Codage de la parole et qualité vocale

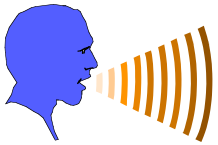
M. Coupechoux, Ph. Martins, Ph. Godlewski
octobre 2008
ENST, Département INFRES

Codage et compression de la parole

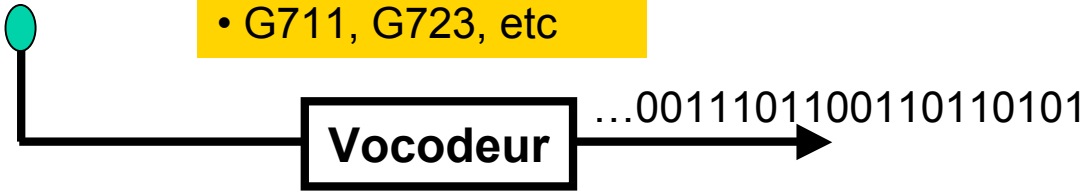
- Introduction
  - Différents types de codage
  - Commutation de circuit et de paquets
- Codecs
  - Caractéristiques et fonctions
  - G711 – Modulation par Impulsion et Codage
  - Principaux codecs voix
  - Aperçu de quelques codecs vidéo
- Qualité vocale
  - Schéma de référence
  - Délai
  - Gigue
  - Perte de paquets
  - Echo
  - Critères de qualité
- Conclusion

Introduction

- **Codage de source** :



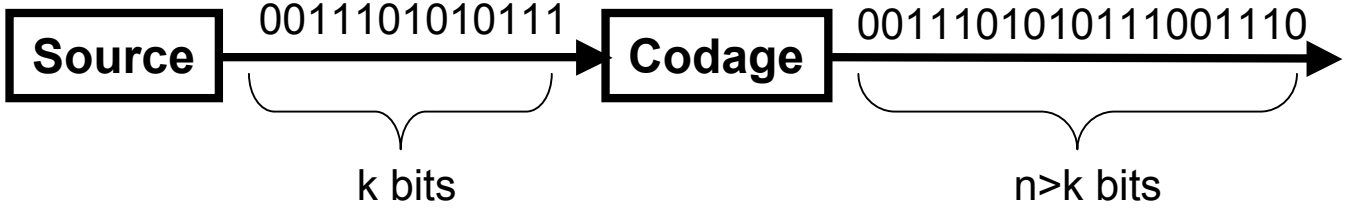
Exemples :
• GSM : FR, EFR, HR
• G711, G723, etc



- **Codage canal** (*channel coding*) :
  - codage correcteur d'erreurs
  - protection contre les erreurs

Exemples :
• Codage convolutionnel (décodage de Viterbi)
• Turbo code
• etc

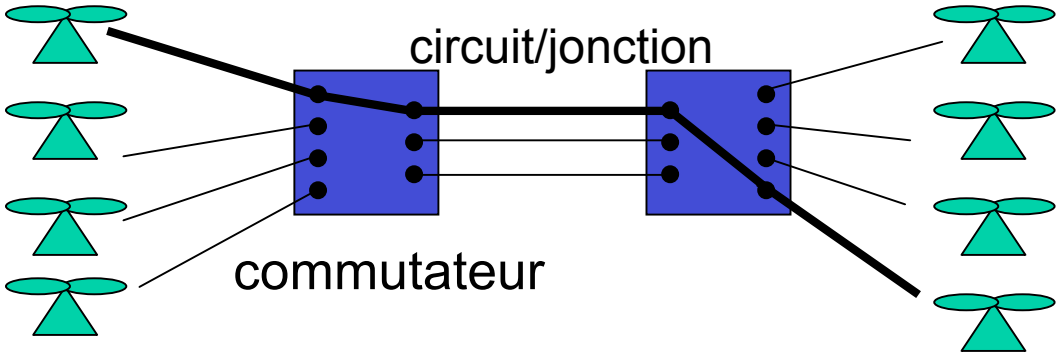
- **Codes d'étalement** (*spreading codes*) :
  - Utilisé en radio-mobiles (UMTS)



Taux de codage (rate) :
R=k/n

Introduction

- **Commutation de circuit** : Le réseau fournit une liaison point-à-point aux terminaux en communication. Les ressources physiques sont réservées le temps de la connexion. Exemple : Réseau Téléphonique Commuté (RTC).

