostrados ponen de manifiesto el retardo introducido por la propagación de dicha señal.

Las señales se pueden clasificar desde diversos puntos de vista. Un primer criterio se refiere a cómo están organizados los ejes de tiempo y valores al representarlas gráficamente. En este sentido, se puede distinguir entre señales analógicas y digitales.

En una **señal analógica**, la magnitud física representada varía de forma continua en el tiempo. Es decir, puede tomar cualquier valor (dentro de unos márgenes), y los cambios de valor de la misma se pueden producir en cualquier instante de tiempo. Por lo tanto, se trata de una variación continua tanto en el eje de valores como en el eje de tiempos. En la **Figura 2.2(a)** se ilustra este concepto. La curva mostrada podría corresponder, por ejemplo, al voltaje de salida de un micrófono cuando detecta las ondas sonoras emitidas por una persona al hablar. Estas ondas sonoras no son más que la propagación de otra señal también analógica, consistente en la variación de intensidad acústica producida al hablar. El micrófono es en este caso el elemento transductor, pues transforma la señal acústica vocal en una señal eléctrica lo más parecida posible.

En el caso de una **señal digital**, la magnitud física sólo puede tomar valores dentro de un conjunto discreto, y los cambios de valor de la misma sólo pueden producirse en instantes regulares de tiempo. Por tanto, a diferencia de lo que ocurre con las señales analógicas, en una señal digital las variaciones son discretas tanto en valores como en tiempo. Las señales digitales no se producen de forma natural, sino que son generadas por los ordenadores y demás dispositivos de computación. La **Figura 2.2(b)** muestra un ejemplo de señal digital. Como se observa, en este caso sólo se utilizan dos niveles de voltaje para representar los dos dígitos binarios, y los cambios de voltaje sólo tienen lugar en instantes que son múltiplos enteros del llamado **tiempo de bit**.

Otro criterio de clasificación de las señales es el que se refiere a si éstas presentan un patrón básico que se repite o no. Bajo este punto de vista, las señales pueden ser periódicas o no periódicas. En una **señal no periódica** no existe ningún patrón que se repita. Por ejemplo, las señales mostradas en la **Figura 2.2** son no periódicas. En cambio, toda **señal periódica**, analógica o digital, satisface la siguiente condición para algún valor mínimo T , llamado **período** de la señal:

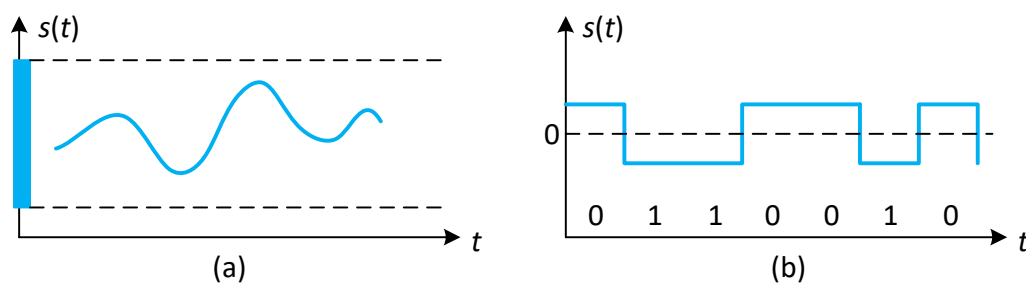


Figura 2.2. Ejemplos de (a) señal analógica y (b) señal digital.

$$s(t + T) = s(t) \quad \forall t \quad (2.1)$$

Lógicamente, si una señal se repite para el valor T (duración del patrón básico), también lo hace para múltiplos enteros de T . Las **Figuras 2.3(a)** y **2.3(b)** muestran, respectivamente, una señal periódica analógica y otra digital. Las señales periódicas son ejemplos de un uso extensivo del término *señal* para denotar también variaciones temporales que no contienen información; en efecto, información es, según vimos en el capítulo anterior, resolución de incertidumbre, pero un patrón repetitivo no resuelve ninguna incertidumbre.

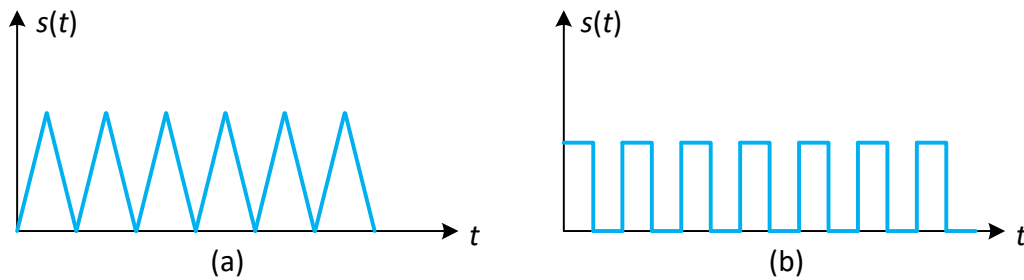


Figura 2.3. Ejemplos de (a) señal periódica analógica y (b) señal periódica digital. Naturalmente sólo se dibujan unos cuantos períodos de cada señal, pues por definición la duración de las señales periódicas es infinita.

2.2. LA SEÑAL SINUSOIDAL

La señal sinusoidal es una señal periódica analógica, y es una de las más importantes en el ámbito de las comunicaciones. Las razones son diversas:

- Es la base del concepto de espectro de frecuencias de cualquier señal.
- Es también la base de técnicas como la modulación y demodulación de señales, utilizadas respectivamente en los transmisores y receptores (**Figuras 1.2 y 1.3**).
- En la actualidad está cobrando importancia el uso de señales electromagnéticas sinusoidales (entre otras) para la transmisión de energía de forma inalámbrica. Se trata de una tecnología puntera (*RF energy harvesting*) con aplicaciones en campos como las redes de sensores, la Internet de las Cosas (IoT - *Internet of Things*) y, en general, las llamadas tecnologías 5G. Observemos que, en este caso, la señal sinusoidal, aunque como señal periódica que es no transporta información, sí transporta energía. Así pues, curiosamente el progreso tecnológico va a validar un uso más ampliado del término *señal*, como variación temporal que puede transportar información y/o energía.

Más adelante trataremos los dos primeros aspectos. Por ahora, es suficiente tener presente que una **señal sinusoidal** se define de la siguiente manera:

$$s(t) = A \cdot \sin(2\pi \cdot f \cdot t + \varphi), \forall t \quad (2.2)$$

De esta definición se desprende que la señal sinusoidal es analógica y, además, periódica, pues existe un intervalo de tiempo para el cual dicha señal se repite. En efecto, combinando las ecuaciones (2.1) y (2.2) se tiene:

$$A \cdot \text{sen}(2\pi \cdot f \cdot (t + T) + \varphi) = A \cdot \text{sen}(2\pi \cdot f \cdot t + \varphi) \Rightarrow 2\pi \cdot f \cdot T = n \cdot 2\pi, n \in \mathbb{Z} \Rightarrow f \cdot T = 1$$

En el último paso se ha tomado $n = 1$ recordando que el período T se define como el valor mínimo del intervalo de tiempo para el cual una señal periódica se repite. Por lo tanto, el parámetro f que aparece en la definición de la señal sinusoidal en (2.2) no es más que la inversa de su período, es decir:

$$f = \frac{1}{T} \quad (2.3)$$

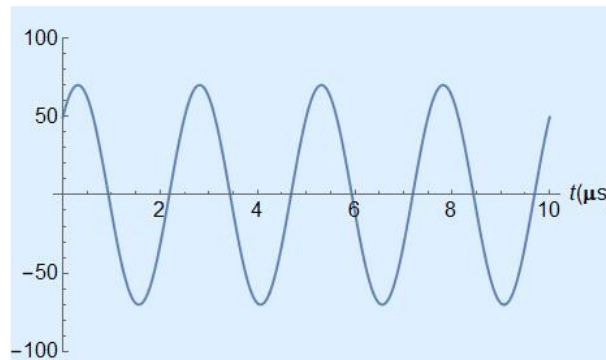


Figura 2.4. Ejemplo de señal sinusoidal con $A = 70$, $f = 400$ kHz y $\varphi = \pi/4$.

Por ello a dicho parámetro se le denomina **frecuencia**, pues no es más que la velocidad con que se repite la señal, es decir, el número de ciclos o patrones básicos que ésta presenta por unidad de tiempo. La frecuencia se mide en Hertz (Hz) o sus múltiplos. El valor de 1 Hz corresponde a 1 ciclo por segundo (1 patrón básico por segundo). A modo de ejemplo, en la **Figura 2.4** se muestra un intervalo de $10 \mu\text{s}$ de una señal sinusoidal cuya frecuencia es de 400 kHz. Su período es, por tanto, de $2.5 \mu\text{s}$, lo que concuerda con las 4 oscilaciones completas durante el intervalo mostrado.

Los otros dos parámetros que caracterizan una señal sinusoidal son la amplitud A y la fase φ . La **amplitud**, también llamada valor de pico, es el valor máximo, es decir, el máximo desplazamiento (en positivo) que presenta la señal con respecto a su valor medio (cero). Tiene las unidades que corresponden al tipo de señal: voltios (o sus múltiplos/submúltiplos) si es una tensión (voltaje), amperios (o sus múltiplos/submúltiplos) si es una corriente, y así sucesivamente. Finalmente, la **fase** indica la posición relativa de la señal sinusoidal con respecto al origen de tiempos. Se puede expresar en grados o radianes. Existen dos convenios para cuantificarla: $[0,360)$ o $[-180,180)$ si se trabaja en grados, o, respectivamente, $[0,2\pi)$ o $[-\pi,\pi)$ si se trabaja en radianes. En estos apuntes se adopta el primer convenio, tanto en grados como en radianes.

Resumiendo, toda señal sinusoidal viene caracterizada por tres parámetros (amplitud, frecuencia y fase) que además son independientes entre sí, ya que se puede modificar cualquiera de ellos sin variar los otros dos. Una muestra de ello es la **Figura 2.5**.

Finalmente, otro parámetro característico de una señal sinusoidal, pero ya no independiente, sino relacionado con la frecuencia, es la **longitud de onda**. En este caso se analiza la señal desde la perspectiva de su propagación. Para ello, consideremos un transmisor y un receptor enlazados directamente mediante algún tipo de medio de transmisión, como se muestra en la **Figura 2.6**. El transmisor, que situamos por conveniencia en el origen ($x = 0$), emite una señal sinusoidal que se propaga a lo largo del medio. Por tanto, el valor de la señal en un punto cualquiera de ese medio corresponderá al valor de la misma en el origen, pero retardada el tiempo que ha requerido para recorrer la distancia que separa ambos puntos. Si la variación temporal en $x = 0$ viene dada por la expresión genérica (2.2), entonces la variación temporal en un punto situado a distancia x corresponderá a un desplazamiento en el tiempo de valor x/v , donde v es la **velocidad de propagación** de la señal en el medio:

$$A \cdot \sin(2\pi \cdot f \cdot (t - x/v) + \varphi) = A \cdot \sin(2\pi \cdot f \cdot t - 2\pi \cdot f \cdot x/v + \varphi) \quad (2.4)$$

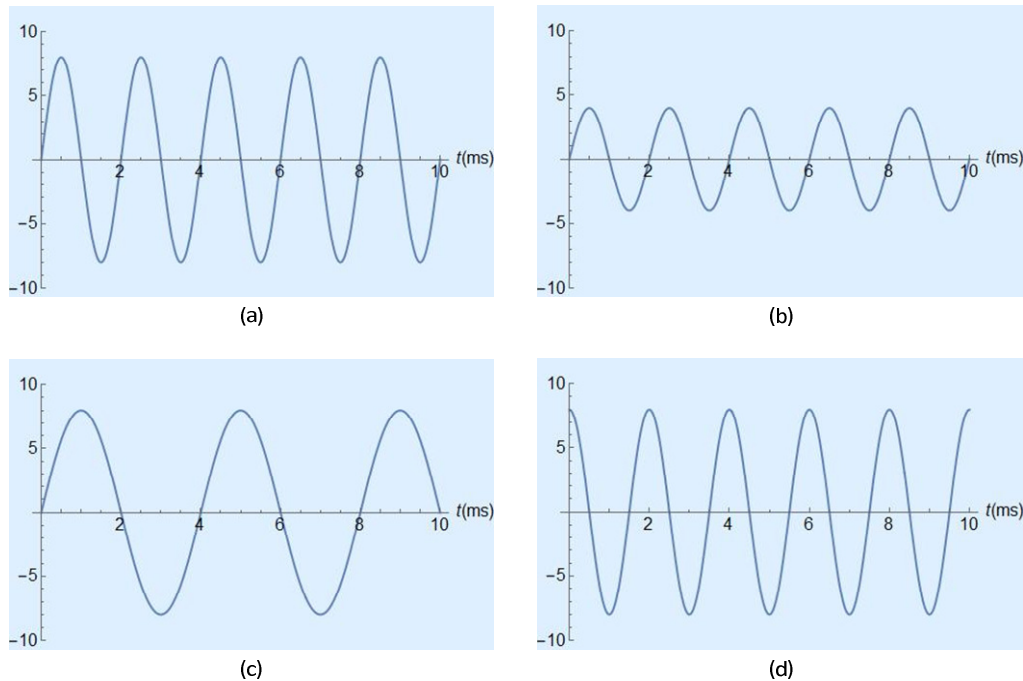


Figura 2.5. Variación de los parámetros de una señal sinusoidal: (a) señal original con $A = 8$, $f = 500$ Hz y $\varphi = 0$, (b) amplitud mitad, (c) frecuencia mitad y (d) fase $\pi/2$.

Puesto que el objetivo está ahora en el concepto de longitud de onda, hemos supuesto por simplicidad que la amplitud de la señal no decrece al alejarse del transmisor, si bien, como veremos más adelante, en realidad esto no es así. La expresión (2.4) define una onda espacio-temporal de tipo sinusoidal, la cual permite conocer el valor de la señal en cualquier punto a distancia x del

transmisor, y en cualquier instante de tiempo t . Si ahora fijamos el instante de tiempo ($t = t_0$) y se toma x como variable, es decir, como si se realizara una fotografía que permitiera ver el valor de la señal a lo largo del medio de transmisión en el instante t_0 , obtenemos una onda sinusoidal en el espacio:

$$A \cdot \text{sen}(2\pi \cdot f \cdot t_0 - 2\pi \cdot f \cdot x/v + \varphi) = A \cdot \text{sen}(2\pi \cdot f/v \cdot x + \varphi') \quad (2.5)$$

En esta última ecuación, el argumento de la función seno se ha manipulado de forma que se obtenga una señal sinusoidal en el formato estándar dado por la expresión (2.2), pero con la variable independiente x en lugar de t . El resultado de esta manipulación (trigonométrica) es una nueva fase φ' , cuya expresión es aquí irrelevante, pero que naturalmente depende del instante de tiempo elegido (t_0), de la fase original (φ) y de la frecuencia (f) (se propone como ejercicio obtener dicha expresión). Como ya hemos señalado, el resultado (2.5) es una señal sinusoidal en el espacio, que también aparece representada en la **Figura 2.6**. Por ser sinusoidal igualmente presenta una periodicidad, aunque en este caso espacial. Al período espacial se le denomina **longitud de onda** (λ), y como ya ocurrió con el período temporal, es el inverso del factor que multiplica a la variable independiente (x en este caso), es decir:

$$\lambda = \frac{v}{f} \quad (2.6)$$

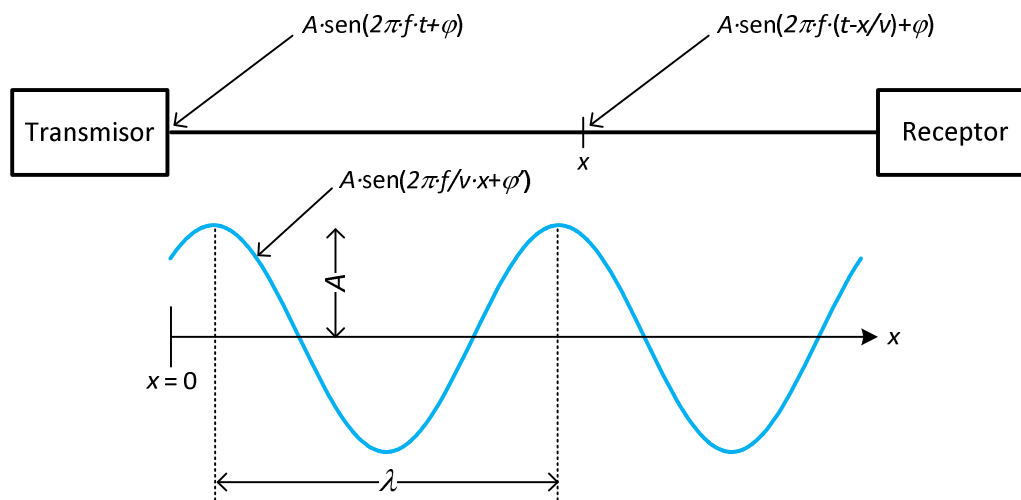


Figura 2.6. Señal sinusoidal espacial y concepto de longitud de onda.

Lógicamente, la longitud de onda tiene unidades de distancia, lo que concuerda con las unidades del numerador y denominador de la expresión (2.6). Una vez conocida la velocidad de propagación, que es una característica del medio de transmisión, tanto la frecuencia como la longitud de onda caracterizan la periodicidad de cualquier señal sinusoidal. Por ejemplo, en el caso de las **Figuras 2.5(a), 2.5(b) o 2.5(d)**, en que la frecuencia es de 500 Hz, si la velocidad de propagación es de 300000 Km/s, la longitud de onda resultante es de 600 Km. Obviamente, al reducir la frecuencia a 250 Hz en la **Figura 2.5(c)**, la longitud de onda se duplica, alcanzando el valor de 1200 Km. Se propone

recalcular estas longitudes de onda para velocidades de propagación de 214000 Km/s y 230000 Km/s.

Un ejemplo típico de la dualidad espacio-temporal que muestran las ondas sinusoidales al propagarse es el del lanzamiento de una piedra en un estanque, si bien en este caso el espacio es bidimensional. Las oscilaciones en cualquier punto a cierta distancia del foco obedecen a una señal sinusoidal en el tiempo (se puede comprobar colocando un tapón de corcho en ese punto). Por otro lado, la onda que se aprecia en todo el estanque en un instante cualquiera obedece a una señal sinusoidal en el espacio (bidimensional). Nótese que la observación de la variación temporal se realiza fijando la coordenada espacial (posición fija), mientras que la variación espacial se observa fijando la coordenada temporal (instante fijo). El primer caso sería equivalente a realizar un vídeo focalizando la cámara en un punto; alternatively, el segundo caso se correspondería con tomar una fotografía de todo el estanque en un instante concreto.

2.3. POTENCIA DE SEÑAL

Como ya hemos indicado, el envío de un mensaje de un punto a otro se lleva a cabo mediante la propagación de una señal. Toda propagación de señal conlleva también un balance de potencias entre el transmisor y el receptor de la misma, puesto que se está emitiendo una energía por unidad de tiempo entre ambos elementos, parte de la cual se pierde por el camino. ¿Cómo se formula, sin embargo, la potencia? Supongamos por ejemplo un transmisor y un receptor conectados directamente mediante un medio metálico (par de hilos trenzados o cable coaxial, según veremos más adelante). Como se observa en la **Figura 2.7**, todo enlace metálico involucra dos conductores a través de los cuales se cierra el circuito que forman el transmisor y el receptor (aunque nunca se dibujan los dos conductores, sino simplemente una línea que une los dos extremos). A efectos del balance de potencias, la conexión transmisor-receptor se puede representar mediante el circuito eléctrico equivalente también mostrado en la **Figura 2.7**, en el que el transmisor se ha substituido por una fuente de voltaje y el receptor por su resistencia (impedancia) equivalente, de valor genérico R . Se trata de un modelo que nos permite evaluar la potencia recibida en el receptor como la potencia disipada en su resistencia equivalente. Concretamente, si $v(t)$ es el voltaje a la entrada del receptor, y por tanto en bornes de la resistencia equivalente, la **potencia instantánea** recibida, $p(t)$, vendrá dada por la expresión $p(t) = v^2(t)/R$, ya conocida de los circuitos eléctricos. También se cumple que $p(t) = i^2(t) \cdot R$, con $i(t)$ la corriente neta instantánea en el circuito del receptor, o corriente que circula por su resistencia equivalente - recuérdese que $i(t) = v(t)/R$ (ley de Ohm). Como vemos en ambas fórmulas, la potencia instantánea es directamente proporcional al cuadrado de la magnitud eléctrica, sea voltaje o corriente.

Si el envío de información se hubiera realizado de forma inalámbrica, mediante antenas en el transmisor y el receptor, la potencia instantánea de la onda electromagnética que alcanzara la antena receptora también hubiera resultado proporcional al cuadrado de la intensidad del campo eléctrico $E(t)$ o del campo magnético $B(t)$, en cada caso con su constante de proporcionalidad específica. Y lo mismo habría sucedido si la transmisión se hubiera realizado mediante pulsos de luz

a través de un enlace de fibra óptica, o mediante ondas acústicas en caso de que la comunicación fuera submarina. Con todo ello queremos decir que, cualquiera que sea el medio de transmisión utilizado, la potencia instantánea en el receptor es siempre proporcional al cuadrado del valor instantáneo de la magnitud física con que se define la señal en dicho medio. A esa constante de proporcionalidad la denotamos genéricamente por k , y es característica del medio de transmisión¹⁰. Por ejemplo, ya hemos visto que cuando la señal es un voltaje, dicha constante es $1/R$. El análisis que sigue a continuación hace uso de esta formulación genérica basada en la constante k .

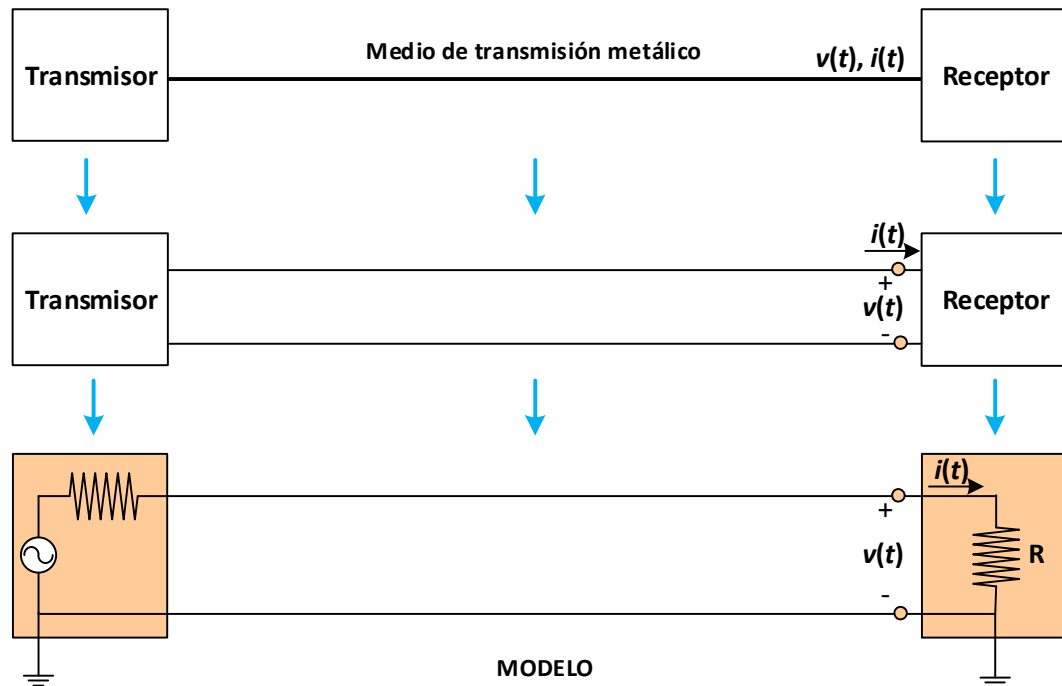


Figura 2.7. Modelo equivalente de un transmisor y un receptor conectados directamente mediante un medio metálico, sólo a efectos del cálculo de potencias (en realidad la señal de voltaje no se disipa directamente en ninguna resistencia, sino que se procesa en el receptor para extraer de ella el mensaje transmitido).

Supongamos una señal $s(t)$ de duración finita τ , es decir, que fuera del intervalo τ el valor de esa señal es nulo, y sea $p(t)$ su potencia instantánea, también nula fuera del intervalo τ . Se define la **potencia media** P de dicha señal en el intervalo τ de la siguiente manera:

$$P = \frac{1}{\tau} \int_{\tau} p(t) \cdot dt = \frac{1}{\tau} \int_{\tau} k \cdot s^2(t) \cdot dt \quad (2.7)$$

¹⁰ En Teoría de Señal se toma $k = 1$ para independizar el análisis y los cálculos de potencia del tipo de señal y del tipo de medio de transmisión utilizado. En todo caso, si fuera necesario, al final se aplicaría la corrección adecuada que tuviera en cuenta la expresión real de la constante de proporcionalidad para el medio de transmisión considerado.

La primera expresión corresponde a la definición de valor medio de cualquier función en un intervalo dado. En el caso de que la duración de la señal sea indefinidamente grande o infinita, su potencia media queda reformulada a través del paso al límite:

$$P = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{\tau} p(t) \cdot dt = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{\tau} k \cdot s^2(t) \cdot dt \quad (2.8)$$

Obsérvese que cuanto mayor sea el nivel de señal en términos absolutos, es decir, sin importar el signo, la potencia media es también mayor, tanto si la duración de la señal es finita como si es infinita. En el caso de señales periódicas, cuya duración es infinita, la potencia media que resulta de aplicar la definición (2.8) coincide con la potencia media calculada en cualquier intervalo de duración igual al período (T):

$$P = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{\tau} p(t) \cdot dt = \frac{1}{T} \int_T p(t) \cdot dt = \frac{1}{T} \int_T k \cdot s^2(t) \cdot dt \quad (2.9)$$

Como hemos visto, el concepto de potencia se ha concretado de dos maneras, potencia instantánea y potencia media. En adelante, salvo que se indique lo contrario, por potencia de una señal entenderemos su potencia media. A continuación, centramos nuestro análisis en dos aspectos implicados en los cálculos de potencia, los efectos de atenuación y amplificación, por un lado, y la definición y uso del decibelio por otro.

Atenuación y amplificación

Cuando una señal se propaga a través de un medio de transmisión, experimenta un debilitamiento progresivo a medida que se aleja del transmisor, fenómeno que se conoce como **atenuación**. Tanto la causa como la caracterización matemática de la atenuación dependen del tipo de medio de transmisión, principalmente de si se trata de un medio guiado o no (distinción que veremos más adelante). En cualquier caso, la atenuación se manifiesta como un factor multiplicativo menor que la unidad sobre la señal de entrada, de modo que, si el receptor está muy alejado, el nivel de señal recibido puede llegar a ser inaceptablemente bajo. Este debilitamiento se compensa, total o parcialmente, mediante la inserción de amplificadores en los extremos de la comunicación (transmisor y receptor) y/o en puntos intermedios de los enlaces. Un **amplificador** no es más que un dispositivo activo (es decir, requiere de una fuente de alimentación) que aumenta el nivel de la señal de entrada. En el caso de un **amplificador ideal**, la relación entre los voltajes de entrada y salida viene dada por la siguiente expresión:

$$v_{out} = \begin{cases} \sqrt{\alpha} \cdot v_{in}, & v_{in} \leq v_{max} \\ v_{sat}, & v_{in} > v_{max} \end{cases} \quad (2.10)$$

Es decir, por debajo de un voltaje v_{max} , el amplificador provoca un efecto multiplicativo superior a la unidad sobre la señal de entrada. Es la llamada **zona lineal**, pues la relación matemática entre el voltaje de salida y el de entrada corresponde en este caso a una función lineal. El factor $\sqrt{\alpha} > 1$ (el uso de la raíz cuadrada lo comprenderemos inmediatamente) es el llamado **factor de amplificación**

de voltaje. Por otro lado, por muy ideal que sea el amplificador, está abastecido por una fuente de alimentación finita, de ahí que a partir de un cierto voltaje de entrada (v_{max}), el voltaje de salida ya no puede seguir aumentando y queda limitado a un valor de saturación v_{sat} . Empieza la llamada **zona de saturación**.

Como la potencia instantánea es proporcional al cuadrado del voltaje, la relación entrada-salida expresada en (2.10) se puede reescribir en términos de potencia de la siguiente manera:

$$p_{out} = \begin{cases} \alpha \cdot p_{in} , p_{in} \leq p_{max} \\ p_{sat} , p_{in} > p_{max} \end{cases} \quad (2.11)$$

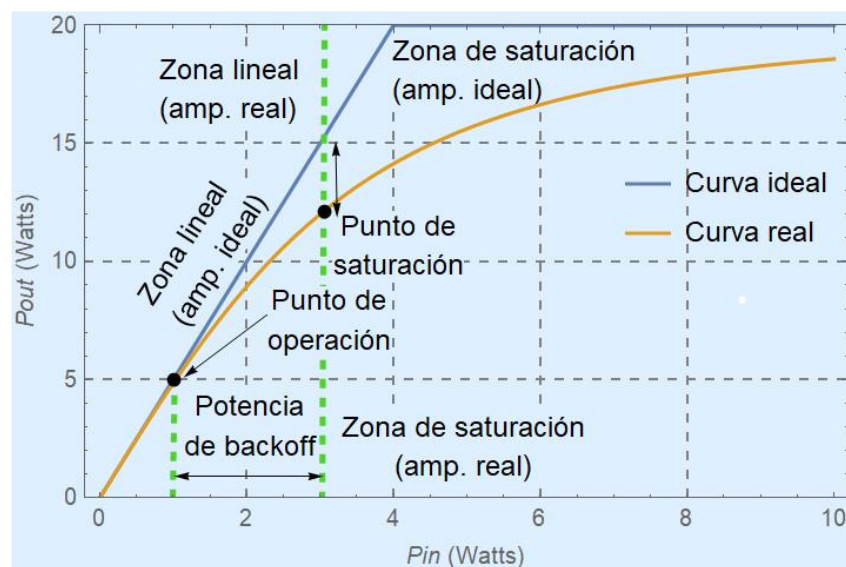


Figura 2.8. Curvas de transferencia de potencia de un amplificador ideal y un amplificador real. Se señalan también las dos zonas, lineal y de saturación, de cada curva. La potencia de *backoff* sería de unos 2 W.

Correspondientemente, el factor $\alpha > 1$ es el llamado **factor de amplificación de potencia** (al haber introducido la raíz cuadrada en (2.10), en (2.11) hemos podido obtener una expresión algo más simple para ese factor, que es el que se utiliza más frecuentemente). La **Figura 2.8** muestra la **curva de transferencia de potencia** del amplificador ideal descrito por la ecuación (2.11), para $\alpha = 5$, $p_{max} = 4$ W y $p_{sat} = 20$ W. A efectos de comparación, también se muestra la que podría ser la curva de transferencia de potencia de un amplificador real que tuviera la misma pendiente inicial y la misma potencia de saturación (en el límite) que el amplificador ideal. Como podemos observar, el comportamiento verdaderamente lineal del amplificador real tiene lugar solamente en una primera zona, donde las dos curvas de transferencia son prácticamente coincidentes. A partir de esa primera zona, la evolución empieza a degradarse, mostrando una transición gradual, no abrupta como en el caso del amplificador ideal, hacia la saturación. Dentro de esta zona de transición, el llamado **punto de saturación** marca el paso de una degradación aceptable, donde la compresión de

la curva real con respecto a la ideal no supera un cierto límite, a una degradación excesiva, y por tanto no permitida. Así pues, la región de trabajo del amplificador real se corresponde con las dos primeras zonas, y permite que la potencia instantánea de la señal de entrada pueda variar entre 0 y la correspondiente al punto de saturación. Dentro de esta región de trabajo, se define el **punto de operación** como la potencia media de entrada recomendada, es decir, la potencia media de la señal a la entrada del amplificador que permite una máxima excursión de su potencia instantánea¹¹; de esta manera, se aprovecha al máximo el amplificador. Si en todos los puntos de la región de trabajo el comportamiento del amplificador real fuera perfectamente lineal, el punto de operación óptimo coincidiría con el punto medio de dicha región; no obstante, como la curva se degrada progresivamente de izquierda a derecha, el punto de operación se rebaja de forma que predominen los valores por debajo sobre los valores por encima. La zona situada entre el punto de operación y el punto de saturación se denomina **zona de backoff**, y la amplitud de la misma es la llamada **potencia de backoff**. Así pues, trabajando en el punto de operación, la máxima subida permitida de potencia instantánea es igual a la potencia de *backoff*.

Es importante resaltar que tanto la atenuación como la amplificación han quedado caracterizadas por un factor multiplicativo, menor que la unidad en el caso de la atenuación, mayor que la unidad en el caso de la amplificación. Esto significa que, tanto en un caso como en el otro, todos los valores instantáneos de la señal entrante quedan multiplicados por el mismo factor, de forma que la señal no se deforma, es decir, no experimenta **distorsión**. Por citar un ejemplo, es como si escuchamos un canal de radio o televisión y reducimos o aumentamos el volumen: la señal acústica tan solo resulta debilitada (atenuación) o reforzada (amplificación), pero no experimenta distorsión alguna: sus características de tono y timbre, esas que nos permiten reconocer el mensaje y quien lo emite, permanecen intactas. Por tanto, en el caso del amplificador, tanto si es ideal como real, la **región de trabajo** apropiada es su zona lineal. Cuando eso ocurre, la relación entre potencias instantáneas de entrada y salida se extiende a las potencias medias, es decir:

$$p_{out} = \alpha \cdot p_{in} \Rightarrow P_{out} = \alpha \cdot P_{in} \quad (2.12)$$

Cuando el nivel de señal entrante es lo suficientemente elevado como para llevar al amplificador a trabajar fuera de su región lineal, la señal de salida resulta distorsionada. A esta distorsión se la conoce como **distorsión no lineal**. Si se conoce la potencia instantánea de pico de una señal, la distorsión no lineal es fácilmente evitable seleccionando el amplificador adecuado.

El decibelio

En los cálculos de potencia es muy habitual hacer uso del **decibelio (dB)**. En realidad, se trata de una unidad de medida de carácter relativo, pues se utiliza para expresar una relación de potencias (instantáneas o medias). Concretamente, si P_1 y P_2 son dos potencias (asumimos, como es habitual,

¹¹ La excursión de potencia instantánea de una señal, es decir, la diferencia entre sus potencias instantáneas máxima y mínima, es lo que se conoce como **margen dinámico** (en potencia) de dicha señal.

que son potencias medias), la relación entre las mismas expresada en decibelios se formula de la siguiente manera:

$$\frac{P_2}{P_1} (\text{dB}) = 10 \cdot \log_{10} \frac{P_2}{P_1} \quad (2.13)$$

Lógicamente, P_1 y P_2 tienen que expresarse en las mismas unidades. Por ejemplo, si $P_1 = 100 \text{ mW}$ y $P_2 = 0.2 \text{ W}$, se tiene $P_2/P_1 = 2$ y se dice que P_2 está 3 dB por encima de P_1 , pues $10 \cdot \log_{10} 2 \cong 3$. Nótese, no obstante, que el resultado en decibelios no implica conocer los valores de P_1 y P_2 por separado, sino solamente la relación entre ellos (por ello decimos que se trata de una medida de carácter relativo). En la **Tabla 2.1** se muestran algunas relaciones típicas expresadas en decibelios.

En el caso particular de que ambas potencias correspondan a la entrada y salida de un amplificador, es decir $P_1 = P_{in}$ y $P_2 = P_{out}$, la expresión (2.13) no es más que su **ganancia** G :

$$G(\text{dB}) = 10 \cdot \log_{10} \frac{P_{out}}{P_{in}} = 10 \cdot \log_{10} \alpha \quad (2.14)$$

Hemos utilizado la expresión (2.12), pues suponemos que el amplificador opera en su región de trabajo (zona lineal). Es decir, la ganancia de un amplificador no es más que su factor de amplificación de potencia expresado en decibelios. Nótese que si $\alpha > 1$, como corresponde a un amplificador, entonces $G > 0$.

Tabla 2.1. Relaciones típicas en decibelios.

P_2/P_1	Valor en dB	P_2/P_1	Valor en dB	P_2/P_1	Valor en dB
1	0	10	10	0.1	-10
2	3	100	20	0.01	-20
4	6	1000	30	0.001	-30
1/2	-3	10^4	40	10^{-4}	-40
1/4	-6	10^n	$n \cdot 10$	10^{-n}	$-n \cdot 10$

Si en lugar de un amplificador consideramos un medio de transmisión, que como ya sabemos produce un debilitamiento en la señal propagada, la relación entre potencias de salida y entrada en decibelios arrojará un resultado negativo. Puesto que resulta más cómodo hablar de atenuación en positivo que de ganancia en negativo, cualquier medio de transmisión se caracteriza mediante una **atenuación** L dada por la siguiente expresión, donde $\beta \geq 1$ es el **factor de atenuación de potencia**:

$$L(\text{dB}) = 10 \cdot \log_{10} \frac{P_{in}}{P_{out}} = 10 \cdot \log_{10} \beta \quad (2.15)$$

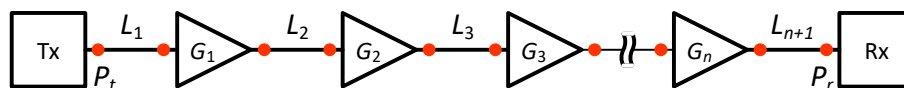


Figura 2.9. Cascada de amplificadores entre un transmisor y un receptor.

Como ya se indicó anteriormente, el enlace directo entre dos dispositivos suele consistir en una alternancia de medios de transmisión y amplificadores, lo que se conoce como **cascada de amplificadores**. Precisamente, estos escenarios sirven para resaltar una de las grandes ventajas que ofrece el decibelio (dB) como unidad de medida relativa. Para comprobarlo, consideremos el enlace genérico mostrado en la **Figura 2.9**, donde un transmisor se conecta a un receptor mediante una cascada de n amplificadores. La expresión de la potencia recibida en función de la potencia transmitida es la siguiente, donde $\alpha_i, i = 1 \dots n$ es el factor de amplificación de potencia del amplificador i -ésimo, y $\beta_i, i = 1 \dots n + 1$ es el factor de atenuación de potencia del tramo i -ésimo de medio de transmisión:

$$P_r = P_t \cdot \frac{1}{\beta_1} \cdot \alpha_1 \cdot \frac{1}{\beta_2} \cdot \alpha_2 \cdot \frac{1}{\beta_3} \cdot \alpha_3 \cdot \dots \cdot \frac{1}{\beta_n} \alpha_n \cdot \frac{1}{\beta_{n+1}} \quad (2.16)$$

Si aplicamos el logaritmo base decimal a ambos miembros de (2.16) y luego los multiplicamos por 10, obtenemos el siguiente resultado:

$$10 \cdot \log_{10} P_r = 10 \cdot \log_{10} P_t - L_1 + G_1 - L_2 + G_2 - L_3 + G_3 \dots - L_n + G_n - L_{n+1} \quad (2.17)$$

Esta expresión es consecuencia de las definiciones (2.14) y (2.15), así como de algunas de las propiedades básicas de la función logaritmo. Lo importante de dicha expresión, comparada con (2.16), es que hemos transformado los productos y cocientes en sumas y restas respectivamente, lo cual hace que la evaluación sea más sencilla. No obstante, un paso definitivo en la simplificación de este tipo de cálculos se da con la introducción de las llamadas **derivaciones absolutas del decibelio**. Éstas consisten en fijar el valor de la potencia P_1 en la expresión (2.13), tomándolo como referencia; de este modo, el cociente P_2/P_1 (dB) se convierte en una evaluación de la potencia P_2 en términos absolutos. Por ejemplo, un valor de referencia habitual es 1 mili-vatio, es decir, $P_1 = 1$ mW, de modo que cualquier potencia $P_2 = P$ se puede expresar como tantos decibelios por encima o por debajo de 1 mili-vatio, lo que se conoce como **dBm** (decibelios con respecto a 1 mW):

$$P(\text{dBm}) = 10 \cdot \log_{10} \frac{P(\text{mW})}{1 \text{ mW}} = 10 \cdot \log_{10} P(\text{mW}) \quad (2.18)$$

Así pues, si la potencia de una señal es de 3 dBm, significa que dicha potencia está 3 dB por encima de 1 mW, o sea, que se trata de una señal de 2 mW (recordemos que una subida de 3 dB equivale a multiplicar por 2 según la **Tabla 2.1**). Análogamente, si la potencia de una señal es de -3 dBm, significa que dicha potencia está 3 dB por debajo de 1 mW, lo que corresponde a 0.5 mW.

