Parole et langage

G. Bailly - ICP

Plan du cours

1. Introduction
2. Langues du monde - systèmes phonologiques
3. Production de parole
4. Perception audiovisuelle de la parole
5. Lecture de spectrogrammes
6. Transcription orthographique-phonétique
7. Synthèse audiovisuelle (1)
8. Synthèse audiovisuelle (2)
9. Reconnaissance de la parole (1)
10. Reconnaissance de la parole (2)
11. Boucles perception/action
12. Interaction Homme-Machine

A lire…

- Traitement de la parole, Boite et al, 1999
- Speech communication - human and machine, O'Shaughnessy, 2000
- An introduction to Text-to-Speech synthesis, Dutoit, 1997
- Computer facial animation Parke & Waters, 1996

Le traitement de la parole

- Naturel et sans effort pour les humains
- Caractéristique fondamentale de l'humanité
  - Débat sur l'origine des langues et du langage
  - Cognition
- Très difficile pour les machines
  - Techniques: synthèse, reconnaissance
- Multimodalité
  - Causes - effets
- Une composante de l'interaction

Les espoirs (1)

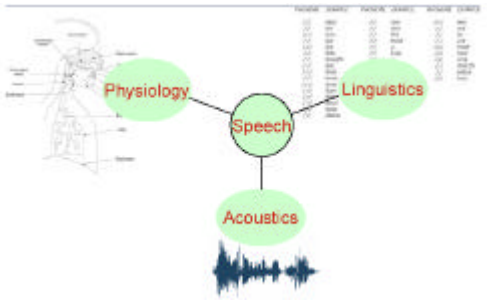
- Ubiquitous computing: *A Space Odyssey* (Kubrick 1968)
  - *"Together, **Hal and Discovery** constitute an essential living organism that symbolizes a hypothetical new humanoid species, humanoid machines. In other words although Hal-Discovery is a single entity - an individual - in the surface story, he symbolizes an entire race of machines in the man-machine symbiosis allegory."* (*Leonard F. Wheat*, 2000)

(*After 2001: A Space Odyssey*, 1968)

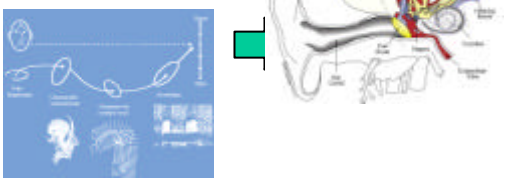
Les espoirs (2)

- Smart objects: *The Talking Toaster* (BBC *Red Dwarf* 1996)
  - *"Instead of fiddling with the toast-quality dial or hitting the down level, the toaster will actually ask you for the settings. Even better, you can simply respond by speaking your reply – no buttons to push, dials to spin, or lights to watch."*

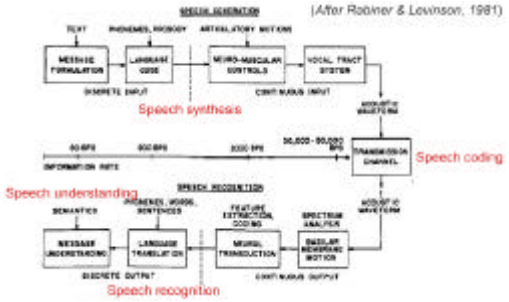
Sciences de la parole



La chaîne de communication

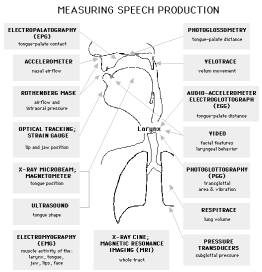


Débits d'information



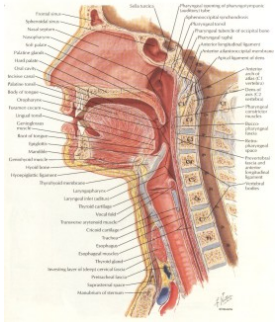
Production de parole

G. Bailly - ICP



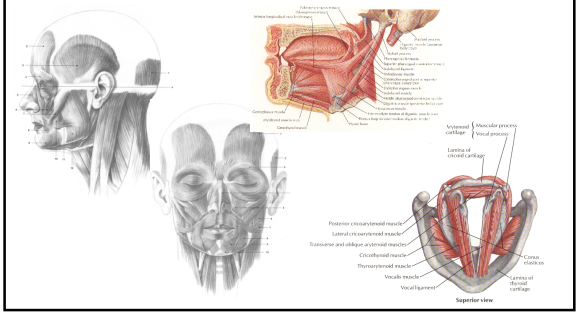
Organes de la phonation...

- Sources
  - Poumons
  - Constrictions
    - Écoulement laminaire/turbulent
    - Mise en oscillation: cordes vocales, trilles bilabiaux/linguaux, ronflement...
  - Obstacles
- Géométrie



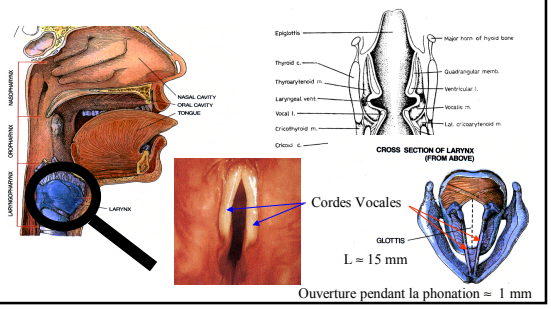
Organes de la phonation...

- Géométrie externe et interne

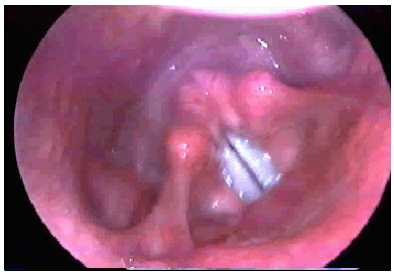


Organes de la phonation...

- Le larynx et les cordes vocales

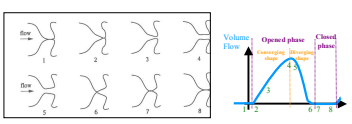


Les cordes vocales en action (1)



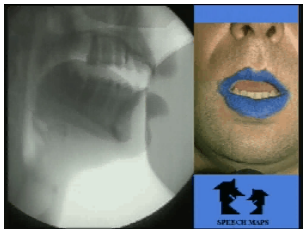
Les cordes vocales en action (2)

- Cycle:
  - Écartement par pression pulmonaire
  - Force de rappel élastique
  - Effet Bernoulli



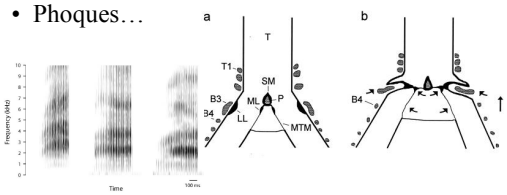
Organes de la phonation...

- Conduit vocal
  - Mise en forme de cavités
- Articulateurs
  - Mâchoire
  - Lèvres
  - Langue
  - Vélum
  - Joues
  - ...



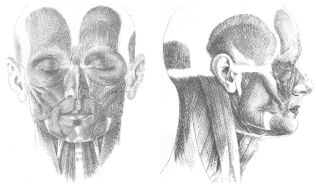
Sons de parole produits par d'autres espèces

- Mainates, perroquets
  - Syrinx: replis de tissus implantés de part et d'autre des voies respiratoires.
- Phoques…



Conséquences visibles

- Mâchoire
- Lèvres
- Joues
- Larynx?



- Débat sur corrélations entre gestes audibles/visibles et sur interprétations

Données sur corrélations (1)

- ATR: 12/18 OPTOTRACK infrared markers vs 4 EMMA coils. Two different sessions. Data alignment procedure: acoustic signal & lip markers (R=0.93).
- UCLA: 18 infrared markers vs 3 EMMA coils. Simultaneous recordings.
- Converge towards correlation between tongue shapes and multilinear predictors from facial data ranging from 0.6 to 0.8.
- Tongue tip less predictable than tongue rear
- Bailly-Badin, ICSLP '2002: Speaker-specific data-driven articulatory models of the vocal tract and the facial deformations have been developed. Both models are fitted to cineradiographic data. Visible movements are accessed by means of the deformations of the speaker's profile whereas VT tract shapes are characterized by mid-sagittal contours of the different speech organs

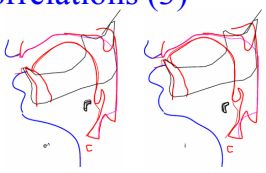
Données sur corrélations (2)

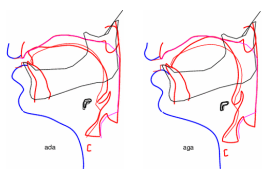
- Mean RMS reconstruction error
  - of the tongue profile is less than 1mm except in the pharyngeal region (1.5mm)
  - Of 14 facial midline fleshpoints is 0.86 mm
- Correlation coefficients
  - for the tongue shape parameters range from 0.37 to 0.74

	j1	l1	l2	l3	j2	s1	LR
LH	.50	-.02	.84	.83	-.08	.11	.99
LP	.13	.96	.34	.02	.09	.33	.98
JH	.99	.19	.44	.40	.04	.15	.99
TB	.24	.07	-.24	.01	.35	-.24	.71
TD	-.11	.20	.22	.18	-.50	-.12	.64
TT	.33	.34	.39	.37	-.01	-.24	.74
TA	.04	-.10	-.01	-.18	-.17	.03	.37
LY	-.00	.57	-.26	-.46	-.13	.25	.84
LV	-.00	.02	-.47	.55	-.26	-.50	.99

Données sur corrélations (3)

- Good … and less good recoveries
  - Configurations associating a jaw/tongue/lips synergy along the axis closed/front (e.g. [i]) vs. open/back (e.g. [a]) are accurately recovered
  - Most configurations requiring constrictions deviating from this synergy are poorly predicted. This later case includes most consonants in open contexts and velars in closed context as well as labialised vowels.





Gestes visibles et peu audibles

- Exploitent la conduction des vibrations par os/peau
  - Microphone gorge
  - NAM
- Application à la téléphonie silencieuse

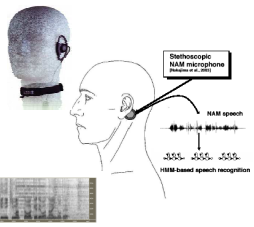


Figure 2: Normal speech waveform - Close-talking microphone

Figure 4: Normal speech spectrogram - Close-talking microphone

Figure 3: NAM speech waveform - NAM microphone

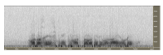
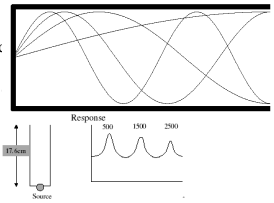
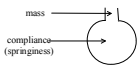


Figure 5: NAM speech spectrogram - NAM microphone

Rudiments d'acoustique linéaire

- Résonances de tubes
  - Fermé: multiples impairs d demi-onde $\frac{kc}{2L}$
  - Ouvert: multiples impairs quart d'onde $\frac{(2k+1)c}{4L}$
  - Helmholtz $\frac{c}{2\pi}\sqrt{\frac{A}{Vl}}$
- Dépend de c
  - Ex: air vs hélium

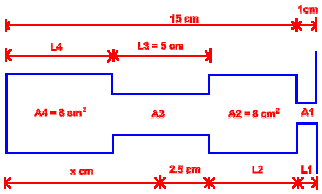


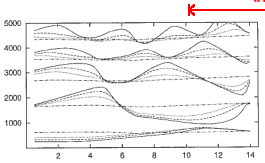


Formula: only bother with it if you want to!
Resonance frequency =
where: c = speed of sound in [air]
A = cross-sectional area of neck
V = volume of bottle area (back cavity)
l = length of neck

Modèle à tubes

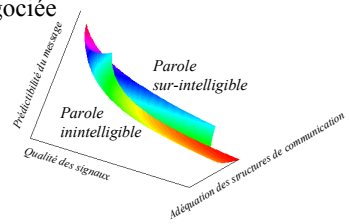
- Fant 1968
  - Deux cavités
  - Deux constrictions





Négociation locuteur-auditeur

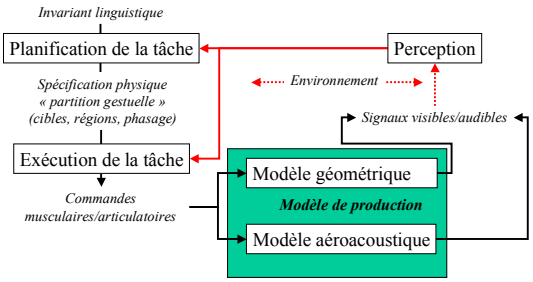
- Gestion "optimale" de l'articulation en fonction de l'espace de croyance mutuel
- Variabilité négociée
- Contenu
- Réalisation
- Nombreuses dimensions...



Sources de variabilité

- Anatomie anthropométrie (ex: longueur cordes vocales, conduit vocal, dimension fosses nasales...)
- Stratégies contrôle (ex: usage mâchoire...) structuration de l'espace (ex:opposition d'ouverture....)
- Coarticulation effet des segments adjacents (ex: réduction, assimilation...
- Prosodie structuration du discours, origines sociolinguistiques et socioculturelles ... et ... état physiologique & psychologique du locuteur, contraintes environnementales...