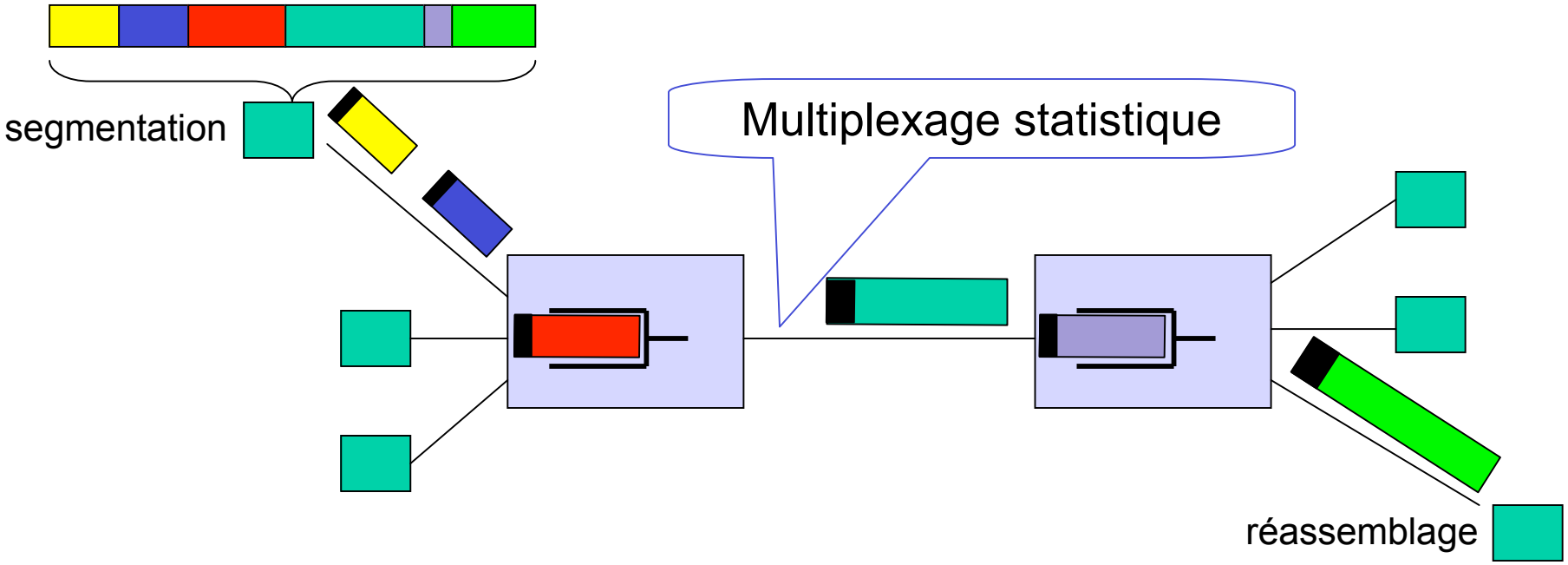


- **Phases** de la communication :

1. Etablissement du circuit : l'ensemble des ressources entre les terminaux sont réservées,
2. Transfert des informations : les ressources sont dédiées au transfert pendant toute la durée de la communication même s'il n'y a pas de transfert de données,
3. Libération : les ressources sont libérées et peuvent être réutilisées par d'autres terminaux.

Introduction

- Commutation de paquets : Les informations sont découpées en petits éléments (**paquets**) au cours de la **segmentation**. A la réception, le message est reconstitué (**réassemblage**).
- Chaque paquet possède un **en-tête** comprenant par exemple les adresses source et destination.
- Avantage : gain statistique.
- Inconvénient : la gestion de la qualité de service est plus difficile.

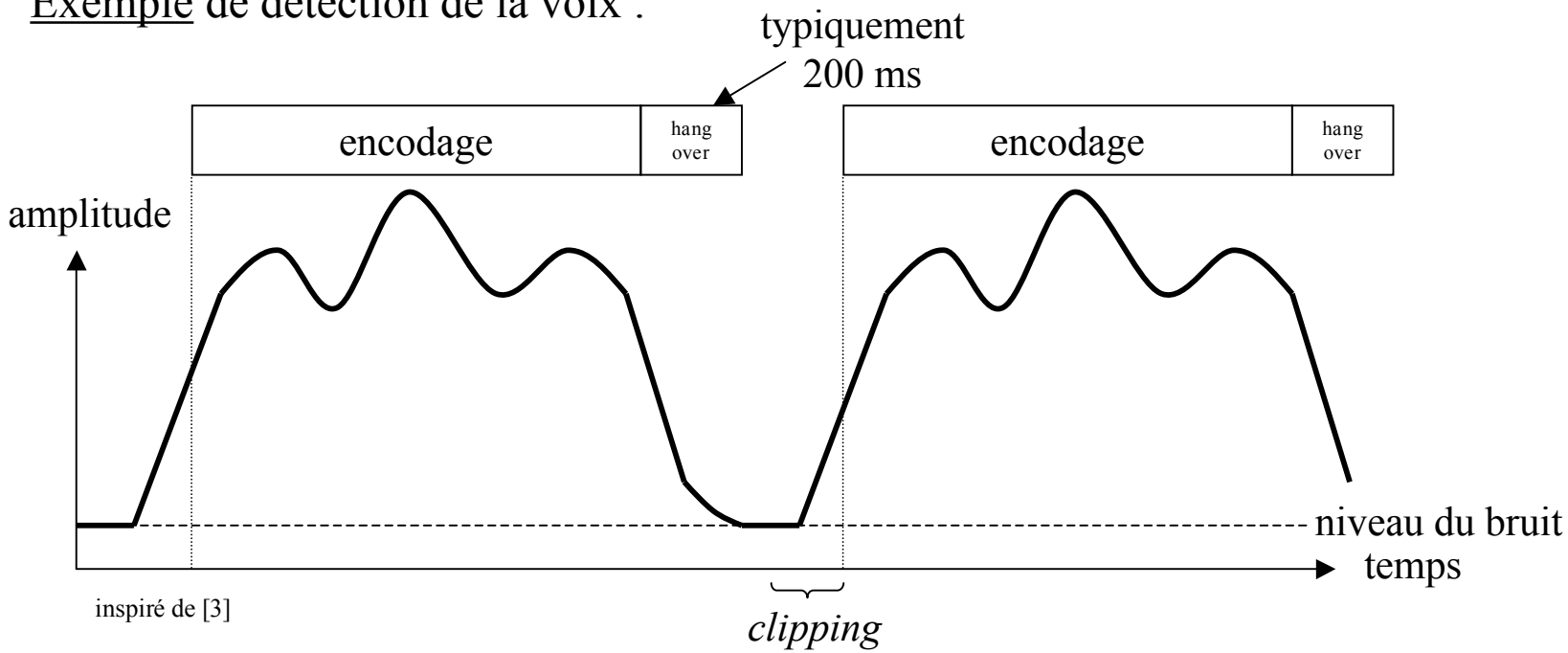


Codecs
Caractéristiques et fonctions

- **Débit** : entre 1.2 Kbps et 64 Kbps
  - Compromis entre la qualité de la voix et la bande utilisée.
  - Réseau Téléphonique Commuté : 64 Kbps.
  - On peut atteindre aujourd'hui une bonne qualité avec 5 Kbps.
  - En radio, le codec AMR adapte son débit à la qualité du canal.

- **Compression de silences** : il n'est pas nécessaire de transmettre des données pendant les périodes de silence.
  - Taux d'activité effectif de la voix dans un sens ~ 45%.
  - VAD (Voice Activity Detection) : détection d'activité de la voix ; il faut éviter le *clipping*.
  - DTX (Discontinuous Transmission) : transmission discontinue ; le codec stoppe la transmission d'informations lorsque le VAD est activé.
  - CNG (Comfort Noise Generation) : génération du bruit de confort ; transmission de faibles quantités d'information destinée à reproduire chez le récepteur l'ambiance sonore de l'émetteur.