Otra derivación absoluta del decibelio consiste en tomar como referencia la potencia de 1 W (útil cuando se trabaja con potencias más elevadas). En este caso hablamos de **dBW** (decibelios con respecto a 1 W):

$$P(\text{dBW}) = 10 \cdot \log_{10} \frac{P(\text{W})}{1 \text{ W}} = 10 \cdot \log_{10} P(\text{W}) \quad (2.19)$$

No resulta difícil determinar la relación entre ambas medidas:

$$P(\text{dBm}) = 10 \cdot \log_{10} P(\text{mW}) = 10 \cdot \log_{10}(P(\text{W}) \cdot 1000) = P(\text{dBW}) + 30 \quad (2.20)$$

Con la ayuda de estas derivaciones absolutas del decibelio, el uso de sumas y restas se extiende a las potencias de entrada y salida de un circuito en cascada. Así, la expresión (2.17) queda reescrita de la siguiente manera cuando tomamos como referencia la potencia de 1 mW:

$$P_r(\text{dBm}) = P_t(\text{dBm}) - L_1 + G_1 - L_2 + G_2 - L_3 + G_3 \dots - L_n + G_n - L_{n+1} \quad (2.21)$$

Y si la referencia es 1 W:

$$P_r(\text{dBW}) = P_t(\text{dBW}) - L_1 + G_1 - L_2 + G_2 - L_3 + G_3 \dots - L_n + G_n - L_{n+1} \quad (2.22)$$

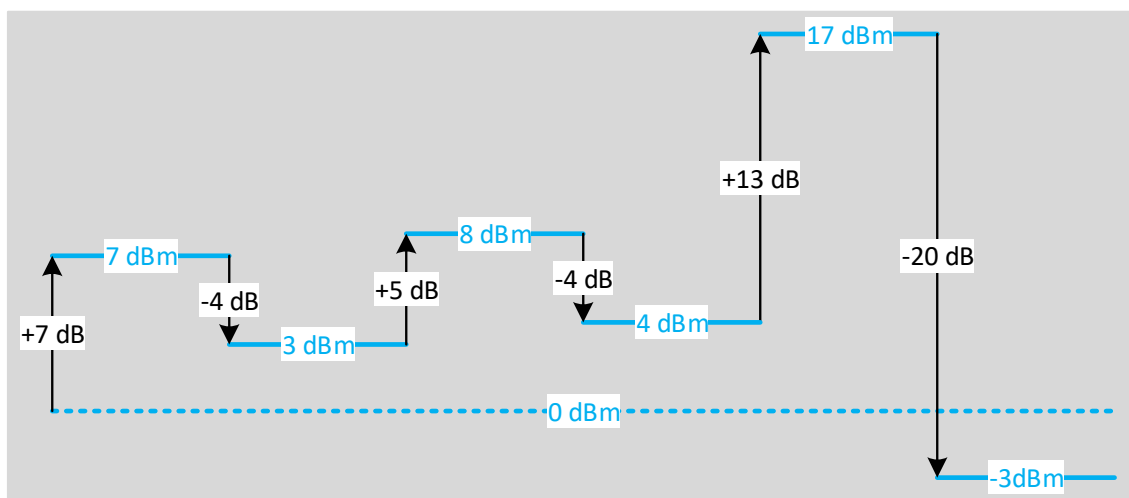


Figura 2.10. Representación gráfica de los cálculos de potencias basados en el decibelio y sus variantes absolutas. El diagrama se lee de izquierda a derecha. Al final todo queda en sumas y restas de decibelios, pues los dBm, al igual que los dBW, no son más que decibelios con respecto a un nivel de referencia, pero en definitiva decibelios.

La aplicación de las fórmulas (2.21) o (2.22) admite una representación gráfica muy sencilla, que ayuda a manejar los decibelios y sus variantes absolutas. Un ejemplo de dicha representación se muestra en la **Figura 2.10**, donde podemos apreciar que consiste en una especie de escalera, con peldaños que suben y bajan. Los peldaños representan los cambios de nivel y se miden en dB, positivos si corresponden a subidas (ganancias), negativos si corresponden a bajadas (atenuaciones). Los rellanos, en cambio, corresponden a niveles absolutos en dBm, pues quedan definidos como tantos decibelios por encima o por debajo de la potencia de referencia, que en este caso es 1 mW, es decir, 0 dBm (línea horizontal discontinua). Si este diagrama correspondiera a una cascada de amplificadores, la potencia transmitida sería de 7 dBm y la recibida sería de -3 dBm. Si la línea de referencia hubiera correspondido a 0 dBW (potencia de referencia igual a 1 W), entonces, para los mismos cambios de nivel, la potencia transmitida sería de 7 dBW y la recibida de -3 dBW. Como vemos, partiendo de un nivel absoluto (expresado en dBm o dBW), podemos aplicarle tantas

subidas o bajadas como queramos, para llegar finalmente a otro nivel absoluto (expresado respectivamente en dBm o dBW). En términos más coloquiales, en los cálculos de potencia siempre partimos de un rellano para llegar a otro rellano, tras atravesar una serie de peldaños.

El valor de la potencia recibida en el ejemplo de la **Figura 2.10** merece dos comentarios adicionales:

- En primer lugar, la constatación de que, al trabajar con derivaciones absolutas del decibelio, podemos obtener resultados negativos (−3 dBm en el ejemplo), que no hacen más que indicar que el valor de potencia alcanzado queda tantos decibelios por debajo del nivel de referencia.
- La segunda constatación es que la potencia recibida puede ser inferior a la transmitida (−3 dBm frente a 7 dBm). En efecto, el objetivo de la conexión no es que la potencia recibida sea igual o superior a la transmitida, sino que la potencia recibida sea igual o superior a la **sensibilidad** (S) del receptor, que es el mínimo valor de potencia de señal que éste puede detectar (definición extensible a cualquier dispositivo que procesa señales); de otro modo, la comunicación no es factible.

Para terminar esta sección dedicada al decibelio, escribamos las fórmulas que nos permiten obtener las potencias en mili-watios o watios a partir de sus valores en dBm o dBW respectivamente. Basta con invertir la ecuación (2.18) en el primer caso:

$$P(\text{mW}) = 10^{\frac{P(\text{dBm})}{10}} \quad (2.23)$$

Y la ecuación (2.19) en el segundo caso:

$$P(\text{W}) = 10^{\frac{P(\text{dBW})}{10}} \quad (2.24)$$

2.4. ESPECTRO DE UNA SEÑAL

La **Figura 2.11** muestra un experimento muy conocido, el que consiste en descomponer la luz blanca en sus múltiples colores a través de un prisma. Puesto que cada color corresponde a una onda electromagnética sinusoidal de una determinada frecuencia, este experimento evidencia que la luz blanca es, en realidad, una onda electromagnética compleja formada por la agregación de esas ondas electromagnéticas más simples. A esta descomposición de la luz blanca en las frecuencias que el ojo humano puede percibir se la conoce como **espectro visible**¹².

Profundizando un poco más, observamos que la **Figura 2.11** ofrece una visión muy simplificada del contenido espectral de la luz blanca, puesto que ésta aparece descompuesta en un conjunto discreto de colores, separados nítidamente entre sí. Si bien dicha figura es aceptable como primera

¹² La luz blanca proveniente del Sol contiene también otras radiaciones electromagnéticas en su espectro completo, pero no son percibidas por el ojo humano y no forman parte, por tanto, del espectro visible: son las radiaciones infrarroja y ultravioleta. En el análisis que realizamos aquí, ignoramos la presencia de estas radiaciones.

aproximación, la **Figura 2.12** responde a un enfoque más científico y realista. En ella podemos observar que el espectro visible está realmente formado por un conjunto continuo de colores, dejándose de apreciar una separación nítida entre los mismos.

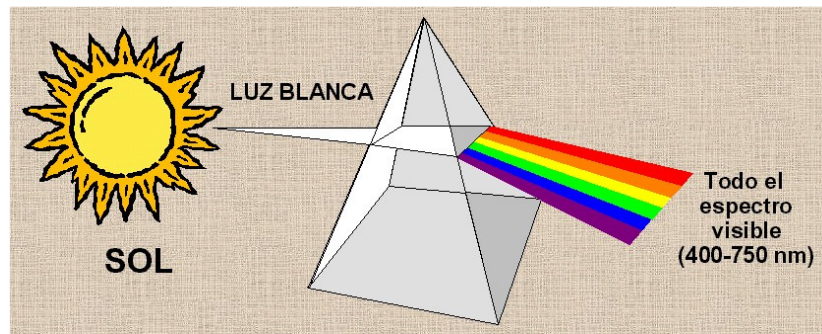


Figura 2.11. Descomposición discreta de la luz blanca a través de un prisma.

Se puede demostrar que la potencia de una señal que es suma de señales sinusoidales de frecuencias distintas es igual a la suma de las potencias de cada señal sinusoidal por separado. Entonces, si admitimos el modelo simplificado de la **Figura 2.11**, la potencia de la luz blanca puede formularse como la suma de las potencias de cada color (recordemos que cada color corresponde a una señal sinusoidal de una determinada frecuencia). Es decir:

$$P_{\text{luz blanca}} = P_{\text{rojo}} + P_{\text{naranja}} + P_{\text{amarillo}} + P_{\text{verde}} + P_{\text{cian}} + P_{\text{azul}} + P_{\text{violeta}} \quad (2.25)$$

Si asociamos cada color con la frecuencia f_i que le corresponde, donde $i = 1 \dots 7$, la fórmula anterior queda reescrita del siguiente modo:

$$P_{\text{luz blanca}} = P_{f_1} + P_{f_2} + P_{f_3} + P_{f_4} + P_{f_5} + P_{f_6} + P_{f_7} \quad (2.26)$$

Alternativamente, si consideramos el modelo real de la **Figura 2.12**, entonces el cálculo de la potencia media de la luz blanca ya no puede formularse como la suma de un conjunto discreto y finito de contribuciones. En este caso necesitamos introducir una función, la llamada **densidad espectral de potencia** o **espectro de potencia**, que refleja cómo se distribuye la potencia de la señal agregada (luz blanca en este caso) a lo largo del conjunto continuo de frecuencias que la componen. Si $S_{lb}(f)$ denota esta función para el caso de la luz blanca, entonces la potencia media de la misma viene dada por la siguiente expresión, donde $f_{lb,min}$ y $f_{lb,max}$ son, respectivamente, las frecuencias más baja y más alta de su espectro de potencia:

$$P_{\text{luz blanca}} = \int_{f_{lb,min}}^{f_{lb,max}} S_{lb}(f) \cdot df \quad (2.27)$$

Como cabía esperar, la suma (2.26) se ha transformado en una integral al pasar de lo discreto a lo continuo.

Aunque la luz blanca no es una señal en el sentido estricto del término, pues se trata de una variación temporal (onda electromagnética) que no es portadora de ningún mensaje, ha resultado útil para

introducir el concepto de espectro de potencia. Y es que este concepto resulta aplicable a cualquier variación temporal, tanto si contiene información y energía, como si solo contiene energía. Específicamente, el **espectro de potencia** es la descomposición de una señal cualquiera (en el sentido amplio del término) en múltiples señales sinusoidales, cada una a una determinada frecuencia. Formalmente, si $S(f)$ denota el espectro de potencia de una señal, y se supone que dicho espectro ocupa una banda de frecuencias entre f_{min} y f_{max} , frecuencias mínima y máxima respectivamente, entonces la potencia media de dicha señal, P , queda expresada de la siguiente manera:

$$P = \int_{f_{min}}^{f_{max}} S(f) \cdot df \quad (2.28)$$

Lógicamente, la densidad espectral se mide en W/Hz (o cualquiera de sus múltiplos/submúltiplos).

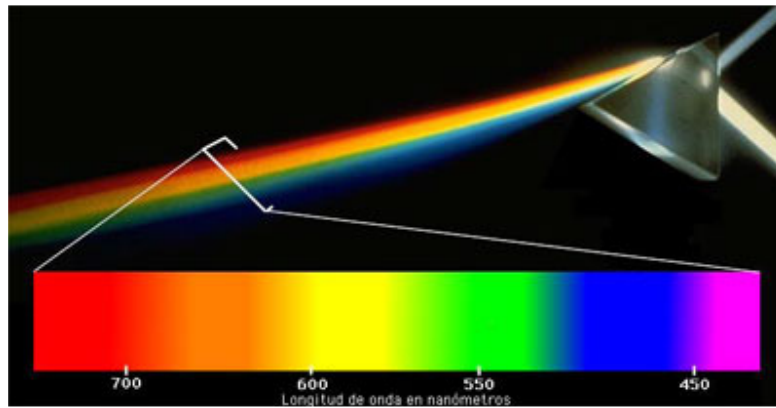


Figura 2.12. Análisis de espectrómetro de la luz blanca. Se observa una transición gradual, no nítida, entre los colores. Corrobora que el espectro de la luz blanca es, en realidad, continuo.

Desde el punto de vista de la ubicación del espectro de potencia, se distinguen dos tipos de señales, tal como indica la **Figura 2.13: señales paso-bajo o en banda base**, aquellas cuyo contenido espectral se concentra en la parte baja del eje de frecuencias (desde $f_{min} \cong 0$ hasta una f_{max} no demasiado elevada), y **señales paso-banda**, cuyo espectro ocupa una franja intermedia del eje de frecuencias, entre dos frecuencias f_{min} y f_{max} que pueden ser arbitrariamente elevadas. Finalmente, otra magnitud relevante es el **ancho de banda absoluto**¹³ de una señal, B_{abs} , que no es más que la anchura de su espectro de potencia:

$$B_{abs} = f_{max} - f_{min} \quad (2.29)$$

Obviamente, para las señales en banda base se tiene $B_{abs} = f_{max}$, puesto que en ellas $f_{min} \cong 0$.

¹³ En contraposición a otro tipo de ancho de banda, el ancho de banda efectivo, que veremos más adelante.

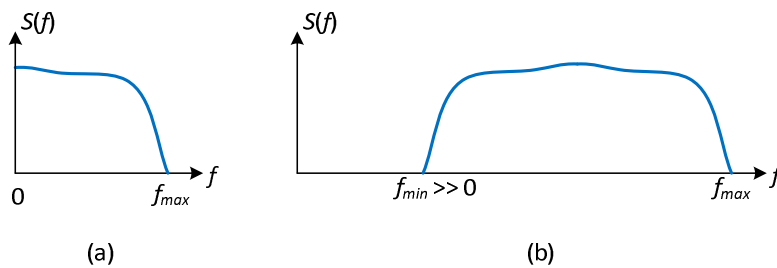


Figura 2.13. Clasificación de las señales según la ubicación de su espectro: (a) señales paso-bajo o en banda base y (b) señales paso-banda.

Las ecuaciones (2.7) o (2.8) permiten calcular la potencia media de una señal a partir de su expresión en el dominio del tiempo, es decir, $s(t)$. De la misma manera, la ecuación (2.28) permite calcular la potencia media de una señal a partir de su expresión en el dominio de la frecuencia, o sea, $S(f)$. De hecho, la expresión en el dominio de frecuencias $S(f)$ resulta tan identificativa de una señal como lo es su expresión en el dominio del tiempo $s(t)$. Ambas son caracterizaciones inequívocas de la misma señal, aunque desde dos perspectivas distintas. Dependiendo del tipo de análisis a realizar, una perspectiva puede resultar más útil que la otra. En particular, el dominio de frecuencias facilita la comprensión y el análisis de los efectos de algunos procesos y dispositivos por los que pasa una señal desde que se transmite hasta que se recibe, y en la misma medida resulta también útil en el diseño de tales procesos y dispositivos.

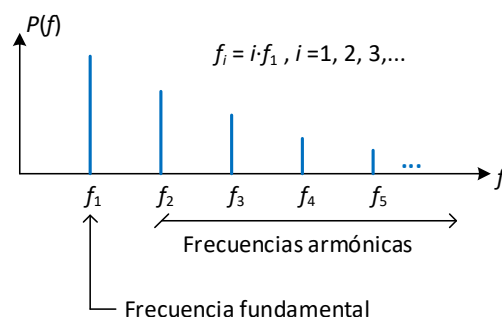


Figura 2.14. Espectro de potencia de una señal periódica. En particular, las barras correspondientes a algunos armónicos (incluyendo el fundamental) pueden ser nulas. La frecuencia fundamental siempre corresponde al máximo común divisor de las frecuencias presentes en la señal (barras no nulas).

La relación entre el dominio temporal y el dominio de frecuencias de una señal se formaliza matemáticamente a través del **análisis de Fourier**, cuyo detalle queda fuera de nuestros objetivos. Tan solo señalar que, en el caso de señales periódicas, el análisis de Fourier establece que éstas se descomponen en un conjunto discreto, finito o infinito, de señales sinusoidales, una fundamental (**primer armónico** o **armónico fundamental**) cuya frecuencia (**frecuencia fundamental**) es igual a la inversa del período de la señal periódica, y las restantes, llamadas **armónicos**, cuyas frecuencias son múltiplos enteros de la frecuencia fundamental (**frecuencias armónicas**). La **Figura 2.14** ilustra el

espectro de potencias de una señal periódica de período $T = 1/f_1$. Como podemos observar, este espectro es una función discreta $P(f)$; la afirmación contraria también es cierta, es decir, si el espectro de potencias de una señal es discreto, entonces la señal es periódica. Para terminar, la potencia media P de una señal periódica cuya frecuencia fundamental es f_1 , viene de nuevo expresada mediante un sumatorio, en consonancia con el carácter discreto de su espectro:

$$P = \sum_{i=1}^{\infty} P(f_i) = \sum_{i=1}^{\infty} P(i \cdot f_1) \quad (2.30)$$

En esta expresión, $P(f_i)$ es la potencia de la señal periódica contenida en el armónico $f_i = i \cdot f_1$.

CAPÍTULO 3. TRANSMISIÓN DIGITAL

En los Capítulos 1 y 2 estudiamos, respectivamente, la información y las señales, estas últimas por su papel como portadoras de la información a través de los medios de transmisión. En este capítulo profundizamos en los mecanismos de imbricación de los mensajes digitales en las señales, de forma que tales mensajes puedan ser enviados a través de una red de computadores. Nos lleva a ocuparnos, en definitiva, de la **transmisión digital**. Trataremos los siguientes aspectos:

- La distinción entre fuente analógica y digital, y entre transmisión analógica y digital.
- Los dos tipos de transmisión digital, conocidos como transmisión digital en banda base y transmisión digital modulada.

El proceso de conversión analógica-digital, que no es más que la digitalización de una fuente analógica.

ÍNDICE

3.1.	FUENTES ANALÓGICAS Y DIGITALES, TRANSMISIÓN ANALÓGICA Y DIGITAL	61
3.2.	TRANSMISIÓN DIGITAL EN BANDA BASE	63
3.3.	TRANSMISIÓN DIGITAL MODULADA	73
3.4.	TRANSCEPTORES	77
3.5.	CONVERSIÓN ANALÓGICA-DIGITAL	78

3.1. FUENTES ANALÓGICAS Y DIGITALES, TRANSMISIÓN ANALÓGICA Y DIGITAL

En la Sección 2.1 establecimos la distinción entre señales analógicas y digitales. En base a estos tipos de señales, podemos clasificar la fuente de información del modelo de comunicaciones de la **Figura 1.3** del siguiente modo:

- **Fuente analógica:** es aquella que produce una señal analógica (Sección 2.1). Un micrófono, una cámara o un sensor de temperatura son algunos ejemplos (obsérvese que en tales ejemplos estamos identificando la fuente con lo que realmente es el transductor en origen, lo cual es práctica habitual, aunque no se corresponda con la definición formal establecida en el modelo de comunicaciones).
- **Fuente digital:** es aquella que produce una señal digital (Sección 2.1). Un computador o cualquier dispositivo de procesamiento y almacenamiento de información binaria actúan como fuentes digitales cuando tienen que transmitir. Esto incluye también la fuente digital equivalente que resulta de un proceso de **codificación de fuente analógica**, según el cual la señal generada por una fuente analógica es transformada en señal digital. Veremos más adelante en qué consiste este proceso, también llamado **conversión analógica-digital**.

Por otro lado, la señal que se transmite por el enlace también puede ser analógica o digital. Parecería lógico afirmar que la señal transmitida por el enlace es analógica cuando la fuente es analógica, y es digital cuando la fuente es digital; no obstante, la casuística no es tan simple. Entonces, ¿qué entendemos por transmisión analógica y digital? La clasificación se establece así:

- **Transmisión analógica:** tiene lugar cuando la fuente es analógica y no se codifica, es decir, no pasa por un proceso de conversión analógica-digital. En este caso, la señal que se transmite por el enlace es también analógica. La **Figura 3.1** representa el escenario de una transmisión de este tipo.
- **Transmisión digital:** tiene lugar cuando la fuente es digital, originalmente o tras un proceso de conversión analógica-digital. La señal que finalmente se transmite por el enlace puede ser, a su vez, analógica o digital. La **Figura 3.2** muestra todos los escenarios posibles. En cualquier caso, un parámetro fundamental en toda transmisión digital es la **velocidad de transmisión R** , cuya expresión en términos de T_b , el tiempo invertido en la transmisión de 1 bit (**tiempo de bit**), es la siguiente:

$$R = \frac{1}{T_b} \quad (3.1)$$

La velocidad de transmisión se mide en bits por segundo (bps) o sus múltiplos¹⁴. La **Tabla 3.1** muestra algunas unidades típicas para la velocidad de transmisión, y su relación con el tiempo de bit. Es importante observar que, cuando se trata de velocidad, los términos kilo, mega, giga,

¹⁴ Propiamente hablando, la velocidad de transmisión debería medirse en dígitos binarios por segundo; sin embargo, en la práctica se asume que la codificación de fuente es óptima o cercana a la óptima, por lo que se pueden identificar dígitos binarios con bits de información. Aceptaremos esta hipótesis de ahora en adelante.

etc. se refieren a potencias de 10, a diferencia de lo que ocurre con esos mismos términos cuando se refieren a cantidad de información o espacio de memoria (en bits o bytes), en cuyo caso representan potencias de 2 (1 K = 2^{10} = 1024, 1 M = 2^{20} = 1024·1024, 1 G = 2^{30} , 1 T = 2^{40} , etc.).

Tabla 3.1. Valores de referencia para el tiempo de bit y la velocidad de transmisión.

Tiempo de bit	Velocidad de transmisión
1 s	1 bps
1 ms = 10^{-3} s	1 Kbps = 10^3 bps
1 μ s = 10^{-6} s	1 Mbps = 10^6 bps
1 ns = 10^{-9} s	1 Gbps = 10^9 bps
1 ps = 10^{-12} s	1 Tbps = 10^{12} bps

La transmisión digital es la que tiene lugar en las redes de computadores, y por tanto a partir de este momento nos centramos exclusivamente en ella. Como indica la **Figura 3.2**, independientemente de cuál sea el origen de la señal digital a la entrada del transmisor, este tipo de transmisión puede ser en banda base o modulada dependiendo de cómo es la señal a la salida de dicho dispositivo. Según veremos, esta distinción es relevante para el tipo de medio de transmisión utilizado en el enlace, y además determina el diseño del transmisor (y por ende del receptor). Antes de tratar con detalle estas dos modalidades de transmisión digital, justifiquemos la razón de su denominación:

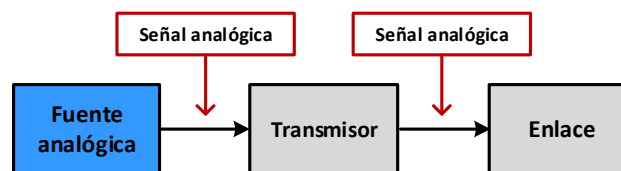


Figura 3.1. Transmisión analógica.

- Toda señal digital presenta un espectro promedio que ocupa la parte baja del eje de frecuencias, es decir, entre una frecuencia mínima cercana a los 0 Hz ($f_{min} \cong 0$) y una frecuencia máxima f_{max} , cuyo valor depende del tipo de señal digital y la velocidad de transmisión. A esta zona espectral se la conoce como **banda base**, y por ello el tipo de transmisión digital en el que la señal a la salida del transmisor es también digital se denomina **transmisión digital en banda base** (**Figura 3.2**).
- A diferencia de lo que ocurre con toda señal digital, el espectro de una señal analógica puede quedar ubicado en cualquier parte del eje de frecuencias. Por ejemplo, la señal generada por una fuente analógica (incluyendo el posible transductor en origen) es típicamente una señal en banda base, pues su espectro se sitúa entre una frecuencia mínima cercana a los 0 Hz y una frecuencia máxima cuyo valor podrá ser mayor o menor dependiendo de cada caso. Uno de estos casos es la señal musical, cuyo espectro promedio abarca la banda de frecuencias audibles

por el oído humano medio (dentro del segmento de población de personas jóvenes y sanas), es decir, desde 20 Hz (prácticamente 0 Hz) hasta 20 KHz. En cambio, la señal analógica a la salida del transmisor en la rama superior de la **Figura 3.2**, es el resultado de un proceso de modulación, que como veremos consiste básicamente en trasladar el espectro de la banda base a una banda de frecuencias superior. Esto resulta interesante por razones que justificaremos más adelante. Hablamos pues, en este caso, de **transmisión digital modulada** (el transmisor es un modulador).

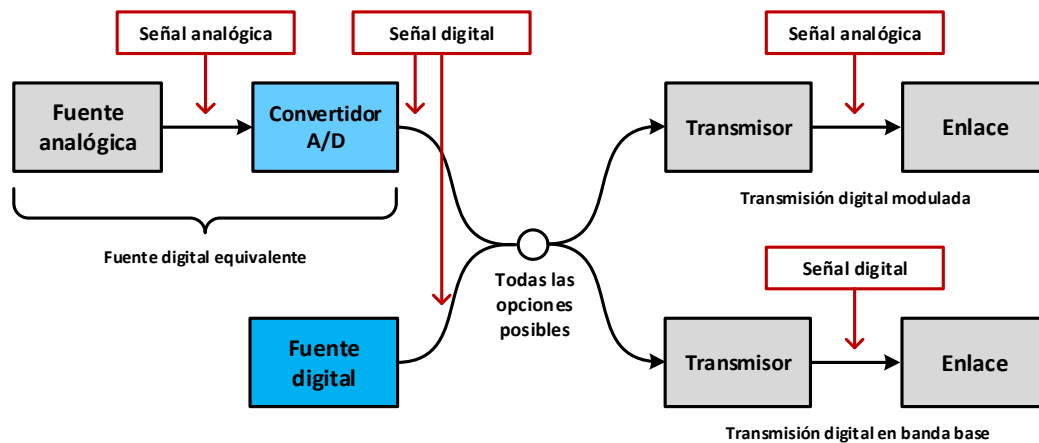


Figura 3.2. Escenarios posibles en transmisión digital. Son los que pueden tener lugar en las redes de computadores, pues cualquier dispositivo que genere información digital o digitalizada puede conectarse a las mismas.

3.2. TRANSMISIÓN DIGITAL EN BANDA BASE

La transmisión digital en banda base exige que el medio de transmisión presente una buena respuesta en la parte baja del eje de frecuencias¹⁵. Si esto se cumple, y dado que la fuente genera una señal digital, ¿por qué utilizar entonces un transmisor, es decir, por qué no transmitir directamente la señal generada por la fuente?

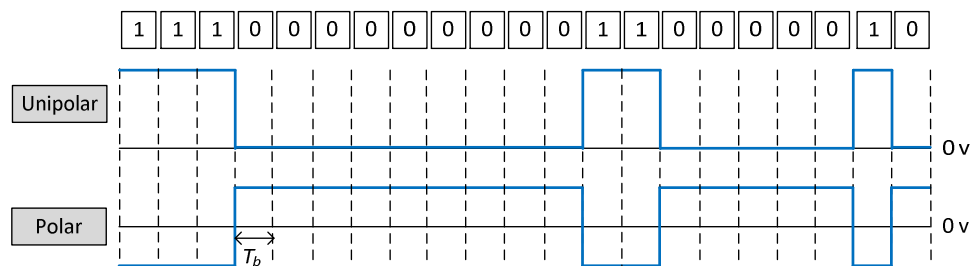


Figura 3.3. La señal digital NRZ, en sus dos variantes.

¹⁵ El significado de “buena respuesta” lo veremos al tratar los medios de transmisión.

El hecho es que las fuentes digitales generan típicamente una señal muy sencilla, llamada NRZ (*Non Return to Zero*). La **Figura 3.3** muestra esta señal en sus dos variantes, unipolar y polar, para un mensaje concreto (la definición formal de ambas variantes se introduce más adelante). Sin embargo, la señal NRZ es solamente una opción dentro de la vasta cantidad de alternativas consistentes en representar cada bit mediante una forma de onda temporal. Además, muchas de estas alternativas ofrecen mejores prestaciones que la señal NRZ como portadoras de mensajes digitales a través del enlace de transmisión. La **codificación de línea** se ocupa precisamente de transformar la señal digital NRZ en otra señal digital con mejores prestaciones para ser transmitida por el enlace. Esa sería pues, la función del transmisor en una transmisión digital en banda base. Su inserción en el modelo de comunicaciones se muestra en la **Figura 3.4**.

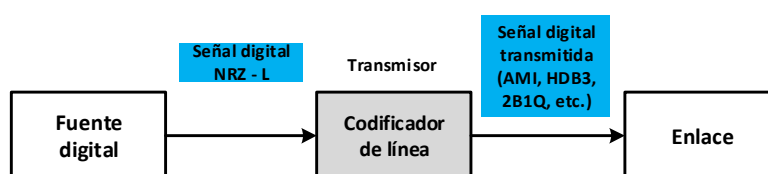


Figura 3.4. Modelo de comunicaciones para una transmisión digital en banda base (lado origen).

Veamos a continuación cuáles son las prestaciones exigidas a un código de línea.

Requisitos de los códigos de línea

Para una transmisión digital en banda base de calidad, es necesario que la codificación de línea genere una señal digital que satisfaga ciertas prestaciones en los siguientes aspectos:

- **Sincronismo.** Dado que los relojes internos del transmisor y el receptor no están sincronizados, es necesario introducir algún mecanismo para garantizar que el receptor muestrea una sola vez todos y cada uno de los símbolos¹⁶ enviados por el transmisor. De lo contrario, la deriva relativa de un reloj con respecto al otro provocaría errores en la secuenciación de tales símbolos en el receptor. Una solución podría ser introducir un enlace de sincronismo entre transmisor y receptor, diferenciado del enlace de datos, que conectara sus relojes y los mantuviera sincronizados; sin embargo, esta solución sería muy cara. La alternativa adoptada en la práctica, mucho más económica, consiste en imbricar el sincronismo en la propia señal digital que transporta los datos, de forma que el enlace de sincronismo pasa a coincidir con el enlace de datos. Esto se consigue forzando frecuentes cambios del nivel de voltaje en la señal de datos, típicamente transiciones a voltaje 0 o pasando por voltaje 0, ya que tales transiciones son fácilmente detectables por el receptor, y por ello utilizables como puntos de anclaje para corregir continuamente la deriva relativa entre relojes y evitar que su desincronización conduzca a errores. Además, es importante que esto pueda garantizarse independientemente de la secuencia de bits transmitidos, lo que se conoce como condición de **transparencia**.

¹⁶ En este contexto, un símbolo equivale a un bit o a una agrupación de bits, puesto que hay códigos de línea que asignan formas de onda elementales a grupos de varios bits. Si un símbolo corresponde a 1 bit, entonces el sincronismo se concreta en el **sincronismo de bit**.

- **Perfil espectral.** En transmisión digital en banda base, un perfil espectral adecuado se refiere a la ausencia de componente continua y de contenido en baja frecuencia. La componente continua (en inglés DC – *direct current*) se corresponde con el promedio a largo plazo de la señal digital, de modo que ésta puede verse como el resultado de superponer (sumar) un nivel constante coincidente con ese valor promedio, y una componente variable de valor medio 0 (en inglés AC – *alternating current*). Dado que el nivel constante consume potencia del transmisor, pero no contribuye a transmitir información, resulta mucho más eficiente generar una señal digital cuyo promedio sea nulo a largo plazo. Con el contenido en baja frecuencia sucede algo parecido: la transmisión de información a una cierta velocidad reside en las variaciones rápidas del nivel de señal, las cuales quedan recogidas en las componentes espectrales de mayor frecuencia. Por tanto, también resulta más eficiente generar una señal digital cuyo contenido espectral en baja frecuencia sea lo más reducido posible, de modo que la potencia de la misma se concentre en la parte más significativa de su espectro. Dicho de otra manera, no solamente interesa que el valor medio de la señal sea nulo a largo plazo, sino que también sea nulo o casi nulo sobre plazos más cortos, es decir, calculado sobre ventanas temporales más reducidas, y cuanto más reducidas mejor. Por otro lado, existen razones de carácter técnico, relacionadas con el uso de acoplamientos capacitivos y transformadores en las transmisiones en banda base, que también aconsejan satisfacer tales requisitos.
- **Métricas de comportamiento.** El término *comportamiento* proviene del inglés *performance*, muy utilizado en el campo de las tecnologías de la información y las comunicaciones. Siempre se refiere a la eficiencia con que un dispositivo, sistema, proceso o algoritmo hace uso de los recursos disponibles, dando por supuesto que ejecuta correctamente la función para la cual fue diseñado. Por tanto, se trata de un concepto que alude más al “cómo” se ejecuta una función (aspecto no funcional), que al “qué” función se ejecuta y si se ejecuta correctamente (aspecto funcional). A modo de ejemplo, un navegador web puede ejecutar correctamente su función, que consiste en descargar las páginas web solicitadas, pero según como esté diseñado, el tiempo de descarga (aspecto no funcional) puede ser mayor o menor, haciendo que la experiencia del usuario sea peor o mejor, respectivamente. No existe una magnitud que cuantifique directamente el grado de “buen” comportamiento, pero en cada caso disponemos de una serie de métricas que conjuntamente reflejan el nivel de prestaciones de un dispositivo o proceso. En el caso de los códigos de línea, las métricas principales son la eficiencia espectral y la tasa de errores de bit. Ambas métricas y los valores que pueden alcanzar se consideran en profundidad más adelante. Por el momento, tan solo una breve descripción de cada una:
 - La **eficiencia espectral** es la relación entre velocidad de transmisión y ancho de banda de la señal digital transmitida, e interesa que dicha relación, medida en bps/Hz, sea lo mayor posible; el motivo, como veremos más adelante, es que el ancho de banda de los medios de transmisión es un recurso escaso, y por tanto interesa consumir el mínimo posible para una determinada velocidad de transmisión)
 - La **tasa de errores de bit** o BER (*Bit Error Rate*) es la probabilidad de que el valor asignado a un bit por el receptor sea opuesto a su valor verdadero, y naturalmente interesa que dicha probabilidad sea lo menor posible.

- **Complejidad.** Cuanto menos compleja sea la codificación de línea, menos complejos son los equipos de transmisión y recepción que deben procesarla y, por tanto, más baratos.

Caracterización de los códigos de línea

Un **código de línea** es una regla de transformación que permite representar las secuencias de símbolos mediante señales digitales. Estas señales se construyen en base a unas formas de onda o patrones básicos predefinidos, en definitiva, elementos o símbolos de un alfabeto (alfabeto de señales elementales, en este caso) como si de una codificación de fuente se tratara. Consideraremos señales de tipo eléctrico, expresadas como voltajes, lo cual es coherente con el hecho de que la transmisión digital en banda base es característica de los medios metálicos. El primer nivel de clasificación de los códigos de línea se refiere a si son binarios o no. Los **códigos de línea binarios** son aquéllos en los que cada forma de onda elemental sólo permite decodificar 1 bit de información. Por el contrario, en los **códigos de línea no binarios**, la cantidad de información decodificada a partir de cada forma de onda elemental es distinta de 1 bit. Consideramos en primer lugar los códigos de línea binarios.

Códigos de línea binarios

Los códigos de línea binarios pueden caracterizarse desde los siguientes tres puntos de vista independientes:

- En primer lugar, pueden ser códigos de nivel o códigos de transición. Un **código de nivel** transporta la información en el nivel de voltaje, mientras que un **código de transición** transporta la información en la variación o cambio de voltaje.
- Por otro lado, también pueden clasificarse en códigos instantáneos o códigos con memoria. En un **código instantáneo**, la codificación de cada tipo de bit (0, 1) se define en términos absolutos; en cambio, en un **código con memoria**, al menos la codificación de un tipo de bit depende de la codificación de uno o varios bits anteriores (generalmente el bit anterior).
- Finalmente, un tercer criterio de clasificación permite distinguir entre códigos unipolares y códigos polares. Un **código unipolar** es aquél que utiliza una sola polaridad de voltaje, positiva o negativa, además del voltaje nulo (por ello los códigos unipolares se denominan también OOK – *ON-OFF Keying*); si las dos polaridades están presentes, con o sin el voltaje nulo, entonces se trata de un **código polar**.

De acuerdo con estos criterios, el código NRZ mostrado en la **Figura 3.3** es un código de nivel, instantáneo o sin memoria, y puede ser unipolar o polar tal como refleja la propia figura. En la versión unipolar, el dígito 1 se representa mediante un nivel de voltaje no nulo, es decir, positivo o negativo (positivo en la figura), y el dígito 0 mediante un voltaje nulo; en la versión polar, el voltaje positivo puede corresponder al 1 y el negativo al 0, o al revés (como en la figura). No obstante, como veremos a continuación, los códigos NRZ están lejos de satisfacer buena parte de las prestaciones exigidas. Por ello, la investigación ha ido desarrollando códigos de línea binarios cada vez más sofisticados, los cuales satisfacen en mayor o menor medida tales prestaciones. El análisis detallado requiere su agrupación por clases o familias, entendidas como grupos de códigos en los que

predomina uno u otro rasgo, funcionalidad o deficiencia. Una taxonomía razonable, coherente con la mayor parte de fuentes consultadas, agrupa los códigos de línea binarios en seis clases o familias, cuya descripción sigue a continuación. El análisis de cada clase se centra principalmente en los aspectos de sincronismo de bit y perfil espectral; por lo que se refiere a la complejidad, en general ésta va aumentando al pasar de una clase a la siguiente, y en cuanto a las métricas de comportamiento, su evaluación se trata al final del presente capítulo. Para facilitar la comprensión de estos códigos, la **Tabla 3.2** especifica las características, desde los tres puntos de vista antes señalados, de algunos ejemplos de cada familia, mientras que la **Figura 3.5** muestra la señal generada por cada uno de ellos para un mismo patrón de bits.

Tabla 3.2. Caracterización de algunos de los códigos de línea más importantes.

Código de línea	Nivel	Transición	Instantáneo	Con memoria	Unipolar	Polar
NRZ unipolar	✓		✓		✓	
NRZ polar	✓		✓			✓
NRZ(I)	✓			✓		✓
RZ Unipolar	✓		✓		✓	
RZ Polar	✓		✓			✓
AMI (NRZ/RZ)	✓			✓		✓
Manchester		✓	✓			✓
Manchester diferencial		✓		✓		✓
B8ZS	✓			✓		✓
HDB3	✓			✓		✓

Los **códigos NRZ** deben su nombre al hecho de que generan formas de onda en las que el nivel de voltaje se mantiene constante durante todo el tiempo de bit, sin regresar al valor 0. Son todos ellos códigos de nivel (de hecho, también se les denomina NRZ (L) o NRZ-L, con L de *Level*), y como ya señalamos, presentan pobres prestaciones en cuanto a imbricación de sincronismo y comportamiento espectral en baja frecuencia. Admiten las dos variantes, unipolar y polar, ya mostradas en las **Figuras 3.3 y 3.5**. También admiten una versión diferencial, con memoria, conocida como NRZ (I) o NRZ-I (I: *Inverted*), en la que los dígitos binarios se codifican manteniendo o invirtiendo la polaridad con respecto al dígito binario anterior¹⁷. Con ello se consigue inmunidad frente a los cambios imprevistos de polaridad en el circuito formado por el transmisor y el receptor: puesto que la información reside en la polaridad relativa de un dígito con respecto al anterior, un cambio súbito del signo del voltaje en todo el circuito no perturba la correcta interpretación de los dígitos. A su vez, el código NRZ (I) admite dos variantes llamadas NRZ (M) y NRZ (S), donde la M se refiere a *Mark* y la S a *Space* (marca y espacio, 1 y 0, respectivamente). En el caso de NRZ (M), el

¹⁷ La definición de los códigos NRZ (I) implica que uno de los dígitos binarios viene representado por una transición de voltaje al final del tiempo de bit, y el otro dígito mediante la ausencia de transición. En este sentido, estos códigos parecen más bien de transición, y no de nivel, como apunta la **Tabla 3.1**. No obstante, si tenemos en cuenta que habitualmente los receptores muestrean la señal en el punto medio de cada tiempo de bit, los códigos NRZ (I) se mantienen en la categoría de códigos de nivel.

dígito 1 se codifica invirtiendo la polaridad con respecto al dígito anterior, y el dígito 0 manteniéndola; en la variante NRZ (S) se aplica el criterio opuesto. Con todo, y a pesar de la mejora aportada por la versión diferencial, los códigos NRZ son los que peores prestaciones ofrecen. Por ejemplo, si el mensaje a enviar contiene una secuencia larga de cualquiera de los dos dígitos, la señalización NRZ genera un voltaje constante durante un intervalo de tiempo largo, lo que puede provocar la pérdida de sincronización entre transmisor y receptor: éste deja de recibir pasos por 0 en el nivel de voltaje, y si esto sucede durante un tiempo excesivamente largo, la desincronización puede superar la frontera de un dígito binario, por defecto o por exceso. Además, la señalización NRZ presenta un contenido espectral promedio muy significativo en baja frecuencia, lo que implica que buena parte de la potencia de señal queda desaprovechada.

