Evaluation de la qualité audio

Gaël Mahé et Sonia Djaziri Larbi

Université Paris Descartes / Ecole Nationale d'Ingénieurs de Tunis

Décembre 2009

Plan

Plan

Quels sons ?

Sons issus de tout système de synthèse ou de transmission :

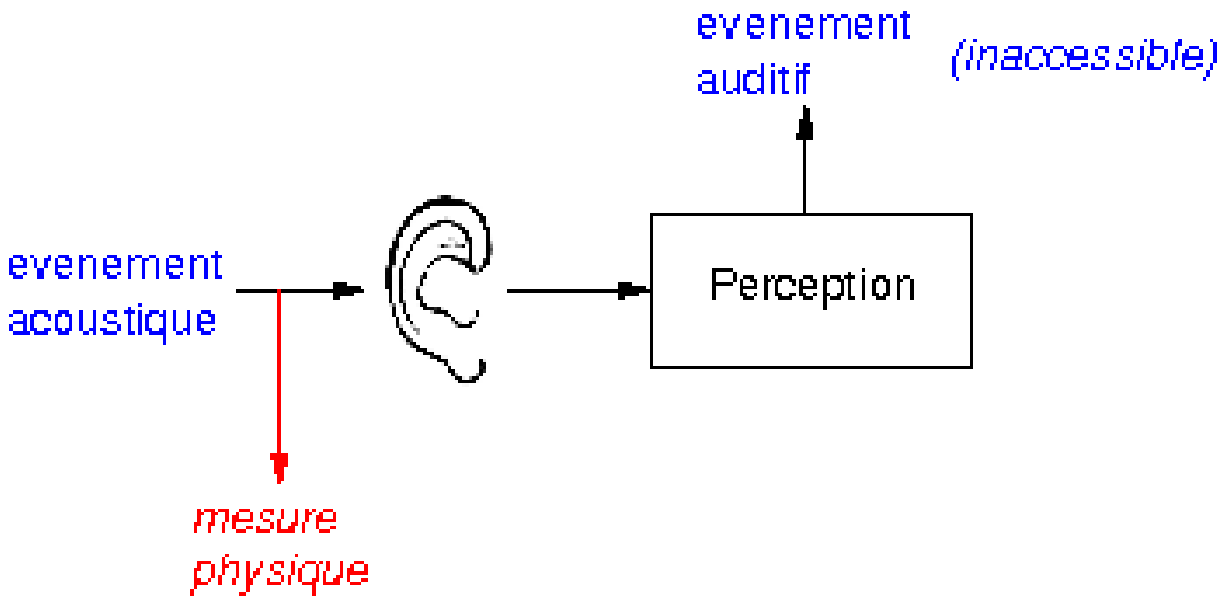
- système de téléphonie (liaison + codec)
- codec audio (MPEG, flac...)
- système acoustique : salle, haut-parleur, casque...
- synthèse sonore (voix ou musique)
- objet : voiture, rasoir électrique, sèche-cheveux...

... on souhaite en évaluer la "qualité" du son, du point de vue de l'auditeur.