- Exemple de détection de la voix :



inspiré de [3]

Phase	Durée [s]	Proportion de temps
Parole simple	1.00	38%
Silence	1.59	61%
Parole double	0.23	6.5%
Silence complet	0.51	22.5%

tiré de [4]

- **Robustesse aux pertes** :
  - Causes des pertes : congestion dans les routeurs, délais trop importants.
  - La répétition des trames est impossible (délais trop importants).
  - Une solution possible : ajouter de la redondance (*Forward Error Correction*).
  - Une autre solution : remplir les trous avec des trames répétées ou interpolées (*Packet Loss Concealment*).
  - Un critère pour le codec : dégradation de la qualité perçue en fonction du taux d'erreur paquets/trames.

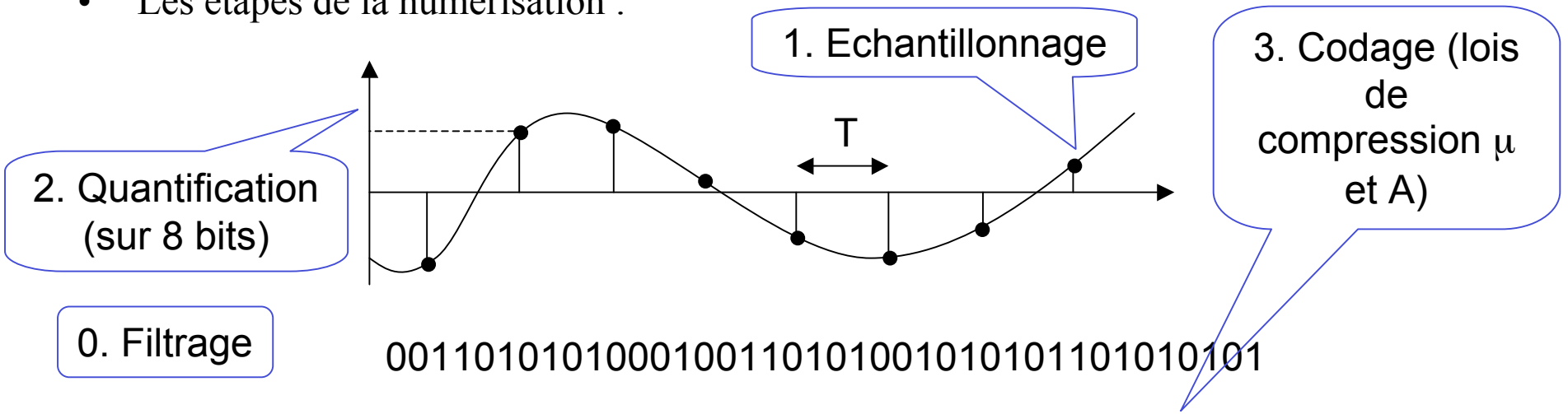
- **Délai de paquétisation** : le codec crée périodiquement des trames de voix ; il lui faut un certain temps pour traiter le signal analogique et créer la trame.
  - Certains codecs améliorent la compression en observant plus que le contenu effectif d'une trame (*look ahead*).
  - Exemple : G723.1 doit mémoriser 37.5 ms d'échantillons voix avant de créer la première trame.

- **Bande étroite ou large bande** :
  - Les codecs bande étroite (ex : G711) échantillonnent la voix sur une bande d'environ 4 KHz (300 Hz – 3400 Hz).
  - Remarque : l'oreille humaine peut capter des sons jusqu'à des fréquences de 16 voire 20 KHz.
  - Les codecs large bande (ex : G722.2 = Wideband AMR) permettent de transmettre des fréquences audio jusqu'à 7 KHz (fréquence d'échantillonnage 16 KHz) ; il en résulte une bien meilleure qualité de voix (clarté).
  - G722.2 a un débit adaptatif entre 6 et 23.85 Kbps.

  - Pour information :
    - La bande Hi-Fi ~ 20 Hz – 15 KHz
    - La qualité CD ~ 20 Hz – 20 KHz

G711 – Modulation par Impulsion et Codage

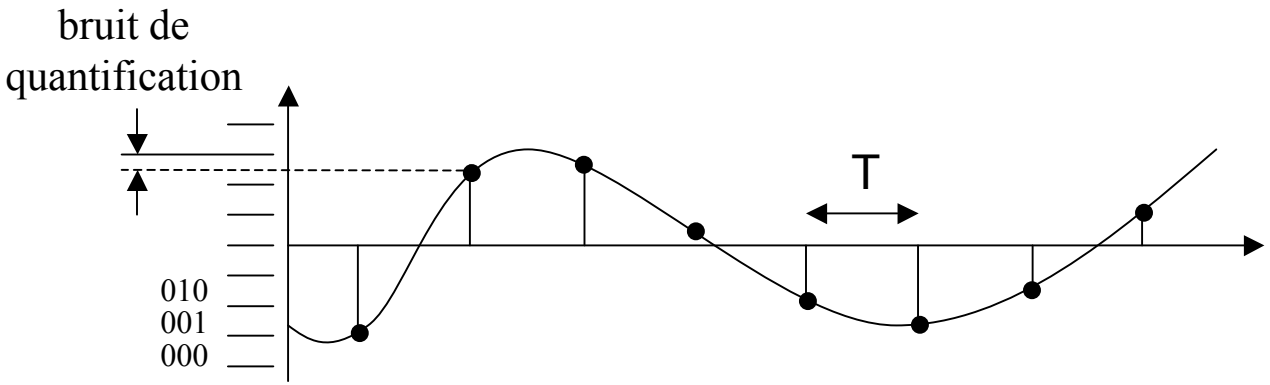
- G711 = MIC = PCM (Pulse Code Modulation)
- Les étapes de la numérisation :



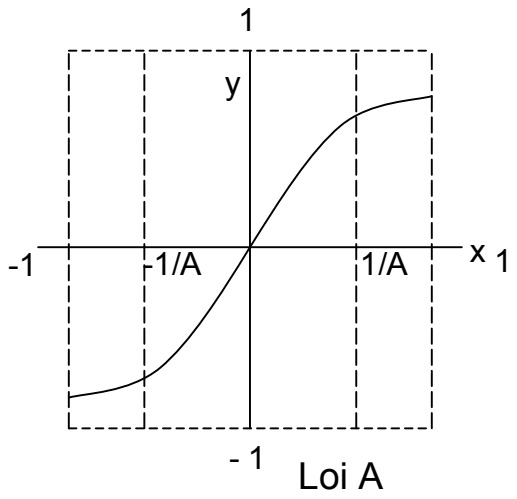
- L'essentiel du spectre de la voix est inclus dans la bande **300-3400 Hz** : on filtre le signal à 3400 Hz de telle sorte qu'il n'y ait plus de résidus à 4000 Hz.