Los **códigos RZ** (*Return to Zero*) generan formas de onda en las que cualquier voltaje distinto de 0 se mantiene constante solamente durante una fracción del tiempo de bit (en general la mitad), volviendo al valor nulo durante el resto de ese tiempo. Dicho de otro modo, en este tipo de códigos la duración de los pulsos de voltaje es inferior al tiempo de bit. En la variante unipolar, el pulso de voltaje (positivo o negativo) corresponde al dígito 1, mientras que el dígito 0 se representa mediante un voltaje nulo durante todo el tiempo de bit. En la variante polar, los pulsos de voltaje positivos corresponden al 1 y los negativos al 0, o viceversa. En general, los códigos RZ representan una leve mejora sobre los códigos NRZ al garantizar el sincronismo durante las secuencias largas de 1's en el caso unipolar, y durante las secuencias largas de ambos dígitos en el caso polar. Sin embargo, la versión unipolar sigue presentando componente continua, y la versión polar consume el doble de ancho de banda que el código NRZ polar, por lo que su eficiencia espectral se ve reducida a la mitad.

Los **códigos de transición**, entre los cuales figuran los **códigos bifase** (Manchester y Manchester diferencial) como ejemplos más representativos. En el **código Manchester**, también llamado *digital biphas*e y *split phase*, se distinguen los dígitos binarios por el sentido de las transiciones de voltaje en el punto medio de cada tiempo de bit. Concretamente, según el convenio original, adoptado por numerosos autores, la transición de nivel bajo (negativo) a alto (positivo) designa el 0, mientras que la transición alto a bajo corresponde al 1. Otro grupo de autores, también numeroso, definen los códigos bifase en base al convenio opuesto (aplicado en la **Figura 3.5**). El código Manchester es vulnerable a los cambios imprevistos de polaridad del voltaje, por lo que también se desarrolló una **versión Manchester diferencial**, que a su vez admite dos variantes. Así, la llamada *split phase* (M), mostrada en la **Figura 3.5**, consiste en que el dígito 1 se codifica invirtiendo el sentido de la transición con respecto al dígito anterior, y el dígito 0 manteniendo el sentido de dicha transición. En la variante *split phase* (S) se adopta el criterio opuesto. La única desventaja de los códigos bifase es que consumen mucho ancho de banda y presentan, por tanto, menor eficiencia espectral que la mayor parte del resto de codificaciones de línea. Además, al tratarse de códigos de transición, son muy robustos frente a interferencias y ruido, pues es más probable que una señal no deseada altere un nivel de voltaje que no el sentido de una transición. Otro conjunto de códigos de transición lo constituyen el código Miller y sus variantes: *biphase* (M), *biphase* (S) y CMI (*Code Mark Inversion*). Este último es, en realidad, un híbrido entre los códigos de transición y los códigos AMI descritos a continuación.

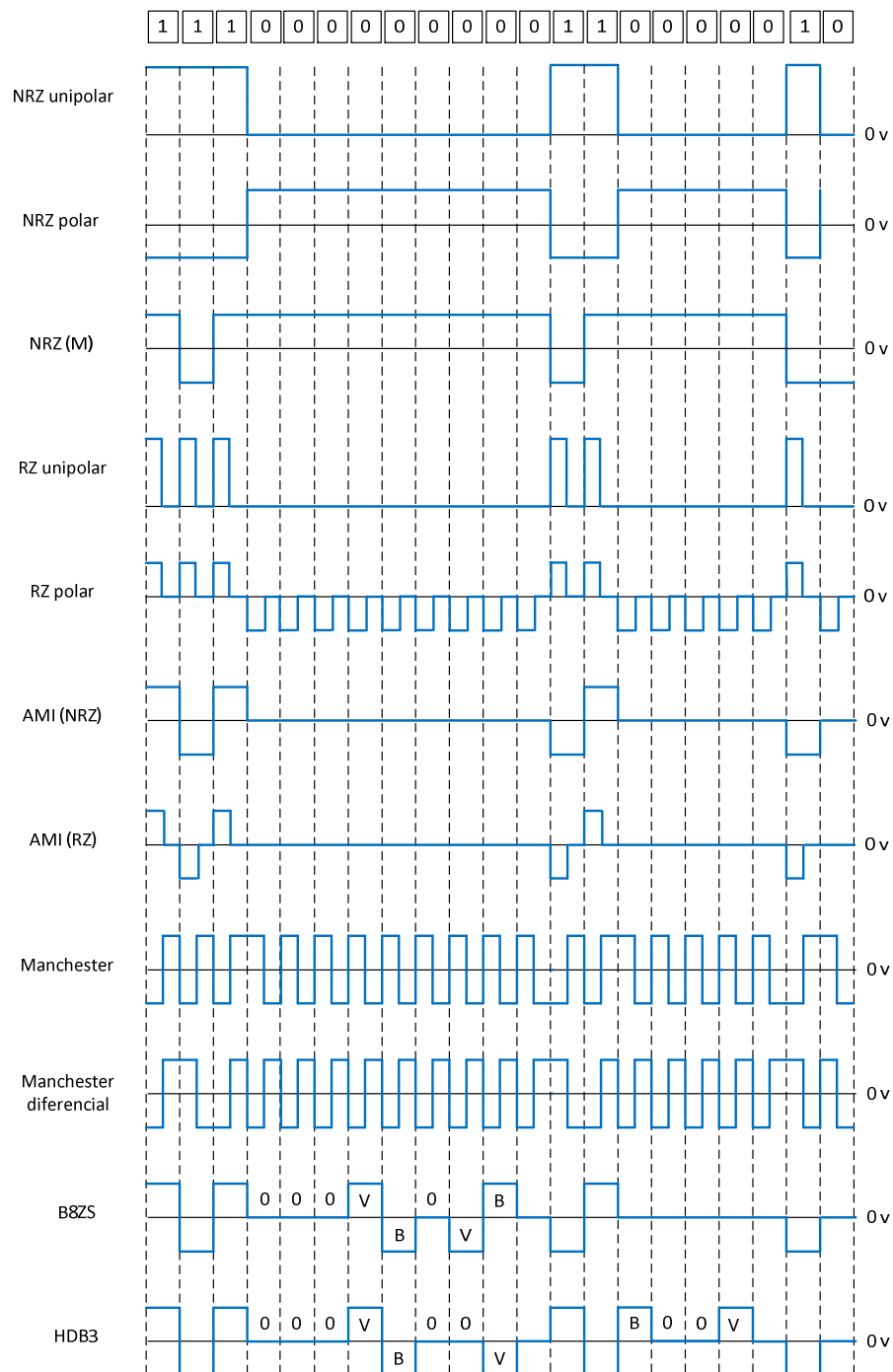


Figura 3.5. Señal generada por diversos códigos de línea para un mismo patrón de bits. En los códigos con memoria, se ha elegido el signo del primer pulso de forma aleatoria.

En los **códigos bipolares**, **códigos binarios multinivel** o **códigos AMI** (*Alternate Mark Inversion*), los 1's se codifican mediante pulsos alternados (son, por tanto, códigos con memoria) y los 0's

mediante voltaje nulo. Con ello se logra eliminar prácticamente el contenido en baja frecuencia de la señal generada y garantizar el sincronismo entre transmisor y receptor en las secuencias largas de 1's, aunque no en las de 0's. La codificación AMI, también llamada pseudoternaria¹⁸, admite dos versiones, con retorno a cero (RZ) y sin retorno a cero (NRZ), conocidas respectivamente como RZ bipolar (BRZ – *Bipolar RZ*) y NRZ bipolar. Al igual que los códigos NRZ (I) y Manchester diferencial, los códigos AMI son inmunes a los cambios imprevistos de polaridad en el circuito formado por transmisor y receptor.

Otro conjunto de códigos binarios lo forman los llamados **códigos multi-transición**, de los cuales el más representativo y probablemente el único utilizado en la práctica es MLT-3 (*Multi-Level Transmit-3 levels*). El **código MLT-3** reúne características de los códigos diferenciales y los binarios multinivel. Más concretamente, MLT-3 puede verse como una extensión de NRZI a 3 niveles de voltaje, +, – y 0, de forma que cada vez que se transmite el dígito 1 se altera el nivel de voltaje con respecto al dígito anterior, pero siguiendo una secuencia cíclica a través de los 3 voltajes anteriores, es decir, de negativo a cero, de cero a positivo, de positivo a cero y de cero a negativo. En total, 4 transiciones distintas¹⁹. Cuando se transmite el dígito 0, se mantiene el voltaje con respecto al dígito anterior. La señalización MLT-3 se utiliza sobretodo en combinación con los códigos 4B5B y RLL que tratamos más adelante.

Los **códigos no alfabéticos**, también llamados **códigos AMI modificados**, se basan en aplicar alguna técnica de *scrambling* sobre la codificación AMI original, a fin de evitar que las secuencias largas de 0's provoquen la pérdida de sincronismo entre transmisor y receptor. En este contexto, *scrambling* consiste en forzar la aparición de pulsos de voltaje durante dichas secuencias, pero de forma que no se alteren otras propiedades favorables del código AMI original²⁰. Hay dos grupos de códigos AMI modificados: los **códigos BNZS** (*Binary N Zero Substitution*) y los **códigos HDBN** (*High-Density Bipolar N*). En ambos casos se introduce un nuevo tipo de pulso, el pulso V (*bipolar violation*), que viola o rompe el criterio de alternancia de polaridad con respecto al último pulso, en contraposición al pulso válido o pulso B (*valid bipolar signal*), que mantiene la alternancia de polaridad con respecto al último pulso. Así, tanto en los códigos BNZS como en los HDBN, la técnica consiste en reemplazar cada grupo de N 0's consecutivos (caso de BNZS) o $N + 1$ 0's consecutivos (caso de HDBN) por un código especial que contiene pulsos B y V. Concretamente, los pulsos V permiten que el código especial sea fácilmente reconocible por el receptor, de modo que pueda recuperar la secuencia original de 0's consecutivos; los pulsos B se utilizan para compensar los pulsos V, y mantener así el

¹⁸ La denominación pseudoternaria proviene del hecho de utilizar tres niveles de voltaje (positivo, negativo y nulo), aunque para representar un único dígito binario, cero o uno. Algunos autores atribuyen dicha denominación al código AMI según el convenio opuesto, es decir, con los 0's representados mediante pulsos alternados y los 1's mediante voltaje nulo.

¹⁹ De esta manera, el ciclo de repetición de la señal cuando se transmiten secuencias largas del dígito 1 es 4 veces mayor, lo que significa que la frecuencia máxima del espectro promedio se divide por 4; esto a su vez favorece la utilización de esta señalización a través de pares de hilos trenzados, cuyo respuesta en alta frecuencia es más pobre. Este tipo de justificaciones se entenderán mejor más adelante, al tratar los medios de transmisión, y en todo caso su relevancia en relación a los objetivos de estos apuntes es menor.

²⁰ El término *scrambling* también se utiliza para designar la alteración de los mensajes binarios con propósitos de encriptación.

nivel medio de señal muy próximo a 0 y, finalmente, ambos pulsos fuerzan transiciones de voltaje que facilitan la sincronización entre transmisor y receptor. Dos ejemplos representativos de estos códigos son B8ZS y HDB3. En la **Figura 3.5** se muestra un ejemplo de **codificación B8ZS**, la cual consiste en substituir cada cadena de 8 0's consecutivos por la secuencia 000VB0VB. Por tanto, si el último pulso que precede a la cadena de 8 0's es positivo (caso de la figura), la señalización adopta la forma 000 + −0 − +, mientras que si el pulso precedente es negativo, la señalización resultante es 000 − +0 + −. En el caso de **HDB3**, no puede haber más de 3 0's consecutivos o, lo que es lo mismo, las secuencias substituidas son de 4 0's. Hay dos opciones: la secuencia 0000 se substituye por B00V o por 000V. El criterio es que los pulsos V consecutivos tienen que tener polaridad alternada a fin de que el valor medio de señal sea lo más próximo a 0; por ejemplo, en la **Figura 3.5** se ha supuesto que el último pulso V que precede a la secuencia mostrada es negativo, y así el primer conjunto de 4 0's se substituye por una cadena 000V, con lo que el nuevo pulso V presenta polaridad positiva (si se hubiera substituido por una cadena B00V, entonces el nuevo pulso V sería también negativo). Los códigos B8ZS y HDB3 tienen prestaciones similares a los códigos Manchester, pero consumen menos ancho de banda, y por ello son preferibles para transmisiones a larga distancia (la relación inversamente proporcional entre alcance y ancho de banda es una consecuencia de la fórmula de Shannon que obtendremos más adelante).

Para completar esta descripción, la **Tabla 3.3** muestra aplicaciones y ejemplos de uso de algunos de los códigos que acabamos de describir.

Tabla 3.3. Aplicaciones de algunos códigos de línea binarios.

Códigos de línea	Aplicaciones
NRZ	Comunicaciones a corta distancia: <ul style="list-style-type: none"> • Comunicaciones internas de los computadores (NRZ polar). • Estándar RS-232 para puertos serie (NRZ polar). • Cables USB (<i>Universal Serial Bus</i>) (NRZ-I).
Manchester	Algunos estándares para redes de área local, en la actualidad obsoletos o en desuso: <ul style="list-style-type: none"> • IEEE 802.4 (<i>token bus</i>). • Primeras versiones de IEEE 802.3 (Ethernet 10BASE2, 10BASE5 y 10BASE-T). • IEEE 802.5 (<i>token ring</i>) (Manchester diferencial <i>split phase</i> (M)).
AMI	<ul style="list-style-type: none"> • Telefonía digital de primera generación. • Líneas serie T1 del estándar <i>T-carrier</i> (primera versión). • Canales B de la RDSI (Red Digital de Servicios Integrados).
MLT-3	<ul style="list-style-type: none"> • Estándar FDDI (<i>Fiber Distributed Data Interface</i>): versión sobre cable coaxial (diccionario predefinido 4B5B + señalización MLT-3). • Estándar IEEE 802.3u (Fast Ethernet) 100BASE-TX (4B5B + MLT-3).
BNZS	Estándar <i>T-carrier</i> para líneas serie: T1 (B8ZS), T2 (B6ZS) y T3 (B3ZS).
HDB3	Estándar <i>E-carrier</i> para líneas serie.

Códigos de línea no binarios

En los códigos de línea no binarios, cada forma de onda elemental deja de corresponder a un solo bit de información. Son los llamados **códigos alfabéticos** o **códigos bloque**, en los que un grupo de

X símbolos binarios (0 o 1) se mapea en un grupo de Y símbolos elegidos de un conjunto de 2 o más niveles. Estos códigos responden al formato “XBYL” (*X Binary symbols mapped on Y L-ary symbols*). Por ejemplo, L puede ser B (*binary*), T (*ternary*) o Q (*quaternary*), entre las opciones más habituales. La selección de los valores de X e Y debe satisfacer la condición $2^X \leq L^Y$, donde $L = 2$ en el caso de símbolos binarios (2 símbolos distintos, por ejemplo pulsos de voltaje + y −, + y 0 o − y 0), $L = 3$ en el caso de símbolos ternarios (3 símbolos distintos, por ejemplo pulsos de voltaje +, − y 0) y $L = 4$ en el caso de símbolos cuaternarios (4 símbolos distintos, por ejemplo pulsos de voltaje −3A, −A, +A y +3A, para un valor dado de A). Observemos que en el caso B cada símbolo del código de línea corresponde a una cantidad de información inferior a 1 bit.

Uno de los más comunes dentro de la primera categoría (XBYB) es el **código 4B5B**, que consiste en mapear cada combinación de 4 dígitos binarios en una palabra de 5 dígitos binarios perteneciente a un conjunto predeterminado o **diccionario**, constituido en este caso por 16 palabras seleccionadas de las 32 posibles. Las palabras seleccionadas satisfacen la condición de que, conjuntamente con la señalización elegida, provocan suficientes transiciones en el estado de la línea como para mantener, independientemente de la combinación de 4 dígitos transmitidos, el sincronismo de bit entre transmisor y receptor. Por ello, la codificación 4B5B es realmente un conjunto de códigos de línea en el que cada uno se identifica por un mapeado concreto en palabras de 5 dígitos binarios, que luego tienen que representarse mediante formas de onda, como por ejemplo NRZI. Cualquiera que sea el mapeado elegido, la codificación 4B5B presenta la ventaja adicional de una mayor robustez frente a errores de transmisión, pues algunos de estos errores pueden dar lugar a palabras no incluidas en el diccionario seleccionado. Una extensión sofisticada de los códigos XBYB la constituyen los **códigos RLL** (*Run-Length Limited*), diseñados con el objetivo principal de garantizar el sincronismo de bit. Puesto que se combinan con señalizaciones como NRZI o MLT-3, que son vulnerables a la pérdida del sincronismo de bit durante las secuencias largas de ceros, la construcción de estos códigos se basa en limitar inferior y superiormente la longitud de dichas secuencias, no sólo a nivel de cada palabra código, sino teniendo en cuenta todas las parejas posibles de palabras código que puedan producirse secuencialmente. De todos modos, la importancia de los códigos RLL para nosotros es relativa, puesto que, si bien se utilizan en el ámbito de las comunicaciones, su uso principal es en dispositivos de almacenamiento tales como discos duros magnéticos y ópticos (CD, DVD, Blu-ray y otros).

Otro conjunto importante dentro de los códigos bloque lo constituyen los **códigos ternarios**, que típicamente se definen sobre los símbolos +, − y 0. Observemos que el código AMI ya manejaba estos 3 símbolos, pero eran binarios, pues cada símbolo sólo se utilizaba para decodificar un solo dígito (de ahí su denominación alternativa como códigos binarios multinivel). Con los códigos ternarios se plantea la posibilidad de aprovechar más eficientemente la disponibilidad de 3 símbolos. Según vimos en el Capítulo 1, la máxima cantidad de información por símbolo, con un alfabeto de 3 símbolos, viene dada por $\log_2 3 = 1.58$ bit/símbolo, es decir, un 58% superior a la cantidad de información por símbolo que se alcanza con los códigos AMI (1 bit/símbolo). Los códigos ternarios se diseñan, pues, con el objeto de obtener una ratio de información por símbolo lo más cercano posible al límite superior dado por 1.58, al tiempo que tratan de satisfacerse el resto de

requisitos de toda codificación de línea. Los ejemplos más representativos son 3B2T, 4B3T, 8B6T y PST (*Pair Selected Ternary*).

Finalmente, entre los **códigos cuaternarios**, los más relevantes son 2B1Q (*polar quaternary NRZ*) y 4D-PAM5 (*4-Dimensional Pulse Amplitud Modulation with 5 levels*). Como su denominación indica, este último código hace uso de 5 niveles, pero uno de ellos se utiliza solamente para propósitos de control de errores, por lo que el código se mantiene en la categoría de los cuaternarios.

Tabla 3.4. Aplicaciones de algunos códigos bloque.

Códigos de línea	Aplicaciones
XBYB	<ul style="list-style-type: none"> Estándar FDDI (<i>Fiber Distributed Data Interface</i>): versión sobre fibra óptica (diccionario predefinido 4B5B + señalización NRZ (I)) y versión sobre cable coaxial (diccionario predefinido 4B5B + señalización MLT-3). Estándar IEEE 802.3u (Fast Ethernet): 100BASE-FX (4B5B + NRZ (I)) y 100BASE-TX (4B5B + MLT-3). Estándar IEEE 802.3ab (Gigabit Ethernet) (8B10B). Estándar Ethernet a 10 Gbps o superior (64B66B).
Ternarios	<ul style="list-style-type: none"> Interfaz primaria de la RDSI (4B3T). Estándar IEEE 802.3u (Fast Ethernet) 100BASE-T4 (8B6T).
Cuaternarios	<ul style="list-style-type: none"> Interfaz básica de la RDSI (2B1Q – <i>polar quaternary NRZ</i>). Estándar IEEE 802.3ab (Gigabit Ethernet) 1000BASE-T (<i>4D-PAM5 – 4-Dimensional Pulse Amplitud Modulation with 5 levels</i>).

Al igual que con los códigos binarios, la **Tabla 3.4** muestra las aplicaciones de algunos códigos de línea no binarios.

3.3. TRANSMISIÓN DIGITAL MODULADA

Puede ocurrir que la transmisión en banda base no sea adecuada por razones prácticas de diversa índole. En tales casos, la **modulación digital** es la técnica que permite trasladar el espectro de una señal digital a una banda de frecuencias intermedia, donde la transmisión pueda tener lugar en mejores condiciones. La señal paso-banda que resulta de este proceso de modulación es una señal analógica, si bien, como ya sabemos, transporta información digital (de ahí la denominación de modulación digital). Este escenario es más frecuente que el de una transmisión digital en banda base y su esquema básico se muestra en la **Figura 3.6**. La diferencia principal con respecto al esquema de la **Figura 3.4** es que el codificador de línea ha sido substituido por un modulador. La **Figura 3.6** también pone de manifiesto la presencia de un oscilador, el cual genera una señal sinusoidal a una cierta frecuencia. Este oscilador es indispensable, pues la modulación digital consiste en variar uno o varios parámetros de una señal sinusoidal (recuérdese de la Sección 2.2 que estos parámetros son amplitud, frecuencia y fase) en función de la señal digital que contiene el mensaje (típicamente una señal NRZ). Así, la señal sinusoidal generada por el oscilador es la **señal portadora**, la señal digital a la entrada del modulador actúa como **señal moduladora** y la señal que finalmente se transmite por el enlace, resultado de la modulación, es la **señal modulada**. La frecuencia de la señal portadora, f_c (c de *carrier*), se denomina **frecuencia portadora**.

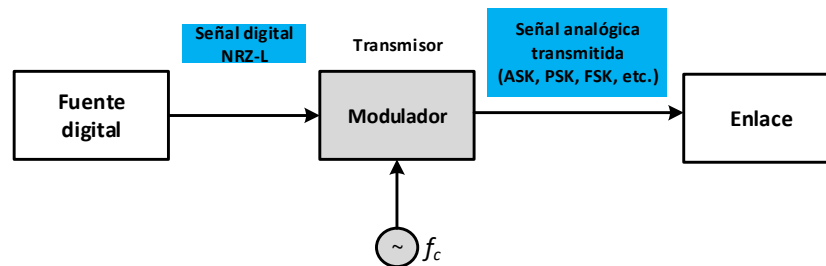


Figura 3.6. Modelo básico de comunicaciones para una transmisión digital modulada (lado origen).

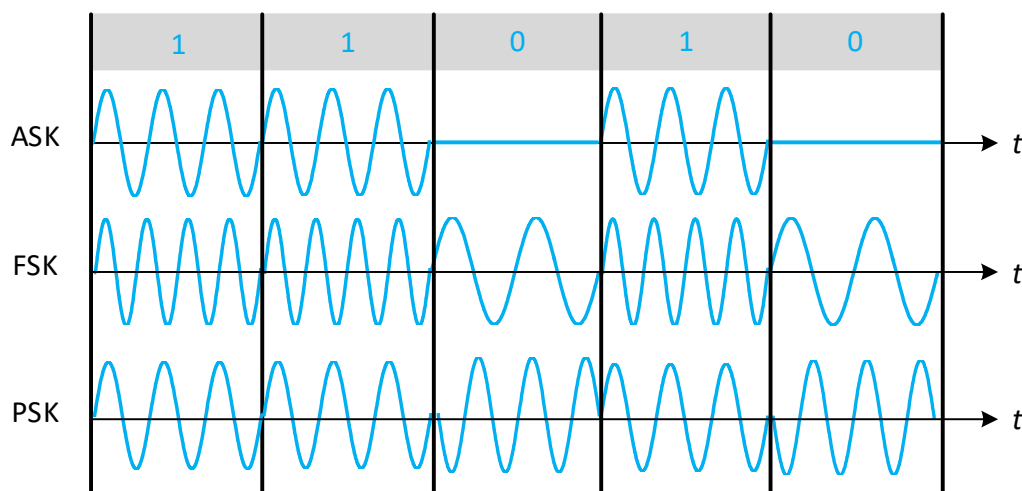


Figura 3.7. Ejemplos de modulaciones ASK, FSK y PSK para el mensaje binario indicado en la parte superior. La duración de cada dígito binario corresponde siempre a un número entero de ciclos de la portadora. Como puede observarse, en todos los casos la señal finalmente transmitida (señal modulada) es analógica.

En la **Figura 3.7** se ilustran las tres modulaciones digitales básicas para una misma secuencia de dígitos binarios. Estas modulaciones son **ASK** (*Amplitude Shift Keying*), **FSK** (*Frequency Shift Keying*) o **PSK** (*Phase Shift Keying*) según se varíe, respectivamente, la amplitud, la frecuencia o la fase de la señal portadora. La implementación de ASK mostrada en la figura, en la que los 1s se codifican con la amplitud de la portadora, y los 0s con amplitud nula, es la habitual y se denomina también **OOK** (*On-Off Keying*). La modulación FSK consiste, como también se ilustra en la figura, en utilizar dos frecuencias distintas, f_0 y f_1 , para representar ambos dígitos binarios. Estas dos frecuencias pueden verse como desviaciones a un lado y otro de la frecuencia portadora, o directamente como dos portadoras. Finalmente, la modulación PSK hace uso de dos fases; el caso de la figura es la versión más simple, en la que la fase 0° representa el 1 y la fase 180° representa el 0.

Los patrones básicos que acabamos de describir son binarios y en cada uno se varía un único parámetro para representar las dos opciones, de forma que los otros dos parámetros se mantienen fijos. No obstante, con el objeto de mejorar las métricas de comportamiento, la investigación ha ido introduciendo patrones más complejos que resultan de extensiones multinivel o combinaciones

híbridas. En el caso de las extensiones multinivel, se trabaja con un número de variaciones del parámetro igual a una potencia de 2, de forma que cada variación corresponde a una secuencia de 2 o más dígitos binarios. Las extensiones multinivel más utilizadas son **MFSK** (*Multilevel-FSK*) y **MPSK** (*Multilevel-PSK*). Como casos particulares de éstas, las modulaciones básicas FSK y PSK se conocen también como **BFSK** y **BPSK** respectivamente (B de *Binary*). Por su parte, las combinaciones híbridas resultan de variar 2 parámetros de la portadora, típicamente amplitud y fase, siendo las modulaciones **QAM** (*Quadrature Amplitude Modulation*) las más relevantes.

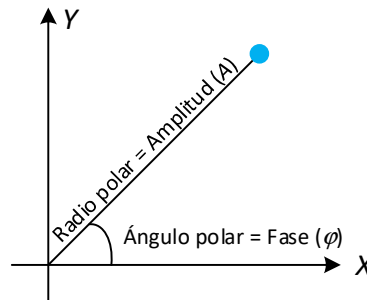


Figura 3.8. Elementos básicos de la representación en forma de diagrama polar.

Una forma gráfica e intuitiva de representar las modulaciones digitales basadas en amplitud y/o fase son las llamadas constelaciones. En nuestro contexto, una **constelación** es un diagrama polar en el que el radio polar se corresponde con la amplitud de la señal portadora, y el ángulo polar con la fase de la misma. La frecuencia se considera fija. La **Figura 3.8** describe esta forma de representación, en la que el punto señalado correspondería a un símbolo concreto de la constelación, es decir, a una determinada combinación de amplitud A y fase φ . Ya en la **Figura 3.9** se muestran las constelaciones correspondientes a las modulaciones básicas ASK y PSK mostradas en la **Figura 3.7**. Y en la **Figura 3.10** se muestran las constelaciones correspondientes a dos modulaciones muy típicas, QPSK (*Quadrature-PSK*) y 16-QAM. Cualquiera que sea la constelación, la asignación concreta de una combinación de dígitos binarios a cada símbolo no es única, pero se suele realizar de forma que los símbolos vecinos difieran en un solo dígito. La modulación QPSK es un caso particular de modulación PSK de 4 símbolos, es decir, una 4-PSK, pero, al igual que todas las QAM, forma parte de las llamadas modulaciones rectangulares o en cuadratura (de ahí la letra Q), pues tienen sus puntos dispuestos en forma de rejilla cuyas direcciones principales son paralelas a los ejes coordenados. Como se desprende de la **Figura 3.10**, todos los símbolos de la modulación QPSK se transmiten con la misma amplitud de portadora y fases respectivas de 45° , 135° , 225° y 315° . Cada uno de ellos transporta la información equivalente a 2 dígitos binarios. En la modulación 16-QAM, cada símbolo codifica 4 dígitos binarios, y corresponde a una determinada combinación de amplitud y fase de la portadora. Como puede observarse, la constelación contiene 3 amplitudes diferentes, una correspondiente a los 4 símbolos interiores (0000, 0010, 1000 y 1010), otra correspondiente a los 4 vértices externos (0101, 0111, 1101 y 1111) y la tercera correspondiente al resto de símbolos. En cuanto a las fases, en total son 12, pero no resultan de la división del arco completo de 360° en 12 partes iguales (la división en partes iguales corresponde a las llamadas modulaciones circulares). Por ejemplo, al símbolo 0100 le corresponde una fase de 18.43° aproximadamente, y al símbolo 1111 una fase de

225°. Las modulaciones QAM son probablemente las más utilizadas, pudiéndose encontrar en los sistemas de televisión digital terrestre (aérea y por cable), en los módems ADSL, en algunas versiones del estándar IEEE 802.11 para redes inalámbricas, y en los nuevos estándares para transmisión de datos a través del cableado doméstico (cable coaxial, líneas telefónicas y líneas eléctricas).

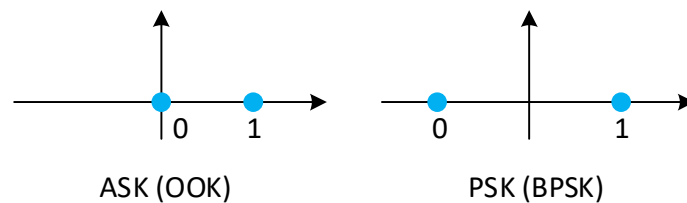


Figura 3.9. Constelaciones correspondientes a las modulaciones digitales básicas ASK y PSK.

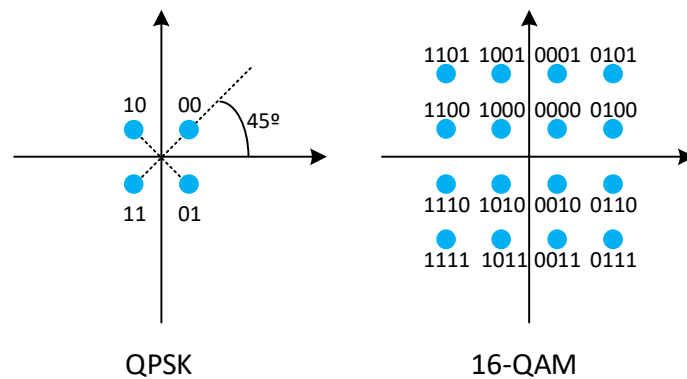


Figura 3.10. Constelaciones correspondientes a las modulaciones QPSK y 16-QAM. La modulación 16-QAM es la mínima de las QAM utilizadas en la práctica que realmente combinan amplitud y fase: las modulaciones 2-QAM y 4-QAM no son más que modulaciones BPSK y QPSK, y la modulación 8-QAM, si bien es híbrida entre amplitud y fase, es poco utilizada por ser solo ligeramente más sencilla que la 16-QAM, pero bastante menos eficiente que ésta desde el punto de vista espectral (el análisis de eficiencia espectral se considera más adelante).

Para terminar esta sección, la **Figura 3.11** muestra un ejemplo de modulación 16-QAM para una secuencia de 20 dígitos binarios generados por una fuente. Se ha supuesto la constelación de la Figura 3.10 con una rejilla definida por las coordenadas $(-3, -1, +1, +3)$ en ambos ejes, de forma que las 3 amplitudes y las 12 fases resultantes son, respectivamente, $A_1 = \sqrt{2}$, $A_2 = \sqrt{10}$ y $A_3 = \sqrt{18}$ por un lado, y $\varphi_1 = 18.43^\circ$, $\varphi_2 = 45^\circ$, $\varphi_3 = 71.57^\circ$, $\varphi_4 = 108.43^\circ$, $\varphi_5 = 135^\circ$, $\varphi_6 = 161.57^\circ$, $\varphi_7 = 198.43^\circ$, $\varphi_8 = 225^\circ$, $\varphi_9 = 251.57^\circ$, $\varphi_{10} = 288.43^\circ$, $\varphi_{11} = 315^\circ$ y $\varphi_{12} = 341.57^\circ$, por otro. Como se observa en la figura, el modem agrupa los dígitos binarios provenientes de la fuente de 4 en 4, y para cada combinación de 4 dígitos produce varios ciclos completos de la señal portadora con la amplitud y fase correspondientes según la constelación (indicadas también en la figura).

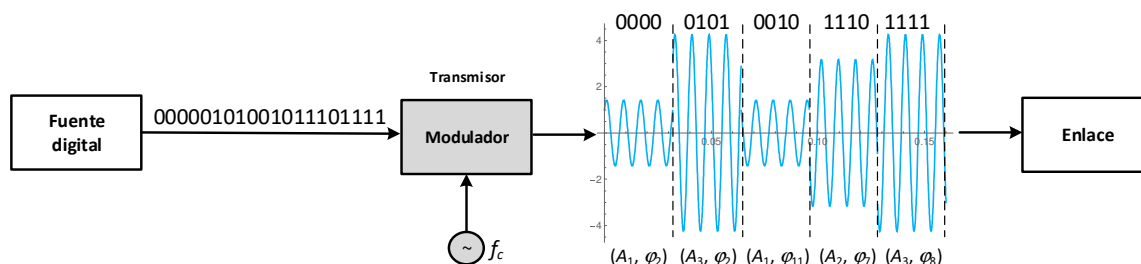


Figura 3.11. Modulación 16-QAM para una secuencia concreta de 20 dígitos binarios.

3.4. TRANSCPTORES

Tanto si la transmisión es en banda base como si es modulada, el receptor deberá ser capaz de recuperar el mensaje original a partir de la señal recibida. Como ya ocurría con la fuente, el destinatario suele estar diseñado para leer señales digitales sencillas del tipo NRZ, de modo que la operación del receptor consistirá en recuperar el mensaje original y presentarlo al destinatario mediante este tipo de señalización. Por tanto, en transmisión digital en banda base, el receptor digital es un **decodificador de línea**, capaz de transformar la codificación de la señal recogida del enlace en una codificación NRZ; en cambio, en transmisión digital modulada, el receptor digital es un **demodulador**. Como generalmente las comunicaciones son bidireccionales, en cada extremo del enlace suelen llevarse a cabo las dos funciones integradas en el mismo dispositivo, llamado **codificador-decodificador de línea**²¹ en el caso de la transmisión digital en banda base, y **modem** (modulador-demodulador) en el caso de la transmisión digital modulada. Ambos equipos son ejemplos de **transceptor**, término que proviene de la contracción de transmisor y receptor (de hecho, algunos autores se refieren al codificador-decodificador de línea como transceptor digital). El transceptor, o bien el transmisor y el receptor por separado, suelen estar integrados en unidades funcionales más complejas, como la sección o cabecera RF (*RF front-end*) de un teléfono móvil o la tarjeta de red de un computador.

Si bien el transceptor es un dispositivo que integra las funciones de transmisión y recepción, esto no significa que dichas funciones compartan parte de la circuitería interna. De hecho, cuando las circuiterías de transmisión y recepción son independientes, el transceptor permite que la comunicación sea bidireccional y simultánea en los dos sentidos. Es lo que se denomina comunicación **full-dúplex**. En cambio, si las dos funciones comparten una parte de la circuitería del transceptor, la comunicación bidireccional es posible pero no de forma simultánea en los dos sentidos. Hablamos entonces de comunicación **half-dúplex**.

²¹ Aunque la contracción de codificador-decodificador se conoce como **códec**, este término se utiliza más en el contexto de la codificación de fuente, concretamente para designar los algoritmos de compresión/descompresión de fuentes de audio o vídeo (códec de audio, códec de vídeo); sería la acepción software del término. Por otro lado, también tiene una acepción hardware, pues algunos autores lo utilizan como sinónimo de conversor analógico-digital y digital-analógico.

3.5. CONVERSIÓN ANALÓGICA-DIGITAL

Con respecto a la **Figura 3.2**, el último escenario que nos queda por considerar es el de la **conversión analógica-digital**, es decir, la transformación de la señal producida por una fuente analógica en una señal digital para que pueda ser transmitida a través de una red de computadores. Esta operación la realiza el llamado **convertidor analógico-digital** (A/D, CAD o, en inglés, ADC – *analog to digital converter*). Como la propia **Figura 3.2** indica, el conjunto formado por la fuente analógica y el convertidor A/D constituyen una fuente digital equivalente, y como tal la transmisión que sigue a continuación puede ser en banda base o modulada. En el extremo receptor, un **convertidor digital-analógico** (D/A o DAC – *digital to analog converter*) permite recuperar una señal analógica lo más parecida posible a la original.

Antes de describir el funcionamiento del convertidor A/D, vale la pena resumir las ventajas de la digitalización de las señales, sobre todo cuando puede parecer un contrasentido transformar una señal analógica en digital para luego tener que transmitirla analógicamente si se elige la opción de modulación de portadora. Las ventajas a las que aludimos son las siguientes:

- Con la expansión de las tecnologías VLSI (*very large scale integration*), los circuitos digitales (utilizados tanto en el lado origen como en el destinatario) resultan muy baratos y consumen muy poco espacio físico.
- La transmisión digital es más inmune al ruido. Aunque trataremos esta perturbación más adelante, podemos anticipar que para que un impulso de ruido altere el valor de un dígito binario, debe ser suficientemente fuerte en el punto de muestreo de ese dígito en el receptor, lo cual ocurre con probabilidad relativamente baja.
- La transmisión digital ofrece también la posibilidad de encriptar la información y preservar así su privacidad.
- La digitalización favorece la integración de redes y servicios, ya que todas las formas de información (datos, voz, imágenes, vídeo y multimedia) quedan igualmente representadas mediante secuencias de unos y ceros que pueden ser transportadas por las redes de computadores.
- Finalmente, la digitalización permite aprovechar el potencial tanto del hardware como del software de computación para procesar la información generada o movida por cualquier servicio, sea de voz, datos, imágenes fijas, vídeo o la combinación de estos formatos.

En la **Figura 3.12** se muestra la estructura interna de un convertidor A/D. Como puede observarse, consta de un dispositivo de muestreo, un cuantificador y un codificador. La operación conjunta de estos tres elementos permite la transformación de una señal analógica en una señal digital, típicamente del tipo NRZ. A continuación, describimos brevemente el papel de cada elemento.