Modèles de production (1)

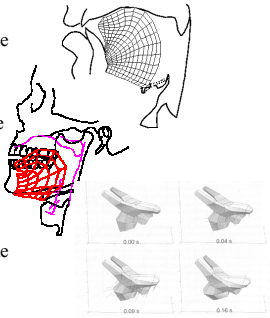


Modèles de production (2)

- Modèles géométriques
  - Tubes
  - Modèles articulatoires
  - Modèles biomécaniques
- Modèles aérodynamiques
  - Modèles fonctionnels
  - Modèles physiques

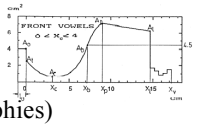
Modèles biomécaniques (1)

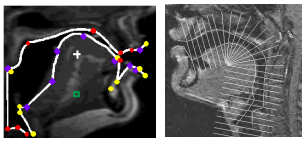
- Payan & al (1996)
  - Modèle élément finis 2D de la langue
- Laboissière et al (1997)
  - Intégration dans un modèle de contrôle du mouvement de la mâchoire et de l'os hyoïde
- Wilhems (1995)
  - Modèle élément finis 3D de la langue



Modèles géométriques (1)

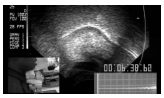
- Modèles de fonction d'aires
  - 2/4 tubes (Fant 68)…
- Modèles 2D (issus radiographies)
  - Fonctions géométriques (Mermelstein 73)
  - Grille/statistique (Maeda 78)
  - Passage à la fonction d'aire

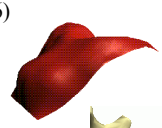


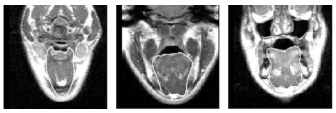


Modèles géométriques (2)

- Modèles 3D
  - Extension de Mermelstein (Engwall 96)
  - Ultrasons (Stone, 1996)
  - IRM (Badin et al, 2002)

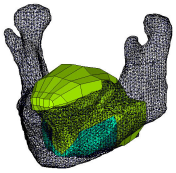






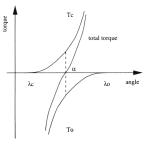
Modèles biomécaniques (2)

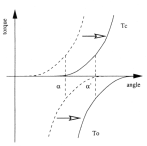
- Gérar & al (2003)
  - Modèle élément finis 3D de la langue

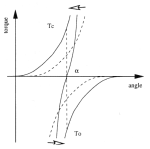


Modèles de contrôle musculaire

- Feldman
  - Seuil de recrutement
  - Point d'équilibre
  - Co-contraction





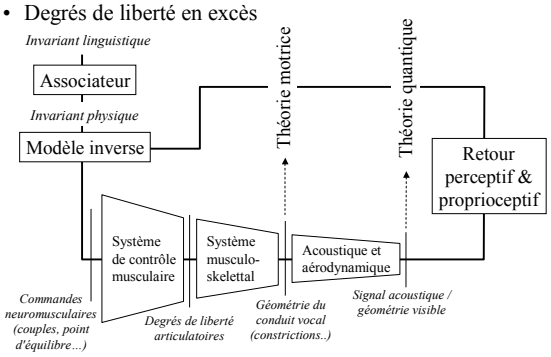


Planification des gestes

- Contrôle direct/rétroaction
- Modèle interne
- Espace de planification

Monnaies d'échange

- Degrés de liberté en excès

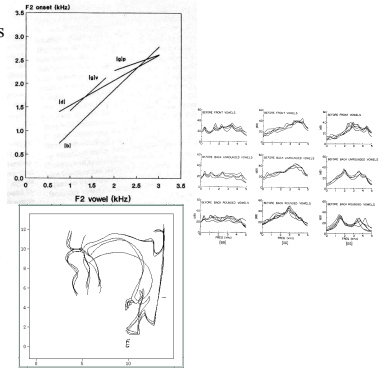


Monnaies d'échange

- Espace de planification de la tâche phonologique
  - Commandes neuromusculaires
    - TH point d'équilibre : cibles + co-contraction
  - Commandes articulatoires
    - degrés de liberté: mâchoire, vélum ....
    - synthèse audio-visuelle
  - géométrie et aérodynamique
    - consonnes, couplage sources-conduit vocal
  - Caractérisation acoustique
    - équifinalité, incomplétude des synthétiseurs articulatoires
- Exécution: modèle direct/inverse

Quels invariants pour les occlusives?

- Equations de locus
- Spectre du burst
- Lieu d'articulation

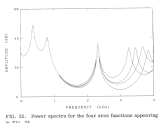


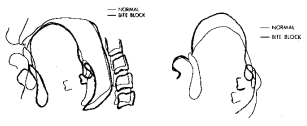
Le débat sur la nature des invariants

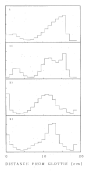
- Summerfield à Klatt: "Suppose that somebody started to implement Ken's view of lexical access on the computer, and simultaneously somebody attempted to form and to implement algorithms to recover the gestures produced in the VT. Would either of them solve the problem? Which of them would succed first?" (Panel discussion: the Motor Theory and alternative accounts *Modularity and the Motor Theory of Speech Perception*, Catford et al, 1991)
- Klatt: Pas de réponse définitive possible car:
  - Pas de relations suffisantes sur les relations articulatoire-acoustique
  - Pas de modèles robustes pour l'extraction d'indices acoustiques invariants

Acoustique inverse

- Atal et al [1978]: "Large changes in the shape of the vocal tract can be made without changing the formant frequencies. These changes are consistent with the hypothesis that compensatory articulation is a possibility… They are also consistent with the art of ventriloquism… It seems worth investigating whether some minimum motion or minimum energy principle is applied in going from one sound to another…"

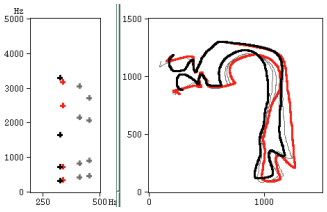






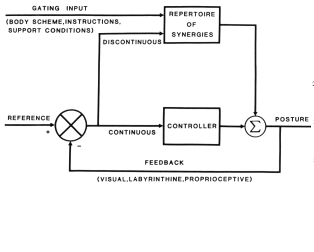
Lip-tubes

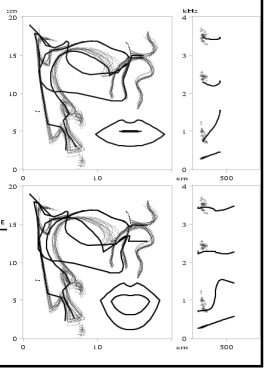
- Savariaux et al [1995]
  - F1 et F2 identiques
  - Solutions articulatoires avec géométries différentes



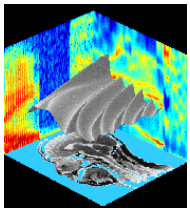
Modèles de contrôle

- Contrôle direct
- Contrôle par rétroaction





Perception de la parole



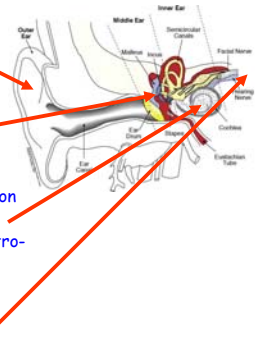
G. Bailly – ICP
(crédits W. Serniclaes - Paris 7 & J.L. Schwartz - ICP)

Contenu

- Physiologie
- Psychoacoustique
- Analyse de scènes auditives
- Perception catégorielle

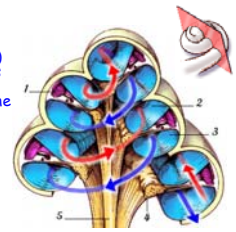
Physiologie

- Oreille externe
  - Protection/amplification
  - Vibration aériennes
- Oreille moyenne
  - Adaptation d'impédance
  - Vibrations mécaniques
- Oreille interne
  - Filtrage/analyse/transduction
  - Vibrations liquidiennes/processus électro-chimiques
- Nerf auditif
  - Traitement de l'information
  - Potentiels d'action



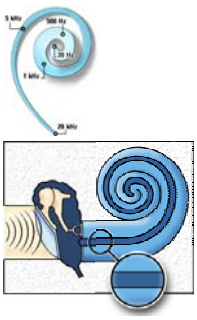
Oreille interne

- Cochlée
  - La section axiale schématise l'enroulement du canal cochléaire (1) contenant l'endolymphe, et celui des rampes vestibulaire (2) tympanique (3) contenant la périlymphe. La flèche rouge vient de la fenêtre ovale et la bleue aboutit à la fenêtre ronde. Au centre, (modiolus) le ganglion spiral (4) et les fibres du nerf cochléaire (5) apparaissent en jaune.
  - Le canal cochléaire (1), contenant l'endolymphe sécrétée par la strie vasculaire (7), est isolé de la rampe vestibulaire (2) par la membrane de Reissner (4). L'organe de Corti est recouvert par la membrane tectoriale (6) flottant dans l'endolymphe ; il repose sur la membrane basilaire (5) au contact de la rampe tympanique (3). La lame spirale osseuse (9) relie l'organe de Corti au ganglion (8).



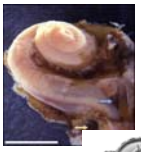
Membrane basilaire

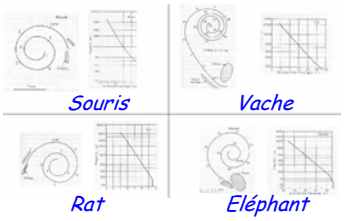
- Membrane de l'oreille interne
  - située dans la cochlée et tendue entre la lame spirale interne et la lame spirale externe.
  - longueur d'environ 25 à 35 mm
  - épaisseur d'environ 0,003 mm.
  - largeur variant selon que l'on considère la base (environ 0,04 mm) ou l'apex (environ 0,36 mm).
  - sert de support à l'organe de Corti, lequel est le récepteur des vibrations.
- Tonotopie
  - Base = HF, Apex = BF
  - Les fibres agissent comme des filtres passe-bandes, leur sélectivité en fréquence dépendant de leur fréquence caractéristique.

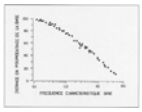


Morphologies et tonotopies

- Nouveau-né
  - 5 mois de gestation: le développement morphologique de la cochlée est achevé
- Morphologie et tonotopie comparée
  - Distribution logarithmique
  - Meilleure discrimination en BF

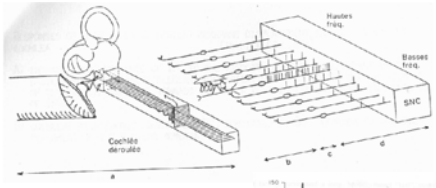




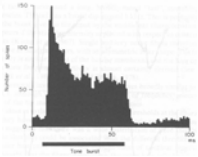


Traitements spatio-temporels (1)

- Neurone primaire, guetteur spectral

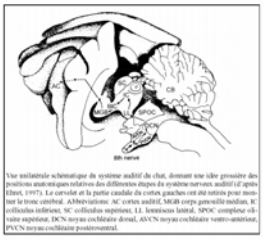


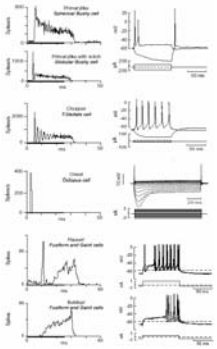
- Adaptation nerveuse



Traitements spatio-temporels (2)

- Etages supérieurs
  - Du noyau cochléaire au cortex auditif
  - A noter neurone « On » dans PVCN





Psychoacoustique

- Sensibilité fonction de la fréquence
- Variables perçues vs. variables physiques
- Masquage
  - Fonctions de masquage employées en codage MP3…

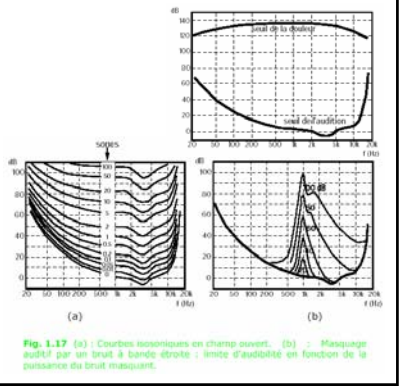


Fig. 1.17 (a) : Courbes isosoniques en champ ouvert. (b) : Masquage auditif par un bruit à bande étroite ; limite d'audibilité en fonction de la puissance du bruit masquant.

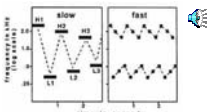
Analyse de scène (1)

- Séparation de mélanges

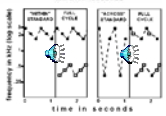
- Principe
  - Décomposition en blocks temps-fréquence
  - Recombinaison
- Principes de regroupement
  - Analogie avec vision

Analyse de scène (2)

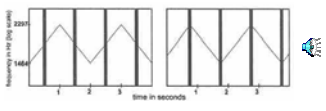
- Streaming



- Shémas

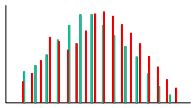


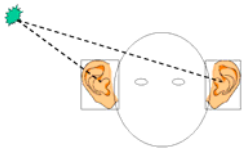
- Pogendorf



Parole et analyse de scène

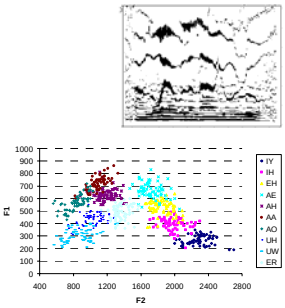
- Regroupement par primitives spécifiques
  - F0
  - Délai interaural: ITD
- Application à la séparation de sources
  - ICA…

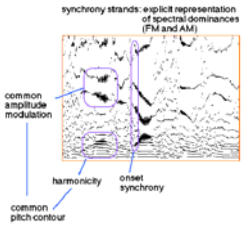




Catégorisation

- Segmentation/regroupement/localisation
- Catégorisation des éléments



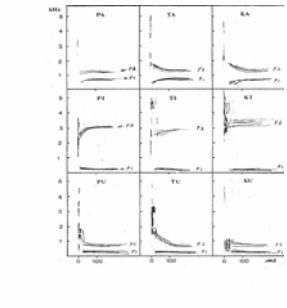


Perception catégorielle

1. Question fondamentale sur la perception de la parole: invariance perceptive des traits phonologiques chez l'adulte et son acquisition chez l'enfant
2. Faits de base: Processus: perception catégorielle et ajustements contextuels
   - Ajustements contextuels
   - Perception catégorielle (PC)
3. Théorie Auditive ou « Quantale » (Stevens)
   - Ajustements contextuels: invariants acoustiques
   - PC: frontières psychoacoustiques naturelles
4. Théorie Phonétique: ou « Motrice » (Liberman)
   - Ajustements contextuels: effets sans changements acoustiques
   - PC: traitement neural spécifique
5. Théorie Phonologique
   - Ajustements contextuels:interactions dans la perception des traits phonétiques
   - PC: frontières perceptives à l'intérieur des catégories phonologiques

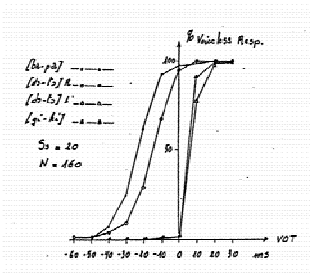
Les faits (1)

- Variations contextuelles dans la production des traits



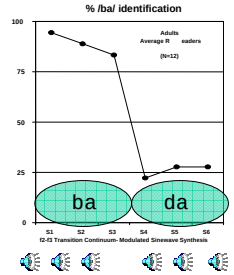
Les faits (2)

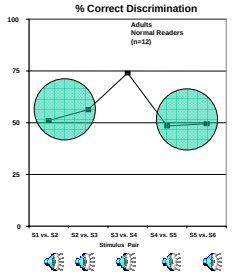
- Ajustements contextuels dans la perception des traits



Perception catégorielle

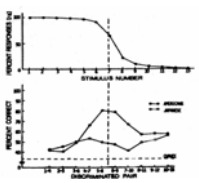
- Définition: un continuum de stimuli sur lequel sont définies des catégories perceptives est perçu de manière catégorielle
- si les différences entre stimuli sont imperceptibles tant qu'ils n'appartiennent pas à des catégories différentes.