- Théorème d'échantillonnage (Nyquist-Shannon) : le signal vocal est échantillonné à 8000 Hz soit un échantillon toutes les **125 μs**.

- **Quantification** :
  - La plage de variation du signal est découpée en intervalles de quantification.
  - On donne un identifiant sur 8 bits à chaque intervalle.
  - Les échantillons de parole obtenus lors de la phase d'échantillonnage sont quantifiés suivant une loi de compression logarithmique : les amplitudes faibles sont codés avec plus d'intervalles que les amplitudes importantes.
  - Loi μ en Amérique du Nord et au Japon, loi A dans le reste du monde.



exemple de quantification linéaire sur 3 bits



- **Débit résultant** : 8 bits toutes les 0.125 ms ➔ 64 Kbps.

Codecs
Principaux codecs voix

Codec	Débit [Kbps]	Inter-arrivées des échantillons [ms]	MOS
G.711 PCM	64	0.125	4.1
G.726 ADPCM	32	0.125	3.8
G.728	15	0.625	3.6
G.729	8	10	3.9
G.729a	8	10	3.7
G.723.1	5.3 – 6.3	30	3.6 – 3.9
iLBC	13.3 – 15.2	30 – 20	3.9

[3-8]

Codecs
Principaux codecs voix

- G.711
  - 64 Kbps PCM utilisé dans le RTC,
  - les paquets VoIP rassemblent plusieurs échantillons par paquet pour atteindre un équivalent de 10 ms de parole (80 échantillons).

- G.726
  - plusieurs débits mais le plus populaire est 32 Kbps (DECT),
  - au lieu d'envoyer le niveau de quantification, il ne code que la différence de niveau avec l'échantillon précédent.

- G.722, G.722.1, G.722.2
  - codecs large bande.

- iLBC
  - libre.
  - RFC3951

Codecs
Principaux codecs voix

- G.723.1
  - France Telecom, NTT, University of Sherbrook, …
  - Trames de 30ms et look-ahead de 7.5ms (délai total de 37.5ms)
  - Echantillonnage 8KHz / 16 bits (240 échantillons / 30ms)
  - Deux débits (débit modifiable de trame en trame) :
    - 6.3Kbps (trames de 24 octets) (MOS=3.9)
    - 5.3Kbps (trames de 20 octets) (MOS=3.6)
  - Le silence est codé dans des trames de 4 octets à 1.1 Kbps [2].
  - Non adapté à la musique, au fax, à la signalisation DTMF
  - Annexe A : trames de 4 octets (Silence Insertion Descriptor) pour CNG
  - Complexité = 25 (G711 = 1, G729a=15)
  - Peu adapté à la musique, au fax, aux signaux modem, aux tonalités DTMF
  - Utilisé pour les visio-conférences en 3G (3GPP et 3GPP2) et obligatoire en H.323.

Codecs
Principaux codecs voix

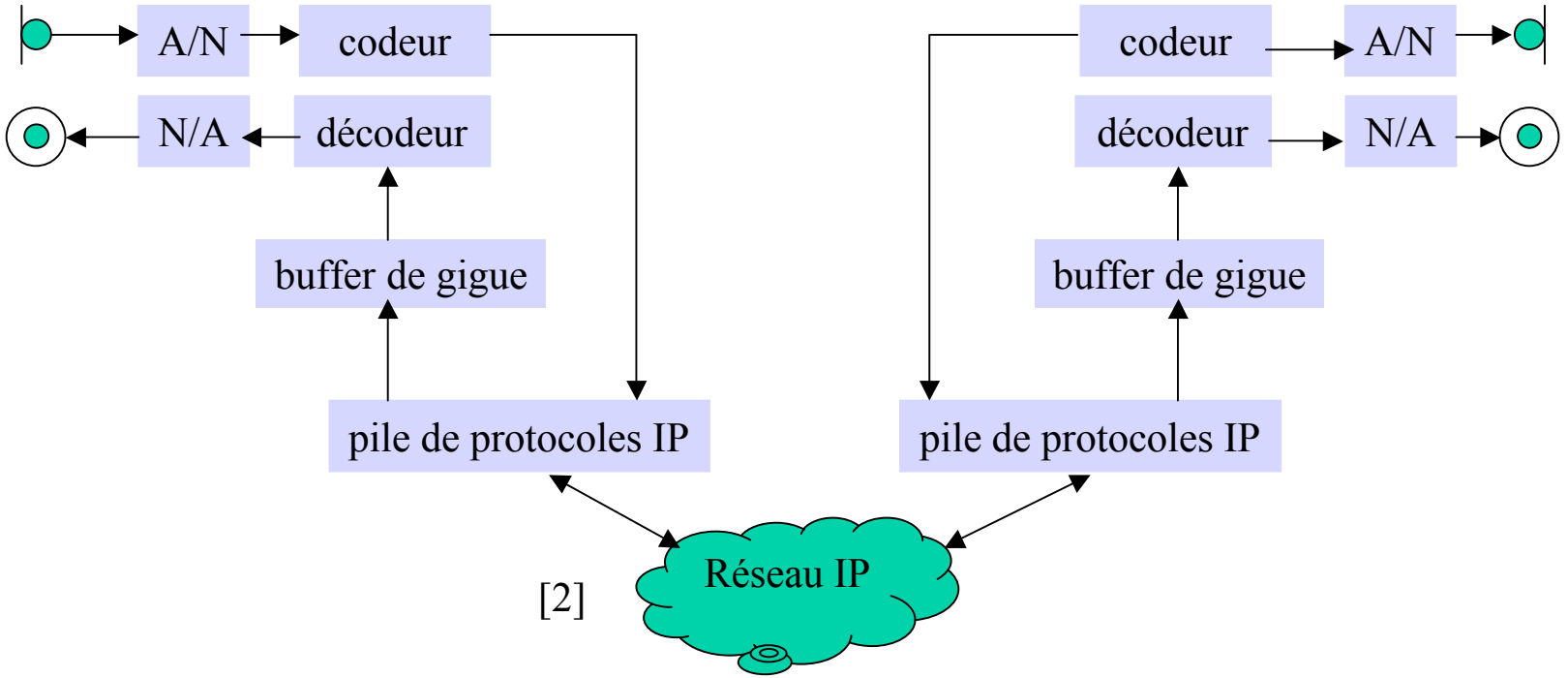
- G.729a
  - France Telecom, Mitsubishi, …
  - Trames de 10ms
  - Echantillonnage 8KHz / 16 bits (80 échantillons / 10ms)
  - Débit 8Kbps (trames de 10 octets)
  - G.729 Annexe B : trames SID de 2 octets pour CNG
  - Complexité=15 (moins complexe que G.729)
  - Peu adapté à la musique, au fax, aux signaux modem, aux tonalités DTMF