Como su nombre indica, el **muestreador** se encarga de tomar muestras de la señal analógica a un ritmo fijo determinado. Cada valor muestreado se mantiene durante un pequeño intervalo de tiempo, como se ilustra en la **Figura 3.13** (de hecho, en inglés, la función de este dispositivo se conoce como *sample & hold*). La señal a la salida del muestreador es la llamada **señal PAM** (*pulse*

amplitude modulation), ya que equivale a un tren de pulsos cuyas amplitudes vienen moduladas por los valores de la señal analógica. Observemos, por tanto, que la señal PAM es una señal discreta en tiempo y continua en valores, representando la primera etapa de transición de la señal analógica original a la señal digital que se obtiene al final del proceso. La frecuencia del tren de pulsos se conoce como **frecuencia de muestreo** f_s , y es la inversa del **período de muestreo** T_s (ver Figura 3.13), es decir, $f_s = 1/T_s$. El valor que debe tomar la frecuencia de muestreo nos lo proporciona el **teorema de muestreo** de Nyquist-Shannon, según el cual, si B_a es el ancho de banda absoluto de la señal analógica (señal en banda base, pues la genera una fuente analógica), la frecuencia de muestreo f_s debe satisfacer la siguiente condición para que la señal analógica original pueda ser recuperada íntegramente, sin distorsión²², a partir de la señal muestreada (obviamos la demostración):

$$f_s \geq 2 \cdot B_a \quad (3.1)$$

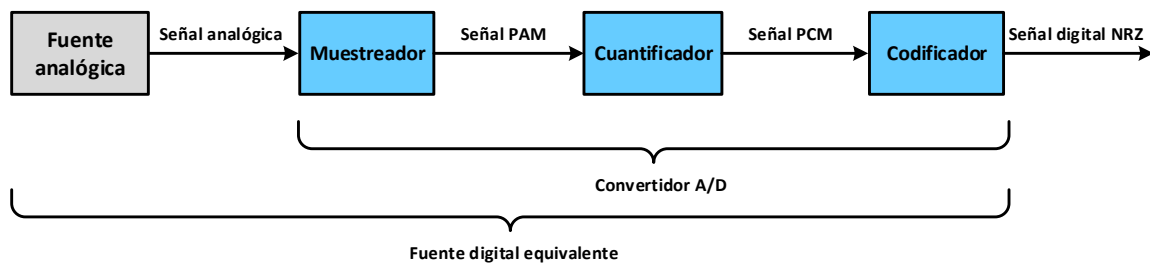


Figura 3.12. Estructura interna de un convertidor A/D, y tipología de señales asociadas.

Esta fórmula no debe ser interpretada como un “cuanto mayor mejor”, sino que una vez satisfecha la condición, aunque sea con el valor mínimo $f_{s,min} = 2 \cdot B_a$, la señal analógica puede recuperarse íntegramente sin distorsión. A esta frecuencia mínima de muestreo se la conoce como **velocidad de Nyquist**, y es característica de cada señal. Por otro lado, la **frecuencia de Nyquist** f_N es una característica del dispositivo de muestreo, y su valor es la mitad de la frecuencia de muestreo a la que trabaja dicho dispositivo ($f_N = f_{s,disp}/2$). Si aplicamos la condición (3.1) sabiendo que el dispositivo muestrea a la frecuencia $f_{s,disp}$, se obtiene que el máximo ancho de banda que puede tener la señal para que el muestreo no produzca distorsión, es precisamente lo que hemos denominado frecuencia de Nyquist. Por tanto, podemos afirmar que la frecuencia de Nyquist de un dispositivo no es más que la frecuencia máxima que puede contener una señal analógica para que, al ser muestreada por dicho dispositivo, no haya distorsión. En conclusión, la transformación de la señal analógica en la señal PAM es realmente un proceso de compresión sin pérdidas: de compresión porque, si consideramos un intervalo de tiempo finito como el mostrado en la **Figura 3.13**, pasamos de una representación consistente en infinitos valores a otra basada en un número finito de muestras (de ahí que la conversión analógica-digital sea conocida también como

²² La distorsión producida por un muestreo que no cumple el teorema de Nyquist-Shannon se denomina distorsión por **aliasing**.

codificación de fuente analógica); y sin pérdidas porque, a pesar de la fuerte compresión impuesta a la señal analógica, ésta todavía puede recuperarse íntegramente si se cumple (3.1).

El muestreo es sólo el primer paso del proceso de digitalización de una señal analógica. El siguiente sería codificar mediante ceros y unos cada valor muestreado contenido en la señal PAM. Pero para representar cualquiera de estos valores habría que utilizar infinitos dígitos binarios, puesto que no debemos olvidar que la señal PAM es continua en valores. Como la representación mediante infinitos dígitos no es viable, la siguiente etapa en el proceso de digitalización es cuantificar la señal PAM. Así pues, el **cuantificador** se encarga de discretizar el eje de ordenadas de dicha señal, de forma que el conjunto de valores posibles a su salida es finito, y por tanto cada uno de ellos representable mediante un número finito de dígitos binarios. La señal a la salida del cuantificador es la llamada **señal PCM** (*pulse code modulation*), llamada así porque la secuencia de pulsos está codificada según unos niveles predefinidos en dicho dispositivo. Por tanto, la señal PCM es una señal discreta en tiempo y discreta en valores, muy cercana ya a la señal digital final.

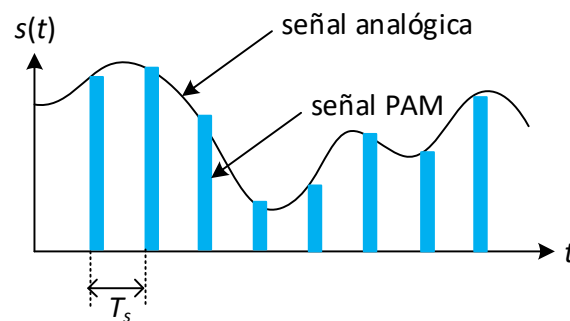


Figura 3.13. Señal PAM a la salida del muestreador.

El proceso de cuantificación se ilustra en la **Figura 3.14**. Concretamente, se ha supuesto una cuantificación uniforme, donde el margen dinámico de la señal de entrada, entre S_{min} y S_{max} , se subdivide en intervalos iguales, cada uno de los cuales queda delimitado por dos valores umbrales y centrado en un nivel cuantificado. El valor PCM de salida no es más que el nivel cuantificado situado entre los dos umbrales entre los cuales queda ubicado el valor PAM de entrada. El número de niveles cuantificados, Q , suele escogerse igual a 2^b , con b el número de dígitos binarios con que se va a representar cada uno.

La **Figura 3.14** también pone de manifiesto que la señal PCM a la salida del cuantificador puede verse como la superposición de la señal PAM y una hipotética señal error. Puesto que esta señal error equivale a una perturbación sobre la señal de entrada, se la denomina ruido, y concretamente **ruido de cuantificación**, pues en realidad se trata de un error provocado por el cuantificador. Precisamente la forma de evaluar el grado de error introducido por el cuantificador es mediante la llamada **relación señal-ruido de cuantificación**, SNR_q , definida como la relación expresada en decibelios entre la potencia de la señal PAM, P_{PAM} , y la potencia de la hipotética señal error, P_{error} . Se puede demostrar que dicha relación, para un cuantificador uniforme como el que estamos considerando, viene dada por la siguiente expresión:

$$SNR_q(\text{dB}) = 10 \cdot \log_{10} \frac{P_{PAM}}{P_{error}} \cong 6.02 \cdot b + 1.76 \quad (3.2)$$

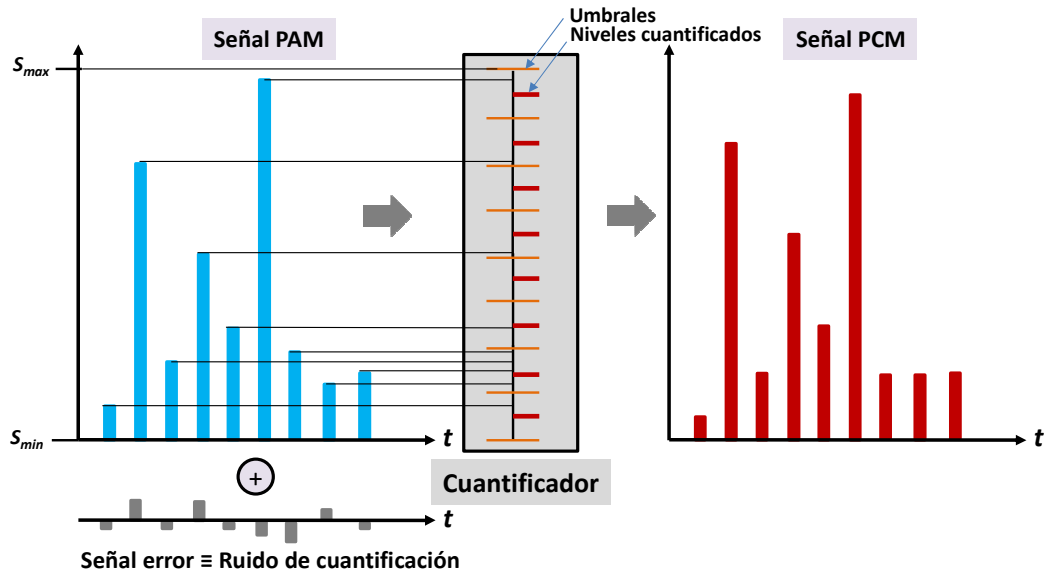


Figura 3.14. Cuantificador uniforme. La señal PCM a su salida puede verse como la suma de la señal PAM a su entrada y una hipotética señal error.

La relación señal-ruido es un concepto frecuente en comunicaciones, aunque en cada caso adopta una expresión específica. No obstante, cualquiera que sea el escenario, siempre interesa que sea lo mayor posible. Cualitativamente, el resultado dado por (3.2) no debe sorprendernos, pues indica que la relación señal-ruido de cuantificación es tanto mayor cuanto mayor es el número de dígitos binarios utilizados para representar cada nivel cuantificado, lo cual significa que el cuantificador trabaja con más niveles y, por tanto, su resolución es mayor.

A diferencia de lo que ocurría con la etapa de muestreo, la etapa de cuantificación introduce un error que hace irreversible la recuperación de la señal PAM a partir de la señal PCM. Evidentemente, cuanto menor sea dicho error, es decir, cuanto mayor sea la relación señal-ruido de cuantificación, mejor, pero en ningún caso el receptor podrá obtener una copia exacta de la señal PAM, y por tanto de la señal analógica original. En otras palabras, la conversión analógica-digital equivale a un proceso de **codificación de fuente analógica con pérdidas** (compresión con pérdidas).

El último elemento del proceso de conversión analógica-digital es el **codificador**, cuya función consiste en asignar un código binario de b dígitos a cada valor PCM y generar la correspondiente señal digital del tipo NRZ. La velocidad de transmisión R a la salida del codificador y, por tanto, del convertidor A/D, vendrá dada por la siguiente expresión:

$$R = f_s \cdot b \quad (3.3)$$

Las expresiones (3.1) - (3.3) conducen a algunos criterios a la hora de seleccionar los parámetros relevantes:

- Al aumentar la frecuencia de muestreo, aumenta la velocidad de la señal digital generada a la salida del convertidor, lo cual exige mayor ancho de banda del medio de transmisión (el concepto de ancho de banda de un medio de transmisión y su relación con la velocidad de transmisión se verán más adelante). Puesto que el ancho de banda de los medios de transmisión es un bien escaso, va a ser suficiente satisfacer la condición (3.1) con la mínima frecuencia de muestreo necesaria, es decir, la velocidad de Nyquist.
- La velocidad de la señal digital y la relación señal-ruido de cuantificación son directamente proporcionales al número de dígitos binarios que maneja el codificador, lo que obliga a fijar este parámetro a un valor intermedio que represente un balance adecuado entre resolución de la señal y consumo de ancho de banda del medio de transmisión (solución de compromiso).

La cuantificación uniforme descrita en la **Figura 3.14** tiene la desventaja de que el error absoluto cometido al cuantificar es independiente del valor de la señal PAM, lo que conduce a mayores errores relativos para valores pequeños de dicha señal que para valores grandes. Para conseguir un error relativo lo más homogéneo posible, en la llamada **cuantificación no uniforme** se impone un paso de cuantificación (separación entre dos umbrales consecutivos) tanto menor cuanto menor es el valor PAM. Una forma de implementar este tipo de cuantificación es comprimiendo la señal PAM antes de su paso por un cuantificador uniforme, de forma que el compresor y el cuantificador uniforme equivalen conjuntamente a un cuantificador no uniforme. En el extremo receptor, se invierte este proceso expandiendo la señal a la salida del convertidor D/A. El proceso completo de comprimir y expandir la señal se conoce como **companding** (contracción de *compression-expanding*). Los algoritmos de *companding* más utilizados son de tipo logarítmico, como es el caso de la **ley-A** (*A-law algorithm*), versión europea del estándar G.711 de ITU-T para codificación de audio, o la **ley-μ** (*μ-law algorithm*), versión de dicho estándar para Estados Unidos y Japón. Otro tipo de conversión analógico-digital es la llamada **modulación delta**, según la cual la señal analógica es muestreada por una función del tipo escalera, que únicamente sube o baja un peldaño en cada punto de muestreo en función de si la señal analógica ha aumentado o disminuido con respecto a la muestra anterior. La magnitud del peldaño es un valor prefijado. De esta forma, a cada muestra le corresponde un solo dígito binario (1 o 0 según se haya subido o bajado un peldaño), simplificándose así el proceso de transformación de la señal analógica en una señal NRZ. No obstante, la relación señal-ruido de cuantificación es peor que en el cuantificador uniforme para una misma velocidad.

CAPÍTULO 4. MEDIOS DE TRANSMISIÓN

La propagación de las señales tiene lugar a través de los **medios de transmisión**, pero estos no actúan como simples puentes transmisor-receptor, sino que, desafortunadamente, producen una serie de alteraciones que pueden llegar a degradar la decodificación de la información contenida en dichas señales. No todos los medios de transmisión producen las mismas alteraciones y en el mismo grado, y cada escenario sugiere el medio más adecuado. Esto también puede venir determinado por criterios prácticos de instalación y coste. El objetivo de este capítulo es poner de relieve estos aspectos de una forma sistematizada.

ÍNDICE

4.1.	CLASIFICACIÓN DE LOS MEDIOS DE TRANSMISIÓN	85
4.2.	PERTURBACIONES INTRODUCIDAS POR LOS MEDIOS DE TRANSMISIÓN	86
4.3.	MEDIOS DE TRANSMISIÓN GUIADOS	98
4.4.	TRANSMISIÓN INALÁMBRICA	110
4.5.	PERTURBACIONES ESPECÍFICAS DE LAS TRANSMISIONES INALÁMBRICAS	130

4.1. CLASIFICACIÓN DE LOS MEDIOS DE TRANSMISIÓN

El sistema de transmisión del modelo básico de comunicaciones puede estar formado por un único medio de transmisión o por una estructura mallada con múltiples medios de transmisión. Un **medio de transmisión** no es más que un camino físico por el que tiene lugar la propagación de señales entre un transmisor y uno o varios receptores. Una primera clasificación de los medios de transmisión permite distinguirlos entre guiados y no guiados. Los **medios de transmisión guiados** son aquellos en los que las señales viajan confinadas dentro de un material sólido. Genéricamente, hablamos en este caso de **transmisión por cable**. A su vez, los medios guiados pueden ser metálicos (típicamente de cobre), como por ejemplo el **par de hilos trenzados** o el **cable coaxial**, o no metálicos, como es el caso de la **fibra óptica**. Por el contrario, en un **medio de transmisión no guiado** la señal se dispersa libremente en todas las direcciones. Hablamos en este caso de **transmisión inalámbrica**, que puede ser a través del **vacío** (comunicaciones entre satélites) o de un fluido, sea **aire** (comunicaciones inalámbricas terrestres) o **agua de mar** (comunicaciones submarinas). La **Tabla 4.1** recoge esta clasificación de los medios de transmisión, junto con el tipo de señales que transportan y la velocidad a la que éstas se propagan. La velocidad de propagación de una señal, v_{prop} , no es más que el cociente entre el espacio recorrido, d , y el tiempo invertido en recorrerlo, t_{prop} :

$$v_{prop} = \frac{d}{t_{prop}} \quad (4.1)$$

Otro detalle significativo de la **Tabla 4.1** es la velocidad de propagación de las señales acústicas en el medio marino, que es superior a la velocidad con la que se propagan estas señales en el aire (unos 340 m/s), y además toma valores dentro de un margen en lugar de uno fijo. Ello es debido a que dicha velocidad depende de tres factores que pueden variar notablemente: temperatura, profundidad y concentración salina. En cualquier caso, el valor resultante es mucho menor que los alcanzados en los demás medios de transmisión, si bien la ventaja de utilizar señales acústicas en las comunicaciones submarinas es que se atenúan mucho menos que las señales electromagnéticas, por lo que su alcance es mucho mayor (con la excepción de las señales electromagnéticas de muy baja frecuencia que utilizaban los submarinos para comunicaciones de largo alcance). También es cierto que la velocidad de propagación en los medios guiados no es única, por lo que los valores que aparecen en la tabla deben interpretarse como típicos. Por ejemplo, en el caso de los medios metálicos, la velocidad de propagación depende de la frecuencia de trabajo y las características físicas del cable, entre ellas, precisamente, si se trata de cable coaxial o par trenzado: en el coaxial se acerca más a 250000 Km/s y en el par trenzado a 200000 Km/s; en el caso de la fibra óptica, la velocidad de propagación depende del índice de refracción del material utilizado, y además el valor de la tabla también se redondea a veces a 200000 Km/s.

La **Tabla 4.1** también indica una forma alternativa de expresar la rapidez con que un medio de transmisión propaga las señales. Consiste en expresar el tiempo que tarda la señal en recorrer la distancia de 1 Km, y se obtiene a partir de la fórmula anterior al calcular t_{prop} para $d = 1$ Km y la

velocidad de propagación que corresponda según el medio considerado. El valor que se obtiene facilita el cálculo del retardo de propagación para cualquier distancia distinta de 1 Km, con solo la realización de un producto en lugar de una división.

Tabla 4.1. Clasificación y características principales de los medios de transmisión.

Medio de transmisión		Tipo de señal	Velocidad de propagación	Tiempo de propagación por Km	VF
Guiado	Metálico	Eléctrica	230000 km/s	4.35 μs/km	0.77
	Fibra óptica	Óptica	214000 km/s	4.67 μs/km	0.71
No guiado	Vacío o aire	Electromagnética ²³	300000 km/s ²⁴	3.33 μs/km	1.00
	Agua de mar	Acústica	[1460 – 1520] m/s		

Finalmente, la última columna de la tabla se refiere al **factor de velocidad** o **VF** (*velocity factor*), que no es más que el cociente entre la velocidad de propagación en el medio de transmisión considerado y la velocidad de propagación de la luz, es decir:

$$VF = \frac{v_{prop}}{c} \quad (4.2)$$

4.2. PERTURBACIONES INTRODUCIDAS POR LOS MEDIOS DE TRANSMISIÓN

La propagación a través de un medio de transmisión está sometida a una variedad de perturbaciones que provocan que la señal recibida difiera de la que fue transmitida. En el caso de una transmisión digital, esto puede traducirse en errores en la decodificación de los dígitos binarios, es decir, en errores de bit. Por lo tanto, el éxito de una comunicación digital va a depender del control de estas perturbaciones, al menos para mantenerlas dentro de los límites tolerados por la aplicación o servicio que genera el tráfico de datos. En términos generales, cualquier medio de transmisión está sometido a cuatro clases de perturbaciones: atenuación, distorsión de atenuación, distorsión de retardo y ruido. En las transmisiones inalámbricas se dan, además, otras perturbaciones específicas por el hecho de que las señales se propagan por un medio abierto en el que tienen lugar muchos fenómenos fuera de control, ya sean naturales o artificiales. Esta sección se ocupa de las perturbaciones generales que afectan a todos los medios de transmisión, de los diferentes mecanismos de compensación o eliminación de estas perturbaciones, y de un dispositivo específicamente utilizado en transmisión digital para contrarrestar el conjunto de todas ellas: el repetidor regenerativo. Las perturbaciones específicas de la transmisión inalámbrica se abordan más adelante, una vez estudiado este tipo de transmisión.

²³ Habitualmente el término “electromagnético” se aplica a las señales emitidas por las antenas, si bien tanto las señales eléctricas como las ópticas son también electromagnéticas.

²⁴ Exactamente, la velocidad de propagación de la luz es 299792.458 Km/s.

Atenuación

Como ya vimos en la Sección 2.3, la **atenuación** (también llamada **pérdida de inserción**) es el debilitamiento que experimenta una señal a medida que se propaga por un medio de transmisión. Es siempre directamente proporcional a la distancia recorrida, pero tanto la causa como su formulación matemática son muy distintas según se trate de un medio guiado o no guiado.

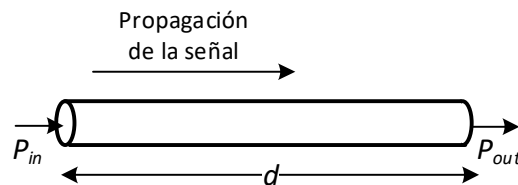


Figura 4.1. Potencias de entrada y salida en un tramo de medio guiado. A medida que se propaga, la señal va perdiendo potencia.

En el caso de un medio guiado, éste va absorbiendo y disipando en forma de calor parte de la energía de la señal a medida que ésta se propaga. Concretamente, si consideramos un tramo de longitud d de un medio de transmisión guiado, cuyas potencias de entrada y salida son, respectivamente, P_{in} y P_{out} (**Figura 4.1**), se cumple que la relación entre ambas es de tipo exponencial:

$$P_{out} = P_{in} \cdot \gamma^{-d} \quad (4.3)$$

En esta expresión, $\gamma > 1$ es un parámetro que caracteriza el medio de transmisión. Cuanto menor es dicho parámetro, mejor. A partir de (4.3) podemos obtener la expresión general de la atenuación en decibelios introducida por cualquier medio guiado:

$$L(\text{dB}) = 10 \cdot \log_{10} \frac{P_{in}}{P_{out}} = 10 \cdot \log_{10} \frac{P_{in}}{P_{in} \cdot \gamma^{-d}} = 10 \cdot \log_{10} \gamma^d = (10 \cdot \log_{10} \gamma) \cdot d \quad (4.4)$$

Obsérvese que se ha calculado la atenuación en positivo, con el numerador igual a la potencia de entrada y el denominador igual a la potencia de salida. El factor $10 \cdot \log_{10} \gamma$, mayor que 0 al ser el parámetro γ mayor que 1, es la atenuación en decibelios por unidad de distancia, pues al multiplicarse por la distancia d como resultado da como resultado la atenuación total en decibelios introducida por el tramo de cable. En efecto, si denotamos dicho factor por Γ , es decir, $\Gamma = 10 \cdot \log_{10} \gamma$, el resultado final mostrado en (4.4) puede reescribirse así:

$$L(\text{dB}) = \Gamma \cdot d \quad (4.5)$$

Como ya ocurría con γ , el parámetro Γ caracteriza el medio de transmisión y también cuanto menor mejor. Típicamente se toma el kilómetro como unidad de distancia, en cuyo caso Γ viene dada en dB/Km (decibelios por kilómetro). La expresión (4.5) pone de manifiesto una ventaja más de trabajar con decibelios, pues el cálculo de la atenuación introducida por cualquier medio guiado ha quedado reducido a un simple producto (en lugar de la potencia negativa de la expresión (4.3)).

Cuando se trata de un medio no guiado (transmisión inalámbrica), la atenuación es debida a la progresiva dispersión de la onda electromagnética a medida que se aleja del foco, es decir, de la antena transmisora. Un caso parecido, aunque en dos dimensiones, es la onda mecánica generada al lanzar una piedra en un estanque. No obstante, volviendo a la onda electromagnética, dado que la potencia recibida depende de las características de las antenas transmisora y receptora y otros parámetros, dejamos su formulación matemática para más adelante, cuando consideremos con más detalle las transmisiones inalámbricas.

Distorsión de atenuación

La distorsión de atenuación surge como consecuencia de que la atenuación introducida por los medios de transmisión es dependiente de la frecuencia. Si el medio de transmisión es guiado, esto equivale a decir que $\gamma = \gamma(f)$ y, lógicamente, $\Gamma = \Gamma(f)$, magnitud que como sabemos viene expresada en dB/Km o similar. Cuando el medio de transmisión es no guiado, veremos que no existe un parámetro equivalente en decibelios por unidad de distancia que lo caracterice, y directamente nos referiremos a su atenuación para una distancia determinada, también dependiente de la frecuencia, es decir, $L = L(f)$ en dB. Tanto $\Gamma(f)$ como $L(f)$ constituyen la llamada **respuesta frecuencial (de atenuación)** del medio. Así, al igual que el espectro de frecuencias caracterizaba unívocamente una señal (Capítulo 2), la respuesta frecuencial caracteriza unívocamente un medio de transmisión. Aunque ambas curvas son cualitativamente indistinguibles, no representan lo mismo: el espectro de frecuencias refleja la descomposición de una señal en múltiples señales sinusoidales a diferentes frecuencias, mientras que la respuesta frecuencial de un medio indica cómo responde ese medio a cada señal sinusoidal en función de su frecuencia. Por lo tanto, el papel de la respuesta frecuencial es determinante a la hora de analizar el efecto de un medio de transmisión sobre la señal que puede estar propagándose a su través. Lo vemos a continuación.

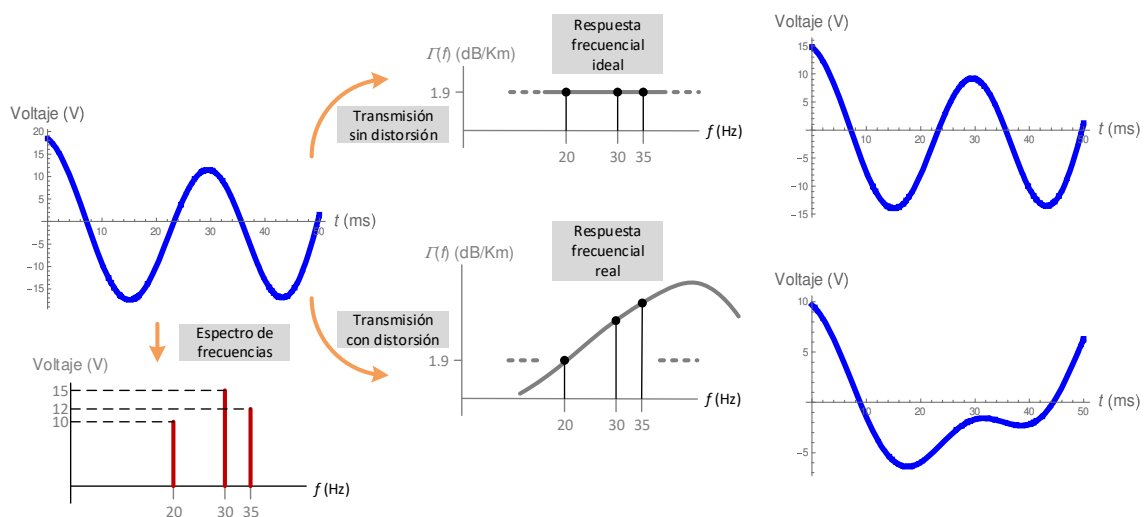


Figura 4.2. Una misma señal, a través de un medio guiado que sólo provoca atenuación o a través de otro que además provoca distorsión de atenuación, se transforma en resultados muy distintos. Estos resultados se han obtenido para 1 Km de distancia.

Si el efecto del medio se redujera a una misma atenuación para todas las frecuencias contenidas en la señal, es decir, si la curva $\Gamma(f)$ o $L(f)$ fuera plana al menos en la banda de frecuencias de dicha señal, ésta experimentaría un debilitamiento, pero no una deformación. Equivaldría a reducir el volumen de una locución de radio o televisión, sin que ello afectara a la inteligibilidad del mensaje y la correcta percepción de quien lo emite. Por el contrario, el hecho de que no todas las frecuencias contenidas en la señal sean atenuadas por igual, conlleva una deformación de la misma. Es decir, la señal transmitida no sólo experimenta un debilitamiento sino también una distorsión, llamada por ello **distorsión de atenuación**²⁵. La **Figura 4.2** ilustra la diferencia entre que se produzca o no distorsión de atenuación, para el caso de un medio guiado (la idea es la misma para un medio no guiado, con su $L(f)$ en dB en lugar de $\Gamma(f)$ en dB/Km). Para ello, se ha considerado una hipotética señal de voltaje cuyo espectro de frecuencias es discreto y está formado únicamente por 3 componentes frecuenciales, concretamente a 20, 30 y 35 Hz, cada una con su amplitud (10, 15 y 12 V respectivamente). Este espectro se muestra en la parte izquierda inferior de la figura, y sobre él se representa un intervalo de tiempo de 50 ms de dicha señal. En la transmisión a través de un medio que no introduce distorsión, todas las componentes frecuenciales de la señal de entrada experimentan la misma atenuación (1.9 dB/Km), y la señal que se obtiene a la salida aparece debilitada, pero conserva la forma original. Si el voltaje de entrada fuera producido por un sensor de temperatura calibrado, bastaría un simple recalibrado para obtener la lectura correcta de temperatura a partir de la señal de salida. Por el contrario, si el medio de transmisión no atenúa por igual las frecuencias de la señal (en el caso de la figura, las atenuaciones son de 1.9, 10.5 y 14 dB/Km para 20, 30 y 35 Hz, respectivamente), a la salida ésta aparece deformada, y no bastaría con un simple recalibrado para obtener la evolución real de temperatura.

La distorsión de atenuación puede conducir a errores en transmisión digital. Consideremos cualquiera de las señales mostradas en la **Figura 3.5**, correspondientes a la transmisión digital en banda base. Todas estas señales se construyen en base al pulso cuadrado como forma de onda elemental. El efecto de la distorsión de atenuación es precisamente un redondeo y un ensanchamiento de estos pulsos, cuyos flancos de subida y bajada se suavizan. Si ambos flancos, especialmente el de bajada, se alargan excesivamente en el tiempo (precisamente, cuanto más distorsión de atenuación, más se alargan), cada pulso pasa a “cabalgar” sobre las colas de los que le preceden, pudiendo llegar a dificultar la identificación y secuenciación correcta de los símbolos en el receptor. Este fenómeno se conoce como **interferencia inter-simbólica** (ISI), ya que son los mismos símbolos de una transmisión digital los que se interfieren mutuamente. Por lo tanto, la distorsión de atenuación no es tanto un problema por la deformación que se produce en la señal (el objetivo de la transmisión digital no es reproducir a la entrada del receptor exactamente la misma forma de onda que a la salida del transmisor, sino detectar correctamente los símbolos), sino por el grado de ISI y la tasa de errores de bit a que puede dar lugar. La transmisión digital modulada no está exenta de ISI, pues en caso de que la respuesta frecuencial no sea plana en la banda de

²⁵ Esta distorsión es distinta de la que producen los amplificadores cuando trabajan fuera de su zona lineal (Sección 2.3). De hecho, a esta última se la conoce como **distorsión no lineal**, mientras que la primera es conocida también distorsión lineal de atenuación.

frecuencias centrada en la frecuencia portadora, los pulsos de la señal NRZ a la salida del demodulador muestran el mismo tipo de redondeo y ensanchamiento comentados más arriba.

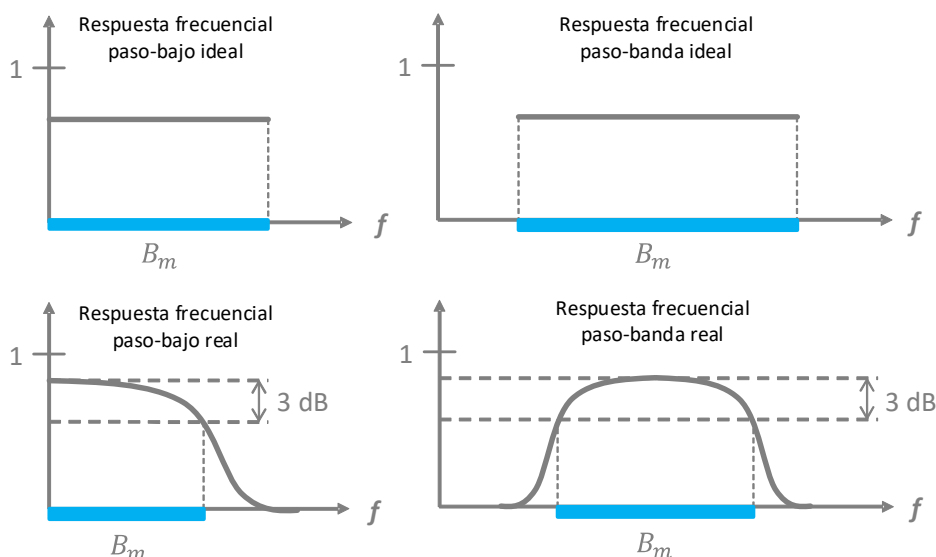


Figura 4.3. Tipos de respuesta frecuencial de los medios de transmisión, en sus versiones ideal y real. Se indica también el concepto de ancho de banda de un medio de transmisión. Los perfiles de las respuestas reales pueden ser más irregulares.

Al igual que ocurría con las señales, cuyo espectro podía ser paso-bajo o paso-banda, la respuesta frecuencial de un medio de transmisión también admite la misma diferencia. Así, una **respuesta frecuencial paso-bajo** es aquella que es plana en la parte baja del eje de frecuencias, desde 0 hasta una frecuencia máxima dada. En el caso ideal, es perfectamente plana; en el caso real, presenta una cierta curvatura, pero a efectos prácticos se considera plana hasta la llamada frecuencia de corte a 3 dB, que es la frecuencia a la que la respuesta frecuencial cae 3 dB por debajo de su valor máximo, que típicamente se da a la frecuencia nula. La **Figura 4.3**, en su parte izquierda, muestra estos dos casos (conviene destacar que todas las gráficas mostradas en esta figura corresponden a la respuesta frecuencial expresada como factor multiplicativo menor que 1, y no como atenuación en decibelios). En la parte derecha de la misma figura se muestra la **respuesta frecuencial paso-banda**, que en el caso ideal es plana en un margen intermedio de frecuencias, entre una frecuencia mínima y una máxima; de forma análoga, se muestra también la versión real de la misma. Los dos tipos de respuestas ponen de manifiesto que un medio de transmisión se comporta como un **filtro**, pues hay frecuencias para las cuales la respuesta es nula o casi nula, y por consiguiente las componentes sinusoidales a estas frecuencias son eliminadas por el medio. De hecho, como dispositivo que es, un filtro puede ser de dos clases: **filtro paso-bajo** y **filtro paso-banda**. Finalmente, un concepto nuevo es el **ancho de banda de un medio de transmisión**, denotado por B_m en la **Figura 4.3**, que no es más que la anchura de la respuesta frecuencial del medio, de la misma manera que el ancho de banda de una señal era la anchura de su espectro. Tratándose de un medio de transmisión (o de un

filtro en general), el ancho de banda es la magnitud de su banda de paso, es decir, del margen de frecuencias que deja pasar.

Distorsión de retardo

La distorsión de retardo solamente se produce en los medios guiados (par trenzado, coaxial y fibra óptica), y se debe al hecho de que la velocidad de propagación a través de un conductor sólido también depende de la frecuencia. Como una señal está formada por múltiples componentes a diferentes frecuencias, unas viajan a mayor velocidad que otras dentro del conductor. Nuevamente, el efecto es una distorsión de la señal, ya que las componentes sinusoidales que la constituyen pasan a superponerse de otra manera, consecuencia de haber recorrido una misma distancia a velocidades distintas. Por ello hablamos de **distorsión de retardo** (también llamada distorsión de fase), la cual provoca un nivel de ISI en transmisión digital incluso más alto que el provocado por la distorsión de atenuación.

Ruido

Se entiende por ruido cualquier señal no deseada que se superpone a la señal deseada en el camino que va del transmisor al receptor. En comunicaciones se dan muchas formas de ruido, incluyendo el llamado ruido inteligible, término que engloba cualquier tipo de interferencia. En efecto, una **interferencia** es una señal no deseada, pero su origen es la fuente de otro sistema de comunicaciones cercano que puede estar generando una señal en la misma banda de frecuencias y del mismo tipo que la señal deseada. De todos modos, sin duda el ruido más importante en comunicaciones es el ruido térmico, y por ello centraremos nuestra atención en esta clase de ruido.

El **ruido térmico** se debe a la agitación térmica de los electrones libres en los conductores y semiconductores, y por ello está siempre presente en los medios de transmisión metálicos y los circuitos electrónicos. Como los electrones libres son cargas eléctricas en movimiento, producen una señal eléctrica, y como ese movimiento es aleatorio, también lo es dicha señal. La **Figura 4.4** muestra una ráfaga típica de ruido térmico, que nos permite comprobar que efectivamente se trata de una sucesión de oscilaciones completamente aleatorias. No obstante, el valor medio de estas oscilaciones es nulo, pues es nulo el desplazamiento neto de los electrones en su agitación puramente térmica.

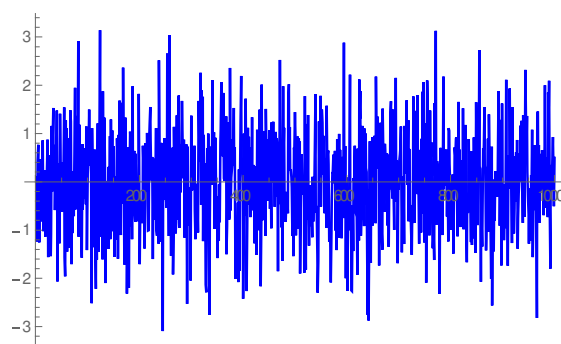


Figura 4.4. Ejemplo de ráfaga de ruido térmico.

Desde el punto de vista espectral, la caracterización del ruido térmico es muy sencilla, pues al estar siempre presente y en el mismo grado en todas las bandas de frecuencia, dicho ruido queda descrito mediante una densidad espectral de potencia plana, tal como se muestra en la **Figura 4.5**. La magnitud N_0 que aparece en la figura solamente varía con la temperatura, siendo linealmente proporcional a ella. Por tanto, la **densidad espectral de potencia del ruido térmico** queda expresada así:

$$\eta(f) = N_0 = k \cdot T, \forall f \quad (4.6)$$



Figura 4.5. Densidad espectral de potencia del ruido térmico. Al ser plana, al ruido térmico se le conoce también como ruido blanco, por similitud con la luz blanca que contiene todos los colores.

En dicha expresión, T es la temperatura ambiente expresada en kelvin (temperatura absoluta) y k es la constante de Boltzmann, de valor $1.38 \cdot 10^{-23} \text{ J/K}$. Si la temperatura viene dada en grados centígrados, basta sumarle el valor 273 para obtener su valor en kelvin.

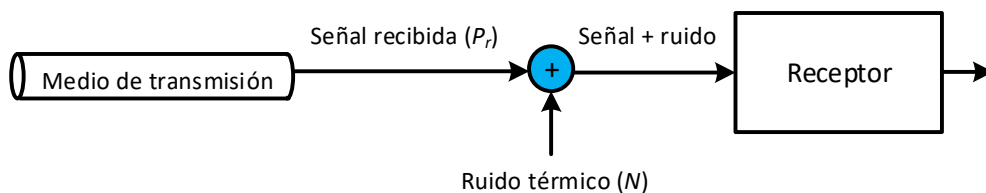


Figura 4.6. Modelo que describe la superposición del ruido térmico sobre la señal recibida (señal deseada en el punto de entrada al receptor).

