Chapitre 2
Introduction à la théorie de l'information

2.1. Introduction

La théorie de l'information est née des préoccupations techniques des ingénieurs de la télécommunication qui voulaient mesurer l'information et étudier à quelle loi elle est soumise (bruit, entropie, etc).
En fait, la théorie de l'information tente de résoudre un problème fondamental qui est :

Réaliser une Communication **Fiable** au travers d'un Canal **non Fiable**

Exemple :

Voix —air→ **Oreille**

Figure 2.1. Exemple emetteur/recepteur

Le problème est donc comment une source d'information peut apporter un message à une destination avec un minimum de distorsion en dépit des interférences.

Pour cela deux solutions sont possibles :

A. Une solution Physique

- utiliser des composants plus fiables dans ses circuits;
- évacuer l'air de l'enceinte du disque pour éliminer la turbulence qui perturbe la tête de lecture de la piste;
- utiliser un patch magnétique plus grand pour représenter chaque bit; ou
- en utilisant des signaux de puissance supérieure ou en refroidissant les circuits afin de réduire le bruit thermique
-

Ces modifications physiques augmentent généralement le coût de la communication canal.

B. Une solution Système

La Théorie de l'information et du codage offre une approche alternative où : on accepte le canal bruité et où on ajoute des systèmes de communication pour qu'on puisse détecter et corriger les erreurs introduites par ce canal.

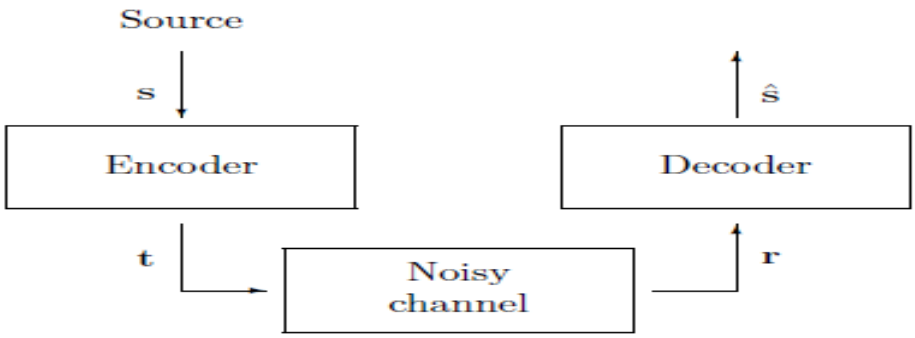


Figure 2.2. Système de transmission

2.2. Théorie de l'information

En se référant à la figure 2.2., on constate que la théorie de l'information adresse deux aspects dans la communication :

Le codage de source (ou compression des données), dont le but est de :

- Augmenter la compacité des signaux (sans ou avec distorsion)
- Éliminer la redondance inutile

Le codage de canal, dont le but est de :

- Accroître la sécurité de la transmission en présence de bruit
- Ajouter de la redondance pour la détection, voire la correction, des erreurs

L'optimisation d'une chaine de transmission peut se faire en optimisant séparément les codeurs de source et de canal.

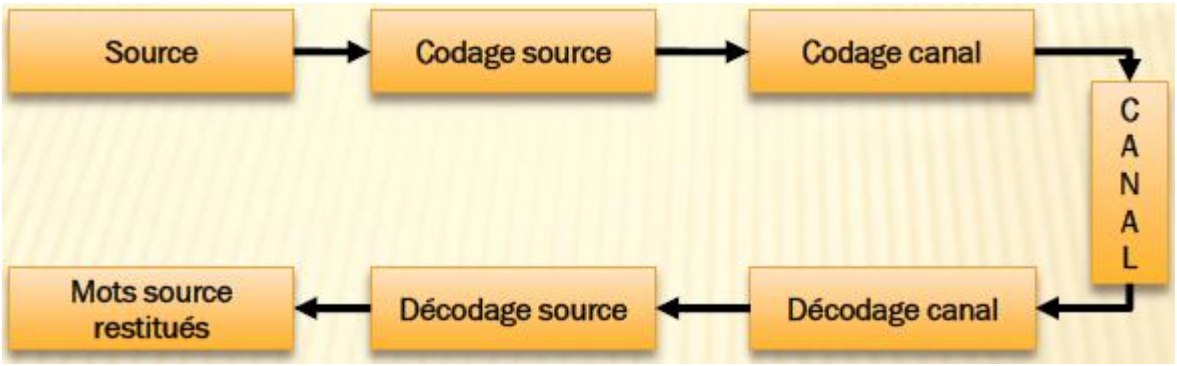


Figure 2.3. Schéma d'une chaîne de communication

La théorie de l'information tente de donner des réponses à deux questions :