Codecs
Aperçu de quelques codecs vidéo

- H.261 : utilisé pour les visio-conférences (entre 40 Kbps et 2 Mbps).
- MPEG1 Part 2: utilisé par les Vidéo CD (VCD), qualité comparable à celle du VHS (1.5 Mbps).
- MPEG2 Part 2: utilisé par les DVD et la diffusion télé.
- H.263 : conçu pour les applications bas débit (vidéo-conférence) ; plus efficace que H.261 pour une même qualité d'image ; cinq résolutions. Une vidéo-conférence de très bonne qualité nécessite 386 Kbps.
- MPEG4 Part 10 ou H.264 ou AVC (Advanced Video Coding) : profil le plus récent ; une vidéo-conférence très bonne qualité nécessite 192 Kbps au prix d'une puissance de calcul plus importante.

[2]

- Réseau IP :



[2]

- Dans le cas d'interconnexion avec le RTC, le RNIS ou un réseau utilisant un codec différent, le transcodage est nécessaire. **Tout cycle supplémentaire de codage/décodage dégrade la qualité de la voix**.

Qualité vocale
Délai

- **Le délai** est la durée mise par la parole pour être transmise de la bouche du locuteur jusqu'à l'oreille de l'auditeur.

- On distingue :

  – Le temps de paquétisation : c'est le temps nécessaire au codec pour assembler les échantillons qui formeront un paquet. Généralement, les codeurs avec une taille de trame plus longue compressent plus.

  – Le temps de traitement (DSP) : c'est le temps nécessaire à l'encodage et au décodage (ordre de grandeur 12 ms [5]).

  – Le temps introduit par le système d'exploitation : la carte son lève des interruptions de l'OS pour qu'il lise les échantillons de parole stockés dans la mémoire tampon (ordre de grandeur 60 ms [2]) ; on peut améliorer les choses en utilisant un système temps-réel.

- Le temps de *look ahead* : certains codecs ont besoins d'échantillons ultérieurs avant de créer le paquet (ordre de grandeur 5 ms avec G.729 [3]).

- Le temps de propagation à travers le réseaux : notamment à cause des temps d'attente dans les files (très variable, ordre de grandeur entre 20 et 100 ms).

- Le temps de déjittérisation : compense la gigue (cf. plus loin).

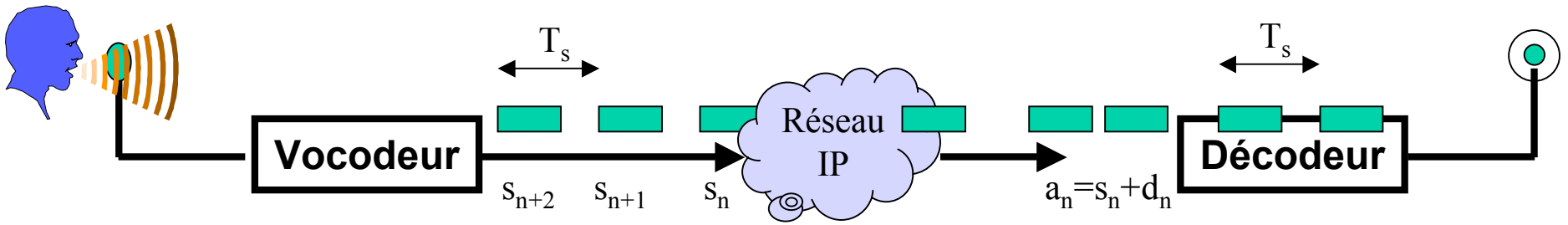
• L'ensemble de ces délais forme le **budget de délai** de bout en bout.

Qualité vocale
Délai

- **Surcharge due à la pile de protocoles IP** :
  - L'en-tête IP est de 20 octets.
  - Le protocole de transport UDP ajoute 8 octets.
  - RTP ajoute 12 octets, soit un total de 40 octets par trame de voix.
- Exemple :
  - G729 crée des trames de 80 bits avec un débit de 8 Kbps.
  - Avec la surcharge (80%), on atteint 40 Kbps !
- Première réponse : la compression d'en-tête.
  - Fondée sur la création de contextes,
  - et sur la non répétition des champs fixes ou peu variables.
  - On peut réduire l'en-tête IP/UDP/RTP à 2 ou 4 octets.
- Seconde réponse : augmenter la charge utile.
  - On transmet plusieurs trames par paquet (*bundling*).
  - Mais : **délai supplémentaire de paquétisation**,
  - Et : sensibilité plus importante aux pertes de paquets.

- **Délai maximal**
- La recommandation G.114 de l'UIT-T [6] spécifie les valeurs maximales de délai T pour les communications vocales :
  - $T \leq 150$ ms : le délai est acceptable.
  - $150 \text{ ms} \leq T \leq 400$ ms : acceptable pour des communications internationales (exemple : via satellite).
  - $400 \text{ ms} < T$: délai inacceptable pour la qualité vocale.