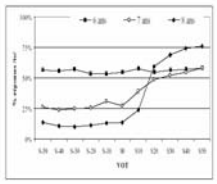


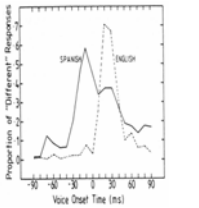


Facteurs affectant la PC (1)

- Influence de la « géographie » phonologique (Wood, 76)
  - Frontières additionnelles
  - Déplacement des frontières
- Influence de la compétence linguistique (Simon & Fourcin ,74)



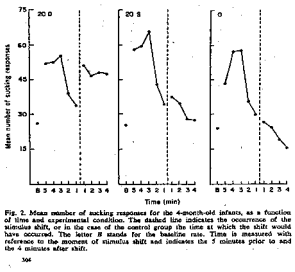




Facteurs affectant la PC (2)

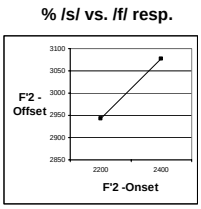
- Les frontières perceptives naturelles, mises en évidence chez l'enfant pré-linguistique (nourrisson, avant 4 mois), sont plus diversifiées que celles de l'adulte
- de plus, ces frontières naturelles ne correspondent pas nécessairement à celles de l'adulte
- suggère que ces frontières inclues dans les prédispositions du nourrisson ne sont pas phonologiques, mais phonétiques (universelles)

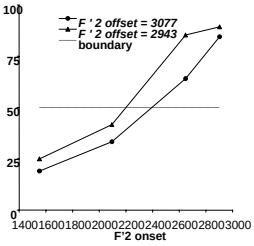
Eimas, Siqueland, Jusczyk, & Vigorito. (1971) ***Science 171***, 303-306.



Modèles d'invariance acoustique

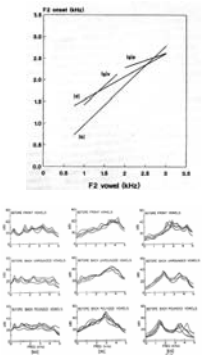
- perceptual tradeoffs between acoustic cues contributing to the perception of the same phonetic feature (Hoffman, 1957) => acoustic cue integration
- productive tradeoffs: cues complement each other: in context in some cues are « weak », others are « strong » (Dorman et al., 1977) => whole is more invariant than parts





Modèles d'intégration

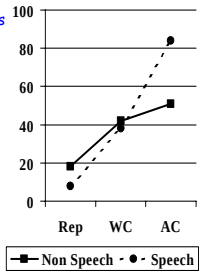
- heterogeneous summation (Tinbergen):
  « cues add up like vectors » (Hoffman, 1957)
- lawful combination:
  Locus theory (Delattre, 1958)
  Linear model (Sussman et al., 1991; 1998)
- holistic properties: short-time spectrum (Quantal theory: Stevens, 1973; 1989)



CP phonétique/phonologique

- Méthode
  - voir si les phénomènes mis en évidence pour la perception de la parole chez l'homme restent présents avec des stimuli qui ne sont pas perçus comme de la parole, tout en ayant des caractéristiques acoustiques les plus proches possibles de celles de la parole
    - Parole sinusoidale (Remez et al., 1981, *Science* 212, 947
    - Stimuli ambivalents d'abord présentés comme des sifflets puis comme des syllabes
    - Réponses corticales mieux discriminées en mode parole
  - comparer les réponses de sujets humains et celles d'animaux à des stimuli de parole.

Dehaene-Lambertz et al. (2005) *NeuroImage*



En faveur de l'invariance acoustique

- Kluender et al. (1987): 'Japanese Quail can learn phonetic categories'
- Learned to discriminate /dis, dus, das…/ from /bis, bus, bas…/ or /gis, gus, gas…/

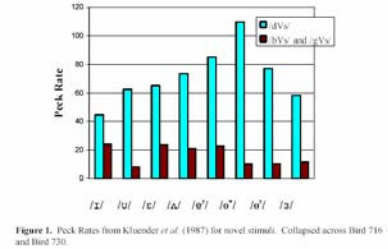


Figure 1. Peck Rates from Kluender *et al.* (1987) for novel stimuli. Collapsed across Bird 716 and Bird 730.

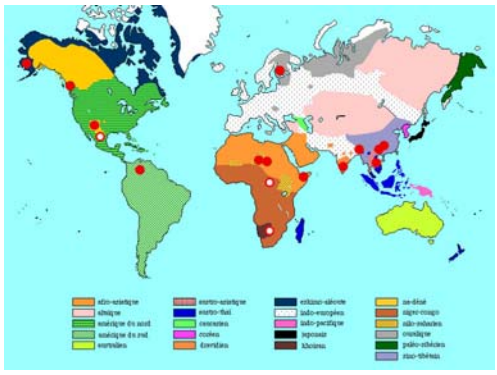
Langues du monde

G. Bailly – ICP
(crédits L.J. Boë, N. Vallée & A. Vilain – ICP)

Les 4000 à 6000 langues du monde sont classées en fonction des ressemblances entre les mots (prononciation)

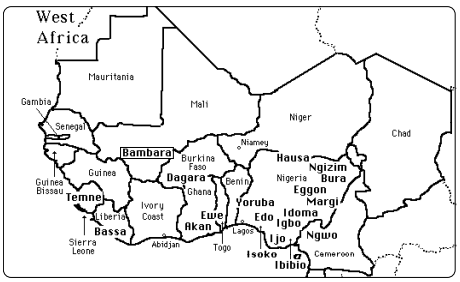
langue	un	deux	tête	oreille	dent
hollandais	en	tve	hoft	or	tant
anglais	wʌn	tuw	hɛd	ijr	tuwθ
français	œ̃	dø	tɛt	ɔʀɛj	dɑ̃
italien	uno	due	tɛsta	orekjo	dɛnte
basque	bat	bi	byry	belari	orts

Les sons des langues du monde



UPSID quelques langues du monde...

indo-européen (23): grec, irlandais, breton, allemand, norvégien, lithuanien, russe, bulgare, français, espagnol, roumain, farsi, pashto, kurde, hindi-urdu, bengalais, cachemirien, cinghalais, albanais, arménien, népali, konkani, ormuri… **ouralo-altaïque (28):** aïnou khanty mari komi finnois hongrois saami nenets nganasan turc azerbaïdjanais chouvache yakoute kirghiz bashkir selkup tuva khalkha even nanai mandchou coréen japonais youkaghir monguor moghol ouzbek dagur...**austro-asiatique (14):** mundari kharia khasi vietnamien sedang cambodgien parauk sre brao khmu? nicobarais pacoh kur bruu… **austro-thaï (39):** thaïlandais lakkia yay soui kam po-ai lungchou atayal javanais malgache cham sama batak tagalog sa'ban chamorro rukai tsou adzera roro kaliai iai hawaïien tigak lenakel lue ivatan mor pohnpeien tiruray lai kwaio paiwan iban gelao tetun fidjien irarutu maranao… **sino-tibétain (21):** mandarin taishan hakka changzhou xiamen fuzhou bai tamang dafla burmese lahu jingpho ao chin bodo karen mien newari hmong phlong naxi… **caucasien (7):** géorgien kabardien lak rutul bats archi avar… **autres familles euro-asiatiques (7):** nivkh kète tchouktchi koryak itelmen basque bourouchaski… **dravidien (6):** telugu kota kouroukh koya tulu brahoui… **niger-congo (55):** moro kadugli kpelle bisa bambara dan wolof diola temne dagbani senadi tampulma bariba ewe akan igbo ga lelemi efik birom tarok amo beembe ogbia ejagham zoulou teke doayo gbeya azande aizi mumuye klao aghem kpan kohumono yoruba bobo-fing noni gwari ewondo jomang sango bete konyagi mbum isoko fe?fe? ndut ijo lua alladian mambila mba-ne dogon… **nilo-saharien (23):** songhaï kanouri maba fur maasai luo nubien nyangi ik sebei tama temein nera tabi mursi lughara yulu berta kunama koma daju dinka nyimang… **afro-asiatique (26):** arabe tigre amharique socotri néo-araméen chleuh tamasheq somali awiya iraqw beja kullo dizi kefa hamer hausa angas margi ngizim kanakuru kera dahalo kotoko dangaleat tera lame… **khoisan (4):** hadza sandawe !xu nama… **na-déné (7):** haida tlingit navaho hupa chipewyan ahtna eyak… **amérique du nord (58):** percé klamath maidu wintu zoque tzeltal totonac k'ekchi mixe huave mazahua mazatec mixtec chatino tseshaht kwakw'ala quileute lushootseed papago luiseno hopi yaqui picuris karok pomo diegueno achumawi yana shasta tol zuni acoma ojibwa tonkawa wiyot seneca wichita dakota yuchi tunica alabama wappo nahuatl kawaiisu amuzgo bella coola tlapanec chehalis tsimshian caddo huasteco yucatec shuswap miwok jacaltec cherokee kiowa chinantec… **amérique du sud (66):** itonama bribri pirahã cayapa paez ocaina muinane caraïbe (galibi) kaingang apinaye amahuaca epena pedee tacana axluxlay abipon nambiquara arabela auca quechua jaqaru tehuelche wapishana caraïbe insulaire amuesha campa guajiro moxo guarani siriono guahibo tucuna barasano siona jivaro cofan mapudungu resigaro yagua cayuvava huari hixkaryana yucuna jebero iranxe cubeo tarascan japreria panare bororo andoke warao akawaio paya shiriana bakairi cuna maxakali qawasqar movima saliba guambiano camsa cacua trumai ache iate **eskimo-aléoute (3):** aléoute inuit yupik… **australien (25):** garawa yanyuwa waray murinhpatha maung tiwi burarra nunggubuyu alawa malakmalak bardi wik-munkan désert occidental arrernte gugu-yalanji ya diyari bandjalang kalkatungu yidiny dyirbal ngiyambaa mbabaram ngarinjin yolngu… **papou (39):** andamanese asmat kwoma sentani nimboran iwam selepet gadsup yagaria kewa chuave dani wantoat dadibi fasu suena dera kunimaipa yareba koiari taoripi nasioi rotokas nambakaengo angaatiha wahgi yawa usan baining yessan-mayo woisika alamblak amele kiwai waris vanimo savosavo makian ekari



Une carte UPSID (Maddieson, 1986) – Le Bambara

Des sons...

- Agoul
- Avar
- Coréen
- Danish
- Dédua
- Éwé
- Igbo
- Kekchi
- Kélé
- Liangshan Yi
- Mandarin
- Melpa
- Mid-Waghi
- Navajo
- Polonais
- Titan
- V'enen Taut
- Xoo

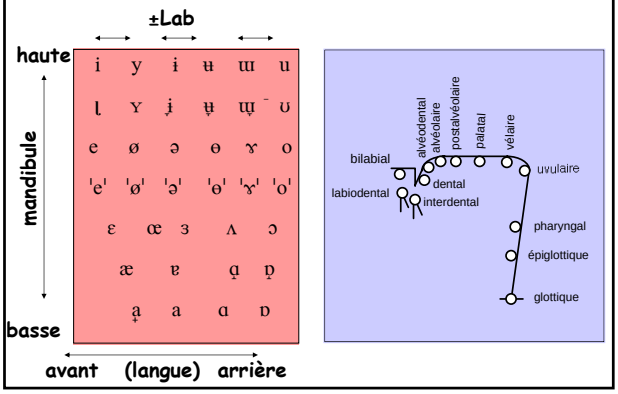
Les éléments sonores des systèmes

une grande diversité

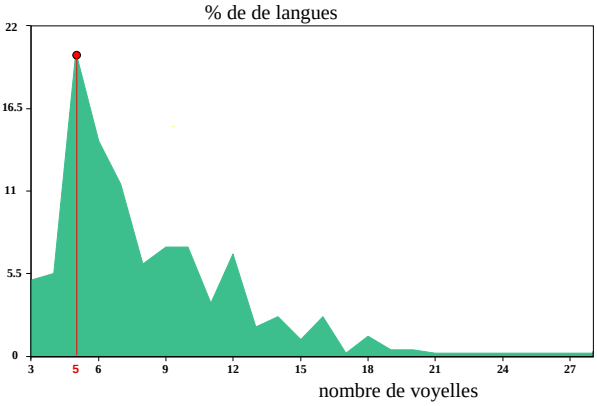
920 sons
dont
177 voyelles (V)
654 consonnes (C)

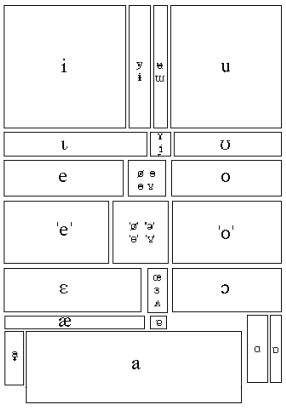
Des tendances générales
non arbitraires ?

Les articulations de base
des voyelles (37) et des consonnes (13 lieux)

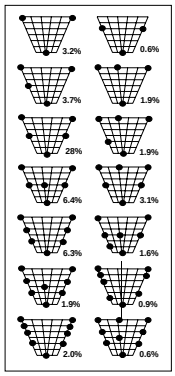


Les voyelles...

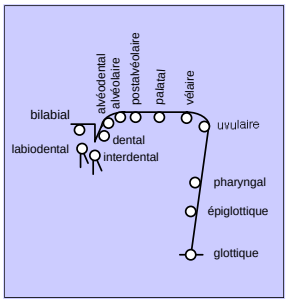




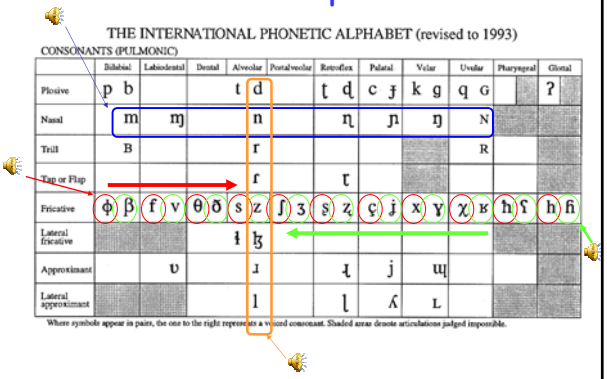
Les systèmes vedettes



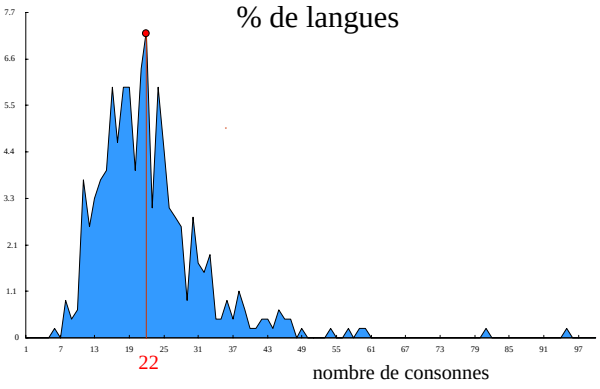
Les consonnes...