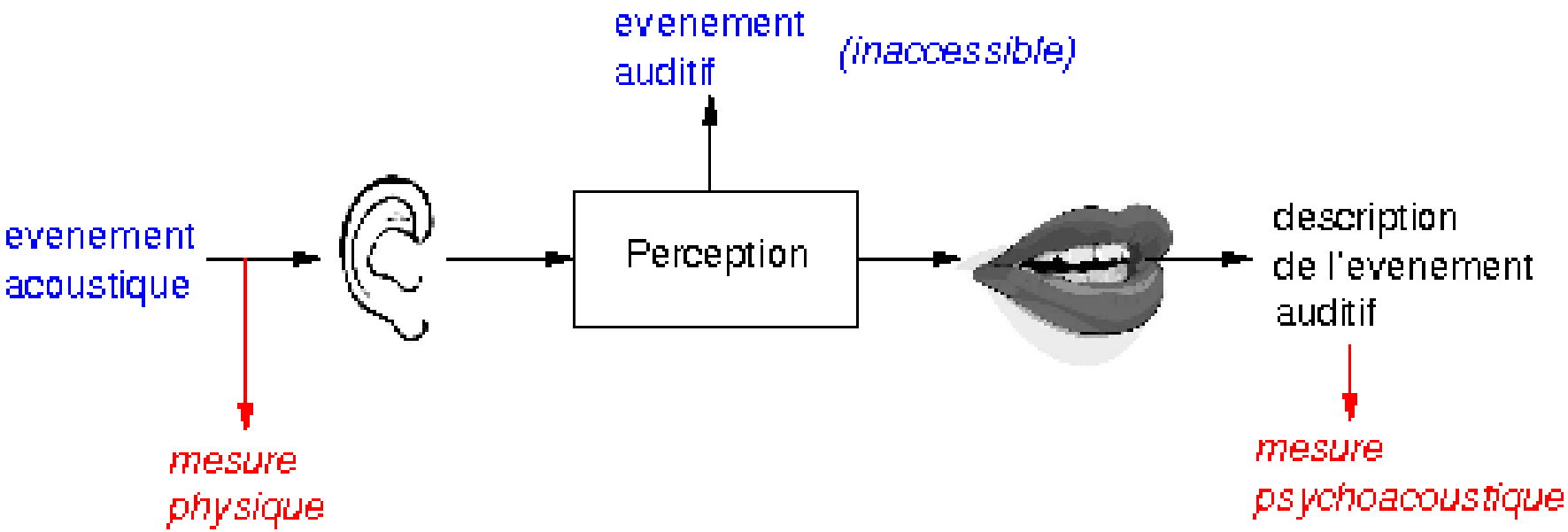
Qu'est-ce que la "qualité" ?

- **Un ensemble d'attributs perceptifs** : rugueux, chaleureux, métallique, aïgu/grave, mat/clair, bruité...
- **niveau de dégradation**
- **niveau de qualité absolue** (échelle de mauvais à excellent) ou préférences. Facteurs cognitifs :
  - culture auditive
  - facteurs sociaux
  - plausibilité du son, attentes de l'auditeur
  - caractère fonctionnel du son