- **Qu'est-ce que la gigue ?** c'est la variation du délai des trames autour de la moyenne
  - Une définition possible : écart-type du délai.
  - Une autre définition parfois rencontrée : différence entre le délai maximal et la moyenne des délais.
- La gigue doit être réduite autant que possible :
  - Les trames sont émises à intervalles réguliers, aux instants $s_n = nT_s$.
  - Côté récepteur, le décodeur « joue » ces trames à intervalles réguliers.
  - Si une trame est en avance, elle est mise en attente.
  - Si une trame arrive en retard, elle est perdue.
  - Le *buffer* de gigue permet de compenser les différences de délai en retenant les paquets pendant un délai D avant de les jouer.



- Hypothèses :
  - Les trames sont émises aux instants $s_n = nT_s$.
  - L'instant d'arrivée de la nième trame est $a_n=nT_s+d_n$.
  - On suppose que la première trame est retenue D secondes dans le buffer.
  - Ensuite, les trames sont jouées toutes les T_s secondes.
- La nième trame est perdue si elle arrive après son instant de jeu :

$$a_n > a_0 + D + nT_s$$
$$nT_s + d_n > s_0 + d_0 + D + nT_s$$
$$d_n - d_0 > D$$

c-a-d si son délai d'acheminement dépasse de D le délai du premier paquet.

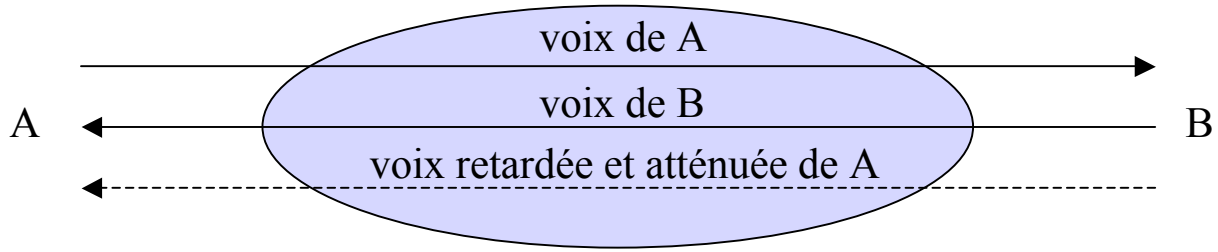
- Si D est grand, la probabilité de perte est faible mais les délais sont augmentés de D.
- Si D est faible, la probabilité de perte est forte mais les délais ne sont pas beaucoup influencés par le *buffer* de gigue.
- D est le **délai de déjittérisation**.
- En pratique, ce délai peut être ajusté dynamiquement en utilisant les périodes de silence.

Qualité vocale
Perte de paquets

- Causes possibles de pertes de paquets :
  - Congestion dans les routeurs du réseau IP,
  - Arrivée trop tardive au décodeur sans que le *buffer* de gigue n'ait pu compenser le retard.
- La retransmission des paquets perdus n'est pas faite car inutile (délais trop importants).
- Une méthode pour amoindrir les effets des pertes :
  - rejouer la dernière trame reçue.
  - L'auditeur ne perçoit pas de silence dans le flux de parole.
  - La perte concerne environ 20 ms de parole, de telle sorte que l'effet n'est presque pas perceptible.
- L'effet sur la qualité vocale de pertes isolées ou en rafales n'est pas le même.
- Ordre de grandeur du taux de pertes acceptable :
  - Codec G711 : **1% de perte maximum** (source : Intel White Paper)

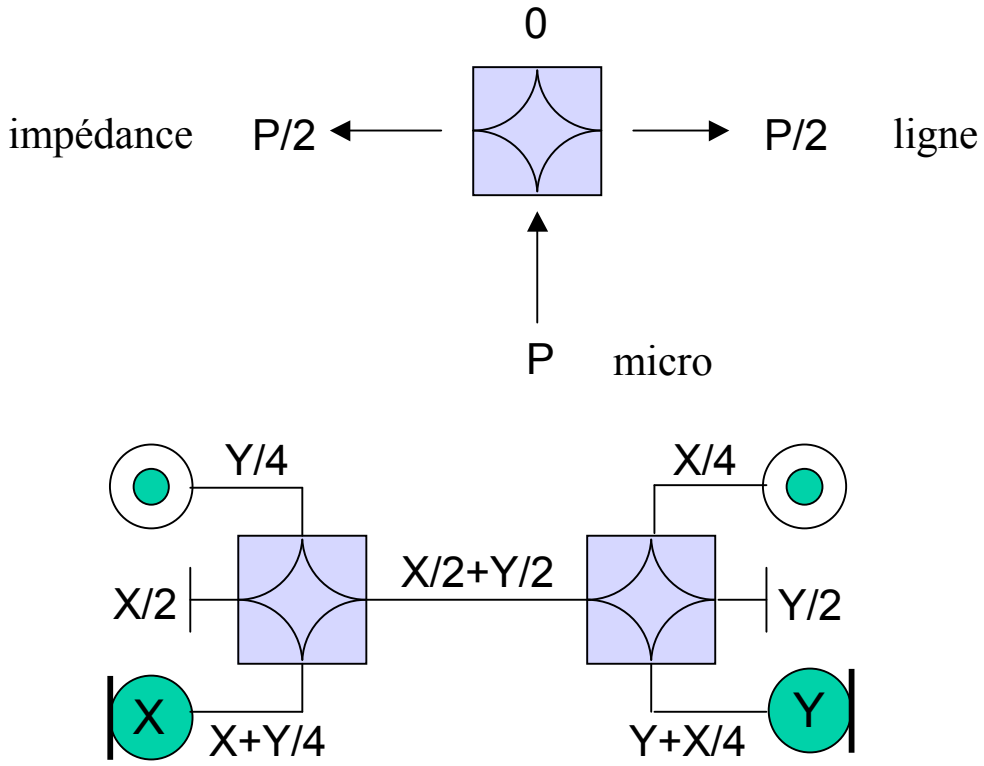
Qualité vocale
Echo

- Il y a **écho** lorsque le locuteur entend une version retardée de sa propre voix.