» Quelle est la complexité limite d'un signal ?
  > *notion d'entropie* : nombre minimum de bit par symbole pour représenter une source.
» Quel est le débit limite pour une communication fiable sur un canal bruité ?
  > *notion de capacité de canal* : débit maximum qui peut être adopté pour un canal.

2.3. L'entropie

Pour définir l'information Shannon et Weaver se basent sur le 2^{e} principe de la thermodynamique (science des machines à feu) énoncé par Carnot : " dans un système physique, l'énergie tend à se dégrader ". Par exemple lorsque l'on met de l'eau chaude dans une baignoire qui se trouve dans une salle de bain froide, au bout d'un moment celle-ci sera tiède mais l'eau du bain aussi. Le tout devient homogène, indifférencié, cela mène à la mort du système, à l'entropie (en thermodynamique principe qui décrit le degré croissant de désordre dans le fonctionnement d'un système ; en communication, degré d'incertitude).

L'information, elle aussi est soumise à l'entropie. Mais, en même temps elle est une lutte contre l'entropie puisqu'elle consiste à imposer un ordre à un message. En effet celui-ci obéit à des règles syntaxiques et lexicales.
De même les machines subissent un échauffement. Il y a donc des parasites que l'on nommera " bruit ". Pour corriger ces bruits on utilise la redondance. En communication écrite ou orale, on usera de répétitions de la phrase, des mots clés, de sur lignages, de gestes, de différence dans le ton de la voix. Mais la redondance exagérée finit par nuire car elle ennuie. Pour éviter cela il faut introduire l'entropie, le désordre pour attirer à nouveau l'attention.
Donc, l'entropie est un concept inventé par Shannon pour décrire la quantité d'information dans un signal.
Exemple : L'entropie de séquence de lettres : Didon dina dit-on du dos d'un dodu dindon.

2.3.1. Quantité d'information

Soit un événement S parmi un ensemble d'évènements possibles.

- Avant l'évènement $S = s_k$, il existe une certaine quantité d'incertitude sur cet évènement,
- Quand cet évènement se réalise, il existe une certaine quantité de "surprise",
- Après la réalisation de l'évènement, il existe un gain en quantité d'information sur cet évènement, cette quantité d'information correspond à la "résolution" de l'incertitude de départ.

La notion d'information est liée à la notion de surprise et d'incertitude.

Définition : la **quantité d'information** gagnée à l'observation de l'évènement $S = s_k$, de probabilité p_k, est définie par : $I(s_k) = -\log p_k$

Propriétés :
$I(s_k) = 0$ pour $p_k = 1$;
$I(s_k) >= 0$ pour $0 <= p_k <= 1$;
$I(s_k) > I(s_i)$ pour $p_k < p_i$;
$I(s_k s_l) = I(s_k) + I(s_l)$ si s_k et s_l sont statistiquement indépendants.

Remarques :
La base du logarithme dans cette définition est arbitraire ;
On utilise par convention la base 2. L'unité correspondante est le *bit* (pour *b*inary un*it*).
Dans ces conditions :
$I(s_k) = -\log_2 p_k$
Si $p_k = 1/2$, alors $I(s_k) = 1$ bit.

Un bit est la quantité d'information gagnée quand un parmi deux évènements équiprobables apparaît.