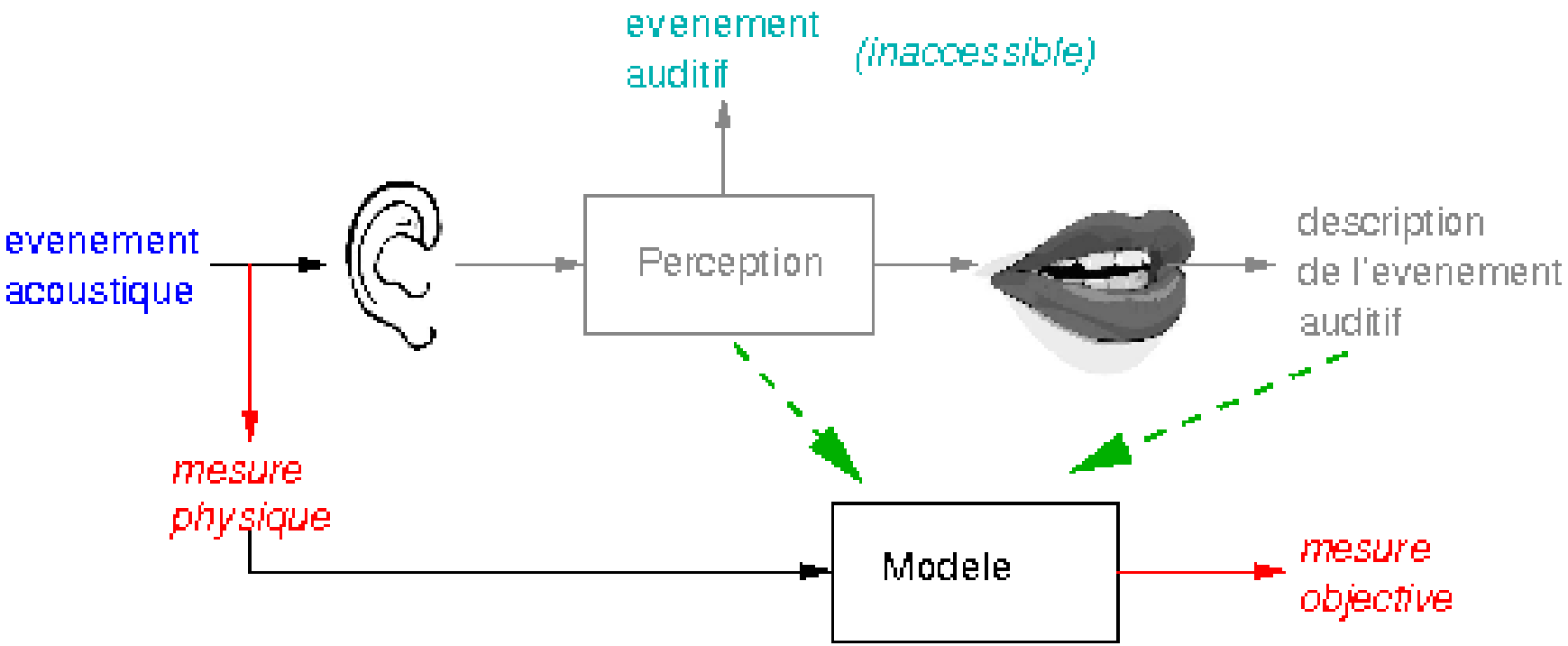
Comment mesurer cette qualité ?



Comment mesurer cette qualité ?



Comment mesurer cette qualité ?

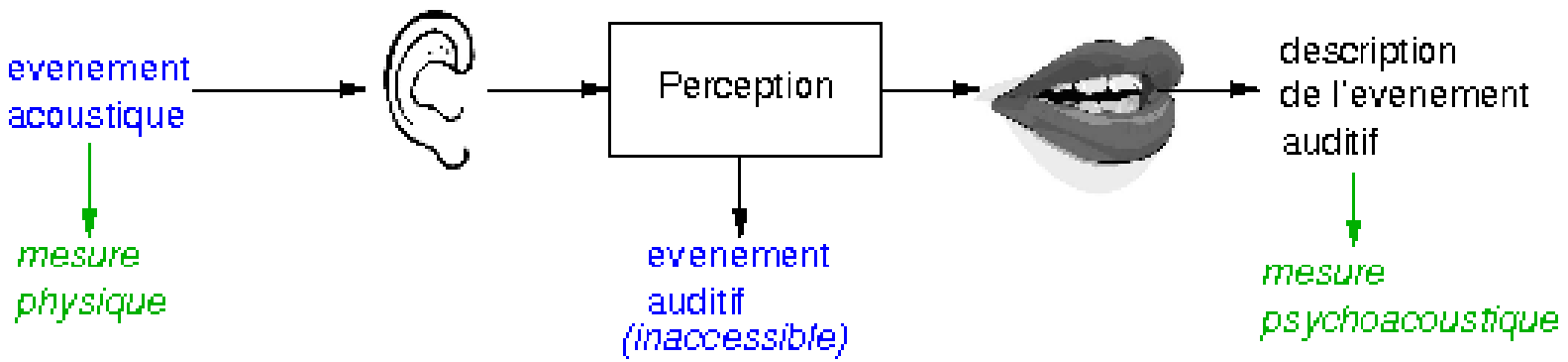


Plan

Variables principales et variables parasites (1)