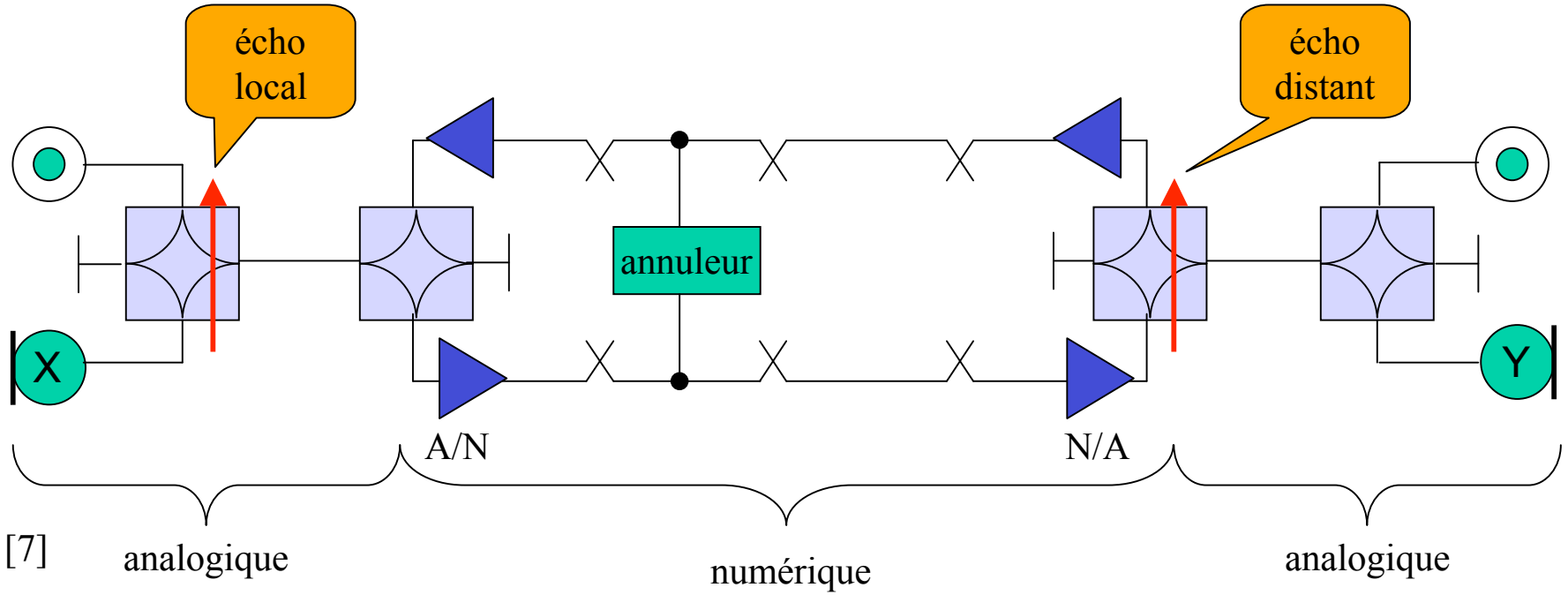
- L'importance de l'écho pour la gène du locuteur dépend :
  - du délai d'aller-retour entre le locuteur et la source de l'écho,
  - de l'amplitude de l'écho.
- Les appels RTC nationaux (délais < 25 ms) ne provoquent généralement pas d'écho.
- Les appels RTC internationaux ou sur réseau IP (délais < 150 ms) nécessitent des annuleurs d'écho.
- Les appels via satellite (délais < 400 ms) nécessitent des annuleurs d'écho.

- Rappel : la transformation 2 fils / 4 fils dans un réseau analogique.
- Pour permettre une communication duplex sur seulement 2 fils, on utilise un circuit 2 fils / 4 fils ou circuit hybride.
- Pour un fonctionnement adéquat, l'impédance de référence doit être adaptée à la ligne.



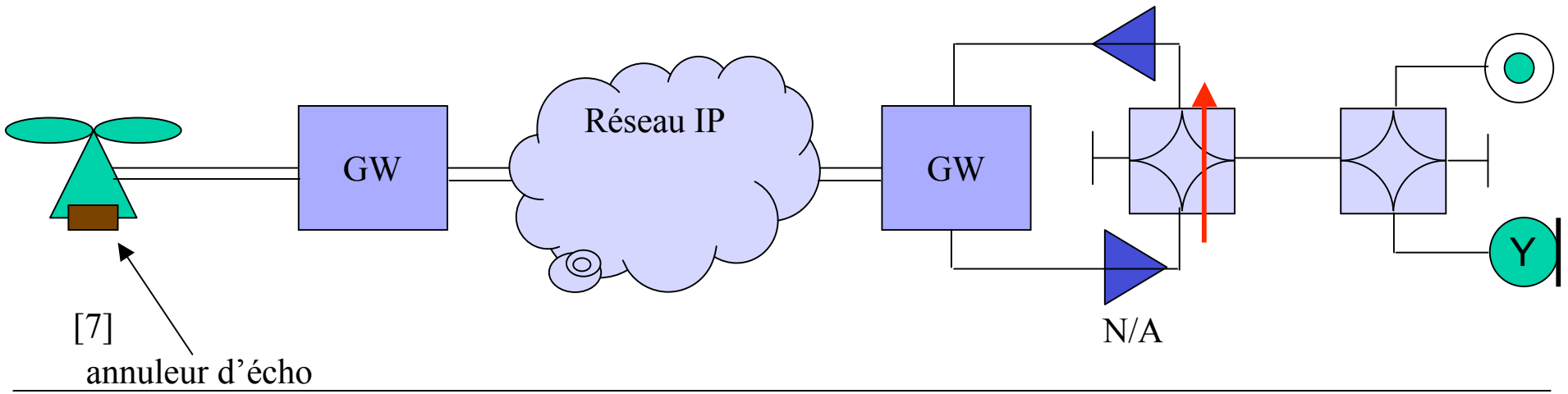
[7]

- **Echo dans le RTC** :
  - Echo local (*sidetone*) : dans le combiné du locuteur ; ce n'est pas un problème, c'est au contraire une source de confort.
  - Echo distant : au moment du passage de 4 à 2 fils ; ce n'est pas un problème si les délais sont faibles ; sinon (traversée d'un réseau IP ou satellite), il faut un annuleur d'écho.