Para cuantificar el efecto del ruido térmico sobre la señal deseada, consideremos el caso sencillo de un transmisor y un receptor conectados directamente mediante un conductor metálico como medio de transmisión (par de hilos trenzados o coaxial). Supongamos además que sólo hay una fuente de ruido, y que ésta es el propio conductor. Aunque el ruido térmico se origina de forma distribuida a lo largo de todo el medio, a efectos de modelado se representa como una señal externa que se suma a la señal deseada a la entrada del receptor. Esto se muestra en la **Figura 4.6**, en la que se denota la señal deseada a la entrada del receptor como **señal recibida** (convenio terminológico que adoptaremos de ahora en adelante, aunque ciertamente el receptor recibe la señal deseada superpuesta con ruido). Sobre la base de este modelo, se define la relación señal-ruido a la entrada del receptor como el cociente, transformado a decibelios, entre la potencia media de señal recibida (P_r) y la potencia media de ruido (N):

$$SNR(dB) = 10 \cdot \log_{10} \frac{P_r}{N} \quad (4.7)$$

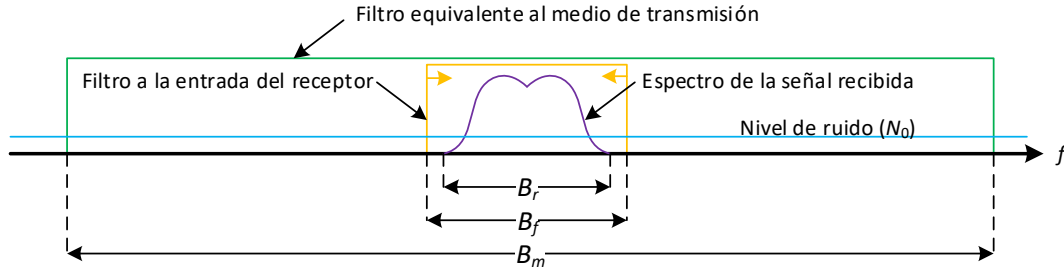


Figura 4.7. El filtro del receptor tiene por objeto eliminar la máxima potencia de ruido superpuesto a la señal recibida. No puede eliminar, sin embargo, la potencia de ruido en la banda de la señal.

El valor de P_r se calcula a partir de la potencia media transmitida y la atenuación introducida por el medio de transmisión. Para calcular la potencia media de ruido, hay que tener en cuenta que el primer elemento que aparece a la entrada del receptor es un filtro centrado en el espectro de la señal recibida (que es igual al espectro de la señal transmitida), tal como se ilustra en la **Figura 4.7**. Este filtro tiene como objetivo eliminar la máxima potencia de ruido posible, al tiempo que debe dejar pasar el espectro de la señal recibida para no distorsionarla. Por tanto, la solución óptima es que el ancho de banda del filtro de entrada al receptor, B_f , sea igual al ancho de banda del espectro de la señal recibida, B_r , es decir, $B_f = B_r$. Este ancho de banda es el llamado **ancho de banda del canal**, el cual se representa por B_c y es generalmente mucho menor que el ancho de banda del medio de transmisión (B_m). De todo ello se deduce que la potencia media de ruido que perturba la señal recibida es la contenida en el ancho de banda del canal, por lo que, aplicando la fórmula (2.28), se tiene:

$$N = \int_{B_c} \eta(f) \cdot df = \int_{B_c} N_0 \cdot df = N_0 \cdot B_c = k \cdot T \cdot B_c \quad (4.8)$$

Al ser la densidad espectral constante en el ancho de banda del canal, el resultado no es más que el área de un rectángulo de lados N_0 y B_c . En unidades logarítmicas:

$$N(dBW) = 10 \cdot \log k + 10 \cdot \log T + 10 \cdot \log B_c = -228.6 + 10 \cdot \log T + 10 \cdot \log B_c \quad (4.9)$$

En la última fórmula, el valor -228.6 proviene de expresar la constante de Boltzmann en formato logarítmico, y la temperatura T y el ancho de banda B_c tienen que estar expresados, respectivamente en kelvin y hertzios. Combinando las expresiones (4.7) y (4.9) tenemos:

$$SNR(dB) = 10 \cdot \log_{10} \frac{P_r}{k \cdot T \cdot B_c} = P_r(dBW) - N(dBW) = P_r(dBW) + 228.6 - 10 \cdot \log T - 10 \cdot \log B_c \quad (4.10)$$

Si la única fuente de ruido es el medio de transmisión, la relación señal-ruido a la salida del receptor es igual a la relación señal-ruido a su entrada - dada por (4.10). El motivo es que la ganancia o pérdida que pueda introducir el receptor afecta por igual a la señal y al ruido. No obstante, es sabido que

cualquier dispositivo degrada la relación señal-ruido, de modo que su valor a la salida es siempre peor (menor) que a la entrada. Esto se debe a que el propio dispositivo genera ruido, y por ello se conoce como **ruido interno**. Su potencia es N_{int} . Para cuantificar la influencia del ruido interno en la relación señal-ruido, supongamos que el receptor se caracteriza por un factor multiplicativo μ_r , ya sea $\mu_r > 1$ (efecto de amplificación) o $\mu_r < 1$ (efecto de atenuación). Trabajando con unidades lineales, formulemos en primer lugar la relación señal-ruido a la entrada del receptor:

$$SNR_{in}|_{lineal} = \frac{P_r}{N} \quad (4.11)$$

A la salida, la relación señal-ruido vendrá dada por la expresión que resulta de considerar la generación de ruido interno por parte del receptor:

$$SNR_{out}|_{lineal} = \frac{P_r \cdot \mu_r}{N \cdot \mu_r + N_{int}} = \frac{P_r}{N + \frac{N_{int}}{\mu_r}} \quad (4.12)$$

Observemos que el término N_{int}/μ_r no es más que la potencia media del ruido interno revertido a la entrada del receptor, es decir, como si se tratara de un ruido entrante adicional. Por lo tanto, cuando el receptor es ruidoso, el modelo de la **Figura 4.6** puede modificarse por el mostrado en la **Figura 4.8**, en el que se ha reemplazado el receptor real por un receptor equivalente sin ruido (receptor ideal) más el ruido interno revertido a la entrada. Finalmente, todo el conjunto se puede substituir por el modelo de la **Figura 4.9**. Conviene recalcar que éste es un modelo útil para el cálculo de la relación señal-ruido a la salida del receptor, pero basado en dos elementos ficticios: el receptor ideal y el ruido interno revertido a la entrada.

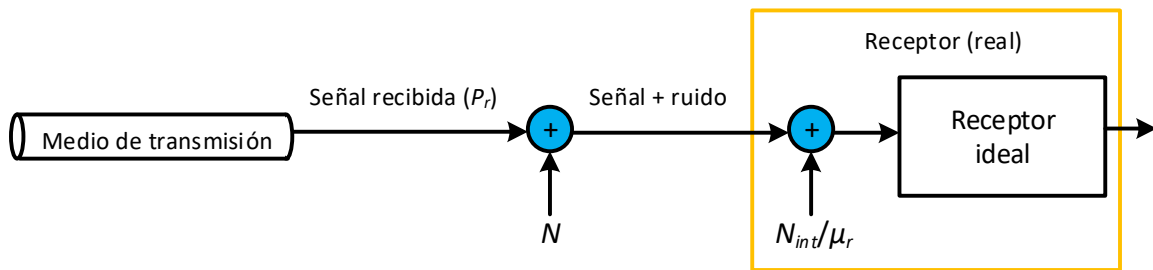


Figura 4.8. Modelo que describe la superposición de ruido sobre la señal recibida en el punto de entrada a un receptor real (ruidoso).

El siguiente paso tendente a facilitar los cálculos consiste en equiparar el ruido interno revertido a la entrada con un ruido térmico a cierta temperatura, llamada **temperatura equivalente de ruido** (T_{eq}):

$$\frac{N_{int}}{\mu_r} = k \cdot T_{eq} \cdot B_c \quad (4.13)$$

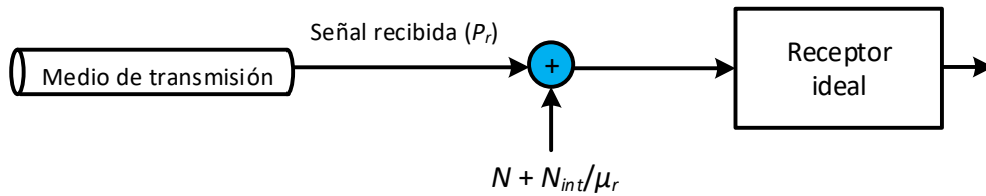


Figura 4.9. Modelo para el cálculo de la relación señal-ruido a la salida de un receptor ruidoso. Se pierde la noción de señal recibida, y no por tanto no permite calcular la relación señal-ruido a la entrada de ese receptor.

La temperatura equivalente de ruido es un dato aportado generalmente por el fabricante del dispositivo. En términos de este parámetro, la relación señal-ruido a la salida del receptor puede reformularse como sigue:

$$SNR_{out}|_{lineal} = \frac{P_r \cdot \mu_r}{(k \cdot T \cdot B_c + k \cdot T_{eq} \cdot B_c) \cdot \mu_r} = \frac{P_r}{k \cdot (T + T_{eq}) \cdot B_c} \quad (4.14)$$

Otro parámetro que también caracteriza la degradación de la relación señal-ruido introducida por un dispositivo es el llamado **factor de ruido** f , definido como sigue:

$$f = \frac{SNR_{in}|_{lineal}}{SNR_{out}|_{lineal}} = \frac{\frac{P_r}{k \cdot T \cdot B_c}}{\frac{P_r}{k \cdot (T + T_{eq}) \cdot B_c}} = 1 + \frac{T_{eq}}{T} \quad (4.15)$$

Como podemos observar, el factor de ruido es siempre mayor que 1, y su valor depende de la temperatura ambiente (T). El fabricante suele proporcionar el factor de ruido para la llamada **temperatura ambiente de referencia**, T_0 , cuyo valor es 290 K. Así pues, salvo que se especifique lo contrario, el factor de ruido que adoptaremos como dato aportado por el fabricante es el siguiente:

$$f = 1 + \frac{T_{eq}}{T_0} \quad (4.16)$$

Por tanto, la temperatura equivalente de ruido en función del factor de ruido queda de la siguiente manera:

$$T_{eq} = (f - 1) \cdot T_0 \quad (4.17)$$

El fabricante también puede proporcionar el factor de ruido en decibelios, en cuyo caso hablamos de **figura de ruido** F :

$$F(\text{dB}) = 10 \cdot \log_{10} f = SNR_{in}(\text{dB})|_{T_0} - SNR_{out}(\text{dB})|_{T_0} \quad (4.18)$$

En definitiva, la temperatura equivalente de ruido, el factor de ruido y la figura de ruido son tres alternativas utilizadas por el fabricante para cuantificar la degradación de la relación señal-ruido introducida por un dispositivo.

Compensación de las perturbaciones

Cada una de las perturbaciones que acabamos de considerar tiene su mecanismo de compensación, aplicable tanto si la señal sujeta a estas perturbaciones es analógica como si es digital. Los mecanismos de compensación son los siguientes:

Amplificación. La amplificación de señal sirve para compensar total o parcialmente su atenuación. El dispositivo que lo implementa es el **amplificador**. Bien con un solo amplificador o con una serie de amplificadores en cascada, el objetivo final es garantizar que el nivel de la señal recibida satisfaga dos condiciones: (a) su valor esté por encima de la sensibilidad del receptor, de forma que la señal pueda ser detectada por éste, y (b) su valor sea lo suficientemente elevado como para satisfacer una cierta relación señal-ruido umbral a la entrada del receptor.

Ecualización de amplitud. La ecualización de amplitud sirve para compensar la distorsión de atenuación. El objetivo es que el conjunto formado por el medio de transmisión y el **ecualizador de amplitud** presente una respuesta frecuencial más o menos plana en la banda de frecuencias de la señal. El ejemplo típico de ecualización de amplitud es el de la línea telefónica para la voz analógica, cuyo espectro promedio abarca la banda que va de 300 Hz a 3400 Hz. Como puede observarse en la **Figura 4.10(a)**, la atenuación introducida por la línea telefónica (par de hilos trenzados) no es plana con respecto a la frecuencia, sino muy curvada, con valores mucho mayores en las frecuencias altas que en las bajas. De no hacer nada, la señal de voz recibida por el abonado (usuario) que escucha sería muy distinta de la generada por el abonado que habla, hasta el punto de comprometer la inteligibilidad del mensaje y/o el reconocimiento de quien lo emite. Para compensar este comportamiento, la técnica implementada en el ecualizador de amplitud puede variar, pero las alternativas son básicamente dos: realzar las frecuencias más perjudicadas o atenuar aquellas más favorecidas. Se obtiene así una respuesta frecuencial más horizontal, como también se pone de manifiesto en la **Figura 4.10(a)**. Observamos que el resultado no es perfecto, pero es suficiente como para que la distorsión de atenuación resultante sea imperceptible por el abonado que escucha.

Ecualización de retardo. La ecualización de retardo sirve para compensar la distorsión de retardo, y en este caso el objetivo es que el conjunto formado por el medio de transmisión y el **ecualizador de retardo** genere más o menos el mismo retardo para todas las frecuencias contenidas en la banda de la señal. Independientemente de cuál sea su implementación, el diseño del ecualizador de retardo se dirige a retrasar artificialmente las frecuencias que viajan a mayor velocidad dentro del cable. Un ejemplo de ecualización de retardo es nuevamente el de la línea telefónica, como se muestra en la **Figura 4.10(b)**. También comprobamos que la ecualización no es perfecta, pero al igual que con la ecualización de amplitud, el resultado es satisfactorio para el oído humano.

Filtrado en recepción. El filtrado en recepción sirve para reducir, aunque no eliminar, el nivel de ruido térmico que se superpone a la señal recibida. En efecto, la **Figura 4.7** nos puso de manifiesto que utilizando un filtro a la entrada del receptor cuyo ancho de banda coincida con el ancho de banda de la señal, reducimos al máximo posible la potencia de ruido térmico entrante al tiempo que preservamos la integridad de dicha señal.

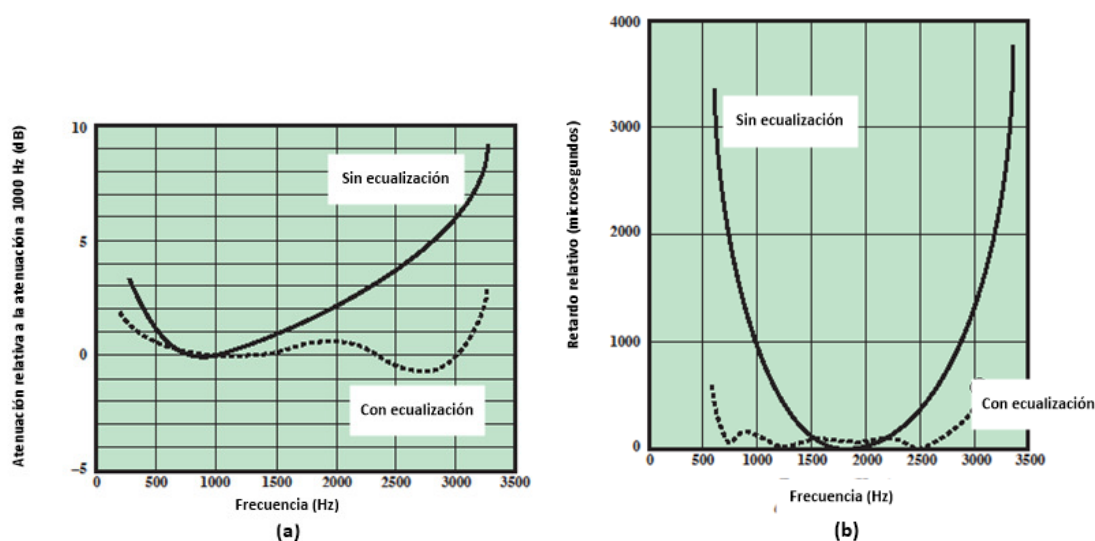


Figura 4.10. Ecualización de la línea telefónica (*voice-grade telephone line*): (a) respuesta frecuencial de atenuación, con y sin ecualización de amplitud, relativa a la frecuencia de 1000 Hz, la más favorecida por la línea telefónica; (b) respuesta frecuencial de retardo, con y sin ecualización, relativa a la frecuencia que viaja más rápido a través del par de hilos trenzados.

El repetidor regenerativo

El **repetidor regenerativo** es un dispositivo exclusivo de la transmisión digital. En una primera etapa ejecuta los mecanismos de compensación y eliminación de las perturbaciones que acabamos de considerar (una especie de “todo en uno”), y a partir de ahí regenera la transmisión. Su estructura básica, para el caso de una transmisión digital en banda base, se muestra en la **Figura 4.11**. En ella podemos apreciar que el proceso que tiene lugar entre la entrada y la salida del dispositivo se puede desglosar en tres etapas clave:

- Según lo indicado, una primera etapa de filtrado, amplificación y ecualización de la señal recibida, con el fin de contrarrestar el ruido, la atenuación y la distorsión respectivamente.
- A partir de la señal depurada resultante de la etapa anterior, el temporizador extrae la señal de reloj imbricada en la misma (recordemos que esta imbricación era uno de los requisitos que debían satisfacer las codificaciones de línea), y la utiliza como base de tiempos para muestrear la señal de datos (señal depurada) y detectar la secuencia de bits.
- A partir de la secuencia de bits, el regenerador reconstruye la señal digital según la codificación de línea utilizada.

En el caso de que la transmisión digital fuera modulada, la estructura del repetidor regenerativo sería la misma que la mostrada en la **Figura 4.11**, pero incluyendo una etapa de demodulación a continuación del bloque de filtrado, amplificación y ecualización. Tanto si la transmisión digital es en banda base como si es modulada, la estructura de un repetidor regenerativo es la de un receptor digital más la etapa de regeneración.

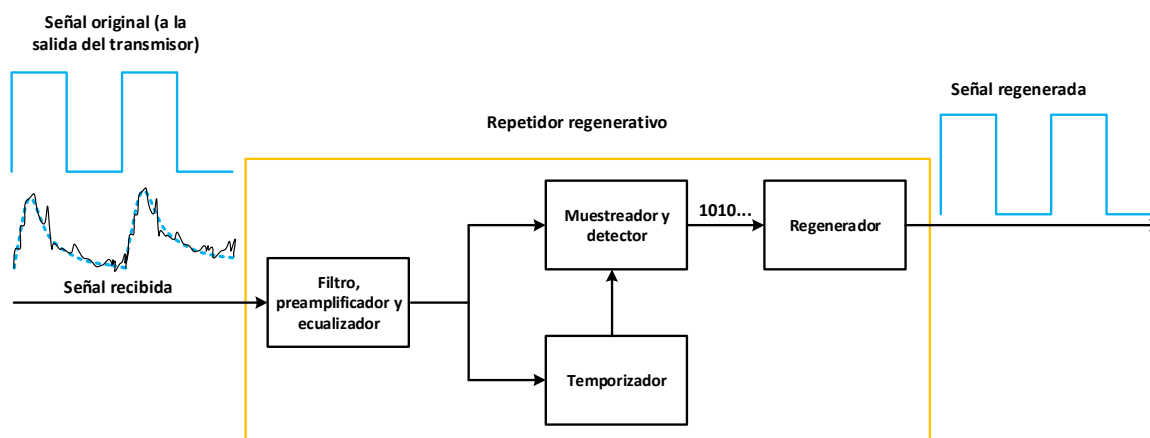


Figura 4.11. Estructura de un repetidor regenerativo para una transmisión digital en banda base. La señal recibida resulta de la atenuación y distorsión de la señal a la salida del transmisor y la superposición de ruido. A la salida del repetidor regenerativo, se tiene de nuevo una señal digital nítida, teóricamente igual a la que había enviado el transmisor.

El repetidor regenerativo tiene la ventaja de que las perturbaciones presentes en la señal de entrada, si conjuntamente no producen una degradación excesiva, quedan completamente eliminadas a la salida, como si la transmisión empezara de nuevo. Es decir, el repetidor regenerativo actúa como un corta-fuegos contra los efectos perturbadores, evitando que estos se acumulen de un repetidor a otro si están conectados en cascada. No obstante, esto también tiene un riesgo: si la distancia entre dos repetidores es demasiado grande y la señal recibida por el segundo es muy débil, la señal regenerada por éste puede contener errores de bit que muy posiblemente se mantendrán hasta el final (destinatario), de modo que el efecto corta-fuegos se torna en desventaja.

4.3. MEDIOS DE TRANSMISIÓN GUIADOS

Los medios de transmisión guiados pueden ser de tres tipos: par de hilos trenzados, cable coaxial y fibra óptica. Sólo los dos primeros son metálicos. A continuación, se describen cada uno de ellos, en términos de sus características físicas y su respuesta en frecuencias.

Par de hilos trenzados

Es el medio de transmisión guiado más barato, manejable y ampliamente utilizado. El **par de hilos trenzados** consiste en dos alambres de cobre aislados, cuyo diámetro varía entre 0.4 y 0.9 mm, que se cruzan entre sí siguiendo un patrón regular de forma helicoidal. Generalmente se agrupan 2, 4 o 25 pares bajo una misma funda o cubierta para formar un cable (aunque puede haber cables con muchos más pares). Los hilos de un mismo par están entrecruzados o trenzados con el objeto de reducir la **interferencia electromagnética** o EMI (*electromagnetic interference*), tanto la que pueda provocar ese par sobre pares vecinos o sistemas de comunicaciones externos, como la que puedan provocar sobre él los pares vecinos o las fuentes externas (motores eléctricos, tormentas con carga eléctrica, etc.). El par de hilos trenzados empezó a utilizarse como medio de transmisión en

telefonía, mediante un cable de 2 pares en el bucle de abonado y uno o varios cables de 25 pares en enlaces de mayor capacidad. Posteriormente, se extendió el uso del par trenzado a la transmisión de datos además de voz, esta vez mediante el cable de 4 pares, que se constituyó en la base del cableado de los edificios de oficinas.

Con el objetivo de proporcionar unas recomendaciones prácticas para el diseño e instalación de los sistemas de cableado en entornos corporativos, de forma que tales sistemas pudieran soportar la amplia variedad de servicios de transmisión de datos existentes y futuros, así como la multiplicidad de fabricantes de los equipos a interconectar, la EIA²⁶ publicó en 1991 el primer estándar de cableado estructurado (EIA-568). El estándar actual, ANSI²⁷/TIA²⁸-568, contiene una de las normalizaciones más extendidas sobre cables de pares trenzados, además de abarcar otros medios de transmisión guiados, como la fibra óptica y el cable coaxial. La **Tabla 4.2** muestra las diversas categorías de cables de pares trenzados desarrolladas hasta la actualidad en el contexto de este estándar americano, que es compatible con el estándar internacional ISO²⁹/IEC³⁰ 11801 sobre sistemas de cableado para telecomunicaciones³¹. Las prestaciones, en términos de ancho de banda y velocidad de transmisión (más adelante veremos la relación directa entre ambas magnitudes), aumentan con el índice de la categoría, por lo que dicho índice es fiel reflejo del estado actual de la tecnología en par trenzado, una tecnología que ha ido dando respuesta a las exigencias cada vez mayores de las aplicaciones y servicios de transmisión de datos.

La **Tabla 4.2** señala dos clases de cables de pares trenzados: UTP (*unshielded twisted pair*), que típicamente presenta una impedancia característica³² de 100 Ω , y STP (*shielded twisted pair*), cuya impedancia característica es de 150 Ω . El cable STP, a diferencia del cable UTP, incluye un envoltorio metálico que rodea el conjunto de todos los pares de hilos, con el objeto de frenar mucho más las interferencias desde/hacia el exterior. Dicho envoltorio puede consistir en una cinta de cobre o aluminio en espiral, una lámina de polímero conductor o una malla metálica. La **Figura 4.12** ilustra la diferencia entre un cable UTP y uno STP. Los cables STP presentan mejores prestaciones que los UTP, pero resultan más caros y difíciles de manejar, de ahí que UTP ha sido la opción dominante en la preinstalación de cableado en los edificios de oficinas cuando ésta se ha basado en el par trenzado. Concretamente, el formato de cable utilizado en estos escenarios es el de 4 pares, que se conectan a los dispositivos y a las tomas de pared mediante el conector RJ-45. Ésta fue también la solución adoptada para el bucle de abonado ADSL, cuando substituyó el antiguo bucle de abonado

²⁶ EIA: Electronic Industries Alliance.

²⁷ ANSI: American National Standards Institute.

²⁸ TIA: Telecommunications Industry Association.

²⁹ ISO: International Organization for Standardization.

³⁰ IEC: International Electrotechnical Commission.

³¹ Algunas categorías sólo están o han estado reconocidas en el estándar ISO/IEC 11801, como es el caso de las categorías 1 y 2, anteriores al estándar americano y actualmente obsoletas, y las categorías 7 y 7A.

³² La impedancia característica de un medio metálico es esencialmente la relación entre voltaje y corriente a través de ese medio. La resistencia equivalente del receptor debe coincidir con la impedancia característica del cable (adaptación de impedancia), pues de lo contrario parte de la señal que llega al receptor se reflejaría sobre el cable (fenómeno de eco).

telefónico basado en el cable de 2 pares (categorías 1 y 2 ya obsoletas) y el correspondiente conector RJ-11.

Tabla 4.2. Categorías de cables de pares según el estándar ANSI/TIA-568. Se especifican los tipos de cables en cuanto a diseño y construcción, el ancho de banda y las aplicaciones típicas que corresponden a cada categoría. La categoría F/UTP también es conocida como ScTP (Sc: *screened*).

Categoría	Tipo de construcción (designación ISO/IEC 11801)		Ancho de banda (MHz)	Aplicaciones
	Cableado UTP	Cableado STP		
1	UTP		0.4	Telefonía (antiguo bucle de abonado)
2	UTP		4	Redes LocalTalk, antiguos terminales IBM, líneas T-1
3	UTP		16	ADSL, LAN
4	UTP		20	LAN
5	UTP		100	LAN
5e	UTP, U/FTP	F/UTP	100	LAN
6	UTP, U/FTP	F/UTP	250	LAN
6A	UTP, U/FTP	F/UTP, S/FTP	500	LAN
7		F/FTP, S/FTP	600	LAN
7A		F/FTP, S/FTP	1000	LAN
8/8.1	U/FTP	F/UTP	2000	LAN
8.2		F/FTP, S/FTP	2000	LAN

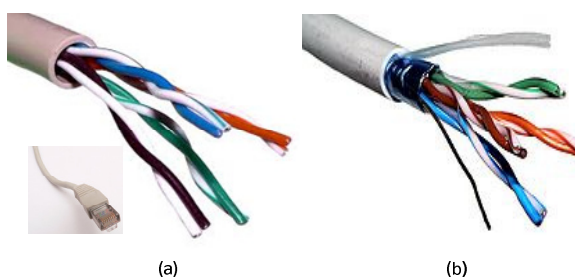


Figura 4.12. Cables de 4 pares: (a) cable UTP y conector RJ-45 y (b) cable STP. Obsérvese el escudo o pantalla metálica en el caso del cable STP. Además, en ambos cables se aprecia que la longitud del trenzado varía de un par a otro, con el objeto de reducir todavía más la interferencia entre ellos.

En la **Tabla 4.2** podemos observar también que tanto el cableado UTP como el STP se dividen en subclases más específicas, dependiendo de la técnica concreta aplicada en su construcción. La nomenclatura de todas las opciones obedece al formato X/YTP del estándar ISO/IEC 11801. Concretamente, la X caracteriza el envoltorio metálico que rodea todos los pares del cable, mientras que la Y se refiere al envoltorio, también metálico, de cada par de hilos individual dentro del cable. Las opciones para X son U (*unshielded*: ausencia de envoltorio), F (*foil shield*: envoltorio del tipo

cinta, capa o papel) y S (*braided shield*: envoltorio del tipo malla metálica); las opciones para Y son las mismas, excepto la S. En general, cada categoría incluye diversas variantes tanto de clase UTP como de clase STP, aunque algunas sólo contemplan una de las dos.

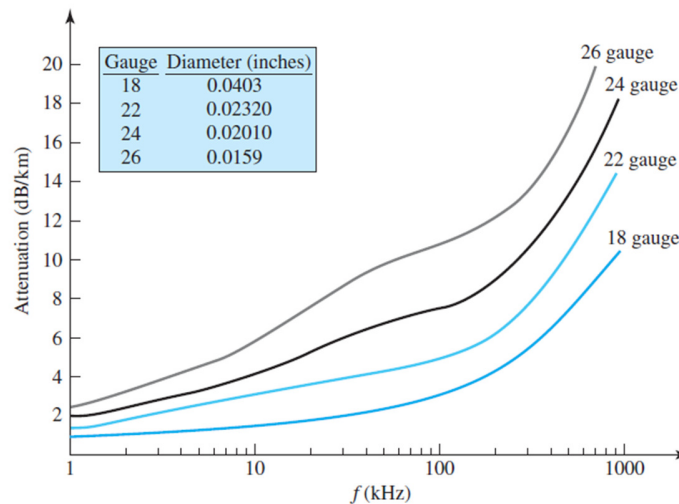


Figura 4.13. Comportamiento de UTP en función de la frecuencia para diversos calibres.

La atenuación por kilómetro introducida por un par de hilos trenzados depende de sus características técnicas, entre ellas el grosor o calibre (de cada conductor), y de la frecuencia. No existe una correspondencia directa entre categorías de par trenzado y calibres, sino que en general cada categoría admite diversas opciones dependiendo del número de pares contenidos en el cable y la distancia a cubrir. La **Figura 4.13** muestra las curvas de atenuación en función de la frecuencia para 4 calibres normalizados en el estándar americano AWG (*American wire gauge*), que data del año 1857, y que asocia índices crecientes a calibres decrecientes. El comportamiento mostrado corresponde al caso en que no se usa ningún tipo de blindaje metálico, es decir, cable UTP. Como podemos observar, la atenuación crece con la frecuencia cualquiera que sea el calibre, experimentando en todos los casos una subida abrupta a partir de los 100 KHz aproximadamente. Fijada la frecuencia, observamos también que la atenuación decrece al aumentar el calibre.

Cable coaxial

El **cable coaxial** consta de dos conductores concéntricos separados por un material dieléctrico, típicamente de polietileno, que ejerce de aislante. El conductor exterior es habitualmente una malla metálica formada por hilos de cobre trenzados, aunque también puede consistir en un envoltorio de papel de aluminio, o en la combinación de ambas opciones. En cualquier caso, el conductor exterior actúa también como pantalla contra las interferencias externas. La **Figura 4.14** muestra la estructura típica de este tipo de cables, donde podemos observar que todo el conjunto está protegido por una cubierta (funda termoplástica). Este tipo de construcción otorga al cable coaxial mejores prestaciones que las del cable UTP, tanto en ancho de banda como en atenuación por unidad de longitud, ventajas que se traducen, respectivamente, en mayor velocidad de transmisión

y mayor alcance. No obstante, su mayor coste y menor manejabilidad, junto al hecho de que la tecnología del cable UTP ha ido avanzando continuamente, han contribuido a que este último haya mantenido su posición dominante durante mucho tiempo (hasta la llegada de la fibra óptica) tanto en el bucle de abonado como en las instalaciones de cableado estructurado.

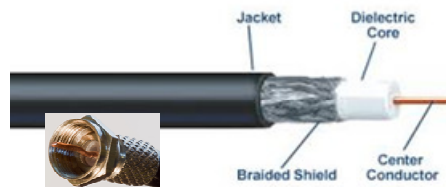


Figura 4.14. Cable coaxial y ejemplo de conector BNC.

Mayoritariamente, los cables coaxiales obedecen a uno de estos dos tipos: cable coaxial de banda base (*baseband coaxial cable*), cuya impedancia característica es de $50\ \Omega$, y cable coaxial de banda ancha (*broadband coaxial cable*), cuya impedancia característica es de $75\ \Omega$. El primero es un cable de un grosor aproximado de 1 cm, ligero y flexible. Las primeras versiones de las redes Ethernet utilizaban cable coaxial de $50\ \Omega$, en dos formatos: cable grueso (especificación 10BASE5 o *thick Ethernet*), y cable delgado (especificación 10BASE2 o *thin Ethernet*), siendo este último formato posterior al primero. En ambos casos se utilizaba transmisión digital sin modulación de portadora, de ahí la denominación *baseband*.

El cable coaxial de $75\ \Omega$ es más grueso, pesado y rígido que el cable de $50\ \Omega$. Su respuesta frecuencial sigue siendo en banda base, pero con un ancho de banda mucho mayor que el de este último (de ahí el uso del calificativo *broadband*). Esto permite dividir dicho ancho de banda en múltiples canales para ser aprovechados por múltiples conexiones al mismo tiempo, sin interferirse mutuamente. De hecho, las principales aplicaciones del cable coaxial de $75\ \Omega$ son la televisión por cable y los enlaces telefónicos a larga distancia.

La televisión por cable tuvo sus orígenes en las comunidades de antena de TV o CATV (*community antenna television*), concebidas inicialmente para ofrecer el servicio de TV a zonas residenciales remotas. Actualmente, aunque ello varía de un país a otro, la distribución de TV por cable combina el cable coaxial y la fibra óptica (infraestructuras HFC – *hybrid fiber coaxial cable*), pudiendo llegar a casi tantos hogares y oficinas como la red telefónica, y transportar hasta cientos de canales de TV a distancias de varias decenas de kilómetros desde cada cabecera local. Además, de la misma manera que el cable UTP sirvió (y sigue sirviendo) para acceder a Internet a través de la red telefónica, mediante los módems tradicionales primero y los módems ADSL después, la tecnología *cablemodem* se desarrolló para proveer acceso a Internet aprovechando la infraestructura de televisión por cable. Así pues, la tecnología HFC constituye una alternativa a la tecnología ADSL en el ámbito del acceso residencial a Internet.

Por lo que se refiere a su uso en la red telefónica, el cable coaxial puede transportar hasta unos 10000 canales en telefonía analógica, y ofrece hasta unos 600 Mbps en telefonía digital. No

obstante, en los últimos años su hegemonía se ha visto truncada por la expansión de la fibra óptica, las microondas terrestres y los enlaces por satélite.

La especificación 10BROAD36 para redes Ethernet a 10 Mbps también se basaba en el cable coaxial de 75 Ω , aunque fue una opción raramente utilizada. Otros usos de estos cables son las conexiones entre transmisor y antena, por un lado, y entre antena y receptor, por otro. También se han utilizado para conexiones entre periféricos y computador, así como en enlaces submarinos, aunque uno y otro uso han sido desplazados por el cable USB y la fibra óptica respectivamente.

El cable coaxial tiene sus orígenes en la industria militar de los EEUU, cuando fue estandarizado mediante la norma MIL-HDBK-216, que asignaba denominaciones del tipo RG a las diversas categorías de este tipo de cable. La versión actual de dicha norma, MIL-SPEC MIL-C-17, especifica cables de mejores prestaciones bajo nuevas denominaciones; no obstante, las denominaciones iniciales fueron tan habituales a lo largo de los años, que todavía hoy perduran, aunque básicamente solo sirven para indicar unas mínimas características, como el grosor y la impedancia característica del cable, así como el tipo de conector a utilizar. De hecho, actualmente la especificación RG-# no garantiza el cumplimiento de ningún estándar de entre todos los existentes.

Tabla 4.3. Algunas categorías relevantes de cable coaxial (RG viene de *radio guide* y U de *universal*).

Categoría	Impedancia característica	Diámetro conductor exterior	Conector	Atenuación a 10/500 MHz (dB/Km)
RG-8/U	50 Ω	10.30 mm (3/8 ")	N	19.4/155.2
RG-58A/U	50 Ω	4.95 mm (1/4 ")	BNC	46.5/346.4
RG-59/U	75 Ω	6.10 mm	F	33.6/156.7
RG-6/U /UQ	75 Ω	8.40 mm /7.62 mm	F	27.0/127.1

La **Tabla 4.3** describe las características, en términos de impedancia, tamaño, tipo de conector y atenuación a las frecuencias de 10 y 500 MHz, de algunas de las categorías RG. Específicamente, la atenuación depende de tres factores: el material dieléctrico utilizado para aislar los dos conductores, el tamaño del cable y la frecuencia. En la **Figura 4.15** se muestran las curvas de atenuación en función de la frecuencia para algunas categorías de cable coaxial. Al igual que ocurría con el par de hilos trenzados, la atenuación crece con la frecuencia, aunque a un ritmo menor, y para una frecuencia dada, la atenuación decrece con el grosor del cable.

Fibra óptica

La **fibra óptica** es el medio de transmisión que mejores prestaciones ofrece. Se trata de un conductor de plástico, cristal multicomponente o cristal de silicio ultra-puro, muy delgado y flexible, que es capaz de conducir una señal óptica. Esta señal consiste en una secuencia de pulsos de luz infrarroja, lo que no es más que una forma de modulación digital. Las modulaciones más utilizadas en fibra óptica son QPSK y QAM.

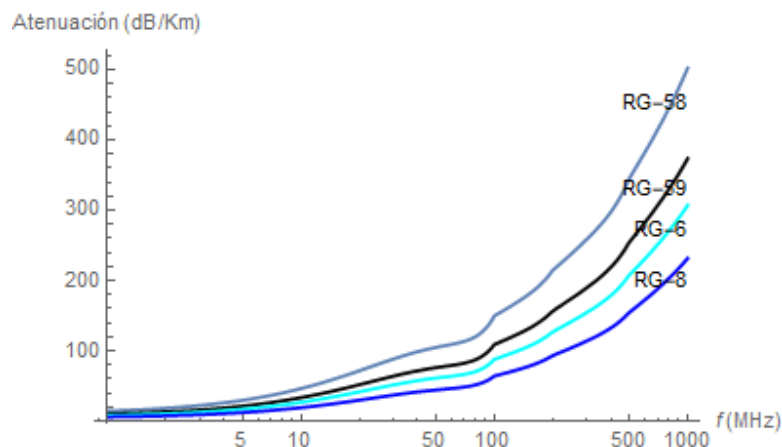


Figura 4.15. Atenuación en función de la frecuencia para diversas categorías de cable coaxial.

Como se ilustra en la **Figura 4.16**, un conductor de fibra óptica tiene forma cilíndrica y consta de tres secciones: el núcleo, el revestimiento y la cubierta. El núcleo es la sección más interna y consiste en un hilo delgado de plástico o cristal, cuyo diámetro puede variar de 9 a 62.5 μm . Está rodeado por un revestimiento también de plástico o cristal, generalmente de 125 μm de diámetro y con distinto índice de refracción. La capa más externa es la cubierta, de un plástico más duro, que sirve para proteger el núcleo y el revestimiento de la humedad, la abrasión o cualquier otro tipo de daño físico que pueda provocar el entorno.

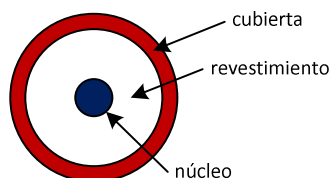


Figura 4.16. Sección transversal de una fibra óptica. La señal óptica se propaga dentro del núcleo.

Los conductores de fibra óptica se agrupan formando cables cuya estructura típica se muestra en la **Figura 4.17**. Como puede observarse, cada cierto número de fibras se disponen en el seno de un material dieléctrico envuelto por una cubierta, formando un primer nivel de agrupación. El material dieléctrico suele consistir en algún tipo de lana, y su función es dar consistencia a ese primer nivel de agrupación. El cable está formado por varias agrupaciones de este tipo, envueltas conjuntamente por una o múltiples capas protectoras, generalmente de PVC. Por el centro del cable viaja además un conducto tubular que puede ser o no metálico, cuya única función es dar consistencia a todo el conjunto. En cuanto a los conectores, en fibra óptica los hay de muchas clases; en la **Figura 4.18** se muestran dos ejemplos.

Dado que la información, antes de ser enviada y una vez recibida, toma la forma de señal eléctrica, el uso de la fibra óptica como medio de transmisión obliga a realizar una conversión del tipo de señal en cada extremo. La **Figura 4.19** muestra la estructura típica de un enlace de fibra óptica para una

dirección (la mayor parte de sistemas operan en ambas direcciones utilizando 2 fibras, de forma que los dos extremos pueden intercambiar información simultáneamente). En el extremo transmisor, se realiza una conversión eléctrico-óptica mediante un **LED (*light emitting diode*)** o un **diodo láser o ILD (*injection laser diode*)**. En el extremo receptor, un dispositivo foto-detector o sensor de luz se encarga de convertir la señal óptica recibida en señal eléctrica. Por tanto, el conjunto formado por el LED o diodo láser, la fibra óptica y el foto-detector constituye el circuito de luz del sistema. El resto sigue siendo, en general, circuitería eléctrica.

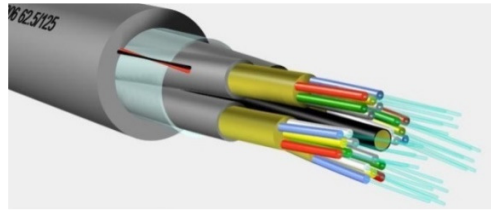


Figura 4.17. Cable de múltiples fibras. Un cable puede contener entre una y varios cientos de fibras.