Objectif

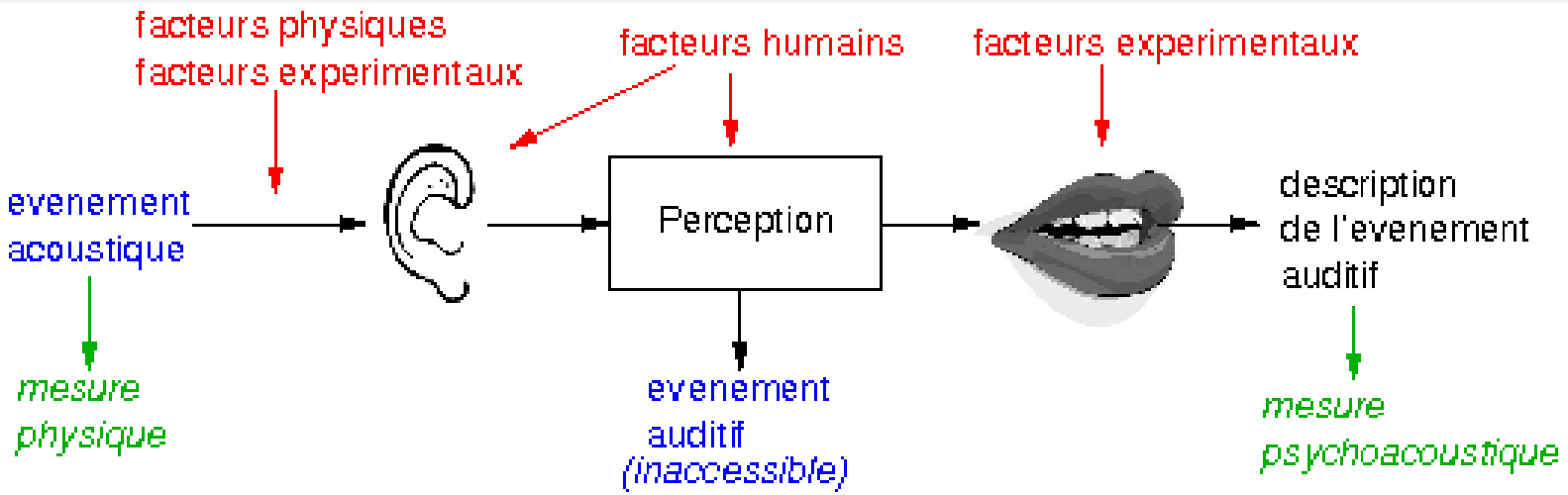
Définir une relation entre les caractéristiques objectives du système testé et la réponse subjective.



Variables principales et variables parasites (1)

Objectif

Définir une relation entre les caractéristiques objectives du système testé et la réponse subjective.



Conclusion

Réduire les sources de variabilité externes au système.

Variables principales et variables parasites (2)

Contrôler les facteurs humains

Variabilité inter- et intra-sujets :

- état physique : fatigue, audition
- état psychologique : motivation, concentration
- facteurs cognitifs (culture, attentes,...)
- entraînement

Inévitable ! Contrecarré par un moyennage :

- sur un grand nombre de sujets (variabilité inter-sujets)
- un nombre raisonnable de tests pour chacun (variabilité interne).

Variables principales et variables parasites (3)

Contrôler les facteurs physiques

Conditions de reproduction du son :

- salle
- hauts-parleurs ou casque
- position sujet / haut-parleurs
- niveau sonore

doivent être constantes et ne pas introduire de dégradations supplémentaires.