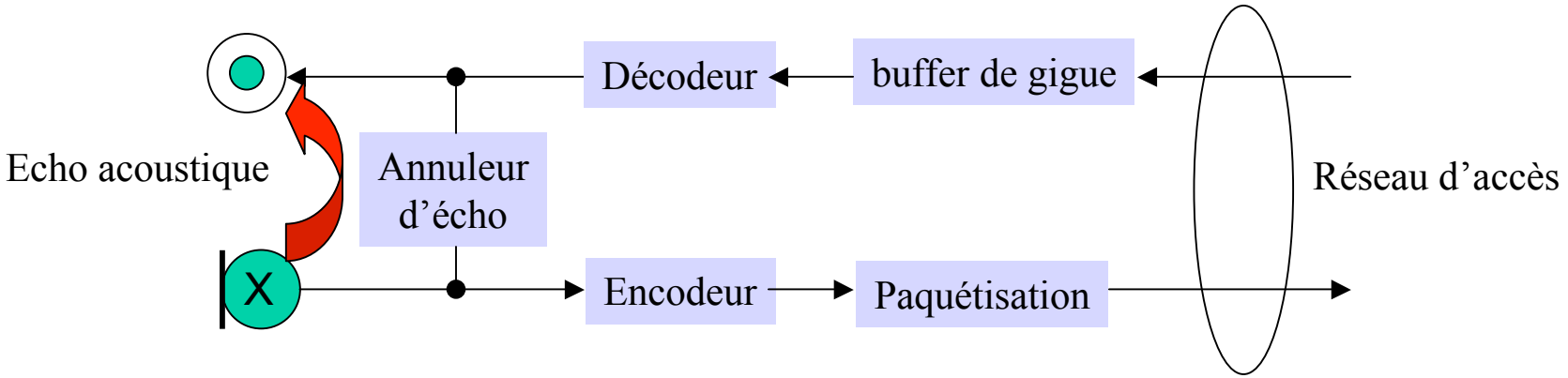


[7]

- **Terminal numérique et réseau IP** :
  - Le réseau IP introduit des délais plus importants que le RTC, rendant nécessaire la présence d'annuleurs d'écho.
  - La conversion analogique – numérique a lieu dans le terminal numérique lui-même.
  - Un annuleur d'écho est nécessaire pour les communications qui impliquent une partie de réseau analogique (schéma), ou pour lutter contre l'écho acoustique.
- Echo acoustique [2] : partie du signal acoustique du haut-parleur capté par le micro. Exemple : terminaux mains libres ou de mauvaise qualité.



- Vue schématique d'un **terminal numérique** et écho acoustique [9] :



- **Mean Opinion Score (MOS) « note moyenne d'opinion »**

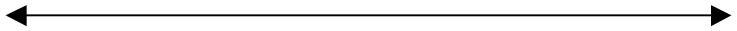
- Il y a deux manières de juger la qualité de la voix :
  - les méthodes objectives fondées sur des métriques objectives (distorsion du signal, rapport signal à bruit, taux de perte des paquets, etc) ;
  - les méthodes subjectives.
- MOS est une méthode subjective :
  - On réalise une batterie de tests MOS,
  - On demande à un groupe d'auditeurs de noter la qualité vocale,
  - On moyenne l'ensemble des notes,
  - 1=mauvais, 5=excellent.
- Parmi les codecs déjà vus (hors large bande), G711 obtient la meilleure note = 4.1.

- Les tests subjectifs sont spécifiés dans les recommandations P.800 et P.830 (UIT)

- **Le modèle E** [10] est une méthode analytique qui prédit la qualité subjective de la voix en considérant les paramètres de la transmission.

- La combinaison de ces paramètres fournit une note **$0 \leq R \leq 100$**. La recommandation fournit ensuite les relations entre R et les autres méthodes d'évaluation comme le MOS.

Valeurs de R	90-100	80-90	70-80	60-70	0-60
MOS	4.34-5	4.03-4.34	3.60-4.03	3.10-3.60	0-3.10
Qualité	Excellente	Bonne	Moyenne	Mauvaise	Très mauvaise



qualité RTC

- **Le facteur R** se décompose de la manière suivante :

$$R = R_0 - I_s - I_d - I_e + A$$

- R_0 représente le rapport signal à bruit de base (prend en compte par exemple le bruit de fond de la pièce dans laquelle se trouve le locuteur),
- I_s représente la dégradation due au signal voix (par exemple le bruit de quantification),
- I_d représente la dégradation due au délai de transmission,
- I_e représente la dégradation due aux pertes de paquets et au choix du codec,
- A représente la dégradation qu'un utilisateur est prêt à accepter s'il sait qu'il utilise une technologie mobile (sinon A=0).

- Remarques :
  - L'effet de la gigue est indirectement pris en compte dans la perte des paquets et le délai,
  - Deux communications ayant la même note peuvent produire des effets subjectifs complètement différents,
  - Le modèle E nécessite la connaissance de nombreux paramètres sur les équipements utilisés qui ne sont pas forcément facilement disponibles.

- Autres méthodes objectives :
  - PSQM (Perceptual Speech Quality Measure) P.861
  - PAMS (Perceptual Analysis Measurement System)
  - PESQ (Perceptual Evaluation of Speech Quality) P.862

Conclusion

- Les principales caractéristiques des codecs :
  - débit,
  - VAD/DTX,
  - durée des trames,
  - robustesse aux pertes,
  - bande étroite / large bande.
- Les principaux codecs : G711, G726, G723.1, G729a.
- Les principaux paramètres qui influencent la qualité vocale :
  - le délai,
  - la gigue,
  - les pertes de trames,
  - les capacités de l'annuleur d'écho.
- Comment noter la qualité ?
  - MOS,
  - Le modèle E (paramètre R).

Références

- [1] W. C. Hardy, VoIP Service Quality, McGraw-Hill Networking, 2003.
- [2] O. Hersent et al., L'essentiel de VoIP, Dunod, 2005.
- [3] J. Davidson et al., Voice over IPO Fundamentals, 2006.
- [4] ITU-T, Recommandation P.59.
- [5] B. Good, Voice over Internet Protocol, Proc of the IEEE, Sept. 2002.
- [6] ITU-T, Recommandation G.114.
- [7] Claude Rigault, The Subscriber Loop, cours ENST.
- [8] L. Madsen et al., Asterisk : The Future of Telephony, O'Reilly, 2005.
- [9] D. De Vleeschauwer et J. Jansen, Voice Performance over Packet-Based Networks, Alcatel White Paper.
- [10] ITU-T, Recommandation G.107.