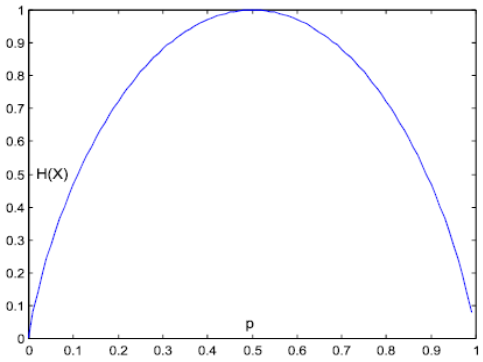
2.3.2. L'entropie

L'entropie d'un ensemble X est définie comme le contenu d'information Shannon moyen d'un évènement:

$$H(X) = \sum_{x \in Ax} p(x) \log_2 \frac{1}{p(x)}$$

Soit un ensemble X de deux évènements avec : P(1)=p et P(0)=1-p

$$H(X) = \begin{cases} -p.\log(p) - (1-p).\log(1-p) & pour \quad 0 < p < 1 \\ 0 & si\ p\ 0\ ou\ 1 \end{cases}$$



Si on reprend l'exemple précèdent, on doit tout d'abord trouver la probabilité d'apparition de chaque lettre dans la séquence : Didon dina dit-on du dos d'un dodu dindon.
Si on calcule la probabilité d'apparition de chaque lettre et on fait le tri par ordre décroissant, on trouve :

D	N	O	I	U	A	T	S
11/32	6/32	5/32	4/32	3/32	1/32	1/32	1/32

$H(X) = 11/32 \log_2 (32/11) + 6/32 \log_2 (32/6) + 5/32 \log_2 (32/5) + 3/32 \log_2 (32/3) + 3 * (1/32 \log_2 (32/1)) = 2{,}56475$

Parmi les propriétés de base de l'entropie c'est que $0 <= H(X) <= \log_2 n$ (le n représente le nombre de lettres).
Quand est ce que $H(X) = 0$?
Réponse : Si la séquence contient une seule lettre. Car $\log_2 1 = 0$.
Si on a une séquence qui contient que des lettres différentes alors $H(X) = \log_2 n$ et dans ce cas, on a un désordre maximal dans la séquence, donc on peut écrire le plus de phrases possibles.
Pour montrer ces deux bornes, on procède comme suit :
La première borne (borne inférieur):
Comme p_i est entre 0 et 1, donc $\log_2 (1/pi) >= 0$
La deuxième borne (borne supérieure) :
La fonction log est une fonction concave et on sait que pour toutes fonction concave f, on a : $(f(x) + f(y))/2 <= f(x+y)/2$ donc pour toute fonction concave :
$\mathbf{\Sigma\ p_i\ f(x_i) <= f(\Sigma\ p_i\ x_i)}$
$\Rightarrow H(X) = \mathbf{\Sigma\ p_i \log_2(1/p_i) <= \log_2(\Sigma\ p_i * 1/p_i)}$ on sait que $\mathbf{\Sigma\ p_i * 1/p_i = n}$
$\Rightarrow H(X) = \mathbf{\Sigma\ p_i \log_2(1/p_i) <= \log_2(n)}$

Un cas particulier de l'entropie :
Si on a tous les p_i qui s'écrivent de la forme
$Pi = 2^{-mi}$ avec $m_i \in N$
$m_i = \log_2(1/p_i)$

Exemple:
Soit le mot : YAWANAWA, si on prend d'une manière aléatoire une lettre appartenant à ce mot et on se pose la question:
Q1 : Est-ce un « A » ? Si oui alors on a la réponse en une seule question sinon si la réponse est Non ⇒ qu'il reste YWNW.

Q2 : est ce un « W » ? Si oui, on a une réponse en 2 questions et si la réponse est non
⇒ YN.
Q3: Est ce un « Y » ? Si oui, on a la réponse en 3 questions sinon c'est forcement un « N » et on a aussi la réponse en 3 questions.
Donc les probabilités des lettres sont comme suit :

A	W	Y	N
P(A) = 1/2	P(w) = 1/4	P(Y) = 1/8	P(N) = 1/8

Selon les questions posées $m_a = 1$, $m_w = 2$, $m_y = 3$, $m_N = 3$
Si on calcule les $m_i = \log_2(1/p_i)$, on trouve la même chose.
Donc : $H(X) = \Sigma\, p_i \log2(1/p_i) = \Sigma\, p_i\, m_i$.

2.3.3. Information mutuelle

Rappel : Un ensemble joint XY est un ensemble dans lequel chaque résultat est une paire ordonnée X, Y avec $X \in Ax = \{a_1, \ldots, aI\}$ et $y \in Ay = \{b1, \ldots, bJ\}$.