Variables principales et variables parasites (4)

Contrôler les facteurs expérimentaux

Le plan d'expérience doit tenir compte de l'influence :

- de l'ordre de présentation des stimuli
- de l'éventail des qualités
- des consignes
- du contexte visuel (cross-modal effects)
- du sens des phrases
- de l'adaptation des sujets aux stimuli

Variables principales et variables parasites (5)

Plan d'expérience en carré latin

Ex : 5 systèmes ABCDE testés par 5 groupes de sujets dans 5 ordres différents :

A	B	C	D	E
B	A	D	E	C
C	E	A	B	D
D	C	E	A	B
E	D	B	C	A

Isole l'effet de 2 variables parasites (sujets et ordre de présentation) de celui de la variable principale (système) si

- pas d'interactions entre les 3 variables
- les 3 variables ont le même nombre d'états

Présentation et évaluation des stimulis

- **Seuls :**
  - **note** sur une échelle de qualité (ACR)
  - **catégorisation**, qualification (tests de netteté)
- **Par 2 ou plus :**
  - **discrimination** : A et B sont-ils différents ?
  - **ordonner** A, B, C... selon préférence ou selon un caractère
  - **quantification** : de combien A et B sont-ils différents ?(DCR) Placer A, B, C,... sur une échelle (MUSHRA)

Dans tous les cas, un ordonnancement ou une quantification doivent se faire selon un caractère à une seule dimension : qualité, dégradation, hauteur, naturel, surroundness, position azimutale...

Plan

La normalisation en qualité audio

Organismes de normalisation

- UIT (ex-CCITT) : Union Internationale des Télécommunications
- ETSI : European Telecommunication Standards Institute
- IEC : International Electrotechnical Commission
- ISO : International Organization for Standards

Intérêt

Les normes

- ne couvrent pas toutes les applications
- ne donnent pas tous les détails des procédures de test
- ne donnent pas forcément la meilleure méthode

mais donnent des méthodes validées par la communauté scientifique, permettant des comparaisons fiables de systèmes.

Tests d'intelligibilité et de netteté

Intelligibilité et netteté

"La netteté concerne les mesures avec des symboles phonétiques dénués de sens ; l'intelligibilité se détermine, au contraire, avec des mots ou des phrases ayant un sens, support ou véhicules d'une idée qu'il faut comprendre... Le premier objet de la netteté est de ne pas faire intervenir, dans la mesure, la capacité de divination que nous acquérons par la pratique de notre langue et qui nous permet de reconnaître un mot, quoique nous ne l'ayons qu'imparfaitement perçu..." Chavasse, 1962, cité par Calliope dans *La parole et son traitement automatique*, Masson, 1989.

Tests d'intelligibilité et de netteté

Intelligibilité et netteté

"La netteté concerne les mesures avec des symboles phonétiques dénués de sens ; l'intelligibilité se détermine, au contraire, avec des mots ou des phrases ayant un sens, support ou véhicules d'une idée qu'il faut comprendre... Le premier objet de la netteté est de ne pas faire intervenir, dans la mesure, la capacité de divination que nous acquérons par la pratique de notre langue et qui nous permet de reconnaître un mot, quoique nous ne l'ayons qu'imparfaitement perçu..." Chavasse, 1962, cité par Calliope dans *La parole et son traitement automatique*, Masson, 1989.

Tests de netteté (1)

Reconnaissance de logatomes (CCITT, 1960)

- Mots de la forme consonne-voyelle-consonne (CVC)
- phonèmes et orthographe issus de l'esperanto (5 voyelles et 21 consonnes)
- 300 listes de 50 logatomes. Ex : stral ; psuc ; fliv ; $\hat{s}$up ; $\hat{c}$ek ; car ; gleg
- opérateurs entraînés
- test précis mais lourd et onéreux

Tests de netteté (2)

Tests de rime

Objectif : réduire la complexité de la tâche en limitant les choix du sujet

Exemple : Diagnostic Rhyme Test (DRT, 1973)

- Le sujet entend un mot mono-syllabique
- Il choisit entre 2 mots qui riment, dont les consonnes initiales ne diffèrent que d'un trait (voisé ou non, nasal ou non, ...)