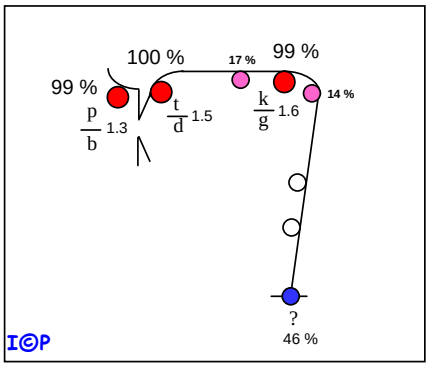
Transcription



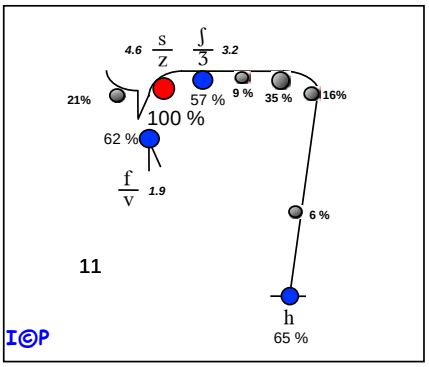
% de langues



Les consonnes occlusives dans les langues



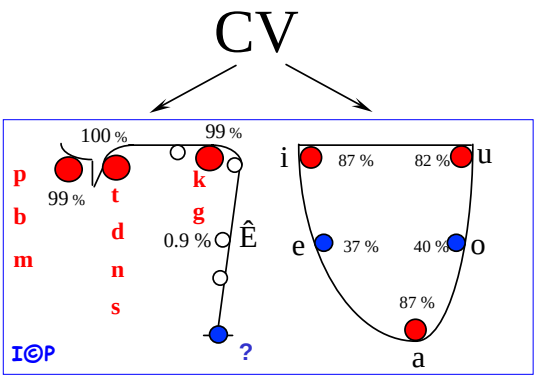
Les consonnes fricatives dans les langues

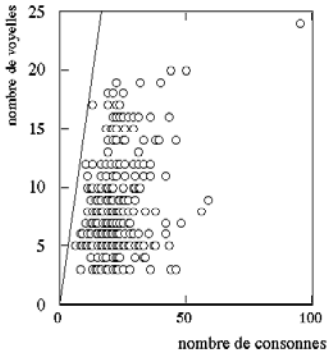


Le babillage et les consonnes les plus fréquentes dans les langues

Langues du monde	Babillage
t	b m
m	p
n k	d
j p	h n
w	t
s	k g
h d b	j w
l g	s

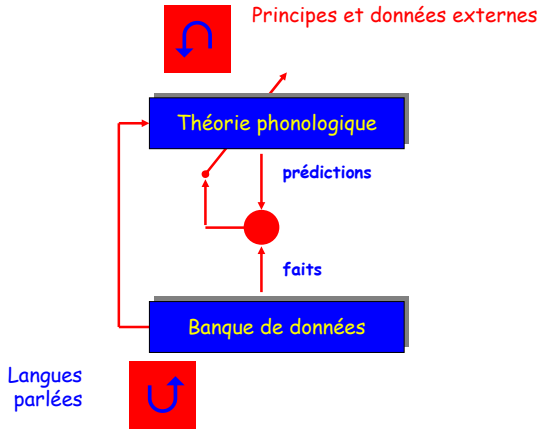
Les syllabes : Consonne Voyelle



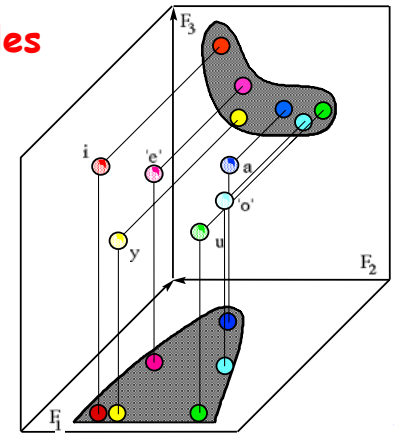


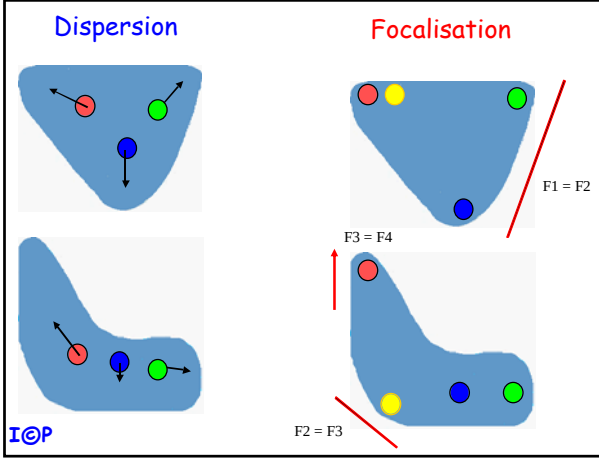
Les grandes tendances des systèmes sonores des langues

- **5 voyelles (V)** [i a u e o]
  et toujours [i a u]
- **22 consonnes (C)**
  - **7 plosives** [p b t d k g ʔ]
  - **4 fricatives** [f s ʃ h]
  - **3 nasales** [m n ŋ]
  - **3 approximantes** [l j w]
  - **2 affriquées** [ts tʃ]
  - **1 vibrante** [r]



5 Voyelles TDF





La théorie de la dispersion focalisation

TDF

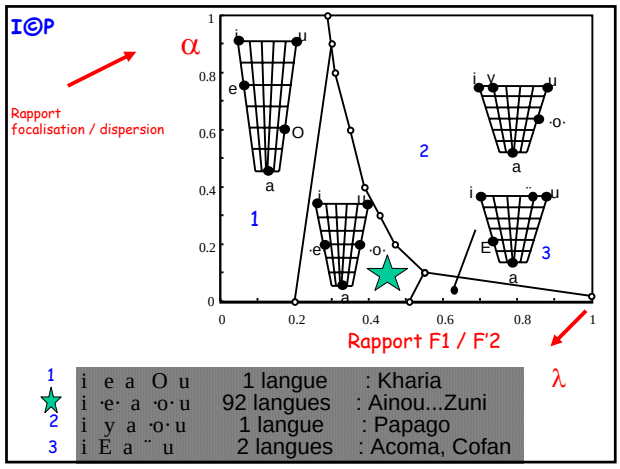
Minimiser $E_{DF} = E_{Dispersion} + E_{Focalisation}$

$E_{Dispersion}$ **dépend d'un paramètre** λ

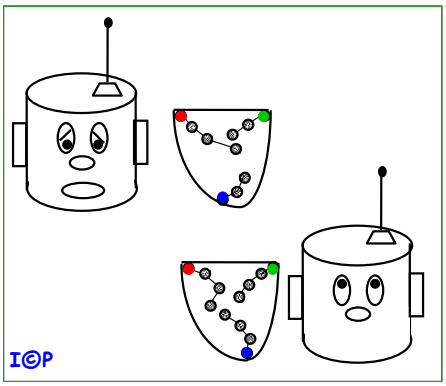
(poids de F1 vs. F'2)

$E_{Focalisation}$ **dépend d'un paramètre** α

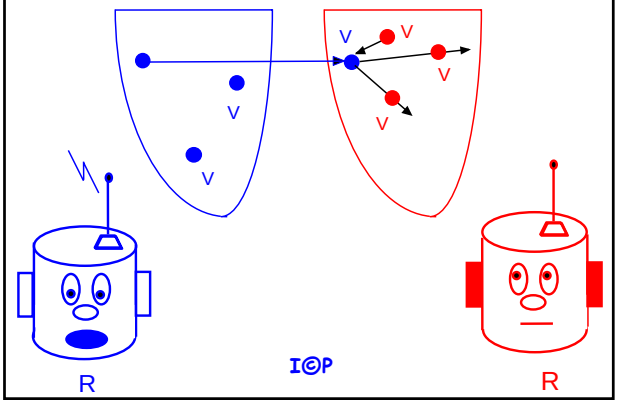
(poids de la focalisation vs. dispersion)



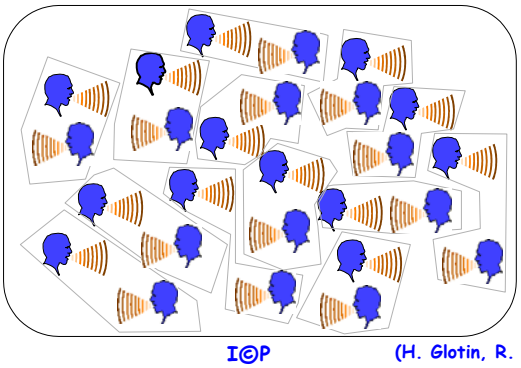
Échanges entre agents



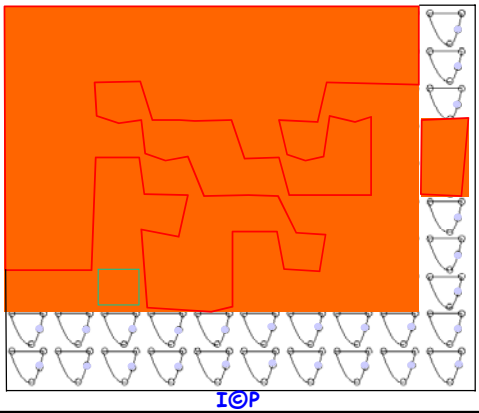
Émergence des systèmes par négociation



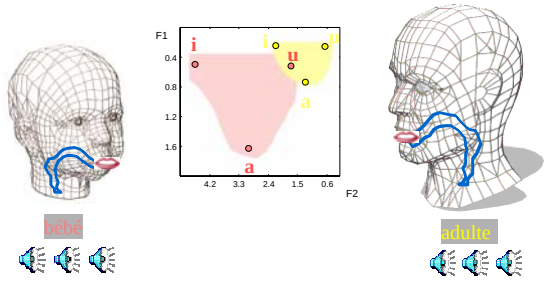
Une société d'agents



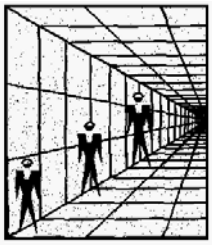
Systèmes à 4 voyelles



[i a u] d'un bébé et d'un adulte

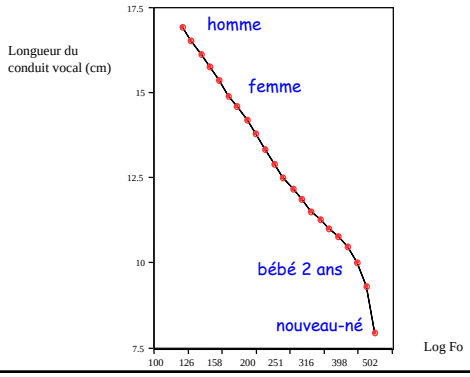


Un problème de normalisation



La distance permet de récupérer la taille

La hauteur de la voix (Fo) permet de récupérer la longueur du conduit vocal et donc de normaliser les formants



Les régimes phonatoires (1)

Voisement: Une consonne est dite voisée si les cordes vocales sont en vibration au moment du relâchement des articulateurs supra-glottiques. Elle est dite non-voisée si les cordes vocales ne vibrent pas, les aryténoïdes étant écartés au moment du relâchement. La notion de voisement caractérise donc une relation temporelle entre vibration des cordes vocales et relâchement de la constriction.
ex: opposition entre plosives voisée et non-voisée en Nepali
plosive bilabiale non-voisée « arrière » vs plosive bilabiale voisée « brûler »

Murmure: ("breathy voice") Les cordes vocales vibrent mais sans faire contact, les aryténoïdes étant écartés. Le débit d'air est plus grand que lors du voisement. Cela correspond à une voix d'hôtesse de l'air !
ex: opposition entre plosives non-voisée aspirée et voisée murmurée en Nepali
non-voisée aspirée « jeter » vs voisée murmurée « front »

Les régimes phonatoires (2)

Aspiration: Une consonne est dite "aspirée" lorsqu'il se produit une période non-voisée pendant et immédiatement après un relâchement articulatoire dans les cavités supra-glottiques.
ex: oppositions entre plosives non-aspirées et aspirées en Quechua

	palato-alvéolaire	vélaire	uvulaire
non-aspiré	« pont »	« bouger »	« langue »
aspiré	« grosse fourmi »	« siffler »	« châle »

Laryngalisation: ("creaky voice" ou "voix craquée") Les aryténoïdes sont étroitement rapprochés, mais seule une petite partie des cordes vocales est en vibration. La tension des cordes vocales est forte. Cela donne une voix "chevrotante". Parfois il en résulte une véritable phase d'occlusion glottale.
ex: opposition entre voyelle non-laryngalisée et laryngalisée, ton bas montant, en Mpi
normal « être pourri » vs laryngalisé « être séché »

Chuchotement: ("whisper") Les cordes vocales sont rapprochées, voire même en contact, à l'exception de la zone entre les aryténoïdes. Il se produit un écoulement turbulent de l'air entre les cordes vocales.

Les régimes aéro-dynamiques

La notion de régime aérodynamique décrit la façon dont est généré le flux d'air que l'on utilise pour produire les sons de parole. On le définit à partir de l'organe qui permet à l'air de bouger, et la direction du flux d'air.
Ce flux d'air peut être généré:
(1) par les poumons: le flux est soit égressif soit ingressif
(2) par le larynx: le flux est soit éjectif (élévation) soit implosif (abaissement)
(3) par le dos de la langue pour produire un click

	égressif	vs	ingressif
Pulmonique (Igbo)			
Glottique (K'ekchi)			
Vélique (!Xu)			

Les types aéro-acoustiques (1)

Les consonnes occlusives

Les organes mobiles (langue, lèvres, mandibule inférieure...) vont s'accoler avec un articulateur fixe (palais dur, incisives...) pour produire un barrage complet au passage de l'air. Le relâchement rapide de cette occlusion provoquera une expulsion brutale de l'air et ainsi un bruit d'explosion: les occlusives sont également appelées des plosives. Si on superpose à ces mouvements d'ouverture/fermeture une vibration des cordes vocales, la consonne est voisée ou sonore, sinon non-voisée ou sourde.