En la **Figura 4.20** se muestran los mecanismos de propagación de la luz en los tres tipos de fibras existentes: la fibra óptica multimodo de salto de índice, la fibra óptica multimodo de índice gradual y la fibra óptica monomodo. En el caso de la **fibra óptica multimodo de salto de índice**, la propagación de la luz se rige por el principio de reflexión total. Cuando un rayo de luz que viaja por un medio incide en la zona de separación con otro medio cuyo índice de refracción es menor, pueden ocurrir dos fenómenos:



Figura 4.18. Conectores LC y ST para fibra óptica.

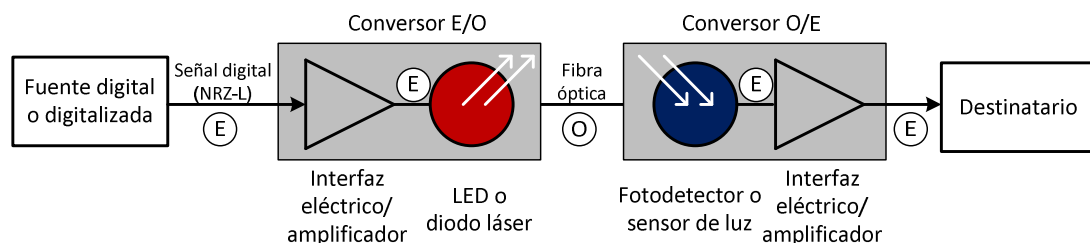


Figura 4.19. Estructura típica de un enlace de fibra óptica. Las letras E y O identifican la naturaleza de la señal en cada tramo, según sea eléctrica u óptica, respectivamente.

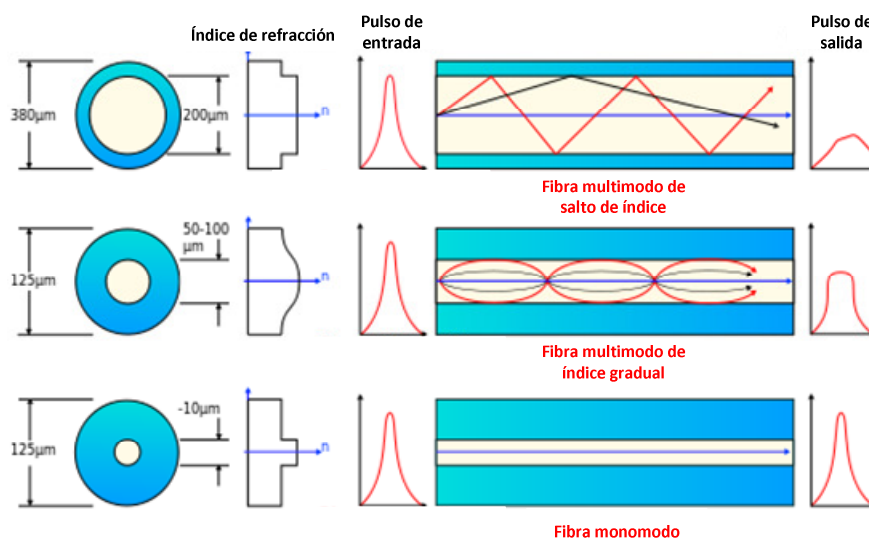


Figura 4.20. Propagación de la luz en los tres tipos de fibra óptica. Obsérvese la reducción del ensanchamiento temporal del pulso de salida a medida que el radio del núcleo se reduce.

- Si el ángulo de incidencia es mayor que el llamado ángulo límite, el rayo se refleja en el propio medio.
- Si el ángulo de incidencia es menor que el ángulo límite, el rayo se refracta y queda absorbido en el otro medio.

En una fibra multimodo de salto de índice, el transmisor de luz es un LED (*light emitting diode*) cuyo haz es lo suficientemente ancho como para que buena parte de los rayos que se emiten en el núcleo incidan con un ángulo superior al límite, de forma que quedan confinados en dicho núcleo alcanzando el foto-detector después de múltiples rebotes. La denominación multimodo proviene precisamente de esa multiplicidad de rayos. No obstante, en realidad esto es una desventaja, conocida como **dispersión modal**: cada rayo representa un camino de propagación con una longitud distinta, por lo que cada uno alcanza el receptor con un retardo también distinto; en consecuencia, los pulsos de luz aparecen en el receptor expandidos en el tiempo, produciéndose nuevamente el fenómeno de ISI. Para contrarrestarlo, es necesario mantener una mínima separación entre los pulsos de salida, lo que implica reducir la velocidad de transmisión y/o la distancia entre el transmisor y el receptor. A fin de evitar estas limitaciones, en la **fibra óptica monomodo** el radio del núcleo se reduce a un valor del orden de la longitud de onda, de modo que sólo puede viajar el rayo axial a su través³³. Con ello se elimina la dispersión modal y se consigue trabajar a velocidades de transmisión y sobre alcances mayores, como requieren, por ejemplo, los enlaces telefónicos de larga distancia y la televisión por cable. La desventaja está en el coste económico, que se incrementa debido al proceso de fabricación de la propia fibra y la necesidad de utilizar un ILD como transmisor,

³³ Tanto la fibra óptica multimodo como la monomodo siguen adoleciendo de las otras causas de ISI, es decir, la distorsión de atenuación y, sobretudo, la distorsión de retardo. En el contexto de la fibra óptica, a la distorsión de retardo se la conoce como **dispersión cromática**.

que es más eficiente que el LED y puede trabajar a velocidades mayores, pero su margen de temperaturas de trabajo es menor y tiene una vida operativa también menor. En una situación intermedia está la **fibra óptica multimodo de índice gradual**, en la que el índice de refracción del núcleo va decreciendo progresivamente en sentido radial, provocando una curvatura de los rayos y una menor dispersión en los retardos que estos experimentan: los rayos que recorren más camino se propagan también a mayor velocidad, de modo que todos alcanzan el extremo receptor en instantes de tiempo muy similares. Este tipo de fibras se utilizan frecuentemente en las redes de área local.

Otro aspecto importante de la fibra óptica es su respuesta frecuencial, que como podemos observar en la **Figura 4.21**, exhibe un patrón muy irregular. No obstante, dentro de esta irregularidad, hay zonas que resultan más apropiadas como “ventanas” o bandas de trabajo, pues en ellas se dan los mejores compromisos entre el comportamiento de la fibra en términos de atenuación, distorsión de atenuación y dispersión cromática, y las prestaciones de la fuente de luz (LED o ILD). De hecho, estas bandas están estandarizadas de acuerdo con la nomenclatura y márgenes de longitudes de onda mostrados en la **Tabla 4.4**, que también incluye el tipo de fibra y fuente de luz utilizados en cada caso.

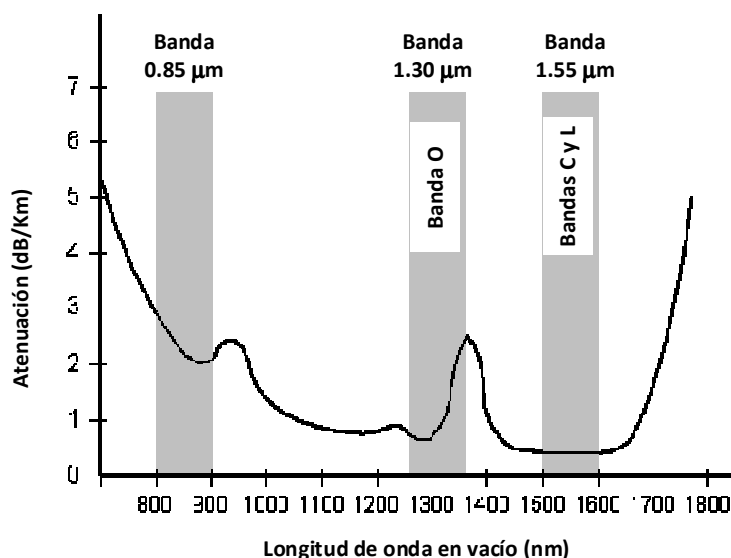


Figura 4.21. Ventanas de trabajo en fibra óptica. Se identifican por su longitud de onda central y, en algunos casos, mediante una letra.

La **Figura 4.21** y la **Tabla 4.4** merecen varios comentarios adicionales:

- El uso habitual, en el caso de la fibra óptica, de la longitud de onda en lugar de la frecuencia tanto para describir la respuesta frecuencial como para definir las bandas de trabajo. No obstante, dado un parámetro se puede obtener el otro, simplemente aplicando la fórmula (2.6).
- El uso de la llamada **longitud de onda en vacío**, es decir, la longitud de onda que le correspondería a la señal sinusoidal si ésta se propagara en el vacío en lugar de la fibra. Si λ

denota la longitud de onda en vacío, la longitud de onda real dentro de la fibra, λ_f , se puede obtener fácilmente aplicando nuevamente la fórmula (2.6) y teniendo en cuenta que la frecuencia es la misma en ambos casos:

$$f = \frac{c}{\lambda} = \frac{v}{\lambda_f} \Rightarrow \lambda_f = \lambda \frac{v}{c} \cong \frac{2}{3} \lambda \quad (4.19)$$

En esta expresión, c y v son, respectivamente, la velocidad de propagación de la luz en el vacío y en la fibra. La velocidad de propagación de la luz en la fibra depende del índice de refracción del material dieléctrico que la constituye, pero se puede redondear a unos 200000 Km/s, con lo que la longitud de onda real en la fibra resulta ser de aproximadamente dos tercios la longitud de onda en vacío.

- Si bien comparadas con toda la respuesta frecuencial de la fibra, las ventanas de trabajo pueden parecer muy estrechas, al calcular su anchura se obtienen resultados enormes. Así, si λ_{min} y λ_{max} son, respectivamente, las longitudes de onda en vacío mínima y máxima correspondientes a una determinada banda, se puede demostrar que el ancho de banda de la misma viene dado por $3 \cdot 10^5 \left(\frac{1}{\lambda_{min}} - \frac{1}{\lambda_{max}} \right)$, donde el resultado se lee directamente en THz si las longitudes de onda están expresadas en nanómetros. Por ejemplo, para la banda C, correspondiente al intervalo entre 1530 y 1565 nm, se obtiene un ancho de banda de casi 4.4 THz, lo cual es un valor muy grande.
- Finalmente, observamos que, contrariamente a lo que parece, el margen de longitudes de onda en vacío que se transmiten por fibra óptica no corresponde al espectro visible (380 – 740 nm), sino que se sitúa dentro de la llamada región cercana al infrarrojo o banda NIR (*near infrared*), que cubre el intervalo desde 750 a 2500 nm.

Tabla 4.4. Bandas de trabajo en fibra óptica. La denominación de las tres grandes bandas (850, 1310 y 1550 nm) corresponde al punto medio del intervalo de longitudes de onda respectivo.

Bandas		Origen de la denominación	Margen de longitudes de onda en vacío (nm)	Tipo de fibra	Tipo de diodo
850 nm			820 – 900	MMF	LED
1310 nm	O	<i>Original</i>	1260 – 1360	SMF	ILD
	E	<i>Extended</i>	1360 – 1460	SMF	ILD
	S	<i>Short wavelengths</i>	1460 – 1530	SMF	ILD
1550 nm	C	<i>Conventional</i>	1530 – 1565	SMF	ILD
	L	<i>Long wavelengths</i>	1565 – 1625	SMF	ILD
	U	<i>Ultralong wavelengths</i>	1625 – 1675	SMF	ILD

Tanto la fibra óptica monomodo como la multimodo se ofrecen en diferentes calidades, medidas en términos del llamado **producto ancho de banda-distancia**. Esta figura de mérito revela el compromiso siempre presente entre ancho de banda y longitud de enlace de un medio de transmisión cualquiera, aunque suele utilizarse solamente para calificar la fibra óptica. Si bien su

justificación la veremos más adelante, la interpretación de este producto es muy sencilla. Por ejemplo, si su valor es igual a 500 MHz·Km, significa que el medio puede transportar una señal de 500 MHz de ancho de banda a una distancia de 1 Km, o, por ejemplo, una señal de 1000 MHz a 0.5 Km (veremos también que un mayor ancho de banda en hertzios conlleva una mayor velocidad de transmisión, de ahí lo significativo de este producto). La **Tabla 4.5** compara las diversas categorías de fibra óptica en base a dicho producto y otros parámetros.

En comparación con los medios guiados metálicos, es decir, el par de hilos trenzados y el cable coaxial, la fibra óptica presenta las siguientes ventajas:

- Menor atenuación. La respuesta en frecuencias de la fibra óptica es mejor que la de los medios metálicos, al presentar una menor atenuación y de valor más o menos constante en múltiples ventanas de gran ancho de banda.
- Capacidad. Esta ventaja es consecuencia de la anterior. Como la fibra óptica ofrece muchísimo más ancho de banda que el cable coaxial y el par de hilos trenzados, también permite alcanzar velocidades de transmisión mucho mayores.
- Aislamiento electromagnético. Los enlaces ópticos no se ven afectados por los campos electromagnéticos externos, y por tanto son inmunes a las interferencias y el ruido impulsivo. Al mismo tiempo, la fibra óptica no radia energía, lo que conlleva una doble ventaja: no afecta a sistemas de comunicaciones cercanos y la información que transporta está más protegida contra el acceso por parte de terceras partes. Este último hecho, unido a la dificultad de “pinchar” una fibra, hacen de ésta uno de los medios de transmisión más recomendables desde el punto de vista de la seguridad y privacidad de las comunicaciones, aparte de ser el medio que mayor velocidad de transmisión proporciona.
- Menor número de repetidores regenerativos. Esta ventaja es también consecuencia de las mejores prestaciones de la fibra óptica. Mientras los enlaces a larga distancia sobre cable coaxial o par trenzado requieren la inserción de repetidores a intervalos de unos pocos kilómetros, en el caso de la fibra óptica el espaciado puede ser de decenas de kilómetros, lo que contribuye a reducir los costes de instalación.
- Menor tamaño y peso. Para capacidades de transmisión similares (en bits por segundo), el cableado de fibra óptica es aproximadamente un orden de magnitud más delgado que el del coaxial o el par trenzado. Por consiguiente, un cable de fibra resulta más ligero y posibilita una mayor simplicidad de los diversos soportes estructurales involucrados en su instalación, lo que contribuye, a su vez, a simplificar y abaratar el despliegue de fibra óptica sobre largas distancias o en entornos urbanos.

Todas estas ventajas, potenciadas además por la evolución tecnológica y la progresiva reducción de su coste económico, hacen de la fibra óptica el medio de transmisión predilecto para transmisiones guiadas de alta velocidad. No debe sorprender, pues, que el uso de este medio se haya extendido al acceso residencial (bucle de abonado) y se haya llegado a materializar el concepto FTTH (*fiber to the home*).

Tabla 4.5. Categorías de fibra óptica (MMF: *multi-mode fiber*; SMF: *single-mode fiber*).

Categoría	Diámetros (μm)	Año	Limitación alcance	Longitud de onda en vacío (nm)
MMF FDDI	62.5/125	1987	160 MHz·Km	850
MMF OM1	62.5/125	1989	200 MHz·Km	850
MMF OM2	50/125	1998	500 MHz·Km	850
MMF OM3	50/125	2003	1500 MHz·Km	850
MMF OM4	50/125	2008	3500 MHz·Km	850
MMF OM5	50/125	2016	3500/1850 MHz·Km	850/950
SMF OM1	9/125	1998	1 dB/Km	1300/1550
SMF OM2	9/125	2000	0.4 dB/Km	1300/1550

4.4. TRANSMISIÓN INALÁMBRICA

El auge de la fibra óptica no ha eclipsado, sin embargo, el uso de la transmisión inalámbrica. La posibilidad de prescindir del cableado, por razones prácticas o de costes de instalación, y la viabilidad que ofrece para las comunicaciones desde equipos móviles o portátiles, la convierten en una opción muy atractiva tanto para los operadores de telecomunicaciones como para los usuarios finales. Todo ello a pesar de las numerosas perturbaciones que puede experimentar una señal radiada, las cuales limitan notablemente las velocidades de transmisión y los alcances asequibles en comparación con los medios guiados. Formalmente, la **transmisión inalámbrica** consiste en la transmisión de señales electromagnéticas a través de un medio no guiado como es el espacio libre. Al no estar confinadas en un conductor físico, estas señales se dispersan en todas direcciones (señales *broadcast*), quedando a disposición de cualquier usuario que disponga de un dispositivo capaz de recibirlas.

Como se describe en la **Figura 4.22**, existen tres grandes modos de propagación de las señales electromagnéticas terrestres: ondas superficiales o GW (*ground waves*), propagación ionosférica o SW (*sky waves*) y visibilidad directa o LOS (*line of sight*). Las **ondas superficiales** siguen la curvatura de la Tierra cruzando la porción más baja de la atmósfera, y pueden propagarse a grandes distancias, mucho más allá del horizonte visual. Este tipo de propagación se da para frecuencias por debajo de los 2 MHz. En la **propagación ionosférica**, la señal se transmite mediante sucesivos rebotes entre la ionosfera y la superficie terrestre. Comparada con las ondas superficiales, la propagación ionosférica permite mayores alcances con menor potencia transmitida. Este modo de propagación se da para frecuencias entre 2 y 30 MHz. Finalmente, a partir de los 30 MHz, la propagación de las ondas electromagnéticas requiere **visibilidad directa** entre las dos antenas.

Las señales electromagnéticas se clasifican por su frecuencia de acuerdo con las bandas y sub-bandas recogidas en la **Tabla 4.6**. Concretamente, esta tabla muestra el espectro electromagnético para las telecomunicaciones, que abarca el amplio margen de frecuencias que va desde 30 Hz hasta casi 800 THz, además del modo de propagación y los principales servicios en cada sub-banda. La demarcación de las grandes bandas de frecuencia (ondas de radio, microondas, etc.) es algo difusa,

pues no todas las fuentes de información consultadas aplican el mismo criterio. En el caso de la **Tabla 4.6**, se han definido las bandas de radio y microondas de forma que encajen perfectamente con grupos completos de sub-bandas de frecuencias. De todos modos, más que la frontera exacta en términos de frecuencia, lo significativo es el comportamiento diferenciado de las ondas electromagnéticas en estas bandas. Así pues, mientras las transmisiones basadas en ondas de radio son isotrópicas, las que hacen uso de microondas pueden ser directivas. Una **transmisión isotrópica** es aquella en la que la potencia transmitida se reparte por igual en todas las direcciones; en cambio, en una **transmisión directiva**, la potencia transmitida se focaliza en una dirección determinada. Las transmisiones isotrópicas son especialmente indicadas para los servicios de *broadcasting*, es decir, servicios dirigidos a un número indefinido de destinatarios geográficamente dispersos; por el contrario, las transmisiones directivas son adecuadas para radioenlaces entre antenas, las cuales deberán estar perfectamente apuntadas.

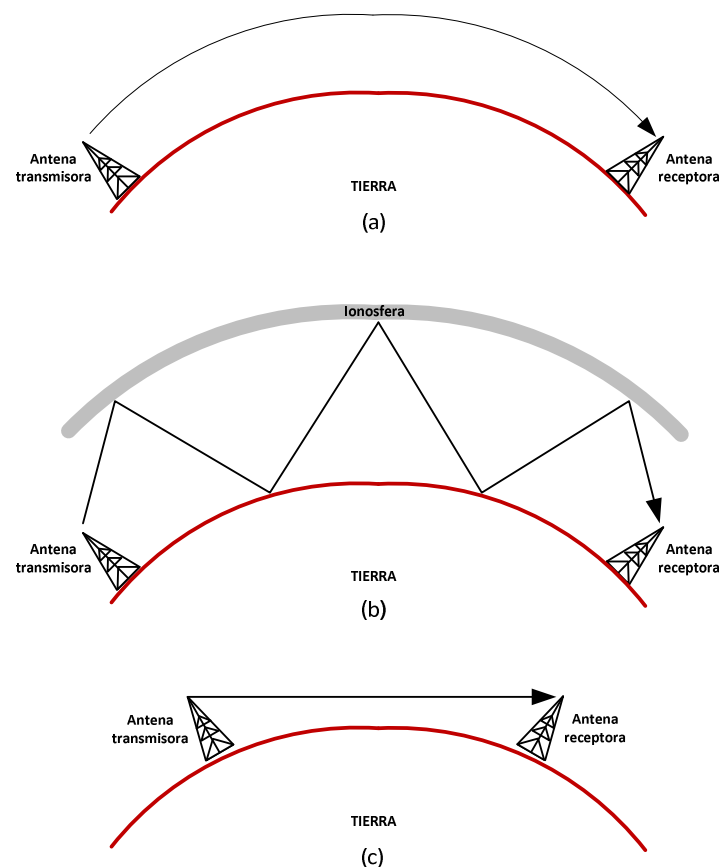


Figura 4.22. Modos de propagación de las ondas electromagnéticas terrestres: (a) ondas superficiales, (b) propagación ionosférica, y (c) visibilidad directa. En el caso de la propagación ionosférica, la figura da entender que se producen múltiples reflexiones en la ionosfera, como si ésta fuera una superficie metálica, aunque en realidad el fenómeno que tiene lugar en cada reflexión es una progresiva refracción dentro de la ionosfera que provoca la curvatura del rayo.

Tabla 4.6. Espectro electromagnético para las telecomunicaciones. Prácticamente todas las bandas de frecuencia están reguladas por la autoridad nacional en cada país, y a nivel internacional por la ITU (*International Telecommunications Union*), que ejerce de organismo coordinador. La nomenclatura mostrada para las sub-bandas dentro de la banda de microondas es del IEEE.

Bandas espectrales		Margen de frecuencias	Propagación	Aplicaciones y servicios
	ELF (<i>extremely low frequency</i>)	30 – 300 Hz	GW	Líneas de potencia domésticas
	VF (<i>voice frequency</i>)	300 Hz – 3 KHz	GW	Telefonía
Ondas de radio	VLF (<i>very low frequency</i>)	3 – 30 KHz	GW	Navegación de largo alcance, comunicaciones submarinas
	LF (<i>low frequency</i>)	30 – 300 KHz	GW	Navegación de largo alcance, sistemas de radiolocalización
	MF (<i>medium frequency</i>)	300 KHz – 3 MHz	SW	Radio marítima, radio AM
	HF (<i>high frequency</i>)	3 – 30 MHz	SW	Radioaficionados, banda ciudadana (CB), radiodifusión internacional (BBC), comunicaciones marítimas, navegación aérea, comunicaciones militares
	VHF (<i>very high frequency</i>)	30 – 300 MHz	SW, LOS	Televisión VHF, radio FM, navegación aérea
Microondas (sub-bandas L, S, C, X, Ku, K, Ka, V, W y mm o G)	UHF (<i>ultra high frequency</i>)	300 MHz – 3 GHz	LOS	TV UHF, telefonía celular, redes inalámbricas, radar
	SHF (<i>super high frequency</i>)	3 – 30 GHz	LOS	Radioenlaces terrestres, enlaces por satélite, telefonía celular, radar
	EHF (<i>extremely high frequency</i>)	30 – 300 GHz	LOS	Radar, enlaces por satélite
Infrarrojos	FIR (<i>far infrared</i>)	300 GHz – 3 THz	LOS	Fibra óptica, LAN infrarrojas
	MIR (<i>mid infrared</i>)	3 – 30 THz	LOS	
	NIR (<i>near infrared</i>)	30 – 300 THz	LOS	
Espectro visible		380 – 790 THz	LOS	VLC (<i>visible light communication</i>)

De la **Tabla 4.6** también se deduce que es la banda de microondas la más implicada en el ámbito de las redes de computadores, por lo que en adelante nos ocuparemos principalmente de las transmisiones inalámbricas en dicha banda. El resto de la sección se ocupa de las antenas, el modelo de propagación en espacio libre, las condiciones de visibilidad directa, las comunicaciones por satélite, los sistemas VSAT y las perturbaciones específicas de las comunicaciones inalámbricas.

Antenas

Una **antena** es un transductor capaz de transformar una señal eléctrica en una señal electromagnética (transmisión) y viceversa (recepción). Se trata, pues, de un **transductor recíproco**, ya que puede operar en los dos sentidos, y además con el mismo nivel de prestaciones. Una antena

está constituida por un conductor o un conjunto de conductores eléctricos, cuyas características en términos de forma, disposición y tamaño le otorgan propiedades especiales para llevar a cabo una conversión eficiente entre las dos energías. En particular, la relación entre el tamaño global de la antena y la longitud de onda de trabajo determina sus características de directividad, de forma que cuanto mayor es esa relación, más directiva es la transmisión. Es por ello que las transmisiones en la banda de ondas de radio son omnidireccionales, pues para que fueran directivas requerirían antenas de tamaño descomunal.

En transmisión, la antena genera una radiación electromagnética cuya potencia se dispersa sobre una superficie esférica cada vez más grande, y no necesariamente de la misma manera en todas las direcciones. En recepción, la antena intercepta parte de una radiación, convirtiendo la densidad de potencia electromagnética en el punto donde está ubicada en potencia eléctrica. La expresión cuantitativa más completa de cómo una antena lleva a cabo estos dos procesos, transmisión y recepción, es su diagrama de radiación (generalmente se asocia este diagrama a la antena como transmisora, si bien la caracteriza igualmente como receptora). El **diagrama de radiación** de una antena no es más que la representación gráfica de su ganancia en cada dirección. A su vez, la **ganancia de una antena** se define como la relación, expresada de forma logarítmica, entre la densidad de potencia que produce la antena en un punto, y la densidad de potencia que produciría, en ese mismo punto, una antena omnidireccional o isotrópica que transmitiera la misma potencia. Al tomar como referencia una antena isotrópica, los decibelios de ganancia de una antena cualquiera se expresan a veces como dBi (decibelios con respecto a la antena isotrópica). A modo de ejemplo, la **Figura 4.23** muestra el diagrama de radiación tridimensional y los cortes azimutal y de elevación para una **antena Yagi**, típicamente utilizada en la recepción de señales de televisión. Como podemos observar, el diagrama presenta un lóbulo de radiación principal en torno a la dirección coincidente con el eje Y, que es la dirección privilegiada de esta antena.

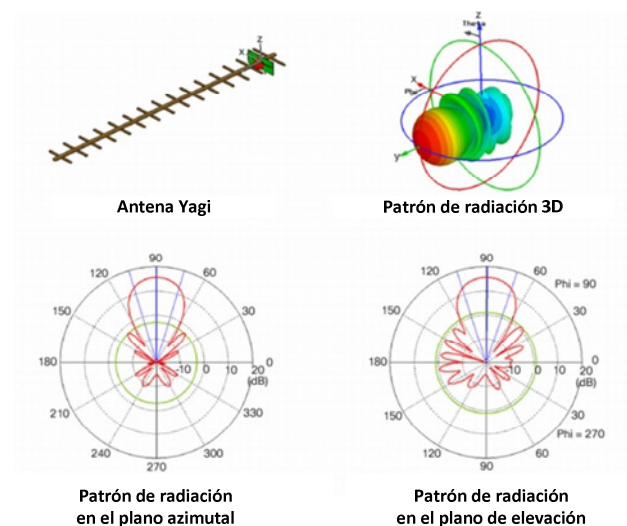


Figura 4.23. Patrón de radiación típico de una antena Yagi. El incremento de potencia radiada en las direcciones privilegiadas (lóbulo principal) es a costa de reducirla en otras direcciones.

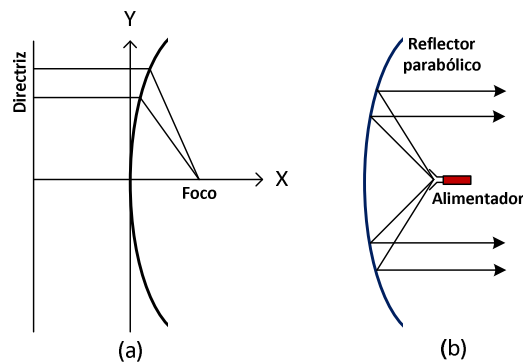


Figura 4.24. Antena parabólica: (a) definición de parábola; (b) principio de funcionamiento en transmisión (en recepción, los rayos que llegan paralelos se reflejan en el paraboloide y se concentran en el foco).



Figura 4.25. Ejemplos de antenas parabólicas. A la izquierda, una típica antena doméstica para la recepción de señales de TV por satélite. En el centro, una antena parabólica en Erdfunkstelle Raisting (Bavaria, Alemania), donde está la mayor instalación del mundo para comunicaciones por satélite. A la derecha, el mayor radiotelescopio del mundo, construido sobre un valle en Puerto Rico (observatorio Arecipo). Tiene 305 m de diámetro y una ganancia de 70 dBi a 2.38 GHz.

Las dos antenas más importantes en la banda de microondas son la antena parabólica y la antena Horn. La antena parabólica es la solución más frecuente para lograr transmisiones muy directivas tanto en enlaces terrestres como por satélite. Recordemos que una parábola es el lugar geométrico de todos los puntos que equidistan de una línea y de un punto fuera de la línea. A la línea se la denomina directriz y al punto foco. Estas definiciones se ilustran en la **Figura 4.24(a)**. La **antena parabólica** es el conjunto formado por una fuente omnidireccional situada en el foco (alimentador), y el paraboloide (reflector), el cual resulta de rotar la parábola alrededor de su eje. En la **Figura 4.24(b)** se muestra una sección de dicha antena. La geometría de todo el conjunto es tal que, si el alimentador fuera un elemento puntual, los rayos generados por éste saldrían perfectamente paralelos después de reflejarse en el paraboloide; no obstante, en la práctica se produce una cierta dispersión, la cual es debida a que el alimentador tiene un determinado tamaño y la radiación que genera no es perfectamente omnidireccional. Cuanto mayor es el tamaño del paraboloide con

respecto al alimentador menor es la dispersión y más directivo es el haz transmitido. Para mayor ilustración, la **Figura 4.25** muestra varios ejemplos de antena parabólica.

La **ganancia de una antena** se define como el valor máximo de su diagrama de radiación. Este máximo se alcanza para una determinada dirección (dirección privilegiada), que va a ser la dirección de apuntamiento de la antena tanto para transmitir como para recibir. Desde la perspectiva de una antena como elemento receptor, otro parámetro importante es su **área efectiva**, que no es más que el cociente entre la potencia eléctrica generada por la antena y la densidad superficial de potencia electromagnética en el punto donde está situada la misma. También puede verse como el área de una superficie equivalente que al interceptar la onda electromagnética captura la misma potencia que la antena real. La **ganancia lineal de una antena** (g) y el área efectiva de la misma (A_e) están relacionadas a través de la siguiente expresión:

$$g = \frac{4\pi \cdot A_e}{\lambda^2} \quad (4.20)$$

Recuérdese que λ es la longitud de onda portadora. Por consiguiente, el área efectiva de una antena omnidireccional se obtiene al igualar a 1 la expresión anterior:

$$1 = \frac{4\pi \cdot A_e}{\lambda^2} \Rightarrow A_e = \frac{\lambda^2}{4\pi} \quad (4.21)$$

Si la antena es parabólica, su área efectiva viene dada aproximadamente por la expresión $0.56 \cdot A$, donde A es el área de la cara frontal de la antena (superficie del disco).

Propagación por el espacio libre

El modelo matemático más simple que relaciona la potencia transmitida por una antena y la recibida por otra antena situada a una distancia d , es el llamado **modelo de propagación por el espacio libre**, pues corresponde al escenario en que no hay ningún obstáculo entre las antenas. Para desarrollarlo, consideremos en primer lugar el caso de un enlace entre dos antenas isotrópicas. Si P_t es la potencia transmitida por una de ellas, la densidad superficial de potencia a una distancia d vendrá dada por $P_t / (4\pi \cdot d^2)$, donde $4\pi \cdot d^2$ no es más que el área de una superficie esférica de radio d centrada en la antena transmisora (téngase en cuenta que la radiación de cualquier antena toma la forma de un frente de onda circular que se abre a medida que nos alejamos de la misma). Por lo tanto, la potencia recibida por una antena a distancia d será el resultado de multiplicar la densidad superficial anterior por el área efectiva de dicha antena. Como se supone que ésta es también isotrópica, el resultado es el siguiente:

$$P_r = \frac{P_t}{4\pi \cdot d^2} \cdot \frac{\lambda^2}{4\pi} = \frac{P_t}{\left(\frac{4\pi \cdot d}{\lambda}\right)^2} \quad (4.22)$$

Esta relación también puede expresarse en decibelios:

$$10 \cdot \log P_r = 10 \cdot \log P_t - 10 \cdot \log \left(\frac{4\pi \cdot d}{\lambda}\right)^2 \quad (4.23)$$

El término que aparece restando representa la pérdida de potencia causada exclusivamente por la dispersión de la señal al propagarse. Se trata, pues, de una atenuación, que podemos denotar por L y expresar de la siguiente manera:

$$L(\text{dB}) = 20 \cdot \log \frac{4\pi \cdot d}{\lambda} \quad (4.24)$$

Obsérvese que esta atenuación depende del número de longitudes de onda contenidas en la distancia que separa las dos antenas. La expresión (4.24) es la llamada **atenuación o pérdida por propagación en espacio libre**, ya que es como si la antena transmisora y la receptora estuvieran suspendidas en el aire sin ningún tipo de obstáculo entre ellas, ni siquiera el suelo. Lógicamente esta situación es hipotética, pero también es cierto que en muchos casos la principal causa de pérdida de potencia sigue siendo la dispersión del frente de ondas, de ahí que en una primera aproximación se acepten las expresiones (4.22) o (4.23) para relacionar la potencia recibida con la transmitida. Ambas expresiones también ponen de manifiesto la relación de proporcionalidad inversa entre la potencia recibida y el cuadrado de la distancia, a diferencia de lo que ocurría en los medios guiados, donde la dependencia era de tipo exponencial. Puesto que un decrecimiento exponencial es más fuerte que uno de tipo polinómico, a igualdad de distancia y sin tener en cuenta las perturbaciones específicas de las comunicaciones inalámbricas, se requieren menos amplificadores o repetidores regenerativos en los enlaces inalámbricos que en los medios guiados. En la **Figura 4.26** se representa gráficamente la pérdida por propagación en espacio libre en función de la distancia, para diversos valores de la frecuencia. Las curvas obtenidas, lineales como consecuencia de la representación del eje de distancias en escala logarítmica, revelan que la atenuación aumenta con la distancia y la frecuencia.

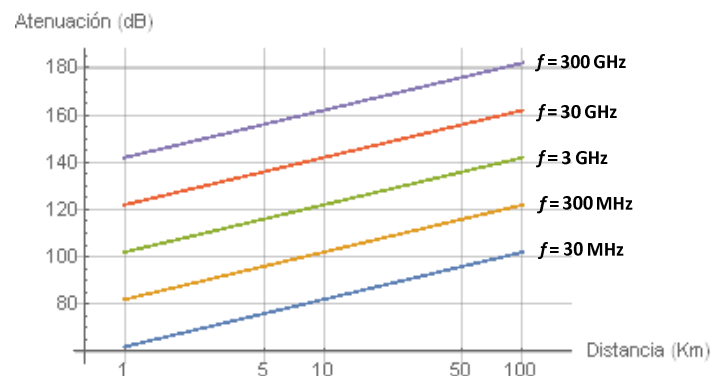


Figura 4.26. Pérdida por propagación en espacio libre en función de la distancia, para diferentes frecuencias.

La ecuación (4.22) corresponde al caso de dos antenas isotrópicas. Si tanto la antena transmisora como la receptora son directivas, con ganancias lineales g_t y g_r respectivamente, la relación entre potencia recibida y potencia transmitida adquiere la siguiente forma:

$$P_r = \frac{P_t}{\left(\frac{4\pi \cdot d}{\lambda}\right)^2} \cdot g_t \cdot g_r \quad (4.25)$$

Ésta es la llamada **fórmula de Friis**. En unidades logarítmicas:

$$10 \cdot \log P_r = 10 \cdot \log P_t - 10 \cdot \log \left(\frac{4\pi \cdot d}{\lambda}\right)^2 + G_t + G_r \quad (4.26)$$

En esta expresión, $G_t = 10 \cdot \log g_t$ y $G_r = 10 \cdot \log g_r$.

Visibilidad directa

El término “visibilidad directa” evoca implícitamente el uso de frecuencias ópticas, que son las que utilizamos los humanos y prácticamente el resto de seres vivos para percibir nuestro entorno. A estas frecuencias, la propagación es en línea recta, y cualquier obstáculo que se interponga en el camino de la señal impide la percepción. Llevado al ámbito de las antenas, aun en ausencia de cualquier objeto en la línea recta que separa dos antenas, la propia Tierra, debido a su curvatura, se convierte en un obstáculo. Si consideráramos dos antenas arbitrariamente próximas y las fuéramos separando progresivamente, llegaría un momento en que la línea recta que las une sería tangente a la superficie terrestre. El punto de tangencia sería precisamente el horizonte (**horizonte óptico**) para cada antena, y es ese el momento en que la Tierra empezaría a ser un obstáculo. La distancia entre una antena situada a una altura h sobre la superficie terrestre y el horizonte óptico, $d_{h.o.}$, obedece a una fórmula conocida, basada en consideraciones geométricas:

$$d_{h.o.} = 3.57\sqrt{h} \quad (4.27)$$

Esta fórmula ya está adaptada para devolver la distancia en kilómetros cuando la altura de la antena está expresada en metros.

La **Tabla 4.6** muestra que el requisito de visibilidad directa (*line-of-sight* o LOS) no es, sin embargo, exclusivo de la banda del espectro visible, sino que también se da en bandas de frecuencias inferiores, como las ondas de radio y las microondas. Como se muestra en la **Figura 4.27**, las ondas electromagnéticas a estas frecuencias experimentan una cierta curvatura que puede aprovecharse para transmitir a un **horizonte radio** mayor que el óptico. Este fenómeno se debe a que la atmósfera presenta un índice de refracción que decrece gradualmente con la altura, provocando la curvatura progresiva de los rayos hacia la superficie terrestre. El alcance que resulta, $d_{h.r.}$, viene dado por la siguiente expresión, en la que K es una constante igual o superior a la unidad que depende de la frecuencia:

$$d_{h.r.} = 3.57\sqrt{K \cdot h} \quad (4.28)$$

En la banda de microondas se tiene $K = 4/3$. El valor $K = 1$ corresponde, naturalmente, al alcance óptico, y la expresión (4.27) no es más que un caso particular de la (4.28).

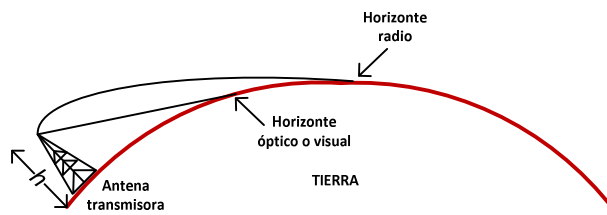


Figura 4.27. Horizonte óptico y horizonte radio. Este último se alcanza apuntando la antena ligeramente por encima de la visual al horizonte óptico.

La expresión (4.28) establece la distancia entre una antena y su horizonte radio. Por lo tanto, la máxima distancia entre dos antenas, d_{max} , para que entre ellas haya visibilidad directa a la frecuencia de trabajo, o **visibilidad directa efectiva**, viene dada por la siguiente expresión, donde h_1 y h_2 son sus alturas respectivas:

$$d_{max} = 3.57\sqrt{K \cdot h_1} + 3.57\sqrt{K \cdot h_2} \quad (4.29)$$

Esta ecuación explica el motivo por el que conviene ubicar las antenas lo más alto posible, además de que con ello se evitan más obstáculos, muy frecuentes en entornos urbanos. Es importante observar, además, que la visibilidad directa efectiva es una condición necesaria pero no suficiente para que el enlace entre dos antenas sea viable; para ello, es también necesario que la antena que transmite lo haga a una potencia suficientemente grande como para garantizar un mínimo nivel de señal en la antena que recibe.

Comunicaciones por satélite

La limitación impuesta por la ecuación (4.29) para el alcance de las microondas terrestres no se da en las comunicaciones por satélite, ya que entre una estación terrena y un satélite hay visibilidad directa sin que la Tierra llegue a constituir un obstáculo. Tampoco lo es la propia atmósfera, ya que las microondas no rebotan en la ionosfera, sino que penetran en ella, y dado que más allá de la ionosfera el medio es prácticamente vacío, la propagación entre la estación terrena y el satélite es casi rectilínea.