Test de conversation (CCITT, 1981

- Deux sujets simulent une conversation téléphonique dans des conditions proches du réel, sur la base d'un "prétexte à conversation".
- Inconvénients :
  - long et coûteux
  - nécessite une implémentation temps-réel des systèmes

Tests d'opinion directe (1)

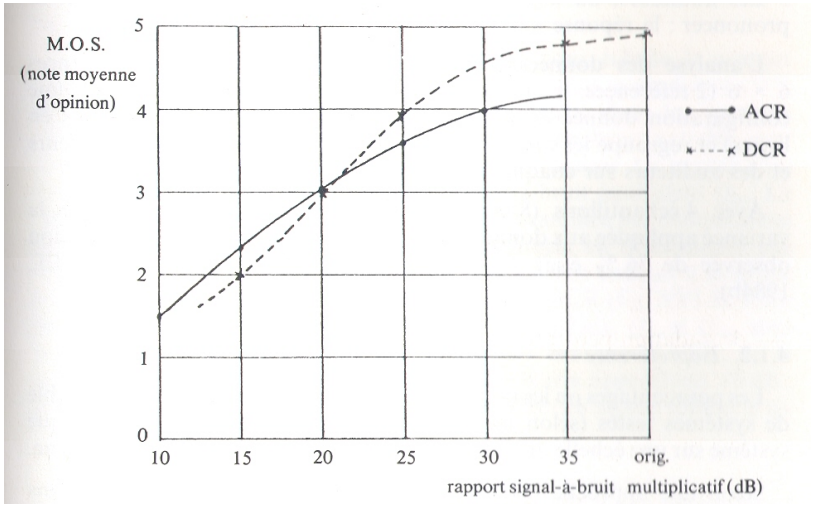
Absolute Category Rating (ACR) (CCITT, 1985)

- Echantillons de 6 à 10s, 2 à 3 phrases
- au moins 2 locuteurs et 2 locutrices
- références = signaux bruités, de RSB instantané constant (MNRU)
- notation de la qualité d'écoute sur échelle à 5 valeurs : excellent / bon / passable / médiocre / lauvais

Tests d'opinion directe (2)

Degradation Category Rating (DCR) (CCITT, 1985)

- Conditions expérimentales similaires
- Présentation des échantillons : ABAB, où A=référence et B = phrase issue d'un système à tester.
- Le sujet note la dégradation de B par rapport à A sur une échelle à 5 valeurs : imperceptible / perceptible mais non gênante / légèrement gênante / gênante / très gênante



Tests d'opinion directe (3)

Comparaison par paires

- utilisé pour choisir entre plusieurs systèmes
- Pour chaque paire de systèmes AB,
  4 comparaison AB + 4 comparaisons BA
- 4 ou 8 échantillons, constitués de doubles phrases
- 16 ou 32 auditeurs
- Le sujet doit toujours indiquer une préférence
- Résultat = matrices de préférences moyennant préférences de tous les auditeurs sur tous les échantillons

Avec 4 échantillons, 32 auditeurs et 4 systèmes, une préférence à 56 % est significativement supérieure à 44 %.

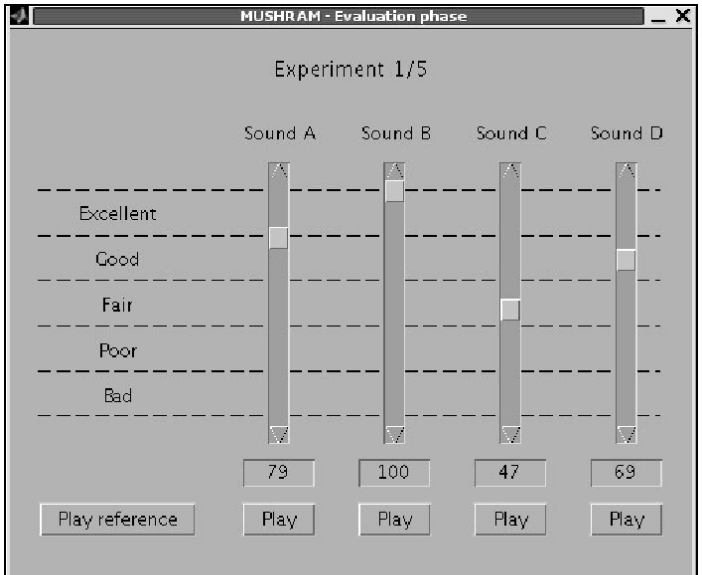
Audio de haute qualité, avec dégradations faibles