Occlusives orales

Les occlusives orales sont produites avec le velum en position relevée, c'est-à-dire que l'air ne passe pas dans les fosses nasales. Ex: Sindhi

	labial	alvéolaire	rétroflexe	palatal	vélaire
impl. voisée	«champ»		«festival»	«illétré»	«manche»
plos. voisée	«forêt»	«porte»	«tu cours»	«illétré»	«qualité»
plos. non-voisée	«feuille»	«fond»	«tonne»	«détruire»	«manger»
plos. sourde asp.	«collerette de cobra»	«nom de district»	«voyou»	«couronne»	«tu lèves»
plos. voisée mur.	«fumier»	«coffre»	«taureau»	«poignée»	«excès »

Les types aéro-acoustiques (2)

Occlusives nasales

La nasalité est réalisée par l'abaissement du voile du palais (= velum), permettant le passage de l'air dans les fosses nasales. Les consonnes nasales sont toujours occlusives et voisées. Ex: Wangurri

bilabiale dentale alvéolaire rétroflexe palatale vélaire

«mère » « là-bas » «assez » « requin » « espèce d'arbre » « voir »

Consonnes constrictives

Elles ne sont jamais nasales. Elles sont produites par un resserrement du canal que l'on appelle une constriction, qui gêne l'écoulement de l'air. Elles peuvent être sourdes ou sonores. Ex: locuteur anglais « polyglotte »

Bilab. Labio-dent. Interd. Alvéol. Rétr. Post-alv. Pal. Vél. Uvul. Phar. Glot. Labio-vél.

Ex: Ewe bilabiale labio-dentale

sourde «il a poli» «il a acheté»

sonore «la langue Ewe» «deux»

Les types aéro-acoustiques (3)

Les latérales

La partie antérieure de la langue s'approche de la voûte palatale pour réaliser une occlusion mais les lames de la langue restent abaissées. L'air s'échappe donc par les côtés. On distingue (1) les latérales fricatives : si l'aperture reste faible, on entend un bruit de friction. La plupart des latérales dans les langues du monde sont réalisées avec une occlusion dans la région dentale/alvéolaire. (2) Les approximantes latérales (liquides) : leur aperture est plus grande que pour les précédentes. Ex: SiSwati approximante latérale sonore «fil » vs fricative latérale sonore «gifler très fort »

Les vibrantes

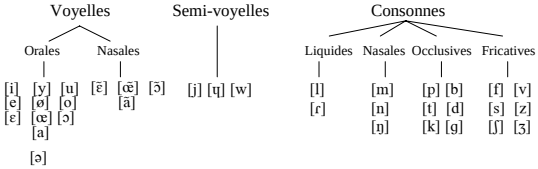
On distingue (1) les vibrantes (trilles) produites avec un articulateur actif (apex ou luette) qui, dans une position de repos, vient frapper un autre articulateur rapidement et de manière répétitive et produisant ainsi des vibrations. (2) Les battues produites avec un articulateur qui est retiré de sa position de repos et qui, en revenant à sa position de repos, frappe un autre articulateur. L'apex touche une seule fois les alvéoles au lieu de plusieurs fois. Ex: Kele vibrante bilabiale vs vibrante alvéolaire (avec attaques nasales)

« vagin » « os »

« visage » « chanson »

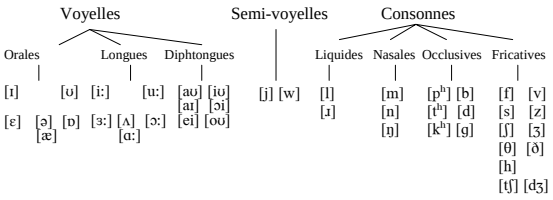
« fruit » « jambe »

Et le français dans tout ça



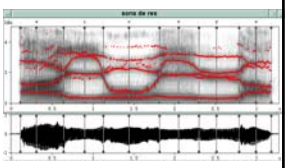
- 14 voyelles
- 20 consonnes et semi-voyelles

Et l'anglais



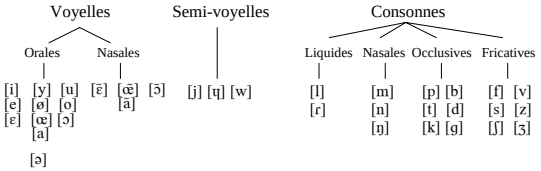
- 10 voyelles + 6 diphotongues
- 24 consonnes et semi-voyelles

Lecture de sonas



G. Bailly – ICP
(crédit D. Archambault – U. Montréal)

Le système phonologique du Français

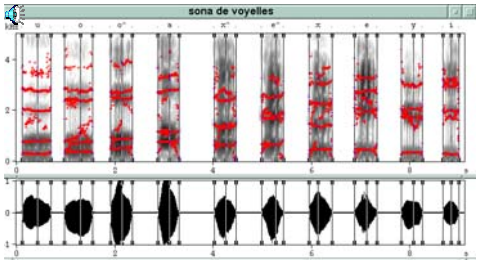


- 14 voyelles
- 20 consonnes et semi-voyelles

Les voyelles...

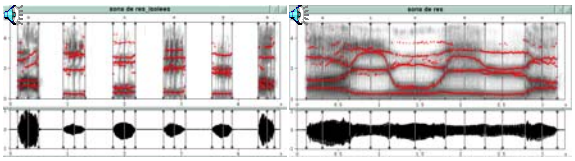
- 14 voyelles
  - 10 orales [i] [e] [ɛ] [a] [œ] [ɔ] [ø] [y] [o] [u]
  - 4 nasales [ã] [ɛ̃] [œ̃] [ɔ̃]
  - schwa [ə]
- Mode de production
  - Généralement voisé
  - Résonateurs:
    - bucco-pharyngal & nasal
    - Appui de la langue au palais/antériorité
    - [+rond]: protrusion
    - [+nasal]: déplacement de la langue (adaptation d'impédance?)
  - Acoustique:
    - F1 inversement proportionnel à la hauteur de la langue/ mâchoire
    - F2 proportionnel à l'antériorité de la langue
    - Labialisation entraîne un abaissement fréquentiel des formants

Les voyelles...

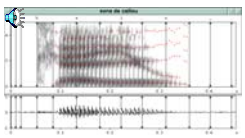


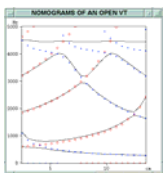
- Énergie fonction de l'ouverture

Les voyelles enchaînées

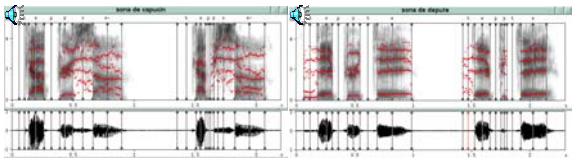


- Suivi d'affiliation…(voir nomogrammes de Fant)
- Description des trajectoires





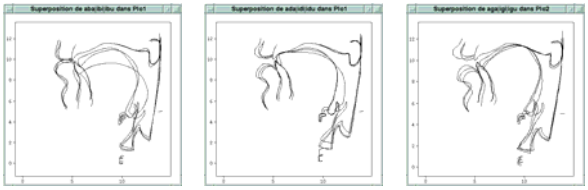
Les voyelles en contexte



- Assourdissement des voyelles hautes
  - capucin /kapysɛ̃/, disputer /dispyte/
- Harmonisation vocalique
  - aider /ede/, déneiger /deneʒe/,...

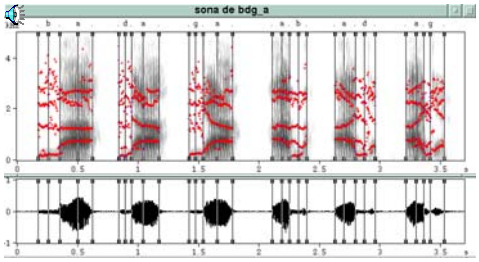
Les occlusives...

- 6 occlusives
  - [p] [t] [k] [b] [d] [g]
- Mode de production
  - sourde [p] [t] [k] et voisée [b] [d] [g]
  - Résonateurs:
    - constriction totale de la cavité buccale/voile du palais fermé
    - bilabiales [p] [b], apico-dentales [t] [d], dorso-vélaires [k] [g]



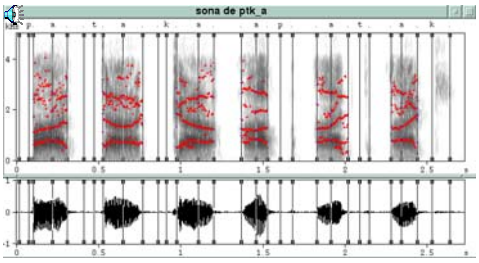
Les occlusives...

- Acoustique
  - 5 états pour les plosives: implosion, occlusion ou tenue (silence ou barre de voisement), explosion: relâchement brusque de la constriction, bruit de constriction, bruit d'écoulement glottique (temps d'établissement correct du voisement)



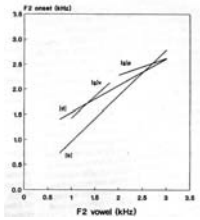
Les occlusives...

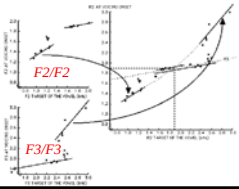
- Acoustique
  - Explosion: burst [p] [b] BF faible, [t] [d] HF, [k] [g] compact vers 2kHz.
  - Transitions: théorie du locus
  - VOT: [k] > [t] > [p]

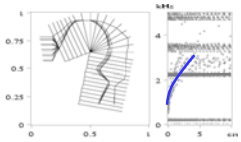


La théorie du locus

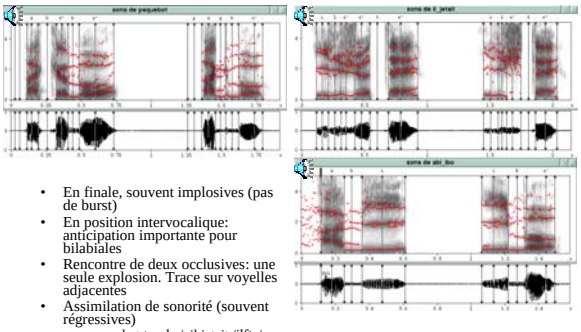
- Invariance relative des formants à l'onset vocalique
  - Fonction de la voyelle support
  - Pour chaque formant (F1 cible à 0, ici F2)
- … Résonances
  - [g] et le changement d'affiliation





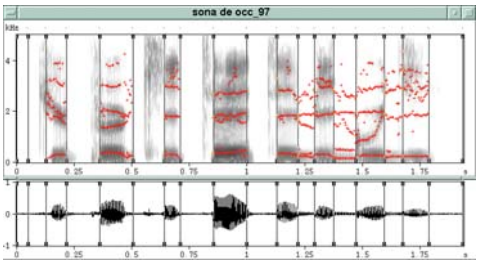


Les occlusives en contexte

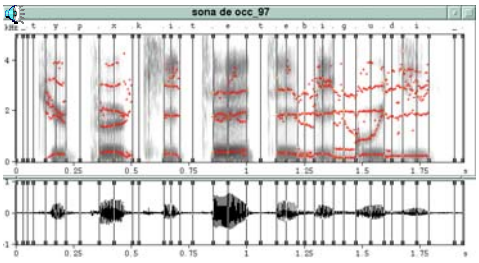


- En finale, souvent implosives (pas de burst)
- En position intervocalique: anticipation importante pour bilabiales
- Rencontre de deux occlusives: une seule explosion. Trace sur voyelles adjacentes
- Assimilation de sonorité (souvent régressives)
  - paquebot /pagbɔ/, il jetait /ilʃte/

Exercice

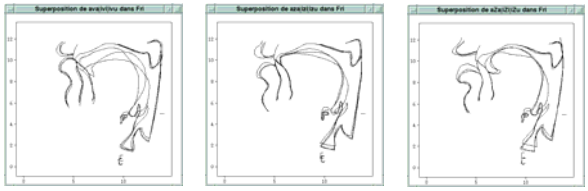


Corrigé



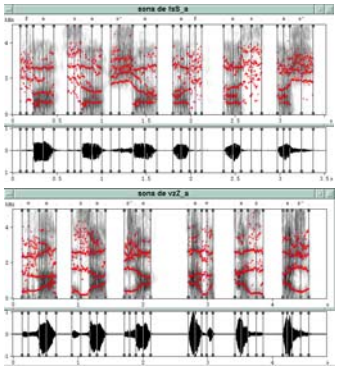
Les fricatives...

- 6 fricatives
  - [f] [s] [ʃ] [v] [z] [ʒ]
- Mode de production
  - sourde [f] [s] [ʃ] et voisée [v] [z] [ʒ]
  - Résonateurs:
    - constriction partielle de la cavité buccale/voile du palais fermé
    - labiodentales [f] [v], apico-alvéolaires [s] [z], dorso-palatales et labialisation [ʃ] [ʒ]

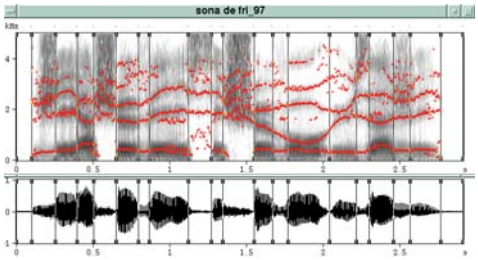


Les fricatives...

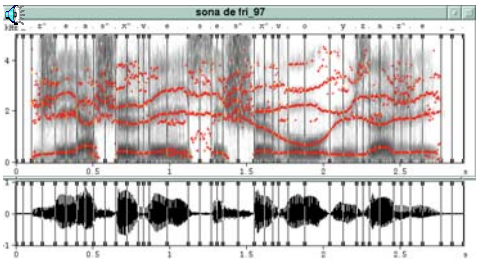
- Acoustique
  - Locus mais transitions souvent très rapides
  - Distribution fréquentielle du bruit de friction
    - [s] [z] HF
    - [f] [v] diffus sur tout le spectre
    - [ʃ] [ʒ] au dessus de 2Kz
- En contexte
  - Même cas de figures que occlusives
    - passe-bande /paz bãd/



Exercice



Corrigé



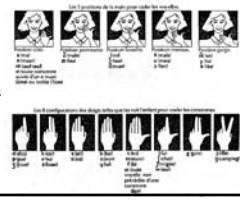
Parole audiovisuelle

G. Bailly - ICP

La parole est multimodale (1)

- On peut voir le conduit vocal…et le toucher *(tadoma)*
  - (40-60% phonemes, 10-20% mots…jusqu'à 60%)

- On peut aisément encoder la parole par le toucher (*prothèses tactiles*)
  ou ajouter de l'information visible sur les gestes invisibles (*Langage Parlé Complété*)



La parole est multimodale (2)

- Lis sur mes lèvres…dans le bruit
- Mais aussi sans…suivre une conversation (*close-shadowing*)!