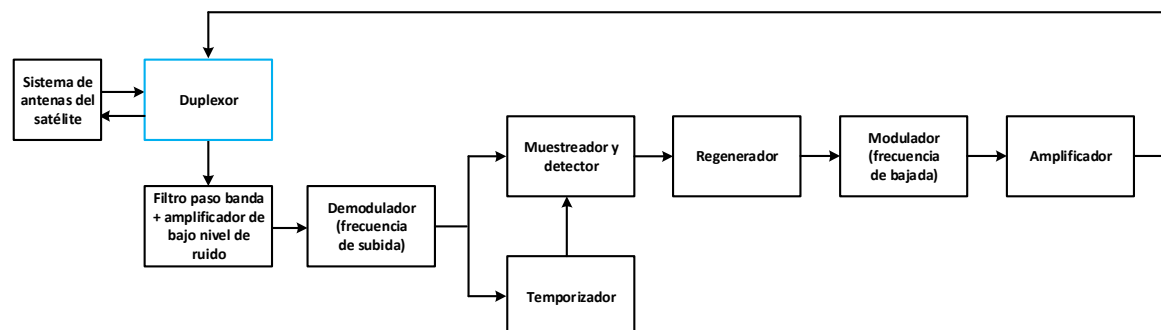


Figura 4.28. Esquema de un transpondedor regenerativo.

Básicamente, un satélite de comunicaciones es un gran repetidor instalado en órbita. En realidad, está formado por un conjunto de subsistemas llamados transpondedores, entendiendo por **transpondedor**³⁴ la serie de unidades funcionales interconectadas que se requieren para transferir una señal recibida por el **enlace ascendente** (*uplink*) al **enlace descendente** (*downlink*). Cada transpondedor trabaja sobre una sub-banda de frecuencias concreta en el enlace ascendente y otra sub-banda de frecuencias concreta en el enlace descendente. Las dos sub-bandas tienen que ser distintas para evitar que la señal de bajada interfiera la señal de subida. En su versión más simple, el transpondedor se limita a recibir la señal de subida, amplificarla, trasladarla en frecuencia y transmitirla como señal de bajada. Esta forma de operación corresponde al llamado **transpondedor de tubo doblado** (*bent pipe*), pues en esencia su función es reconducir la señal del enlace ascendente al enlace descendente (vía amplificación y cambio de frecuencia), como si de una tubería doblada se tratara. Podríamos afirmar que se trata de un transpondedor analógico. Una de sus principales desventajas es que también transfiere todo el ruido captado en el enlace de subida al enlace de bajada. No obstante, si la transmisión es digital, el uso de un repetidor regenerativo en lugar de uno convencional o analógico conlleva numerosas ventajas, entre ellas su capacidad para aislar la salida de la entrada en lo que a ruido se refiere. Su inclusión en la infraestructura de comunicaciones de un satélite da lugar al segundo tipo de transpondedor, el **transpondedor regenerativo**, cuyo esquema se muestra en la **Figura 4.28**. El único elemento nuevo es el **duplexor**, una puerta de microondas bidireccional que recibe la señal de subida de la antena del satélite y transmite la señal de bajada hacia dicha antena. En general, un mismo satélite puede contener múltiples transpondedores de anchos de banda diversos. Una forma de expresar el ancho de banda total de un satélite (en el enlace de subida o en el de bajada, pues son iguales) es mediante el número de transpondedores equivalentes de 36 MHz, pues éste es un valor bastante habitual para el ancho de banda de un transpondedor. Así pues, una nueva unidad de ancho de banda es el **TPE** (*transponder equivalent*), equivalente a 36 MHz. Por ejemplo, un ancho de banda total de 12 TPE equivale a 432 MHz, ya que $12 \cdot 36 = 432$, lo cual no significa que el satélite contenga 12 transpondedores de 36 MHz cada uno, sino que puede consistir en un conjunto heterogéneo de transpondedores cuyo ancho de banda total es de 432 MHz. Otra apreciación importante acerca del transpondedor es que su ancho de banda, sea cual sea, suele distribuirse en múltiples canales mediante técnicas de acceso múltiple del tipo FDMA o TDMA, que se estudian más adelante.

Un aspecto esencial del diseño de un sistema de comunicaciones por satélite es la definición de su órbita, que en general es elíptica. Ésta determina el llamado **período orbital**, es decir, el tiempo que tarda el satélite en dar una vuelta a la Tierra. Concretamente, según la tercera ley de Kepler, el cuadrado del período orbital es directamente proporcional al cubo del semieje mayor de la elipse que define la órbita y, por tanto, cuanto mayor es la altura del satélite, mayor es dicho período. En

³⁴ El término “transpondedor” proviene de la contracción entre transmisor y respondedor, y aunque conceptualmente es similar a “transceptor” (contracción entre transmisor y receptor), no debe confundirse con éste. El transceptor es el resultado de implementar dentro de una misma carcasa el transmisor y el receptor de un sistema de comunicaciones bidireccional, en el que las señales que viajan en uno y otro sentido son diferentes e independientes. El transpondedor, en cambio, es un dispositivo unidireccional que recibe una señal en una banda de frecuencias y de forma automática la reenvía en otra banda.

particular, si la órbita es ecuatorial a una altura constante de $35786 \cong 36000$ Km sobre la superficie terrestre (es decir, es una órbita circular), el período orbital coincide con el período de rotación de la Tierra, que es exactamente de 23 horas, 56 minutos y 4 segundos. Es la llamada **órbita geoestacionaria** o **GEO** (*geostationary Earth orbit*). Los satélites en la órbita geoestacionaria (satélites GEO) ofrecen algunas ventajas importantes, como su posición aparente fija con respecto a la Tierra, y el hecho de que con sólo 3 de ellos separados 120° se pueden cubrir las comunicaciones de todo el planeta, salvo las zonas polares. Ello es gracias a que, al situar el satélite a tanta altura sobre la superficie terrestre, la apertura de su haz puede llegar a cubrir aproximadamente un tercio de la misma.

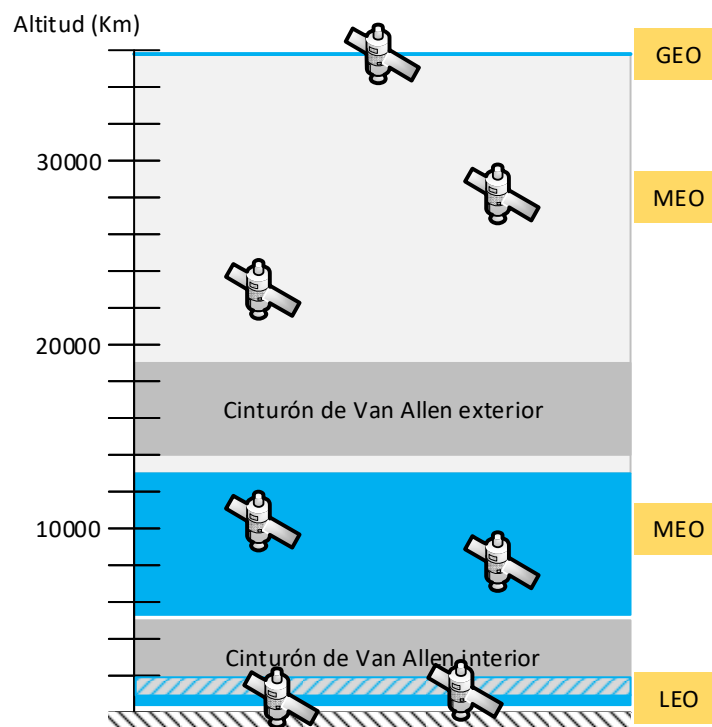


Figura 4.29. Estructura orbital de las comunicaciones por satélite. La zona sombreada más clara es parte del cinturón de Van Allen exterior, con menor densidad de partículas, pero capaz de interferir las comunicaciones con los satélites GEO.

No obstante, las comunicaciones vía satélite GEO presentan la desventaja de su elevado retardo de propagación. Por ésta y otras razones, se desarrollaron también sistemas de comunicación por satélite sobre órbitas más bajas, concretamente los llamados sistemas de **órbita baja** o **LEO** (*low-Earth orbit*) y los de **órbita mediana** o **MEO** (*medium-Earth orbit*). No obstante, la determinación de estas órbitas inferiores no está solamente condicionada por el retardo de propagación y el período orbital, sino también por la presencia de los llamados **cinturones de Van Allen**, capas densas de partículas cargadas resultantes de la interacción de los vientos solares con el campo magnético terrestre. Realmente, la distribución de estas partículas en función de la altitud es continua, pero sobresalen dos zonas donde la concentración de las mismas es más elevada. Se trata del **cinturón**

de Van Allen interior (*lower Van Allen belt*) y del cinturón de Van Allen exterior (*upper Van Allen belt*). Ambos tienen forma de toroide en torno al eje magnético de la Tierra, y en su interior las partículas cargadas se mueven en espiral a alta velocidad entre los polos magnéticos del planeta. El cinturón de Van Allen interior está formado por protones y su espesor es bastante estable, extendiéndose desde unos 1000 Km por encima de la superficie terrestre hasta una altitud en torno a 5000 Km. El cinturón de Van Allen exterior está formado por electrones, y es más inestable. Situado a una altura de unos 13000 Km, su espesor puede alcanzar los 38000 Km, si bien la mayor densidad de partículas se da en la franja localizada entre 14000 y 19000 Km aproximadamente. Estos cinturones son a priori zonas de exclusión para los satélites, ya que los altos niveles de radiación presentes en ellos, como resultado de las elevadas concentraciones de partículas cargadas y en movimiento a velocidades relativistas, pueden interferir significativamente las comunicaciones. No obstante, y a pesar de ello, en la actualidad hay unos 800 satélites operativos dentro de estas capas de radiación. Para mayor detalle, la **Figura 4.29** muestra la ubicación de los satélites GEO, MEO y LEO con respecto a los cinturones de Van Allen. Aunque la franja correspondiente a los satélites LEO parece muy estrecha y se sitúa muy por debajo del resto de órbitas satelitales, ocupa un espacio considerable y bastante por encima de la superficie terrestre, habitualmente entre 160 y 1000 Km, pero pudiendo alcanzar los 1500 – 2000 Km. Por otro lado, los satélites MEO se localizan a cualquier altitud entre las órbitas LEO y GEO (evitando en lo posible los cinturones de Van Allen), si bien la mayoría orbitan a una altura que varía entre 5000 y 12000 Km.

Un subconjunto de órbitas LEO lo constituyen las **órbitas polares**, que definen una rotación de norte a sur en lugar de oeste a este. Su altitud puede variar entre 200 y 1000 Km, y su desviación con respecto al eje de rotación de la Tierra puede llegar a los 20 – 30 grados (es decir, los satélites en estas órbitas no tienen por qué pasar exactamente por encima de los polos geográficos). Finalmente, un caso particular de órbita polar es la órbita síncrona con el Sol o **SSO** (*Sun-synchronous orbit*), que pasa justo por encima de los Polos, de tal manera que el satélite se posiciona siempre sobre cada punto a la misma hora solar local. La altitud de una órbita SSO oscila entre 600 y 800 Km.

Tabla 4.7. Valores típicos para los principales parámetros orbitales.

Órbita	Retardo de propagación	Período orbital	Número de satélites para cobertura global
LEO	20 – 25 ms	90 – 120 min	50 – 200
MEO	70 – 200 ms	3 – 7 horas	8 – 20
GEO	240 – 300 ms	24 horas	3 (sin cobertura polar)

La **Tabla 4.7** muestra los valores típicos de los principales parámetros de cada tipo de órbita. El retardo de propagación se refiere al tiempo de subida más el de bajada, lo que se conoce también como tiempo de un salto (*one hop time*). Su valor varía dependiendo de la altura del satélite sobre la superficie terrestre y de su elevación sobre el horizonte. En el caso de los satélites geoestacionarios, dependiendo de la latitud en la que se encuentra la estación terrestre, el retardo de propagación puede variar entre 240 y 300 ms, por lo que es habitual adoptar un valor intermedio

de 270 ms para dicho retardo. La última columna de la **Tabla 4.7** hace referencia al concepto de **cobertura global**, que no es más que la situación en que cualquier punto del planeta está cubierto por al menos un satélite en todo momento. Naturalmente, la velocidad angular de un satélite es inversamente proporcional a la altura de su órbita, por lo que, cuanto menor es ésta, mayor es el número de satélites que se requieren para alcanzar la cobertura global.

Tabla 4.8. Aplicaciones de los sistemas LEO, MEO y GEO.

Órbita	Aplicaciones
LEO	Comunicaciones: <ul style="list-style-type: none"> • Mensajería de baja velocidad. • Comunicaciones de voz. • Acceso a Internet de banda ancha (aplicaciones interactivas): navegación por Internet (<i>Internet browsing</i>), servicios de chat, VoIP y juegos online (<i>gaming</i>).
	Escaneo de la superficie terrestre (<i>imaging</i>). En el caso de órbitas SSO, se puede observar la evolución temporal de algún fenómeno terrestre (patrón meteorológico, deforestación, aumento del nivel del mar, incendios o inundaciones, etc.) a intervalos perfectamente regulares (días, y por extensión, semanas, meses, años, etc.).
MEO	Comunicaciones
	Sistemas de navegación (terrestre, marítima, aérea, espacial) o GNSS (Global Navigation Satellite System): son los sistemas de posicionamiento, como GPS, Galileo, etc.
	Observación de la Tierra y el espacio
GEO	Comunicaciones
	Meteorología

La selección de la órbita es un aspecto estrechamente ligado a la aplicación para la que se despliega la constelación de satélites. En la **Tabla 4.8** se recogen las aplicaciones típicas de los sistemas LEO, MEO y GEO, cualquiera que sea el ámbito (comercial, científico o militar) para el que se haya podido desarrollar cada ejemplo concreto de tales sistemas. Por otro lado, en la **Tabla 4.9** se constatan las ventajas e inconvenientes de las tres opciones.

El organismo internacional encargado de definir y asignar las órbitas de los satélites es la ITU. También define y asigna las bandas de frecuencia en las que operan estos, un aspecto especialmente crítico en lo que se refiere a los enlaces descendentes, ya que pueden interferir los enlaces de microondas terrestres. Para llevar a cabo la asignación de frecuencias, la ITU define los siguientes servicios ofrecidos por los satélites:

- **Servicio fijo** o **FSS** (*fixed satellite service*). Es el servicio que tiene lugar entre estaciones ubicadas en puntos fijos de la superficie terrestre. Estos puntos pueden corresponder a posiciones prefijadas o a cualesquiera posiciones seleccionables dentro de las zonas de cobertura del sistema. El servicio FSS abarca comunicaciones punto a punto o punto a multipunto entre dos o más estaciones terrenas con capacidad de transmisión y recepción. Tales estaciones pueden corresponder a equipos terminales o a equipos intermedios de entrada a redes LAN o WAN. Un ejemplo es la interconexión de las redes telefónicas públicas mediante un enlace por satélite

(enlaces telefónicos de larga distancia para llamadas internacionales transoceánicas, por ejemplo). En la **Figura 4.30** se muestra la versión punto a punto de este servicio FSS.

- **Servicio de difusión** o **BSS** (*broadcast satellite service*). Éste es el servicio utilizado cuando se trata de la difusión de señales de radio o televisión vía satélite. Se ilustra en la **Figura 4.31**. La transmisión es unidireccional entre una estación terrena y las múltiples estaciones terrenas que pueden estar desplegadas dentro del área de cobertura del sistema. Típicamente, la estación terrena que transmite forma parte de la infraestructura de un centro de producción de programas (estudio) de televisión, desde donde se emite la señal hacia el satélite para su posterior difusión y recepción directa por el público en general. Por público en general se entiende, aquí, comunidades de usuarios (**CATV** o *community antenna television*) o usuarios individuales (sistema **DBS** o *direct broadcast service*). En el primer caso, la señal se envía desde la estación terrena receptora al usuario final mediante una red de distribución por cable (coaxial). Por su parte, el desarrollo de los sistemas DBS tuvo lugar gracias a la progresiva reducción del tamaño y coste de las antenas parabólicas, lo que permitió un servicio de distribución directo al hogar o **DTH** (*direct to home*).

Tabla 4.9. Ventajas e inconvenientes de los sistemas LEO, MEO y GEO.

Órbita	Ventajas	Inconvenientes
LEO	<ul style="list-style-type: none"> • Bajo retardo de propagación y <i>throughput</i> elevado, necesario para las aplicaciones de carácter interactivo. • Alta resolución de las imágenes gracias a la relativa cercanía de los satélites a la Tierra (aplicaciones de <i>imaging</i>). • Gran variedad de órbitas (polares, ecuatoriales, inclinadas). • Atenuación de la señal en los enlaces de subida y bajada relativamente baja. 	<ul style="list-style-type: none"> • El servicio permanente requiere regular el apuntamiento de las estaciones terrenas (<i>tracking</i>). • En el caso de las comunicaciones, el servicio permanente también requiere transferir las conexiones de un satélite a otro (<i>handover</i>). • Constelaciones con muchos satélites.
MEO	<ul style="list-style-type: none"> • Posición intermedia entre los sistemas LEO y GEO con respecto a la mayoría de características. • Al igual que en los sistemas LEO, el servicio permanente requiere <i>tracking</i> y <i>handover</i> (este último en el caso de las comunicaciones). 	
GEO	<ul style="list-style-type: none"> • Servicio permanente sin necesidad de <i>tracking</i> ni <i>handover</i>. • Posibilidad de observación continuada de una misma zona de la Tierra. • Constelaciones de tamaño mínimo. 	<ul style="list-style-type: none"> • Retardo de propagación elevado y menor <i>throughput</i>. • Una sola órbita disponible, sin cobertura sobre las regiones polares. • Elevada atenuación de la señal en los enlaces de subida y bajada.

- Servicio móvil o **MSS** (*mobile satellite service*). Es el servicio utilizado para comunicaciones de voz y datos entre usuarios móviles. Es equiparable al servicio de telefonía móvil terrestre, con la particularidad de que la estación base está orbitando en el espacio. En el caso de zonas remotas que están fuera de la cobertura móvil terrestre, y dónde tampoco llega la fibra óptica, la única opción de conexión es precisamente el servicio MSS. El tamaño de los equipos

terminales para acceder a este servicio puede variar en un rango similar al que hay entre un teléfono móvil y un dispositivo portátil (tableta u ordenador). Es el caso del **servicio móvil terrestre** o **LMS** (*land mobile service*). Los equipos terminales también pueden estar montados en vehículos, como es el caso del **servicio móvil marítimo** o **MMS** (*maritime mobile service*) y el **servicio móvil aeronáutico** o **AMSS** (*aeronautical mobile-satellite service*) (**Figura 4.32**).

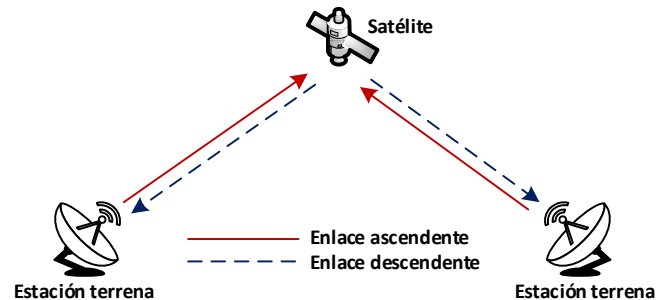


Figura 4.30. Servicio FSS en la versión punto a punto. En general, la comunicación a través del enlace satelital es bidireccional y simultánea en ambos sentidos. En el caso punto a multipunto, la comunicación, también bidireccional con simultaneidad, tiene lugar entre una estación terrena primaria y un grupo de estaciones terrenas secundarias, es decir, de una a varias y de varias a una.

La **Tabla 4.10** muestra las diferentes bandas de frecuencia en las que operan los satélites. Como puede observarse, estas bandas se identifican mediante una o dos letras (nomenclatura IEEE), y la mayor parte de ellas se sitúan en la franja SHF del espectro electromagnético (**Tabla 4.6**). Dentro de cada banda, la gestión del espectro obedece a dos criterios básicos:

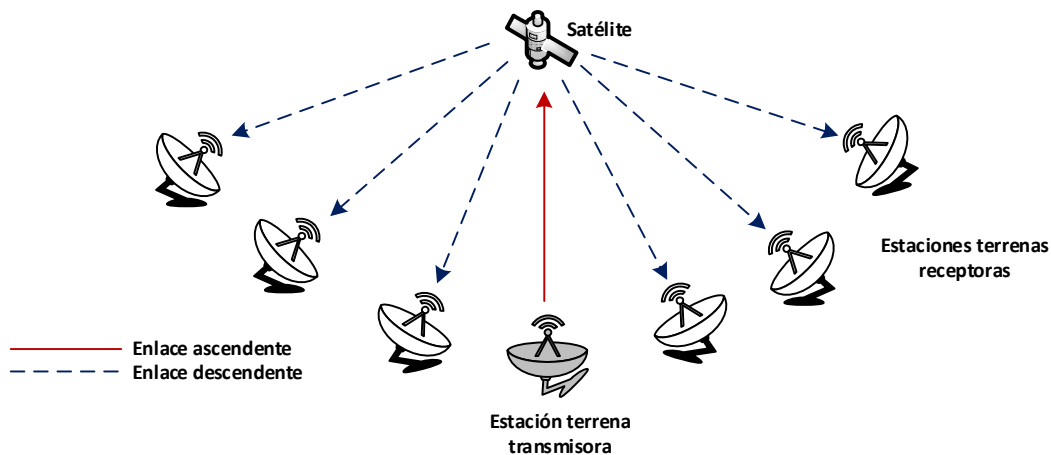


Figura 4.31. Servicio BSS. Obsérvese que básicamente se trata de un servicio fijo unidireccional.

- A fin de evitar que la señal transmitida por un satélite interfiera la señal recibida por éste, es necesario que los enlaces ascendente y descendente operen en sub-bandas distintas.
- La frecuencia portadora asignada al enlace descendente es inferior a la del enlace ascendente. Esto es debido a que la atenuación por propagación en espacio libre es directamente

proporcional a la frecuencia, y a que resulta menos costoso dotar de más potencia de transmisión a la estación terrena que al satélite.

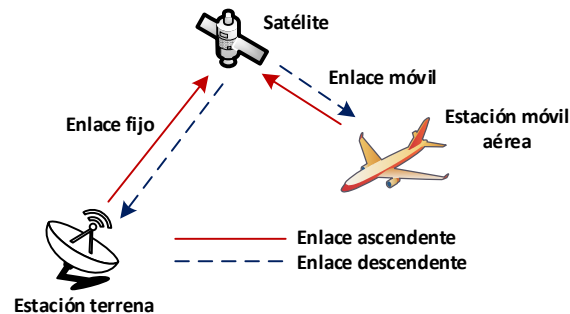


Figura 4.32. Servicio AMSS, suponiendo que el satélite es GEO (enlace estación terrena-satélite fijo).

Sobre la base de estos criterios, la asignación de los rangos concretos de frecuencias de subida y bajada depende del tipo de servicio y/o aplicación y de la región ITU, que puede ser 1, 2 o 3 tal como se muestra en la **Figura 4.33**. No obstante, algunos rangos son de uso generalizado; por ejemplo, la mayor parte de satélites que operan en la banda C, utilizan la sub-banda que va de 3.700 a 4.200 GHz en sus enlaces de bajada y la que va de 5.925 a 6.425 GHz en sus enlaces de subida. Es la llamada banda C estándar (*standard C-band*), a partir de la cual se han ido introduciendo ligeras variaciones dependiendo de los países. En cada satélite que opera en esta banda, el ancho de banda total de 500 MHz se distribuye entre sus múltiples transpondedores, cuyo ancho de banda se reparte, a su vez, en múltiples canales.

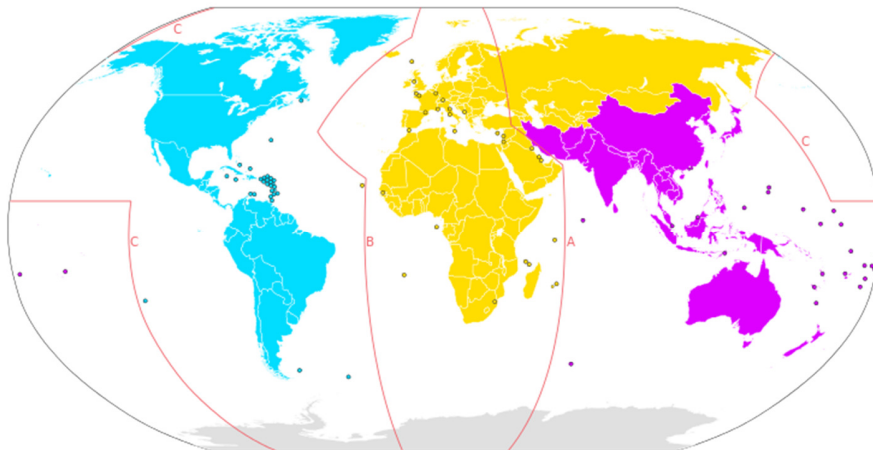


Figura 4.33. Regiones ITU para la asignación de frecuencias y líneas divisorias. La Región 1 comprende Europa, África, la antigua URSS, Mongolia y la parte occidental de Oriente Medio, incluyendo Irak; la Región 2 está formada por todo el continente americano, Groenlandia e islas del Pacífico oriental; finalmente, la Región 3 engloba Asia excepto la antigua URSS, incluyendo Irán, y la mayor parte de Oceanía. [Fuente:

https://commons.wikimedia.org/wiki/File:International_Telecommunication_Union_regions_with_dividing_lines.svg]

Tabla 4.10. Bandas de frecuencia utilizadas en los satélites.

Banda	Margen de frecuencias (GHz)	Aplicaciones
L	1 – 2	Sistemas de posicionamiento y navegación, MSS, enlace de bajada para datos e imágenes de alta resolución meteorológicos.
S	2 – 4	Sondeo remoto (<i>remote sensing</i>), satélites meteorológicos militares (<i>US Defense Meteorological Satellite Program</i>), comunicaciones NASA – Estación Espacial Internacional (ISS) y NASA-transbordador espacial, FSS, MSS y BSS (sólo en Asia y Oriente Medio).
C	4 – 8	FSS, BSS (en especial, muchos canales de TV internacionales – <i>international TV broadcast</i>).
X	8 – 12	Aplicaciones fundamentalmente militares y gubernamentales: exploración del espacio, radar meteorológico, control del tráfico aéreo y marítimo, etc.
K _u	12 – 18	FSS, DBS (canales de TV nacionales) y MSS.
K	18 – 27	
K _a	27 – 40	BSS.
V	40 – 75	Futuros servicios de comunicación comerciales.

En general, las bandas de trabajo más altas permiten mayores anchos de banda y antenas de menor tamaño para alcanzar los niveles de ganancia deseados, pero también están sujetas a una mayor degradación de la señal debido a la atenuación provocada por el vapor de agua, la lluvia o la nieve. Esta atenuación es especialmente importante en la franja que va de los 18 a los 27 GHz, y en concreto a la frecuencia de 22.24 GHz se produce un pico de resonancia de las moléculas de agua que absorbe prácticamente toda la energía de la señal. Por ello, la banda K original se subdividió en tres bandas, la banda K_u (*K-under*), la banda K propiamente dicha, no apta para las comunicaciones de largo alcance, y la banda K_a (*K-above*). De las dos bandas operativas, la banda K_a es más susceptible a la atenuación por lluvia que la banda K_u, si bien tanto una como otra presentan problemas en las zonas ecuatoriales. A pesar de ello, estas bandas han sido cada vez más utilizadas en las comunicaciones por satélite, por las ventajas, ya señaladas, que se derivan de trabajar en franjas más altas del espectro electromagnético, y también por la saturación de las bandas C (la primera banda de uso comercial introducida), S y L. En el otro extremo de esta última banda está la banda V, que permite elevadísimos anchos de banda, a costa de exigir visibilidad directa sin ningún tipo de obstáculo. En general, es una banda poco explotada en el ámbito de los satélites, si bien existen algunos ejemplos de constelaciones para comunicaciones militares donde los enlaces entre satélites utilizan una portadora de 60 GHz. En esta frecuencia se produce un importante pico de absorción provocado por las moléculas de oxígeno, lo cual proporciona protección contra interferencias e interceptaciones procedentes de estaciones terrenas enemigas. Finalmente, en la actualidad existen algunas propuestas de constelaciones en la banda V para fines comerciales.

Sistemas VSAT

El campo de los satélites de comunicaciones no ha sido una excepción dentro del proceso evolutivo de las tecnologías de la información y las comunicaciones hacia equipamientos más reducidos y

menos costosos. Los primeros satélites geoestacionarios proyectaban un único rayo que cubría casi una tercera parte de la superficie terrestre. Con el desarrollo de la microelectrónica y la ocupación de bandas de frecuencia cada vez mayores, fue posible integrar en un solo satélite múltiples transpondedores y múltiples antenas de menor tamaño, pudiendo proyectar simultáneamente varios rayos sobre zonas geográficas mucho más reducidas, del orden de unos pocos cientos de kilómetros de diámetro (*spots*). Gracias a ello, se pueden mantener varias comunicaciones simultáneas entre estaciones terrenas a través del mismo satélite. También las propias estaciones terrenas han evolucionado en esta dirección. De las grandes estaciones terrenas, dotadas de antenas de 10, 20 o 30 metros de diámetro, a los receptores GPS miniaturizados presentes en los teléfonos móviles, hay un trecho de unos 30 años. En un estadio intermedio están los **sistemas VSAT** (*very small aperture terminal*), constituidos por estaciones terrenas de pequeño tamaño, con diámetros de antena del orden de 1 m o menos. La potencia transmitida por estas estaciones es limitada (del orden de 1 W), así como la sensibilidad de sus receptores. Por todo ello, las velocidades de transmisión que pueden alcanzarse con estos sistemas no son muy altas, a lo sumo unos cuantos Mbps. No obstante, todas estas características propician que su coste sea reducido, al alcance de muchas empresas e incluso algunos usuarios domésticos, dependiendo de la velocidad contratada. Más específicamente, la tecnología VSAT representa una alternativa rentable para cualquier empresa que quiera establecer una red privada entre sus diversas delegaciones, al margen de cualquier operador público, o a veces la única alternativa posible para aquellos usuarios ubicados en zonas muy remotas, fuera del alcance de cualquier operador terrestre, sea por cable o de forma inalámbrica. En cualquiera de los casos, el sistema VSAT permite interconectar diversos tipos de equipos terminales para tráfico de voz, datos o vídeo.

Los satélites de los sistemas VSAT son geoestacionarios, de forma que las estaciones terrenas no requieren realizar ningún tracking; esto también contribuye a que el coste de las mismas sea reducido. Además, los transpondedores de tales satélites suelen ser los menos sofisticados, es decir, los de tubo doblado, por lo que en ellos la conexión entrada-salida es casi directa, consistiendo simplemente en una amplificación y una traslación de frecuencia.

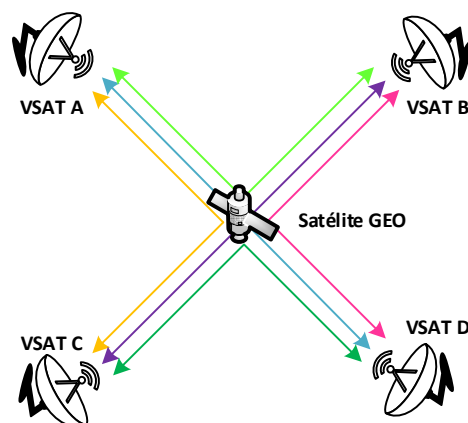


Figura 4.34. Red VSAT con topología mallada (caso de 4 estaciones VSAT).

Los sistemas VSAT se pueden presentar en dos topologías posibles: topología mallada (*mesh topology*) y topología en estrella (*star topology*). En la **topología mallada**, representada en la **Figura 4.34**, las estaciones VSAT se conectan directamente a través del satélite, es decir, mediante un solo salto (*one hop*). No obstante, a veces las limitaciones de estas estaciones no permiten su conexión directa, sino que dicha conexión tiene que realizarse a través de una estación intermedia, denominada **hub**, con mejores prestaciones en términos de potencia transmitida, ganancia de antena y sensibilidad de recepción. La estación intermedia puede ser origen o destino de las comunicaciones, o actuar como intermediaria entre dos estaciones VSAT. Ésta es la **topología en estrella** o de doble salto (*two hop*); se muestra en la **Figura 4.35** para una pareja de estaciones VSAT. Como puede observarse, la comunicación entre estas dos estaciones tiene lugar en dos fases o saltos: de una estación VSAT al *hub*, y del *hub* a la otra estación VSAT. El enlace correspondiente a la primera fase se denomina **enlace entrante** (*inbound link*), mientras que el enlace correspondiente a la segunda fase es el **enlace saliente** (*outbound link*). Así pues, tanto el enlace entrante como el saliente se componen de un enlace ascendente o de subida y un enlace descendente o de bajada. Este doble salto tiene la desventaja de duplicar el tiempo de propagación entre estaciones VSAT con respecto al caso de la topología mallada, pero cuenta con la ventaja de un mayor nivel de sensibilidad de recepción en el enlace entrante y una mayor potencia transmitida en el enlace saliente.

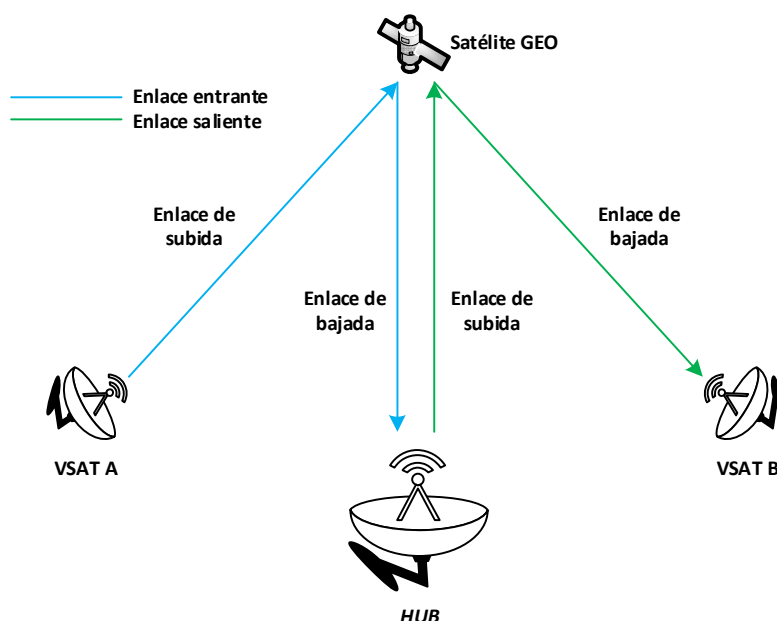


Figura 4.35. Red VSAT con topología en estrella. El *hub* es el elemento central de dicha topología.

Los sistemas VSAT con topología en estrella pueden ser, a su vez, de dos clases:

- **Bidireccionales** (*two-way*). En este caso, cada estación VSAT puede transmitir y recibir, y por tanto estos sistemas soportan aplicaciones interactivas. En el caso de una red VSAT de empresa

(red privada), el *hub* suele ubicarse en la sede central, y las estaciones VSAT en las delegaciones restantes.

- **Unidireccionales** (*one-way*). En este caso, el tráfico tiene lugar exclusivamente del *hub* a las estaciones VSAT, las cuales sólo tienen capacidad receptora. El ejemplo típico es el de los sistemas DBS descritos anteriormente.

Tabla 4.11. Algunas constelaciones relevantes (los datos proporcionados pueden variar). Se puede encontrar un listado muy completo de constelaciones en <https://www.newspace.im/>.

Constelación	Órbita	Nro. de satélites	Servicio	Operador
OG2	LEO (825/750 Km)	50	IoT/M2M, AIS ³⁵	Orbcomm
Inmarsat L-band, Global Xpress y EAN	GEO	14	IoT/M2M, telefonía y acceso a Internet (LMS, MMS y AMSS - WiFi aeronáutica)	Inmarsat
Iridium	LEO (780 Km)	66	IoT/M2M, telefonía y acceso a Internet (LMS, MMS y AMSS)	Iridium Communications Inc.
Globalstar	LEO (1410 Km)	48	IoT/M2M, telefonía y acceso a Internet (LMS y MMS)	Qualcomm, Alcatel
Starlink	LEO/SSO (múltiples altitudes)	1318 → 41493	Acceso a Internet	SpaceX
Eutelsat	GEO	35	DBS, acceso a Internet (LMS, MMS y AMSS), VSAT	Eutelsat ³⁶
Astra	GEO	16	DBS	SES S.A.
Hispasat	GEO	9	DBS, acceso a Internet, Hotspot WiFi, IoT/M2M	Grupo HISPASAT
O3b/O3b Mpower	MEO (\cong 8000 Km)	20 → 70	Backhaul Internet (para clientes corporativos multisectoriales)	SES S.A.
OneWeb	LEO (1200 Km)	146 → 648	Acceso a Internet	OneWeb
DMC3/TripleSat, KOMPSAT, SuperView, Vivid-i	SSO (500 – 685 Km)	24	Observación de la Tierra	Earth-i
Meteosat	GEO	4	Meteorología y monitorización del clima (Europa, África y Océano Índico)	EUMETSAT
Metop	SSO (817 Km)	3	Meteorología y monitorización del clima	EUMETSAT
GPS	MEO (20180 Km)	24	Navegación	USSF
Galileo	MEO (23222 Km)	24	Navegación	GSA, ESA
GLONASS	MEO (19130 Km)	24	Navegación	Roscosmos
BeiDou	3 GEO, 3 GSO ³⁷ , 24 MEO	30	Navegación	CNSA

³⁵ AIS es el acrónimo de *automatic identification system*. El servicio AIS de Orbcomm se utiliza principalmente para el seguimiento de flotas de barcos.

³⁶ Eutelsat también proyecta lanzar hasta 25 satélites LEO (flota ELO) para aplicaciones de IoT/M2M.

³⁷ Una **órbita GSO** es una órbita geosíncrona, es decir, su período coincide con el período de rotación de la Tierra, pero no tiene por qué ser una órbita GEO. En general, las órbitas GSO se sitúan en un plano inclinado con respecto al plano ecuatorial, y tienen una cierta excentricidad, es decir, son elípticas, si bien todas ellas tienen un semieje mayor igual a $35786 + 6371 = 42157$ Km (6371 Km es el radio de la Tierra). El movimiento

Constelaciones relevantes

Para concluir la descripción de los sistemas satelitales, en la **Tabla 4.11** se constatan las características básicas de algunas de las constelaciones más relevantes. Esta tabla representa tan solo un porcentaje muy reducido del total de constelaciones, que además va aumentando de año en año. En particular, cada vez son más las constelaciones que ofrecen servicios de comunicación para aplicaciones IoT/M2M, lo que indudablemente permitirá extender la cobertura de las futuras redes 5G. Por otra parte, la **Tabla 4.11** parece indicar que la tecnología satelital está solo al alcance de las grandes multinacionales o agencias gubernamentales, pero actualmente se está produciendo una evolución paralela que permite el acceso a esta tecnología de aquellas organizaciones con presupuestos más restringidos. Es la revolución de los satélites en miniatura, en particular los llamados **CubeSats**, unos satélites modulares que se desarrollan en unidades de 10 cm x 10 cm x 10 cm y pesan menos de 1 Kg, y que se lanzan como carga secundaria aprovechando otras misiones espaciales de mayor envergadura. El sitio <https://www.nanosats.eu/> constituye la base de datos más completa sobre satélites en miniatura.

4.5. PERTURBACIONES ESPECÍFICAS DE LAS TRANSMISIONES INALÁMBRICAS

Las transmisiones inalámbricas no adolecen de algunas de las perturbaciones que afectan a los medios cableados, como es el caso de la distorsión de retardo, pero están sujetas a numerosas perturbaciones de carácter específico. Este es el motivo por el cual normalmente los servicios inalámbricos presentan peor calidad. Al tratarse de transmisiones no guiadas, el grado de exposición a fenómenos externos es mucho mayor. Fenómenos como la presencia de obstáculos o la captación de fuentes de ruido del espacio exterior, dan lugar a señales indeseadas que se superponen a la señal que se pretende recibir. El propio medio atmosférico y las condiciones meteorológicas también provocan perturbaciones que alteran directamente la señal emitida por la fuente. A continuación, describimos una primera clasificación de todas estas perturbaciones.

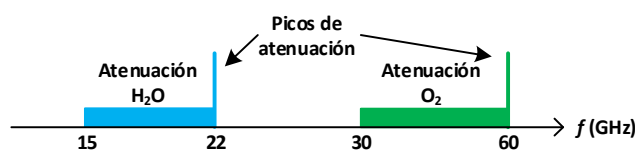


Figura 4.36. Franjas y picos de absorción atmosférica causados por el oxígeno y el agua.