Norme ITU.R BS.1116-1 :

- Sujets :
  - 20 à 30 sujets experts
  - entraînés à détecter les artefects à bas-débit
  - post-sélection des sujets
- Condition expérimentales :
  - échantillons = différents types de musique
  - "matériaux critiques"
  - casques, HP et salle précisément définis
- Test :
  - de dégradation à 3 stimuli : référence, A, B
  - échelle de dégradations comme DCR
  - normalement, notes > 4

Audio de qualité intermédiaire : ITU.R BS.1534

- Contexte : diffusion musique sur internet, à bas débit
- MUSHRA : MUlti-Stimulus Hidden Reference and Anchor
  - Multi-stimulus : accès instantané aux signaux issus de tous les systèmes.
  - Hidden Reference : un des signaux tests est une l'original
  - Anchor : un des signaux tests est l'original filtré passe-bas
- échelle continue de 0 à 100, parallèlement aux appréciations ACR



Plan

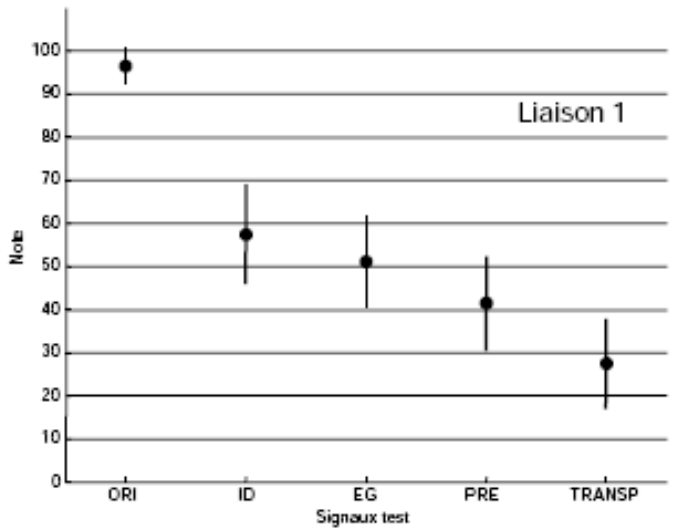
notes ACR, DCR, MUSHRA...

Résultats =

Pour chaque système,

- note moyenne μ
- intervalle de confiance à 95 % = $\mu \pm 1.96 \frac{\sigma}{\sqrt{N}}$

$\rightarrow$ significativité des écarts de moyennes ?