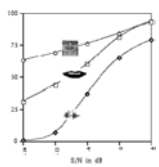
La parole est multimodale (3)

- On ne peut pas s'empêcher d'utiliser la multimodalité
  - Effet Mc Gurk …

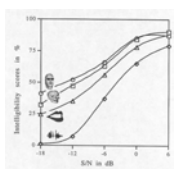
- Ontogenèse
  - Aveugles ont des difficultés à apprendre les contrastes bien visuels mais faiblement audibles [m] vs [n], [f] vs [ð]
  - Aveugles ont des dynamiques labiales réduites pour le même résultat acoustique
  - Bilabiales majoritaires dans les premières phases du développement, plus pour les sourds, moins pour les enfants aveugles
- Phylogenèse
  - [m] vs [n]: 94% des langues du mondes exploitent ce faible contraste acoustique

La parole est multimodale (4)

- Renforcement
  - A+V>A & >V
    - perception des gestes d'anticipation
    - réduction du bruit (perception & codage)

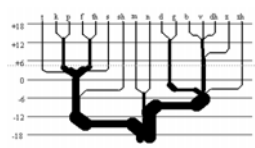


- analyse/synthèse de scènes audiovisuelles
  - téléconférence
- Handicap
- Convivial...

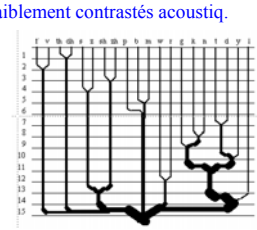


Complémentarité audiovisuelle

- Auditif: mode d'articulation, lieu d'articulation linguale
- Vision: lieux d'articulation faiblement contrastés acoustiq.



Confusions auditives
(Miller & Nicely, 1955)



Confusions visuelles
(Montgommery & Walden, 1971)

Langage parlé complété (1)

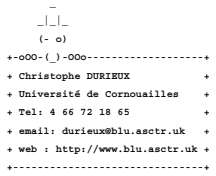
Side	Mouth	chin	Eye	throat
a (m**a**) o (m**aux**) œ (t**euf**) (*)	i (m**i**) ɔ̃ (**on**) ɑ̃ (r**ang**)	ɛ (m**ais**) u (m**ou**) ɔ (f**o**rt)	ɛ̃ (m**ain**) ø (f**eu**)	œ̃ (**un**) y (t**u**) e (f**ée**)

Handshape 1	Handshape 2	Handshape 3	Handshape 4
p (**p**ar) d (**d**os) ʒ (**j**oue)	k (**c**ar) v (**v**a) z (**z**ut)	s (**s**el) R (**r**at)	b (**b**ar) n (**n**on) ɥ (l**ui**)

Handshape 5	Handshape 6	Handshape 7	Handshape 8
t (**t**oi) m (a**m**i) f (**f**a) (*)	l (**l**a) ʃ (**ch**at) ɲ (vi**gn**e) w (**oui**)	g (**g**are)	j (fi**ll**e) ŋ (camp**ing**)

Langage parlé complété (2)

Transcription orthographique -phonétique



G. Bailly – ICP
(crédits F. Yvon - ENST & A. Black - CMU)

Plan

- Noter l'écrit / noter l'oral
- La transcription des formes isolées
- La transcription des séquences de mots
- Phonétisation, variantes et prosodie
- La transcription de textes / documents

Systèmes de notation de l'oral

- Trois types idéaux:
  - Les systèmes *idéographiques* ou *logographiques*: chaque symbole graphique représente un signifié différent (ex: le chinois)
  - Les systèmes *sémiographiques* : chaque symbole graphique représente un signe (i.e. un signifiant et un signifié)
  - Les systèmes *phonographiques* : chaque symbole graphique représente un fragment de la forme sonore (signifiant)
    - Notation syllabique (ex japonais)
    - Notation consonantique (ex. langues sémitiques)
    - Notation alphabétique (ex langues romanes, etc)

Graphème, phonogramme…

- Graphème (Mounin, 1974)
  - Unité graphique minimale entrant dans la composition de tout système d'écriture, le graphème peut représenter un concept (écriture idéographique) ou un élément de la réalisation phonique (écriture phonographique: syllabique, consonantique, alphabétique). Dans l'écriture alphabétique, il est couramment appelé lettre.
  - Signe substitutif au phonème dans le système graphique de représentation d'une langue: le groupement des deux lettres « q » et « u » constitue en français le graphème « qu » représentant le son /k/ (phonogramme chez N. Catach)
- Le graphème au sens B est une abstraction (comme le phonème), dont la compréhension est nécessaire pour passer du graphème (au sens A) au phonème.

Phonème, phone...: écrire les sons

Le **phonème** est également une abstraction, qui ne se comprend qu'au travers la spécification d'un ensemble *cohérent* d'oppositions systématiques (traits distinctifs). Selon une certaine conception, le phonème représente un ensemble de propriétés différentielles (les traits distinctifs), qui caractérisent de façon *conjointe* une unité différentielle minimale.

Un même phonème est susceptible de recevoir de multiples réalisations acoustiques, dépendant à la fois de son environnement phonématique (expression de contraintes de production), de sa position dans la chaîne, mais aussi des propriétés physiques de l'appareil vocal, du milieu social, de l'origine géographique du locuteur. On appelle **phone** une classe de réalisations particulières d'un même phonème (**allophones**).

Le **phonème** de transcription graphème-graphème est intermédiaire entre phonème ``linguistique'' et le phone, plus ou moins proche de l'un ou de l'autre suivant le procédé de synthèse. D'un point de vue très pragmatique, on peut le voir comme l'expression d'une commande permettant de sélection la bonne forme d'onde sonore à produire.

Les différents types de variabilité

On catégorise généralement les sources de variabilité

- variabilité inter locuteur (physiologique, socio-géographique): /œ̃/ vs /ɛ̃/, /a/ vs /ɑ/, comportement de /ə/, des voyelles à deux timbres, consonnes finales, gémination, liaisons...
- variabilité intra-locuteur (émotion, débit, intention, contexte d'élocution au sens large): élision de /ə/, assimilations, synérèse-diérèse, liaisons
- variabilité contextuelle (conditionnée phoném(t)iquement, syntaxiquement, prosodiquement, sémantiquement): eg.\ élisions de /ə/, liaisons; *six*, *plus*, y (*il y a* vs *vas-y vite*!), séquences de chiffres, noms propres.
- variabilité « libre »: réalisations allophoniques, /ananas/ vs. /anana/, emprunts.

Gérer des variantes, élaborer une norme

Trois stratégies de prise en compte de la variabilité:

1. production directe d'une variante unique
2. production de plusieurs variantes, parmi lesquelles une est sélectionnée ultérieurement (par les autres modules)
3. production de solutions « abstraites » (formes profondes? formes de base?), génération « au choix » des variantes.

La question de la variabilité (au niveau phonétique) est toutefois le plus souvent éludée en synthèse (utilisation de références au niveau du mot), au risque d'une perte de naturel (méthode 1)).

Définition d'un alphabet phonémique, choix d'un niveau d'abstraction pour les transcriptions, choix d'une norme (dépendant de l'application), répartition des tâches (en particulier de la réalisations des variantes combinatoires allophoniques) entre le module linguistique et le module acoustique.

Dictionnaires vs. règles

Le pour et le contre:

d + mémoire rapide et pas chère; représentations et algorithmes permettant stockage efficace et un accès rapide (machines à états finis); accès au lexique indispensable pour les pré-traitements linguistiques, et déjà réalisé dans les applications de synthèse à partir de concepts

d + les dictionnaires sont déjà là ! Bon marché !

r + traitement de la création lexicale, des emprunts, des items extra-lexicaux (sigles, noms propres, etc).

r - coût de développement (l'expertise) et de maintenance des « grosses » bases de règles

Les systèmes à base de règles demeurent majoritaires, mais la distinction s'estompe.

La règle de récriture contextuelle

De manière abstraite:

G → P / **Gg** + **Gd**

Quelques exemples élémentaires:

- **ai** → ɛ (maison, saine...)
- **g** → ʒ / + **e** (mangeais,vengeance...)
- **a** → ɛ / + **i** (maison, saine...)
- **i** → ɛ / **a** + (maison, saine...)

Il y a toujours plus d'une façon de faire les choses! Donc penser à la productivité / pertinence linguistique; « maintenabilité » (conflits/interactions entre règles); souplesse de gestion; efficacité du traitement.

Contextes, méta-contextes et classes

- **an** → ɑ̃ / + **[bcdfgjklmpqrstvwxz#]** (manger,danser...)
- **an** → ɑ̃ / + {C,**#** (manger,danser...)
  où C = [**bcdfgjklmpqrstvwxz**] (mais que fait **p** dans cette liste ?)
- **ill** → ij / C + (bille,fillette…)
  où C = [**bcdf(gu)jklmp(qu)rstvwxyz**] (les consonnes *phonétiques*)
- **e** → e / + {C,CD V + **o** (banderillo, azulejo...)
  où C représente les consonnes, CD les consonnes doubles, V les voyelles
- **ea** → i / + Leng (stream, dealer...)
  où Leng liste les lexies d'origine anglaise
- **gin** → dʒin / # + # (gin)

Gérer une collection de règles

Principe de base (motivé par des raisons d'efficacité): ordonnancement strict du plus spécifique au plus général, la première règle applicable s'applique, **il n'y a pas de retour arrière**.

- **aill** → aj / + V (travailleur)
- **ain** → ɛ / **+** {C,**#** (vain)
- **ai** → ə / **f** + **s{ae** (faisan, faiseur)
- **ai** → ɛ (aigle, faisceau)
- **a** → a (pas)

Formellement, la condition d'ordonnancement strict rend le processus de récriture, au delà de son aspect **contextuel**, équivalent à une transduction **régulière** (Kaplan & Kay 94), implantable dans un transducteur d'état fini, et donc virtuellement réalisable en temps linéaire.

Récriture complète d'une graphie

Forme graphique	Règles			Sortie
chasseurs	**ch**	→ ʃ		ʃ
cha**ss**eurs	**a**	→ a		ʃa
cha**ss**eurs	**ss**	→ s		ʃas
cha**ss**e**u**rs	**eu**	→ œ	/ + [**rlpbvfiy**]	ʃasœ
cha**ss**e**u**r**s**	**r**	→ ʁ		ʃasœʁ
cha**ss**e**u**r**s**	**s**	→ +z	/ + # [**+Plu**]	ʃasœʁ+z

Compiler un ensemble de règles

Il existe différentes manières d'organiser ou de « compiler » un tel ensemble de règles pour réaliser la transduction Graphème - Phonème:

- parcours linéaire
- Indexation
- organisation arborescente
- pré-expansion des contextes vs. appariement dynamique
- compilation de la machine à états finis correspondante
- système mono vs. multicouche (maintenabilité)

Organisation par couche (détail)

Succession de récritures en trois étapes principales:

1. règles graphème - graphème (réforme généralisée de l'orthographe, gommage des accidents historiques, exceptions, assimilation orthographique des emprunts). Nombreuses, et évolutives. Exemples: Fresnes→Frêne, choriambe→koriambe.
2. règles graphème - phonème (phonogrammes + variantes contextuelles). Nombre limité de règles, extrêmement stables.
3. règles phonème - phonème (modélisation des phénomènes morpho-phonologiques intra (et inter-) lexicaux)

Organisation linguistiquement fondée, exprimant les généralisations au niveau approprié (nasalisation, chute des consonnes finales), aisément « maintenable » (relative indépendance des couches), hautement paramétrable (différents styles, débits...)

Dictionnaires

- Très grosses bases de données
- Avantages
  - Disponibilité des dictionnaires informatisés (Robert, etc…) avec transcriptions validées
- Désavantages
  - Redondance
  - Noms propres: annuaire UK ~ 5000000 entrées
  - Mots nouveaux, termes techniques (ex. liposuccion) emprunts (ex. email)
- Solutions
  - Prétraitement morphologique avant parcours
  - Recours aux règles si pas trouvé
  - Identification du pays d'origine
  - Prononciation par analogie

Transcription par analogie

- Problème de classification
- Alignement: **chanson** → ʃ _ ã _ s ɔ̃ _
  - Alignements silencieux **ch** → ʃ _
  - Alignements diphones **tablier** → t a b l ij e _

  boticelli→ b ɔ t i tʃ ɛ l _ i, **exact**→ ɛ gz a k t, **voyage**→ v wa j a ʒ _
- Apprentissage automatique de la correspondance graphème-phonème en contexte
  - Alignement, recombinaison, dérivation de segments: [Yvon, 96]
  - Approche connexionniste: Nettalk [Sejnowski & Rosenberg, 1987]
  - Arbres de décision: ID3 [Quinlan, 1993]
  - Approche probabiliste: Chaînes de Markov, multigrammes…
- Problèmes
  - Codage des entrées et des sorties
  - Sens de l'analyse

Arbres de décision (1)

- Breiman, Friedman, Olshen, Stone. 1984. Classification and Regression Trees. Chapman & Hall, New York.
- Description/Usage:
  - Arbre binaire de décisions, nœuds terminaux déterminent la prédiction ("20 questions")
  - Si les variables dépendantes sont catégorielles, "arbre de classification",
  - Si continues, "arbre de régression"
- Méthode très répandue, rapide et disponible (ex: ID3, C4.5: http://www.rulequest.com/Personal/)

Arbres de décision (2)

- La construction à la main n'est possible que pour des domaines réduits et des variables en petit nombre
- Recherche exhaustive dans l'ensemble des arbres possibles impossible
  - exponentiel en fonction du nb. d'attributs : **d** et du nb. moyen de valeurs par attributs : **a**

D	A	Arbres possibles
4	2	30
6	2	72385
8	2	18.10^{18}

$$\sum_{i=0}^{d-1} (d-i)^{a^i}$$

- Beaucoup d'algorithmes pour induction d'AD
- Principe
  - règles de séparation: Doit-on créer deux branches
  - règles d'arrêt : quand doit-on déclarer un nœud terminal
  - assignation: quelle classe/valeur assigner à un nœud terminal

Règles de séparation

- Séparations candidates considérées:
  - Coupures binaires: pour variables continues (-inf < x < inf) considérer:
    - X <= k **vs.** x > k ∀K
  - Partitions binaires: pour catégorie x ∈ {1,2,…} = X considérer :
  - x ∈ A **vs.** x ∈ X-A, ∀A ∈ X
- Choisir le meilleur partage
  - Méthode 1: Choisir k (continu) or A (catégoriel) qui minimise l'erreur de classification (régression) après le partage
  - Méthode 2 (pour classification): Choisir k ou A qui minimise l'entropie après le partage

Règles d'arrêt

- Elagage
  1. Croissance maximale de l'arbre
  2. Former tous les sous-arbres, T0…Tn depuis l'arbre complet jusqu'à la racine
  3. Estimer un taux d'erreur "honnête" pour chaque sous-arbre
  4. Choisir la taille de l'arbre avec le taux d'erreur "honnête" minimum
1. Pour estimer le "taux d'erreur "honnête" , tester sur données différentes des données d'apprentissage (i.e. estimer sur 9/10 des données, tester sur 1/10, répéter 1à fois et moyenner (validation croisée).