Absorción atmosférica. El fenómeno de la absorción atmosférica se debe principalmente a la presencia de oxígeno y agua (en sus múltiples formas) en la atmósfera. Dependiendo de la frecuencia (portadora) de la señal, estas moléculas absorben una mayor o menor parte de la energía de la misma, lo que se traduce en una componente adicional de atenuación. En la **Figura 4.36** se muestran las franjas de frecuencia donde la atenuación provocada por el oxígeno y el agua es alta

aparente de un satélite en órbita GSO es un número 8 en el cielo, cuya forma exacta depende de la inclinación y excentricidad de la órbita. Este número 8 se dibuja una vez cada día sideral. Por lo tanto, podemos afirmar que una órbita GEO es una órbita GSO con inclinación nula.

o incluso presenta un fuerte pico. La influencia de estos picos en las comunicaciones inalámbricas puede ser muy significativa, como ya hemos visto en el caso de las comunicaciones por satélite.

Obstrucción (*shadowing*³⁸). Este fenómeno se debe a la presencia de obstáculos en la línea que une el transmisor con el receptor. Un obstáculo absorbe parte de la energía de la señal, pudiendo provocar incluso el bloqueo de la misma cuando la atenuación resultante es muy grande. Si realizamos el experimento de mantener fija la posición del transmisor y variar la posición del receptor, ya sea en sentido radial (acercando o alejando el receptor en una determinada dirección), ya sea en sentido circular (manteniendo fija la distancia al transmisor), observaríamos variaciones en el nivel de señal recibida precisamente sobre distancias proporcionales a los tamaños de tales objetos (10 – 100 metros en entornos exteriores, y menos en entornos interiores). Por lo tanto, son **efectos de propagación a gran escala** (*large-scale propagation effects*).

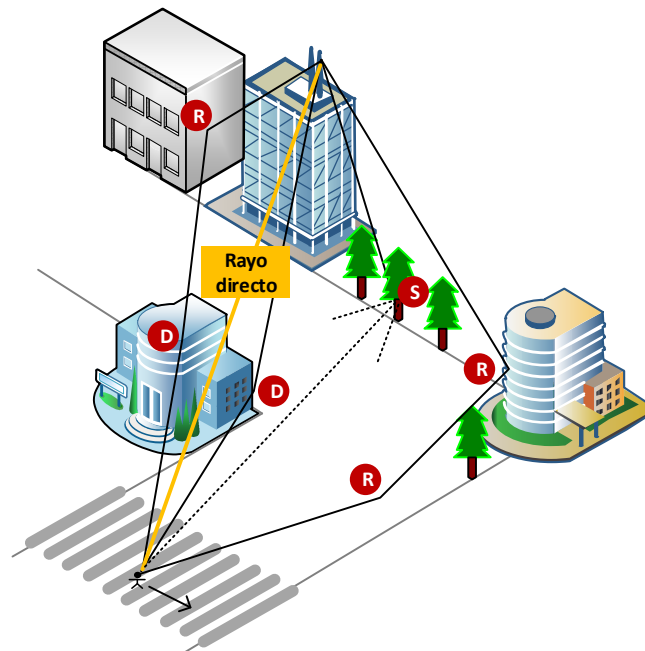


Figura 4.37. Propagación multicamino en un entorno exterior urbano. Aparte del rayo directo (si no resulta bloqueado), los rayos que alcanzan un determinado punto pueden deberse a reflexión (R), difracción (D) o *scattering* (S).

Propagación multicamino. La propagación multicamino se debe a la presencia de obstáculos en el entorno del camino que une el transmisor con el receptor, de manera que este último recibe, además de la señal deseada correspondiente al rayo directo (si no ha sido bloqueado), múltiples réplicas de dicha señal con amplitudes y retardos diversos. Estas amplitudes son generalmente menores que la amplitud de la señal directa, y los retardos son mayores que el de ésta. Tanto si el rayo directo ha sido bloqueado como si no, el fenómeno que tiene lugar es la recepción de una señal

³⁸ El término utilizado en inglés para designar la obstrucción es *shadowing*, pero en ocasiones este término designa el debilitamiento de la señal provocado por cualquier perturbación, no sólo la obstrucción.

que es suma (superposición) de múltiples señales; dependiendo de las amplitudes de estas señales y de sus diferencias de retardo en el punto donde se encuentra el receptor, la señal suma puede presentar una amplitud mayor o menor. Si repitiéramos el experimento anterior consistente en mover radial o circularmente el receptor, comprobaríamos que las variaciones del nivel de señal se producen en este caso sobre distancias cortas, del orden de la longitud de onda correspondiente a la frecuencia portadora. Son, por ello, **efectos de propagación a pequeña escala** (*small-scale propagation effects*).

Como se ilustra en la **Figura 4.37**, las réplicas de la señal directa pueden provenir de tres orígenes: reflexión, difracción y *scattering*. La **reflexión** tiene lugar cuando una señal electromagnética incide en una superficie que es grande en comparación con la longitud de onda portadora de la señal. La **difracción** se produce cuando la señal incide en el flanco (borde, esquina, etc.) de algún objeto grande en comparación con la longitud de onda portadora de la señal. En este caso, la señal rebota en múltiples direcciones, como si el propio flanco fuera la fuente. Cuando la señal incide en un objeto cuyas dimensiones son del orden de la longitud de onda portadora o inferior, entonces es dispersada por el objeto en forma de múltiples señales débiles en todas las direcciones. Es el fenómeno de **scattering**.

Salvo situaciones concretas en que una o varias réplicas de la señal directa se superponen favorablemente con ella, en general el efecto de todas las perturbaciones anteriores es un debilitamiento del nivel de señal llamado desvanecimiento o **fading**, que se añade a la atenuación debida a la pérdida por propagación en espacio libre. En general, el **fading** no sólo varía de forma espacial, sino también temporalmente, pues tanto los obstáculos como el propio receptor pueden estar en movimiento. Ejemplos de obstáculos en movimiento son el tráfico rodado, las hojas de los árboles cuando hay viento o la superficie del mar; el ejemplo más claro de receptores en movimiento nos lo proporciona a diario la telefonía móvil. Además, todas estas variaciones dependen de las características específicas de cada entorno, y por tanto son aleatorias y a lo sumo solo predecibles estadísticamente. Por todo ello, la caracterización del nivel de **fading** es enormemente compleja y la forma de abordarla es mediante **modelos de propagación**. Como modelos que son, resultan de simplificar la realidad, pero ofrecen la ventaja de ser manejables y proporcionar estimaciones por lo general aceptables. Todos incluyen la pérdida por propagación en espacio libre, de modo que el resultado que proporcionan es la atenuación total experimentada por la señal. Fundamentalmente hay tres clases de modelos:

Modelos deterministas. Son los llamados **modelos de rayos** (*ray tracing*), que resultan aplicables a escenarios de geometría sencilla en los que solamente están presentes la señal directa y un conjunto de réplicas producidas por reflexión. Estos modelos se obtienen aplicando las leyes básicas de reflexión de rayos y proporcionan el nivel de **fading** en función de parámetros geométricos y dieléctricos del entorno y de las características de radiación de las antenas transmisora y receptora. Para distancias relativamente grandes, los modelos de rayos dan lugar a una expresión aproximada que decrece monótonamente en función de la distancia elevada a una cierta potencia, conocida como **exponente de pérdida por propagación**. Su valor suele estar entre 2 (caso ideal, cuando el único rayo es el directo) y 6. A estos modelos se los denomina **modelos de pérdida por propagación**

(*path-loss models*), e incluyen el modelo de pérdida por propagación en espacio libre como caso particular. Al igual que la obstrucción, los modelos de pérdida por propagación dan lugar a variaciones de señal a gran escala (de hecho, a una escala mayor que la correspondiente a la obstrucción).

Modelos empíricos. Son modelos cuya expresión matemática no es el resultado de un procedimiento analítico, sino de intensas y extensas mediciones en escenarios representativos de diferentes entornos, tanto de interior (*indoor models*) como de exterior (*outdoor models*): natural, semiurbano, urbano, edificio de oficinas, doméstico, fábrica, almacén, etc. Suelen estar formulados en términos de parámetros geométricos y de factores de corrección para tener en cuenta, por ejemplo, diversos tipos de terreno.

Modelos estadísticos. Ni los modelos deterministas ni los empíricos tienen en cuenta la aleatoriedad del *fading*. Por el contrario, los modelos estadísticos permiten hacer estimaciones probabilísticas del nivel de señal recibida, las cuales permiten a su vez la obtención de métricas de calidad de servicio, como por ejemplo la probabilidad de que el nivel de señal recibida en un punto sea inferior a un umbral durante un cierto porcentaje de tiempo. Se utilizan mucho en diversos ámbitos (investigación, planificación de redes móviles e inalámbricas, etc.). Una categoría de modelos estadísticos la constituyen los *modelos de shadowing* (*shadowing models* o *shadow fading models*), válidos cuando los fenómenos dominantes son la pérdida por propagación (cualquiera que sea el valor del exponente) y la obstrucción, en definitiva, los fenómenos que provocan variaciones de señal a gran escala. Estos modelos son útiles, por ejemplo, para planificar la cobertura de las redes celulares de telefonía móvil. La categoría más completa la forman los *modelos de fading multicamino* (*multipath fading models*), que son capaces de capturar las variaciones temporales, a cualquier escala, del nivel de señal detectado por un receptor al moverse, ya que tienen en cuenta todas las perturbaciones, es decir, pérdida por propagación, obstrucción y propagación multicamino. A su vez, estas variaciones temporales pueden ser rápidas (*fast fading*) o lentas (*slow fading*).

CAPÍTULO 5. LIMITACIONES DE LA TRANSMISIÓN DIGITAL

Vistas las bondades de la transmisión digital, este capítulo se ocupa de sus limitaciones, las cuales son consecuencia de las perturbaciones consideradas en el capítulo anterior. Las **limitaciones de la transmisión digital** se reflejan en unos parámetros o métricas de comportamiento, fundamentalmente la eficiencia espectral y la tasa de errores de bit, que conjuntamente determinan la calidad de un canal de transmisión. Es posible mejorar esta calidad trabajando con codificaciones de línea o modulaciones digitales espectralmente más eficientes, utilizando medios de transmisión con mayores prestaciones, y robusteciendo la comunicación vía codificación de canal. No obstante, uno de los resultados más importantes de la Teoría de la Información, el Teorema Fundamental de Shannon, establece un límite inalcanzable para la velocidad de transmisión a través de un canal caracterizado por unos valores finitos de ancho de banda y relación señal-ruido. Por tanto, este capítulo cierra un ciclo que comienza y acaba en la Teoría de la Información.

ÍNDICE

5.1.	ANCHO DE BANDA EFECTIVO DE UNA SEÑAL DIGITAL	137
5.2.	CRITERIO DE NYQUIST Y FÓRMULA DE SHANNON	138
5.3.	MÉTRICAS DE COMPORTAMIENTO	142

5.1. ANCHO DE BANDA EFECTIVO DE UNA SEÑAL DIGITAL

Una de las características comunes a todas las codificaciones de línea estudiadas en el Capítulo 3 es su fundamentación en el pulso rectangular como elemento básico de construcción de los símbolos. Este pulso es, pues, el llamado **elemento de señal**. Dependiendo del tipo de codificación, la duración de un elemento de señal, T_s , puede corresponder a una fracción del tiempo de bit, a un tiempo de bit o a múltiples tiempos de bit. Se define la **velocidad de señalización**³⁹ R_s como la velocidad a la que se transmiten los elementos de señal, es decir, $R_s = 1/T_s$. Su unidad es el **baudio**, de manera que 1 baudio equivale a 1 elemento de señal por segundo. Lógicamente, la relación entre la duración de un elemento de señal y el tiempo de bit es la relación entre velocidad de transmisión y velocidad de señalización. Si r denota dicha relación, se tiene:

$$r = \frac{T_s}{T_b} = \frac{R}{R_s} \quad (5.1)$$

La **Figura 5.1** muestra varios casos para este cociente. El caso $r = 1$ podría corresponder, por ejemplo, a una codificación NRZ polar, el caso $r = 1/2$ a una codificación Manchester, el caso $r = 2$ a un código 2B1Q y, finalmente, el caso $r = 3/2$ a un código 3B2T.

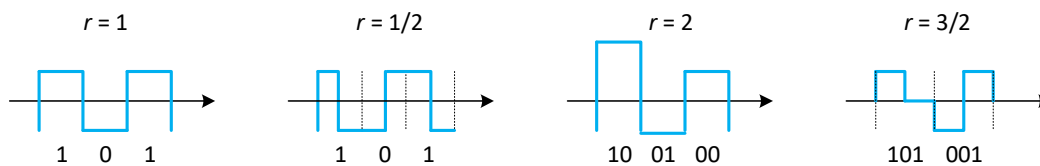


Figura 5.1. Diversas relaciones entre velocidad de transmisión y velocidad de señalización.

La importancia de la velocidad de señalización radica en su influencia en el ancho de banda de la señal digital; de hecho, es la velocidad de señalización y no la de transmisión la que determina directamente dicho ancho de banda. No obstante, estrictamente hablando, el espectro promedio de cualquier señal digital abarca todo el rango de frecuencias, por lo que el ancho de banda absoluto de cualquier codificación de línea resulta ser infinito, y por tanto no caben distinciones entre ellas. Ahora bien, en todos los casos sólo una parte del espectro promedio es significativa, pudiendo considerarse nula la parte restante, de modo que el ancho de banda de las codificaciones de línea es, a efectos prácticos, finito. Éste es el llamado **ancho de banda efectivo** de una señal digital, y es el que consideraremos a partir de ahora. Si tenemos en cuenta que en transmisión digital la distorsión no representa un problema mientras el receptor pueda recuperar la información binaria contenida en la señal, la necesidad de introducir un ancho de banda efectivo que contrasta con el ancho de banda absoluto de valor infinito, queda todavía más justificada.

³⁹ También se la conoce como **velocidad de modulación**. En inglés, la variedad de denominaciones es todavía mayor: *signal rate*, *modulation rate*, *pulse rate* y *baud rate*.

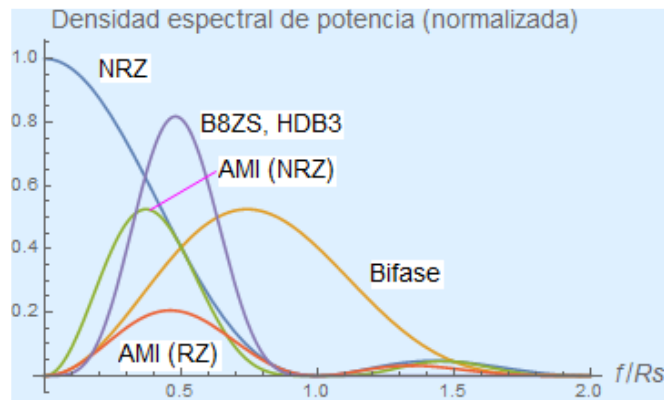


Figura 5.2. Densidad espectral de potencia (normalizada) en función de la frecuencia normalizada con respecto a la velocidad de transmisión, para diversas codificaciones de línea de uso común.

La **Figura 5.2** muestra el espectro promedio de algunos códigos de línea habituales sobre un eje de frecuencias normalizado con respecto a la velocidad de señalización. El ancho de banda efectivo de una señal digital está centrado precisamente en la llamada **frecuencia central** de su espectro, que denotamos por f_{bbc} , pues viene a ser la frecuencia “portadora” en banda base (*baseband carrier*). Por otra parte, una estimación del ancho de banda efectivo obedece a la siguiente expresión:

$$B_{efec} = \frac{1}{2}R_s = \frac{1}{2} \cdot \frac{R}{r} \quad (5.2)$$

Esta expresión pone de manifiesto que, fijada la codificación, cuanto mayor es la velocidad de transmisión, mayor es el ancho de banda de la señal. Esta conclusión no resulta difícil de intuir, ya que cuanto mayor es la velocidad de transmisión, más rápidas son las variaciones de la señal digital (cualquiera que sea el tipo de codificación), y por tanto mayor es su contenido en alta frecuencia.

Si consideramos la frecuencia normalizada de la **Figura 5.2**, el ancho de banda efectivo es siempre 0.5. A modo de ejemplo, consideremos dos codificaciones de línea, NRZ y Manchester (bifase), ambas a la velocidad de transmisión de 200 Kbps. Aplicando la expresión (5.2) y observando la **Figura 5.2**, para la señal Manchester obtenemos $R_s = 400$ KBaud (recordemos que $r = 1/2$ en una señal Manchester), $B_{efec} = 200$ KHz y $f_{bbc} = 320$ KHz, pues la frecuencia central normalizada se sitúa alrededor del valor 0.8; los valores correspondientes para la señal NRZ son: $R_s = 200$ KBaud ($r = 1$ en una señal NRZ), $B_{efec} = 100$ KHz y $f_{bbc} = 0$ KHz. La señal Manchester consume, por tanto, el doble de ancho de banda que la señal NRZ. Ésta es su principal desventaja.

5.2. CRITERIO DE NYQUIST Y FÓRMULA DE SHANNON

Dos resultados ineludibles en transmisión digital son el criterio de Nyquist y la fórmula de Shannon. Ambos reflejan algo que también cabía suponer: cuanto mayor es el ancho de banda de un canal, mayor es la velocidad de transmisión que permite alcanzar (de hecho, hoy en día se intercambian ambos términos). No obstante, las condiciones asumidas en ambas formulaciones son distintas. Empezamos con el criterio de Nyquist.

Criterio de Nyquist

Formalmente, el **criterio de Nyquist** establece lo siguiente (la traducción no es literal):

Dado un canal paso-bajo ideal de ancho de banda B_c , es posible transmitir a una velocidad de señalización $R_s \leq 2 \cdot B_c$ sin interferencia inter-simbólica (ISI). No es posible transmitir a una velocidad de señalización $R_s \leq 2 \cdot B_c$.

Recordando la expresión (5.2) para el ancho de banda efectivo de una señal digital, la condición dada por el criterio de Nyquist no es más que $B_c \geq B_{efec}$, algo que no debe sorprender si tenemos en cuenta que el canal no tiene que distorsionar la señal digital más allá de la distorsión tolerada implícitamente en el concepto de ancho de banda efectivo. Como el ancho de banda de los medios de transmisión es un recurso escaso, de ahora en adelante asumiremos que $B_c = B_{efec}$ (de hecho, este ancho de banda B_{efec} no es más que el ancho de banda B_r de la **Figura 4.7**).

Para alcanzar la cota superior definida por el criterio de Nyquist ($R_{s,max} = 2 \cdot B_c$), el elemento de señal debería consistir en un tipo especial de pulso, no rectangular y de duración infinita, cuyos valores en los instantes de muestreo de los símbolos vecinos fueran nulos, de forma que no provocara ISI. No obstante, el uso de un elemento de señal de duración infinita no es viable, por lo que la velocidad de señalización real siempre será inferior al máximo teórico. De todos modos, en la práctica pueden alcanzarse valores muy próximos a ese máximo.

Teniendo en cuenta la expresión (5.1), podemos reformular el criterio de Nyquist de la siguiente manera:

$$R_s = \frac{R}{r} \leq 2 \cdot B_c \Rightarrow R \leq 2 \cdot B_c \cdot r \quad (5.3)$$

En el caso de que la duración de un elemento de señal corresponda a un número entero de bits, es decir, $r = \log_2 M$, con M una potencia de 2 igual al número de símbolos que maneja el canal, tenemos:

$$R \leq 2 \cdot B_c \cdot \log_2 M \quad (5.4)$$

Fórmula de Shannon

Aparentemente, según la expresión (5.4) podemos aumentar indefinidamente la velocidad de transmisión del canal a base de aumentar el número de símbolos (M), con la única limitación de la complejidad del receptor: dada una potencia de señal, cuanto mayor es el número de símbolos, mayor es el grado de sofisticación exigido al receptor, pues debe ser capaz de discernir entre símbolos cada vez más próximos entre sí. Sin embargo, en la práctica no es éste el único factor limitador, pues el canal también introduce ruido, elemento no considerado en el criterio de Nyquist. Fue Shannon quien, en 1944, desarrolló una fórmula para la capacidad teórica de un canal de ancho de banda limitado en presencia de ruido.

Recordemos que el problema de maximizar la cantidad de información transferida por un canal quedaba recogido en las ecuaciones (1.27) del Capítulo 1. Dichas ecuaciones, juntamente con la ecuación (1.24), ponían de manifiesto que, para maximizar la información mutua media, era necesario optimizar la fuente, por un lado, y el canal, por otro. En lo que respecta a la fuente, el objetivo se concretaba en maximizar su entropía teniendo en cuenta las restricciones básicas sobre las probabilidades de los símbolos, es decir:

$$\max_{p(x_i)} H(X) \quad (5.5a)$$

tal que:

$$p(x_i) \geq 0 \quad (5.5b)$$

$$\sum_{i=1}^S p(x_i) = 1 \quad (5.5c)$$

De hecho, ya conocemos la solución a este problema, pues en el Capítulo 1 vimos que la entropía máxima era $\log_2 M$, y que ésta se alcanzaba para el caso en que todos los símbolos tenían la misma probabilidad. En esto consistía precisamente la codificación de fuente (véase la **Figura 1.15**). No obstante, todo esto es cierto cuando la fuente es discreta, es decir, toma valores dentro de un conjunto discreto, lo que no deja de ser una restricción. Si bien esto es inevitable en transmisión digital, la contribución de Shannon fue más allá, pues consistió en determinar la máxima entropía de la fuente con las mínimas restricciones posibles, es decir, sin imponer siquiera que la fuente fuera discreta. Así, el problema fue formulado del siguiente modo:

$$\max_{p(x)} H(X) \quad (5.6a)$$

tal que:

$$p(x) \geq 0 \quad (5.6b)$$

$$\int_{-\infty}^{+\infty} p(x) \cdot dx = 1 \quad (5.6c)$$

Observamos que la función de probabilidad $p(x_i)$ se ha convertido en una función de densidad $p(x)$, y la fuente discreta se ha transformado en una fuente continua en valores. Sin embargo, ¿tiene sentido una fuente continua en valores en un contexto digital? La respuesta es afirmativa si recordamos que toda fuente analógica con un ancho de banda limitado queda completamente representada por una secuencia de muestras si el ritmo de generación de éstas satisface el teorema de muestreo. Por tanto, a los efectos de encontrar la máxima entropía de una fuente, el trabajo de Shannon se desarrolló para una hipotética fuente que produjera una señal “digital” del tipo PAM, con las únicas restricciones dadas por (5.6b) y (5.6c). En estas condiciones, la función de densidad que se obtiene como solución al problema de optimización dado por (5.6) viene dada por la siguiente expresión:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{x^2}{2\sigma^2}} \quad (5.7)$$

Esta función de densidad corresponde a una variable aleatoria normal (gaussiana) de media 0 y varianza σ^2 . Es lógico que la media sea 0, pues lo contrario significaría que la señal PAM presentara

una componente continua, que ya sabemos que no aporta información alguna. Por otro lado, se puede demostrar que la varianza σ^2 coincide con la potencia media de dicha señal.

Para la función de densidad dada por (5.7), la entropía máxima que se obtiene es la siguiente:

$$H(X)|_{max} = \frac{1}{2} \log_2(2\pi \cdot e \cdot \sigma^2) \quad (5.8)$$

En lo que al canal se refiere, recordemos que el objetivo era minimizar el término de equivocación. En este caso, la formulación del problema es necesariamente continua, pues como mínimo el ruido que se superpone a la señal de entrada es continuo en valores. Específicamente, si $x = x(t)$ es la señal generada por la fuente (señal de entrada) y $n = n(t)$ es el ruido introducido por el canal, entonces la señal de entrada al receptor (señal de salida) es $y = y(t) = x(t) + n(t)$, y el problema de minimizar la equivocación puede expresarse de la siguiente forma:

$$\min_{p(y|x)} H(X|Y) \quad (5.9a)$$

tal que:

$$p(y|x) \geq 0 \quad (5.9b)$$

$$\int_{-\infty}^{+\infty} p(y|x) \cdot dx = 1 \quad (5.9c)$$

Shannon encontró la solución a este problema partiendo de las siguientes hipótesis:

1. El canal tiene un ancho de banda limitado de valor B_c , y dentro de este ancho de banda no produce distorsión ni ganancia o atenuación, es decir, equivale a un efecto multiplicativo de valor 1 dentro de su ancho de banda.
2. El canal impone que la señal generada por la fuente tenga un ancho de banda finito y una potencia media finita de valor $P_t = \sigma^2$. Además, puesto que el canal equivale a un efecto multiplicativo de valor 1, este valor corresponde también a la potencia recibida P_r .
3. La señal a la entrada del receptor está contaminada con **ruido blanco gaussiano** (en inglés, **additive white Gaussian noise** o **AWGN**) de media 0 y potencia $N = N_0 \cdot B_c$. Ambos calificativos, “blanco” y “gaussiano”, describen el comportamiento del ruido en los dominios de frecuencias y tiempo respectivamente. En particular, el término “blanco” se refiere a que la densidad espectral de potencia de ese ruido es plana, mientras que el carácter gaussiano indica que el valor del ruido en cualquier instante de tiempo obedece a una distribución normal:

$$p(n) = \frac{1}{\sqrt{2\pi N}} \cdot e^{-\frac{n^2}{2N}} \quad (5.10)$$

El ruido térmico obedece a todas estas características.

4. La señal generada por la fuente, $x = x(t)$, y el ruido introducido por el canal, $n = n(t)$, son independientes, y por tanto la potencia de la señal a la entrada del receptor, $y = y(t)$, viene dada por la suma de potencias, es decir, $P_r + N$ (si se trabaja con unidades lineales).

Estas cuatro condiciones definen un modelo de canal conocido como **canal AWGN**. Considerando este modelo, la solución que se alcanza para el problema (5.9) es la siguiente:

$$H(X|Y)|_{min} = \frac{1}{2} \log_2 \frac{2\pi \cdot e \cdot \sigma^2 \cdot N}{\sigma^2 + N} \quad (5.11)$$

Por lo tanto, la solución al problema (1.27) en la versión continua en valores queda de la siguiente manera:

$$C_s = H(X)|_{max} - H(X|Y)|_{min} = \frac{1}{2} \log_2(2\pi \cdot e \cdot \sigma^2) - \frac{1}{2} \log_2 \frac{2\pi \cdot e \cdot \sigma^2 \cdot N}{\sigma^2 + N} = \frac{1}{2} \log_2 \frac{2\pi \cdot e \cdot \sigma^2}{\frac{2\pi \cdot e \cdot \sigma^2 \cdot N}{\sigma^2 + N}} = \frac{1}{2} \log_2 \frac{\sigma^2 + N}{N} = \frac{1}{2} \log_2 \left(1 + \frac{\sigma^2}{N}\right) \quad (5.12)$$

Combinando este resultado con la velocidad de señalización máxima según el criterio de Nyquist ($R_{s,max} = 2 \cdot B_c$), la capacidad del canal de transmisión dada por la ecuación (1.28) queda de la siguiente manera (recuérdese que $\sigma^2 = P_r$):

$$C = B_c \cdot \log_2(1 + SNR|_{lineal}) \quad (5.13)$$

Esta expresión definitiva de la capacidad de un canal de transmisión, conocida como **fórmula de Shannon**, culmina la formulación del teorema fundamental de Shannon para un canal ruidoso. Es importante observar que el valor proporcionado por la fórmula de Shannon es inalcanzable, tanto por la parte referente al criterio de Nyquist como por las condiciones en que se ha optimizado la entropía de la fuente. No obstante, es posible alcanzar velocidades de transmisión cercanas a dicha capacidad combinando codificaciones de línea que den lugar a tasas de errores de bit reducidas y codificaciones de canal. En la siguiente sección, dedicada a las métricas de comportamiento, se constatan las tasas de errores de bit de algunas codificaciones de línea habituales.

5.3. MÉTRICAS DE COMPORTAMIENTO

El teorema fundamental de Shannon para un canal ruidoso establece que, si la velocidad de transmisión no supera la capacidad del canal, se puede encontrar un sistema de codificación tal que la tasa de errores de bit sea arbitrariamente pequeña. Este sistema de codificación engloba definitivamente sus tres variantes, es decir, codificación de fuente, de línea y de canal, pues todas contribuyen a optimizar la cantidad de información que se transmite por unidad de tiempo en un canal de comunicaciones. La tasa de errores de bit viene determinada directamente por la codificación de línea, que luego puede complementarse con la codificación de canal que veremos más adelante. De hecho, recordemos que un elemento esencial en la caracterización de los códigos de línea lo constituían las llamadas métricas de comportamiento, que englobaban la tasa de errores de bit y la eficiencia espectral. Estas métricas también caracterizan las modulaciones digitales, pero trataremos ambos casos por separado.

Tasa de errores de bit y eficiencia espectral de las codificaciones de línea

Se puede demostrar que para una señalización polar de M símbolos mapeados en bits según un **código Gray** (de forma que los símbolos vecinos sólo difieran en 1 bit), un canal AWGN y una relación

señal-ruido a la entrada del receptor suficientemente grande, la **probabilidad de error de bit** generada por una codificación de línea, p_b , viene dada por la siguiente expresión:

$$p_b = 2 \frac{M-1}{M \cdot \log_2 M} Q \left(\sqrt{\frac{6}{M^2-1} \cdot \frac{P_r}{N_0 \cdot R_s}} \right) \quad (5.14)$$

Recuérdese que M es una potencia de 2. La función $Q = Q(x)$ es la probabilidad de que una variable aleatoria Gaussiana de media m y varianza σ^2 , es decir, $N(m, \sigma^2)$, tome un valor mayor o igual que $m + \sigma \cdot x$. Su cálculo obedece a la siguiente expresión:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} \cdot du \quad (5.15)$$

Desafortunadamente esta integral no tiene solución cerrada, pero puede calcularse con la ayuda de tabla, gráficas o programas que la calculen. Combinando las expresiones (5.14) y (5.1), tenemos:

$$p_b = 2 \frac{M-1}{M \cdot \log_2 M} Q \left(\sqrt{\frac{6}{M^2-1} \cdot \frac{P_r}{N_0 \cdot R} \cdot r} \right) \quad (5.16)$$

Además, se cumple la siguiente relación:

$$\frac{P_r}{N_0 \cdot R} = \frac{P_r \cdot T_b}{N_0} = \frac{E_b}{N_0} \quad (5.17)$$

E_b es la **energía por bit**, es decir, la energía contenida en 1 bit, pues es el resultado de multiplicar la potencia media de señal por 1 tiempo de bit. El parámetro E_b/N_0 aparece siempre en los cálculos de las tasas de errores de bit, incluso en el caso de las modulaciones digitales.

Combinando las expresiones (5.16) y (5.17) se tiene:

$$p_b = 2 \frac{M-1}{M \cdot \log_2 M} Q \left(\sqrt{\frac{6 \cdot r}{M^2-1} \cdot \frac{E_b}{N_0}} \right) \quad (5.18)$$

Si además suponemos que el elemento de señal corresponde a un número entero de bits, es decir, $r = \log_2 M$, la expresión anterior se particulariza de la siguiente forma:

$$p_b = 2 \frac{M-1}{M \cdot \log_2 M} Q \left(\sqrt{\frac{6 \cdot \log_2 M}{M^2-1} \cdot \frac{E_b}{N_0}} \right) \quad (5.19)$$

En la **Figura 5.3** se representa gráficamente la expresión (5.18) para una codificación Manchester ($M = 2, r = 1/2$) y la expresión (5.19) para diversos valores de M . Naturalmente, al aumentar el parámetro E_b/N_0 , directamente proporcional a la relación señal-ruido, la tasa de errores de bit disminuye.

La otra métrica de comportamiento importante es la llamada **eficiencia espectral**, η , que obedece a la siguiente expresión:

$$\eta = \frac{R}{B_c} \quad (5.20)$$

Se mide en bps/Hz y mide el grado de aprovechamiento de un cierto ancho de banda de canal (recuérdese que el ancho de banda de los medios de transmisión es un recurso escaso). A diferencia de otras muchas eficiencias, la eficiencia dada por (5.20) puede ser mayor que la unidad, y, además, interesa que sea lo mayor posible. Combinando la expresión de la eficiencia y (5.17) obtenemos una relación entre la relación señal-ruido convencional y el parámetro E_b/N_0 :

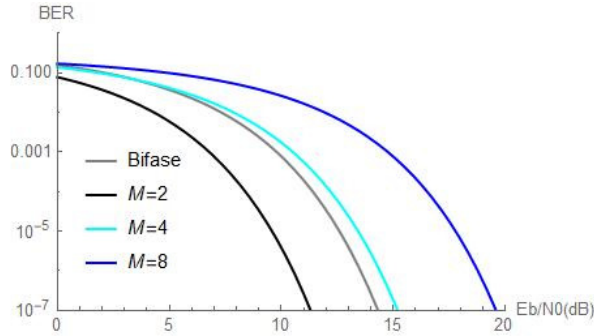


Figura 5.3. Tasa de errores de bit para diversas codificaciones de línea.

$$SNR|_{\text{lineal}} = \frac{P_r}{N} = \frac{P_r}{N_0 \cdot B_c} = \frac{P_r}{N_0 \cdot B_c} \cdot \frac{R}{R} = \frac{P_r \cdot T_b}{N_0} \cdot \frac{R}{B_c} = \frac{E_b}{N_0} \eta \quad (5.21)$$

En unidades logarítmicas:

$$SNR(dB) = \frac{E_b}{N_0}(dB) + 10 \cdot \log_{10} \eta \quad (5.22)$$

Las expresiones (5.18) y (5.19) se pueden combinar con la expresión (5.21) para obtener la tasa de errores de bit en función de la relación señal-ruido convencional. En el caso general:

$$p_b = 2 \frac{M-1}{M \cdot \log_2 M} Q \left(\sqrt{\frac{6 \cdot r}{M^2-1} \cdot \frac{SNR|_{\text{lineal}}}{\eta}} \right) \quad (5.23)$$

Y en el caso de que la duración de un elemento de señal corresponda a un número entero de bits:

$$p_b = 2 \frac{M-1}{M \cdot \log_2 M} Q \left(\sqrt{\frac{6 \cdot \log_2 M}{M^2-1} \cdot \frac{SNR|_{\text{lineal}}}{\eta}} \right) \quad (5.24)$$

Recordemos que hemos ubicado siempre la relación señal-ruido en el punto de entrada al receptor, aunque, para ser más precisos, es la que tiene lugar en el punto de pre-detección de la señal, tal como se indica en la **Figura 5.4**. La razón es que ese es el punto justo antes del proceso donde tiene lugar la decodificación de la señal en ceros y unos, y por tanto donde tienen lugar los posibles errores de bit. Si la primera etapa (compensación de las perturbaciones) no genera ruido (ruido interno nulo), entonces la **relación señal-ruido en pre-detección** coincide con la relación señal-ruido a la entrada del receptor.

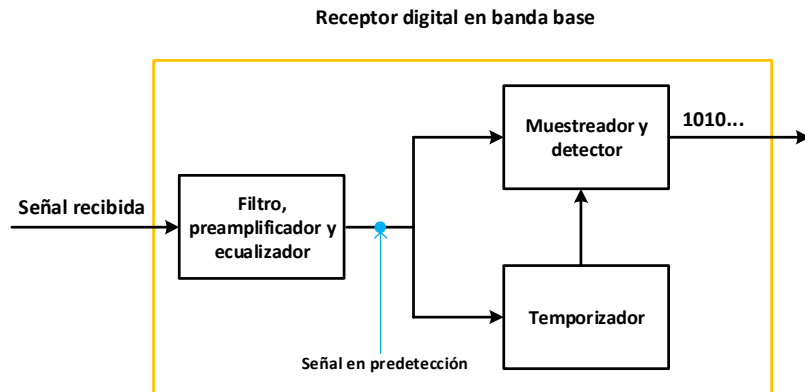


Figura 5.4. Punto de pre-detección en un receptor digital. Obsérvese que el receptor digital tiene la misma estructura que un repetidor regenerativo, pero sin la etapa de regeneración.

Tasa de errores de bit y eficiencia espectral de las modulaciones digitales

El ancho de banda de canal consumido por las modulaciones digitales es mayor que el correspondiente a las codificaciones de línea. Para paliar este aumento, a la salida del modulador se suele colocar un filtro que reduce el ancho de banda de la señal modulada, sin que se degrade la detección de símbolos en el lado destinatario, aunque a costa de aumentar la complejidad del receptor. La presencia de este filtro a la salida del modulador se refleja en el llamado **factor de filtrado** ρ , que satisface la condición $0 \leq \rho \leq 1$. Entonces, para el ancho de banda exigido por las diversas modulaciones se admiten las siguientes expresiones:

- Modulaciones basadas en combinaciones híbridas multinivel de ASK y PSK:

$$B_c = (1 + \rho) \cdot R_s \quad (5.25)$$

- Modulaciones basadas en extensiones multinivel de FSK:

$$B_c = M \cdot (1 + \rho) \cdot R_s \quad (5.26)$$

Observamos que las modulaciones basadas en FSK consumen mucho más ancho de banda que las restantes. La eficiencia espectral queda en ambos casos de la siguiente manera:

- Modulaciones basadas en combinaciones híbridas multinivel de ASK y PSK:

$$\eta = \frac{R}{B_c} = \frac{R}{(1+\rho) \cdot R_s} = \frac{\log_2 M}{1+\rho} \quad (5.27)$$

- Modulaciones basadas en extensiones FSK multinivel:

$$\eta = \frac{R}{B_c} = \frac{R}{(1+\rho) \cdot R_s} = \frac{\log_2 M}{M \cdot (1+\rho)} \quad (5.28)$$

El mayor consumo de ancho de banda de las modulaciones basadas en FSK se traduce en una menor eficiencia espectral.

La probabilidad de error de bit de las modulaciones digitales depende significativamente del mecanismo de detección utilizado en el receptor, que en general admite diversas opciones. La **Tabla 5.1** recoge dicha probabilidad para diversas modulaciones digitales con los mecanismos de detección señalados. Nuevamente observamos que las expresiones para la probabilidad de error de bit dependen directamente del parámetro $\frac{E_b}{N_0}$, que sigue estando relacionado con la relación señal-ruido a través de las fórmulas (5.21) y (5.22).

Tabla 5.1. Expresiones matemáticas de la probabilidad de error de bit de algunas modulaciones digitales.

Modulación	Detección	BER
OOK, FSK	Envolvente	$\frac{1}{2} e^{-\frac{E_b}{2N_0}}$
DPSK	Comparación de fases	$\frac{1}{2} e^{-\frac{E_b}{N_0}}$
BPSK	Coherente	$Q\left(\sqrt{2 \frac{E_b}{N_0}}\right)$
4-QAM, QPSK	Coherente en cuadratura	$Q\left(\sqrt{2 \frac{E_b}{N_0}}\right)$
M-PSK ($M \geq 8$)	Coherente en cuadratura	$\frac{2}{\log_2 M} Q\left(\sqrt{2 \cdot \log_2 M \cdot \frac{E_b}{N_0} \cdot \sin^2\left(\frac{\pi}{M}\right)}\right)$
M-QAM	Coherente en cuadratura	$\frac{4}{\log_2 M} \left(1 - \frac{1}{\sqrt{M}}\right) Q\left(\sqrt{\frac{3 \cdot \log_2 M}{M-1} \cdot \frac{E_b}{N_0}}\right)$

En la Figura 5.5 se muestran las gráficas correspondientes a las modulaciones de la Tabla 5.1, considerando 16 niveles en el caso de M-PSK y M-QAM. Nuevamente podemos observar el decrecimiento de la tasa de errores de bit al aumentar el parámetro E_b/N_0 .

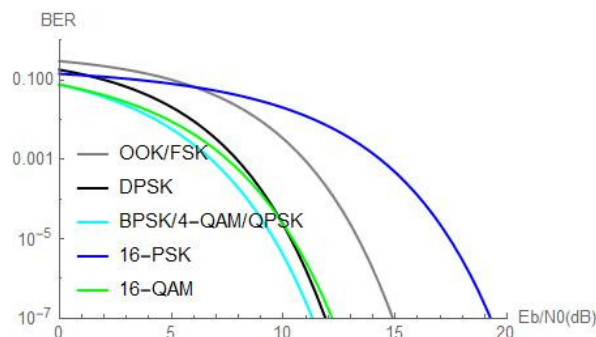


Figura 5.5. Tasa de errores de bit para diversas modulaciones digitales.