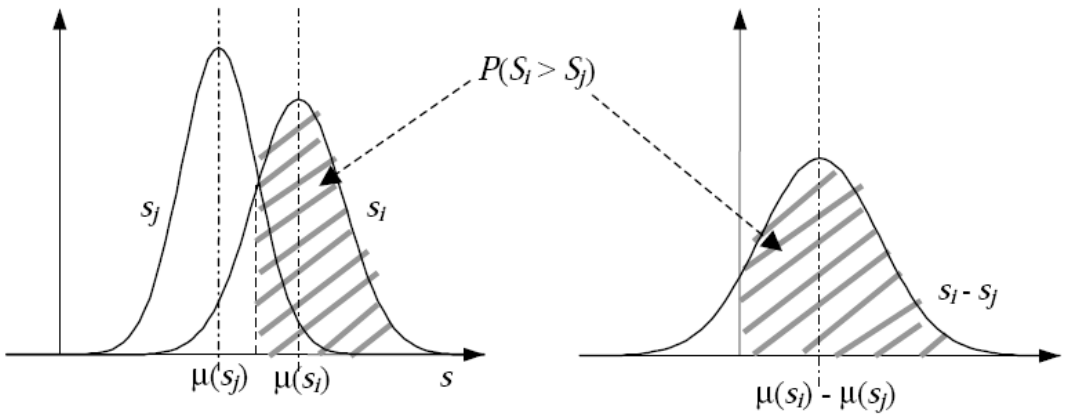
Expression de préférences

Objectif

Pourcentages de préférences par paire AB, AC, BC...
$\rightarrow$ positionnement des stimuli sur une échelle de préférence

Echelle de Thurstone

Postulat : stimulus $S_i \rightarrow$ jugement s_i = variable aléatoire gaussienne de moyenne $\mu(s_i)$
Un sujet préfère S_i à S_j si $s_i > s_j$: $\Pr(S_i > S_j) = \Pr(s_i - s_j > 0)$
Connaissant $\Pr(S_i > S_j)$, que vaut $\mu(s_i) - \mu(s_i)$?



Plan

Motivations

Les tests d'écoute...

- permettent l'évaluation la plus fiable de la qualité
- mais ils sont chers et chronophages

Objectif

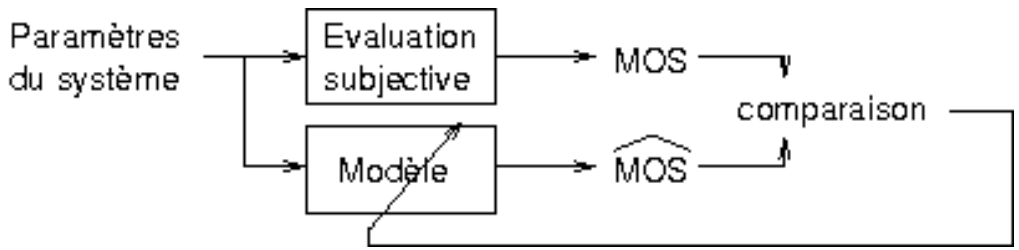
Prédire la qualité perçue à partir des paramètres physiques du son ou du système qui le transmet, grâce à la modélisation de la perception.
Mais la perception n'est pas parfaitement modélisable (aspects cognitifs non maîtrisés)

Modèles paramétriques

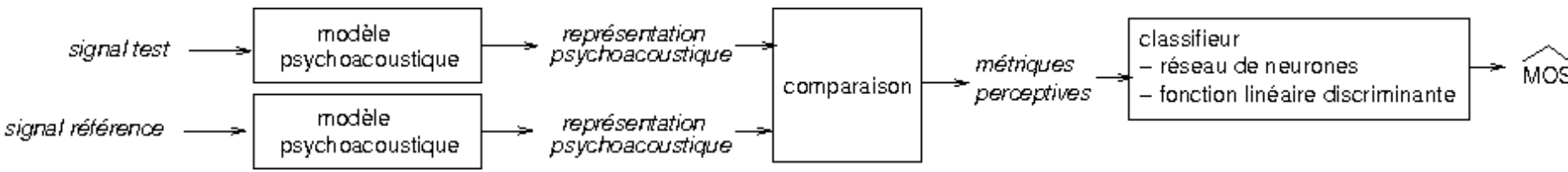
Principe

Paramètres physiques du système $\rightarrow$ note MOS
Paramètres physiques = délai, écho, bruit, atténuation...
Ex : norme ITU-T G.107

Apprentissage du modèle :



Modèles fondés sur les signaux



Ex :

- Pour la parole : PESQ (Perceptual Evaluation of Speech Quality) = ITU-T P.862
- Pour l'audio : PEAQ (Perceptual Evaluation of Audio Quality) = ITU-R BS.1387

Validité des modèles

- Modèles valables dans certaines conditions précisées par les normes
- Fiabilité du modèle dépend de l'apprentissage