Evaluation

- Entrées lexicales vs. textes
- Apprentissage/test
- Comptage des erreurs
  - Alignement (outil Sclite: http://ww.nist.gov/speech/tools/).
  - Poids
- Transcripteurs « français » sur liste de noms propres
  - Boula de Mareuil et al (Interspeech 2005)

%Error	Lab1	Lab2	Lab3	Lab4
First names	8.4	17.4	10.5	23.8
Surnames	12.7	21.7	13.6	25.0
Total	12.9	17.1	17.2	19.3
Phonemes OK	95.9	96.1	94.3	94.4

Les homographes hétérophones (1)

Problème 1: la forme graphique ne suffit pas toujours à calculer la prononciation, à cause des multiples « collisions » graphiques, le plus souvent entre une forme verbale et une autre forme. Quelques exemples bien connus: *est*-V vs *est*-N, *couvent*-V vs *couvent*-N, *bus*-V vs *bus*-N, *violent*-V vs *violent*-A, *portions*-V vs *portions*-N

Solution:

1. désambiguiser la prononciation par le biais d'un pré-traitement (eg. l'assignation plus ou moins fin des parties du discours par exploration du contexte, pas toujours suffisant).
2. introduire des *méta-contextes* (ou séparer les jeux de règles) qui exploitent cette information~:

si MC: **G** → P / **Gg** + **Gd** où MC prend la forme cat == X.

Les homographes hétérophones (2)

Problème 2: Hétérophonie dans la même classe lexicale

Quelques exemples bien connus: *fils*-N vs *fils*-N

Idem en anglais: *saw*-V vs *saw*-V, *read*-V vs *read*-V

Solutions:

(a) Identifier

(b) Trouver occurrences dans des textes (thésaurus/ontogenèses)

(c) Caractériser le contexte pour trouver traits contextuels distinctifs: classes des mots environnants, capitalisation,

(d) Apprentissage: arbres de décision, etc

Exemples:

Henry V: Part I act II Scene XI: Mr X is I believe, V Lenin and not Charles I. The madness of King George III. William Gates III.

De l'utilité de l'analyse morphologique

Problèmes: les frontières internes (de morphèmes dérivationnels) ne sont pas marquées graphiquement, mais peuvent influer sur la prononciation. Quelques exemples: *a-social*, *anti-américain*, *co-occurence*, *parasol*, *Mont-rouge*, *bons-hommes*. Nécessité d'identifier les morphèmes flexionnels (liaison), pour les variantes contextuelles, en particulier les liaisons (*souris*-sing vs *souris*-plur).

Solutions:

(a) réaliser une analyse morphologique complète qui réintroduit les frontières internes

(b) utiliser ces frontières dans les règles: **s** → z / V + V ne se déclenche pas sur *a-social*

(b') gèrer le problème (somme toute assez localisé en Français) directement dans les règles: **s** → s / #{**a,anti,pro** +V

Analyse morphologique

Problèmes: décomposer les frontières internes

Solutions:

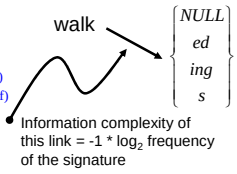
(a) Approche formelle: lexiques de bases, affixes et désinences et grammaires (automates d'états finis, grammaires à validation-saturations, etc)

(b) Apprentissage automatique

- Décompositions possibles

  Français: *.ité.s (nom) ou *.ement (adverbe)

  Anglais: *.ed.ing.s (verb) ou *.er.ly (adjectif)
- Fréquence de la signature
- Grammaire qui force à ré-utiliser bases et signatures
- Minimisation de la taille de la grammaire



La phonétisation des noms propres (1)

Les noms propres fournissent une proportion non-négligeable des inattendus rencontrés dans les textes. De leur bonne prononciation dépend également l'acceptabilité d'un certain nombre d'applications (annuaire inverse, synthèse de mel...).

Ces items posent des problèmes difficiles aux systèmes de conversion graphème-phonème:

- noms propres d'origine française: énorme variation orthographique (pas de réforme de l'orthographe), d'où rémanence de graphies/phonogrammes « archaïques » dans: *Fresnes*, *Aulnay*, *Gautherault*, *Borzeix*, *Goudaillier*...; hétérogénéité des langues sources régionales. Multiplication des ambiguïtés (au niveau des graphèmes, mais aussi au niveau des mots), qu'il faut prendre en compte au prix d'une extension importante des règles de transcription.

La phonétisation des noms propres (2)

- noms propres d'origine étrangère: représentent une partie non-négligeable des noms propres (actualité internationale, immigration). Les règles de transcription des noms propres qu'utilisent les locuteurs français changent selon l'origine présumée du nom propre, ce qui pose le double problème de l'identification de l'origine et de la construction de jeux de règles pour chaque langue (en particulier pour la prononciation des voyelles).
- d'une manière générale, manque de référence (dictionnaire) et de description: problème des variantes de prononciations et de la sélection de la meilleure variante (qui peut varier selon l'identité du porteur du nom, mais aussi suivant d'un utilisateur à l'autre: prononciations locales...)

Solutions: techniques à base de lexiques, méthodes automatiques d'identification de la langue.

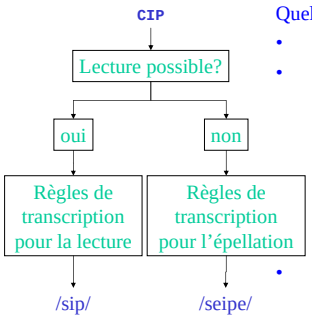
La phonétisation des sigles

Les sigles représentent une proportion non-négligeable des formes imprévisibles dans les corpus de textes, spécialement les corpus journalistiques.

En supposant en première approximation le problème de l'identification des sigles (sur la base de leurs caractéristiques typographiques) résolu, leur phonétisation automatique pose trois types de problèmes:

- certains sigles sont lus (eg. *cnet*, *aupelf*, *uref*, *enserg*); d'autres sont épelés (*dea*, *enst*, *taln*); d'autres enfin acceptent les deux modes (*onu*, *cip*)
- les règles de transcription des sigles lus sont différentes de celles qui s'appliquent pour le vocabulaire commun (*enset*, *maif*)
- l'épellation aussi peut être non standard (*ssii*, *ieee*...)

Sigles: l'approche classique



Quelques règles de sélection:

- Les sigles de deux lettres sont lus
- Les sigles de trois lettres sont épelés, sauf les sigles en VCV et CVC. Parmi ces derniers, ceux qui comportent un **h** ou dont la deuxième ou troisième lettre est un **e** ou dont la seconde lettre est dans [**bcdgpt**] sont épelés (eg. **epo**, **uta**)
- les sigles de cinq lettres qui contiennent une voyelle sont lus;

Les « vraies » contraintes

Contraintes sur la structure syllabique: la forme oralisée doit être phonotactiquement satisfaisante:

- analysable en une suite de syllabes ``légales'' (syllabicité)
- sans séquence /VV/ (prohibition de l'hiatus dans les formes lues)

- Contraintes de gabarit prosodique: la forme oralisée doit constituer (minimalement) un mot minimal du français:
  - elle contient au moins 2 mores
  - elle contient au moins une syllabe /CV/
  - elle contient moins de quatre syllabes
- Contraintes de fidélité~: la forme oralisée doit permettre de reconstituer aussi précisément que possible la forme lue

Contradictions et violations

Quelques paradoxes:

La sélection porte sur la forme graphique, les « vraies » contraintes sur la forme phonétique: il faut connaître le résultat du calcul pour l'effectuer correctement

La forme épellation est évaluée, mais constitue également l'oralisation *par défaut*;

Les oralisations exceptionnelles sont ignorées

Quelques exceptions:

- **drlav**, **crlao** sont lus, mais **patc** et **ricm** épelés;
- **af**, **aes**, **sa** et **ffi** sont épelés
- **pao**, **maif**, **seita** comportent des hiatus « inutiles »;

les formes orales de **sncf**, **anpe** ont quatre syllabes;

ddass, **ensma**, **seepu** sont lus infidèlement;

Les séquences de E-muets

Le e-muet est un phonème instable, pouvant se réaliser ou pas suivant son contexte. Dans le mot isolé, il s'élide (pour faire simple) en fin de mot, ou dans les séquences CVCeCV, parfois à l'initiale (*petit*, *renard*, mais pas *pelote*, *vedette*). Ces règles se généralisent (pour faire simple) au syntagme, ce qui pose 4 problèmes nouveaux:

- connaître le « futur » phonétique pour décider du maintien ou non (complexification des règles, gestion d'une couche phonologique)
- possibilité d'insertions (ou épenthèses) *un film très intéressant*
- sélection d'une seule variante parmi la combinatoire (*je ne te le redirai pas* !)
- effets contextuels importants, surtout sur les mots outils (impact sur le naturel: *je ne suis* → ʃɥi).

Autres facteurs: sociolectes, style, débit et contexte d'élocution, facteurs rythmiques...

La question de la liaison

Le français dispose d'un mécanisme favorisant d'une part la réalisation de syllabes CV au détriment des hiatus, permettant d'autre part dans certain cas de désambiguiser les énoncés: la liaison, ie. La réalisation *contextuelle* de segments consonantiques en fin de mots. On oppose généralement:

- liaisons obligatoires (det+nom, adj+nom..., vrb-cli...), locutions (*de temps en temps*) , idiomes...
- liaisons « facultatives » (vrb+prep, prep+vrb...)
- liaisons interdites (nom+adj, nom+vrb)

Pour compléter la description mentionnons: une grande variabilité (inter/intralocuteur), tendance diachronique à à la réduction du nombre de liaisons (norme en évolution), nombreuses erreurs en parole spontanée (« liaisons mal-t-à propos »), opposition stylistique: avec vs. sans enchaînement.

Prise en compte des liaisons

Le « calcul » des liaisons mobilise trois sources de connaissances:

1. un étiquetage catégoriel au niveau de chaque mot
2. des règles d'actualisation de la liaison, distribuant différents types de séparateurs dans l'entrée (attention aux exceptions: court/petit, divin/malin, assez/suffisamment...)
3. des règles de transcription identifiant les phonèmes « latents »:

   #les → $le+z_l$ / + $\#_l V$ (les enfants)

   ein → $\tilde{\varepsilon}$ / **pl** + $\#,\#_l C$ (plein camion)

   ein → $\varepsilon+n_l$ / **pl** + $\#_l V$ (plein air)

Pour le naturel, il semble meilleur d'oublier des liaisons « facultatives » que de surgénérer. Attention toutefois aux liaisons « désambiguisantes », et aux possibles pertes de compréhensibilité.

« Nettoyages » et Normalisation (1)

Les textes « réels » (corpus journalistiques) ne se présentent pas immédiatement sous la forme de phrases, et demandent une mise en forme préalable:

- balises, des marques et variations typographiques (les filtrer ou les utiliser ?)
- partie ``non-linguistiques'' (tableaux de résultats, météo, figures, formules, cours de bourse): filtrage, reformulation orthographique littérale (*Bordeaux-Toulouse 3-0*)

« Nettoyages » et Normalisation (2)

Les textes « réels » (corpus journalistiques) ne se présentent pas immédiatement sous la forme de phrases, et demandent une mise en forme préalable:

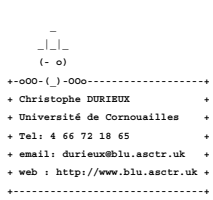
- chiffres numéraux et cardinaux (comment les lire: scores, durées, dates (romaines !), numéro de tel, carte bleues, cours, montants...): récriture contextuelle
  - In 1996 she sold 1995 shares and deposited $42 in her 401(k)
  - 1776 date: seventeen seventy six.
  - 1776 phone number: one seven seven six
  - 1776 quantifier: one thousand seven hundred (and) seventy six
  - 25 day: twenty-fifth
- traitement des abréviations (éventuellement ambiguës) et des « ponctuations » lues (%, $, '...) - utilisation de dictionnaires et de règles contextuelles
  - Dr. North lives on Maple Dr. South.
- traitement des incises, citations, etc (pour la prosodie).

La synthèse des mels

Une application à la mode: la consultation téléphonique de boîtes aux lettres (messagerie intégrée):

1. segmentation du mail : en-tête (RFC822, format MIME), corps du message (textes, inclusions, segments non-textuels), attachements, signature; identification de la langue (plusieurs langues dans un même mail) → une stratégie de lecture.
2. texte synthétisé : prise en compte des spécificités linguistiques des mels (syntaxe relâchée, nombreuses fautes de frappe, omission des accents, jargon, utilisation de « conventions » typographiques d'emphase, abus de sigles (BTW,ASAP...)...) → test de la robustesse des outils linguistiques, ajout de mécanismes de pré-traitement dédiés.
3. certains champs demandent des traitements (récritures) spécifiques: décodage des noms dans les adresses, mise en forme des adresses, des URL, des signatures...
4. Smileys (☺, ;), :<)..) voire les bannières et dessins réalisés avec des caractères alphanumériques

Synthèses des e-mails: problèmes et solutions



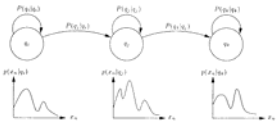
Vu l'extrême imprévisibilité de l'entrée, les traitements linguistiques utilisent massivement des techniques d'apprentissage:

- classification automatique des portions de texte (segmentation)
- modèles de langage pour l'identification des langues
- apprentissage de règles de réaccentuation
- modèles de langage pour l'identification des noms propres ou de leur origine...