Nous appelons P (X; Y) la probabilité jointe de X et Y.

Avant de donner la définition de la mesure de l'information proposée par Shannon, nous allons essayer de décrire le concept d'information. En suivant le modèle probabiliste, fournir une information à un utilisateur consiste à choisir un événement parmi plusieurs possibles. Qualitativement, fournir une information consiste donc à lever une incertitude sur l'issue d'une expérience aléatoire. La notion d'information est déjà inhérente à celle de probabilité conditionnelle. Considérons les événements {A = a} et {B = b}. La probabilité p(a|b) peut être interprétée comme la modification apportée à la probabilité p(a) de l'événement {A = a} lorsque l'on reçoit l'information que l'événement {B = b} s'est réalisé. Ainsi

- si $p(a|b) \leq p(a)$, l'incertitude sur a augmente,
- – si $p(a|b) \geq p(a)$, l'incertitude sur a diminue.

Pour mesurer la variation de l'incertitude, il faut choisir une fonction dé- croissante de la probabilité. On choisit également une fonction continue. Le logarithme permet d'exprimer commodément les variations d'incertitude. On notera I(a) l'incertitude sur a, encore appelée information propre de a :

$$I(a) = -\log_2 p(a).$$

Ainsi l'information «b est réalisé» diminue l'incertitude sur a de la quantité :

$$I(a) - I(a|b) = \log_2 (p(a|b) / p(a)).$$

Cette dernière quantité est appelée information mutuelle de a et b.

On peut aussi écrire : $I(a; b) = \log_2 (p(a|b) / p(a))$.

Par définition $p(a, b) = p(a|b)p(b) = p(b|a)p(a)$. Donc

$$I(a; b) = I(b; a) = \log2(p(a, b) / p(a)p(b))$$

Nous allons discuter le signe de I(a; b).

- I(a; b) > 0 signifie que si l'un des deux événements se réalise, alors la probabilité de l'autre augmente ;
- – I(a; b) < 0 signifie que si l'un des deux événements se réalise, alors la probabilité de l'autre diminue ;
- – I(a; b) = 0 signifie que les deux événements sont statistiquement indépendants.

On peut également définir dans l'espace probabilisé joint l'information propre conditionnelle de a sachant b qui est la quantité d'information fournie par l'événement {A = a} sachant que l'événement {B = b} est réalisé : $I(a|b) = -\log2\ p(a|b)$.

Problème : Y étant une version bruitée de X, et H(A) mesurant l'incertitude a priori sur X, comment mesurer l'incertitude sur X après avoir observé Y ?

Définition : l'*entropie conditionnelle* (à l'observation de *Y*) de *X*, sachant $Y=Y_k$ vaut :

$$H(A|B) = \sum_{k=0}^{K-1} H(A|Y=y_k)p(y_k)$$
$$= -\sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j|y_k)p(y_k)\log_2 p(x_j|y_k)$$
$$= -\sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j,y_k)\log_2 p(x_j|y_k)$$

H(*A* | *B*) représente la quantité d'incertitude restante sur l'entrée après que la sortie a été observée.

- *H*(*A*) est l'incertitude sur l'entrée *avant* l'observation de la sortie ;
- *H*(*A* | *B*) est l'incertitude sur l'entrée *après* l'observation de la sortie ;
- *H*(*A*) – *H*(*A* | *B*) représente l'*incertitude sur l'entrée résolue par l'observation de la sortie*

On peut aussi écrire : $H(X|Y) = -\mathbf{\Sigma}_{x,y} P(X = x,Y = y)\log(P(X = x|Y = y))$

Elle vérifie donc la propriété : H(X, Y) = H(Y) + H (X|Y).
Qui signifie que l'information apportée par les 2 variables X et Y vaut l'information apportée par Y seule plus l'information apportée par X connaissant déjà la valeur de Y.
H(X,Y) désigne l'entropie de la variable aléatoire (X,Y).

Quelques propriétés :

- $H(x, y) \leq H(x)+ H(y)$;
- $H(x, y) = H(x)+ H(y|x)$;

- $H(x)+ H(y) \geq H(x, y) = H(x)+ H(y|x)$ donc $H(y) \geq H(y|x)$.