Marquage du texte (SPEECHML)

Reconnaissance de parole

G. Bailly - ICP

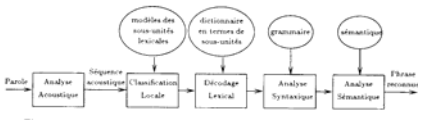


Plan de l'exposé

- Shéma-bloc
  - Analyse acoustique
  - Classification locale et décodage
  - Modèle de langage
- Modèles du signal et techniques de la RdF
  - Anamorphose temporelle (DTW)
  - Chaînes de Markov cachés (HMM)
  - Réseaux de neurones (ANN)
  - Approches hybrides (HMM-ANN)
- Fusion de données
  - Analyse multi-bandes/reconnaissance audiovisuelle

Schéma-bloc

- Analyse acoustique
  - Codage de l'enveloppe spectrale + cinématique
- Classification locale et décodage
  - Modèle de variabilité
- Modèle de langage
  - Contraintes lexicales et morpho-syntaxiques



Pré/post-traitements

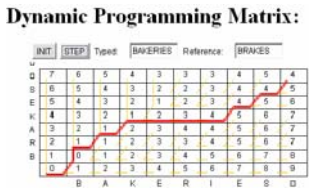
- Pré-traitement
  - Détection de parole: classification parole/non-parole, séparation de sources (matrice de microphones)
- Attention!!!
  - Signal = voix + environnement + microphone + échantillonnage + convertisseur analogique/numérique
- Post-traitement
  - Modèles de langage pas suffisants:
    - « How to recognize speech with a new display » vs. « How to wreck a nice beach with a nudist play »
  - Scores de reconnaissance, adaptation
  - Traitement de la prosodie…

Comparaison de deux mots

- Difficultés
  - Le débit caractérise le locuteur
  - Le locuteur peut varier son débit volontairement (cf. hypo/hyper)
  - Calculer
    - Distance entre deux occurrences d'un même mot
    - Distance entre deux mots
    - Aligner deux occurrences d'un même mot pour caractéristiques moyennes (robustesse)

Chemin d'alignement

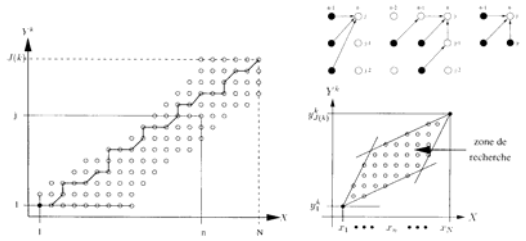
- Principe
  - Correspondance optimale entre séquences de vecteurs
  - On cherche une relation R(i,j) ssi x_i est aligné avec y_j: axe des temps commun



  - Anamorphose temporelle (« time warping »)
    - Contraintes: limites, monotonie, continuité locale & globale (cinématiques possibles), pondération des vitesses

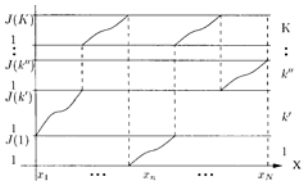
Programmation dynamique

- Distance locale entre deux vecteurs acoustiques
- Distance globale entre séquences
  - Anamorphose temporelle: $D(n,j) = d(n,j) + \min_{p(n,j)} \{D(p(n,j)\}$ (*antécédents de*)
  - Contraintes locales p(n,j) / globales



Mots enchaînés

- Détection de début et fin de mot
  - Avant analyse spectrale
  - Modèle de silence
- DTW à deux niveaux
  - Contraintes à l'intérieur du mot / entre mots

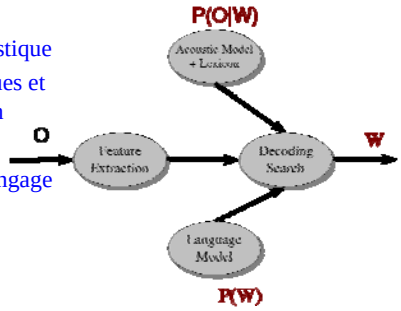


Modèles de Markov (1)

- Objectif
  - P(mot|observation) = P(W|O) ⇒ Maximiser P(W|O)
- Loi de Bayes
  - P(W|O) = P(O|W) P(W) / P(O)
  - max{P(W|O)} = max {P(O|W) P(W)} (comme P(O) est le même pour tous les mots)
  - P(W) = probabilité a priori (e.g., fréquence) "prior"
  - P(O|W) = probabilité d'émission "likelihood"
- Apprentissage
  - Maximiser P(O|W)
- Composantes
  - P(O|W) modèle acoustique; P(W) modèle de langage

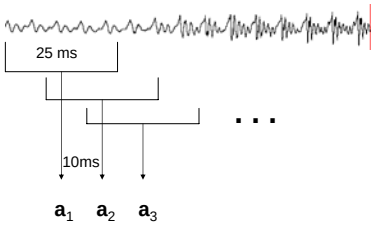
Composantes

- Extration de primitives
- Modèle acoustique
- HMM, lexiques et prononciation
- Décodage
- Modèle de langage



Extraction de primitives

- Trames
- Signaux à court-terme
- Contenu fréquentiel
- MFCC
  - FFT
  - Rééchantillonnage Mel
  - DCT



Modèle acoustique

- 39 paramètres tous les 10 ms
  - 12 MFCC
  - 12 Delta MFCC
  - 12 Delta-Delta MFCC
  - 1 (log) énergie trame
  - 1 Delta (log) énergie trame
  - 1 Delta-Delta (log) énergie trame

Modèle de Markov (ordre 1)

- Un ensemble d'états
  - Q = $q_1, q_2 \ldots q_N$; l'état au temps t est q_t
- L'état courant ne dépend seulement de l'état précédent

$$P(q_i \mid q_1 .. q_{i-1}) = P(q_i \mid q_{i-1})$$

- Matrice des probabilités de transition A

$$a_{ij} = P(q_t = j \mid q_{t-1} = i) \quad 1 \le i, j \le N$$

- Vecteur spécial de probabilités initiales π

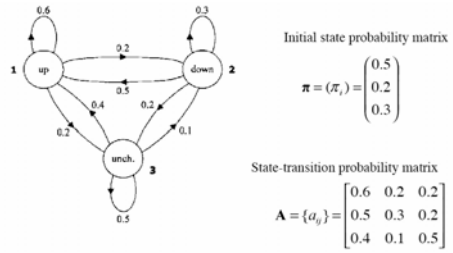
$$\pi_i = P(q_1 = i) \quad 1 \le i \le N$$

- Contraintes:

$$\sum_{j=1}^{N} a_{ij} = 1; \quad 1 \le i \le N \qquad \sum_{j=1}^{N} \pi_j = 1$$

MM pour CAC40

- Probabilité de 5 jours de hausse consécutifs
  - Séquence est (up-up-up-up-up) i.e. (1-1-1-1-1)
  - P(1,1,1,1,1) = $\pi_1 a_{11} a_{11} a_{11} a_{11}$ = 0.5 x $(0.6)^4$ = 0.0648



Modèle de Markov Caché (1)

- Un ensemble d'états
  - Q = $q_1, q_2 \ldots q_N$; l'état au temps t est q_t
- Matrice des probabilités de transition A= $\{a_{ij}\}$

$$a_{ij} = P(q_t = j \mid q_{t-1} = i) \quad 1 \le i, j \le N$$

- Matrice de probabilités d'observation B=$\{b_i(k)\}$

$$b_i(k) = P(X_t = o_k \mid q_t = i)$$

- Vecteur spécial de probabilités initiales π

$$\pi_i = P(q_1 = i) \quad 1 \le i \le N$$

- Contraintes:

$$\sum_{j=1}^{N} a_{ij} = 1; \quad 1 \le i \le N \qquad \sum_{k=1}^{M} b_i(k) = 1 \qquad \sum_{j=1}^{N} \pi_j = 1$$

Modèle de Markov Caché (2)

- Hypothèse de Markov

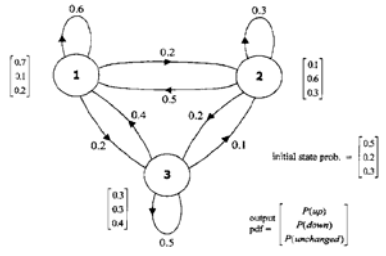
$$P(q_i \mid q_1 .. q_{i-1}) = P(q_i \mid q_{i-1})$$

- Hypothèse d'indépendance des observations

$$P(o_t \mid O_1^{t-1}, q_1^t) = P(o_t \mid q_t)$$

HMM pour CAC40

- Mesure indirecte de caractéristiques de l'état
  - Ex: climatologues en 3004 veulent savoir temps en 2004 à Grenoble. Seules données sont le nb. de glaces mangées chaque jour par une adolescente rapporté sur son journal intime. Observations: nb. de glaces. Etat: jour chaud/froid.



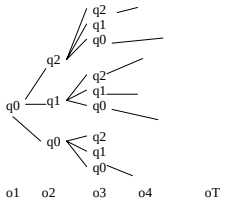
Trois pbs de base pour HMM

- Problème 1 (Evaluation): Etant donnés la séquence d'observation O=($o_1 o_2 \ldots o_T$), et un HMM Φ = (A,B,π), comment calculer P(O| Φ), la probabilité de la séquence d'observation, étant donné le modèle
- Problème 2 (Décodage): Etant donnés la séquence d'observation O=($o_1 o_2 \ldots o_T$), et un HMM Φ = (A,B,π), comment choisir la séquence d'états Q=($q_1 q_2 \ldots q_T$) qui est optimale (i.e., explique de la meilleure manière les observations)
- Problème 3 (Apprentissage): Comment ajuster les paramètres du modèle Φ = (A,B,π) pour maximiser P(O| Φ)?

Evaluation

- Etant donnés la séquence d'observation O et HMM Φ, calculer P(O| Φ)
- Pourquoi est-ce difficile? Cumuler pour toutes les séquences possibles!

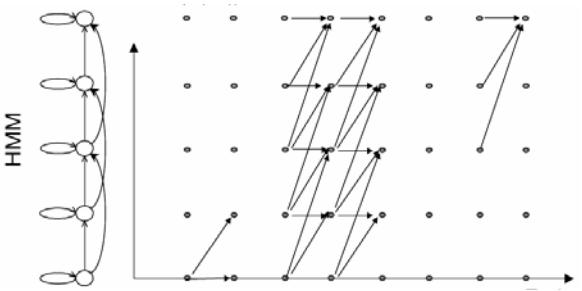
$$P(O|\Phi) = \sum_{all\ S} P(S|\Phi)P(O|S,\Phi) = \sum_{all\ S} a_{s0,s1} b_{s1}(o_1) a_{s1,s2} b_{s2}(o_2) \ldots a_{sT-1,sT} b_{sT}(o_T)$$



P(o1o2o3|q0q0q0)
+
P(o1o2o3|q0q0q1)
+
P(o1o2o3|q0q1q2)
+
P(o1o2o3|q0q1q0)
…

Evaluation/décodage

- Viterbi: idem que Programmation dynamique
- Décodage par retour arrière



Viterbi (1)

- The Idea: Just like Forward, fold exponential paths into a simple trellis, so that all possible paths will remerge into N states at every time slice.
- We define the *viterbi probability* as follows:
  $v_t(i) = P(o_0 o_1 \cdots o_t, Q_1^{t-1}, q_t = i|\Phi)$
- $v_t(i)$ is the probability that the HMM Φ is in state i at time t having generated partial observation O_1^t by passing through the most likely state sequence Q_1^{t-1}.
- We again compute it by induction:
  - Initialization:

$$v_1(i) = \pi_i b_i(o_1), 1 \le i \le N \quad (4)$$

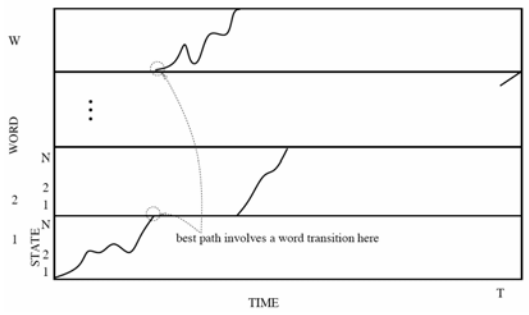
$$bt_1(i) = 0 \quad (5)$$

  - Induction:

$$v_t(j) = [\max_{1 \le i \le N} v_{t-1}(i) a_{ij}] b_j(o_t),$$

Viterbi (2)

- Viterbi: peut inclure des transitions entre mots

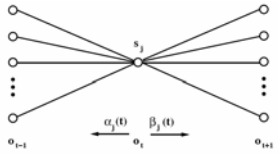


Apprentissage HMM (1)

- « avant-arrière » (Baum-Welch)
  - Cas spécial de l'algorithme EM Expectation-Maximization (Dempster, Laird, Rubin)
  - Estimation itérative: espérance mathématique des fréquences relatives des transitions et émissions
  - Maximisation: mise à jour des paramètres de façon à maximiser la fonction de vraisemblance (ex: paramètres de distributions gaussiennes pour p(x|q))
- Initialisation par alignement

Apprentissage HMM (2a)

- La structure des états est toujours imposée
- EM converge vers maximum local
- Principe:
  - Calcul de $\xi_t(j)$, la probabilité d'être dans l'état j à l'instant t.

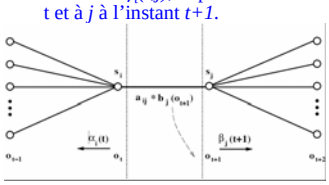


- $\xi_t(j) = \frac{P(q_t = j, O|\Phi)}{P(O|\Phi)}$
- $\xi_t(j) = \frac{\alpha_t(j)\beta_t(j)}{P(O|\Phi)}$

$$\hat{b}_j(v_k) = \frac{\sum_{t=1 s.t. O_t = v_k}^{T} \xi_j(t)}{\sum_{t=1}^{T} \xi_j(t)}$$

Apprentissage HMM (2b)

- La structure des états est toujours imposée
- EM converge vers maximum local
- Principe:
  - Calcul de $\gamma_t(i,j)$, la probabilité d'être dans l'état *i* à l'instant t et à *j* à l'instant *t+1*.



$$\gamma_t(i,j) = P(q_t = i, q_{t+1} = j \mid O, \Phi)$$

$$\gamma_t(i,j) = \frac{\alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j)}{\alpha_T(N)}$$

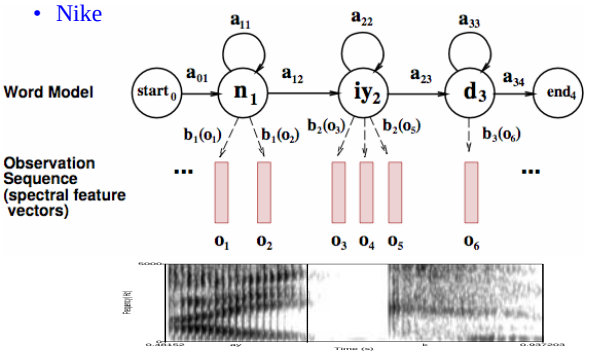
$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \gamma_t(i,j)}{\sum_{t=1}^{T-1} \sum_{j=1}^{N} \gamma_t(i,j)}$$

Apprentissage HMM (3)

- Résumé
  1) Initialiser $\Phi=(A,B,\pi)$
  2) Calculer α, β, ξ
  3) Estimer new $\Phi'=(A,B,\pi)$
  4) Remplacer Φ avec Φ'
  5) Si non convergence aller à 2

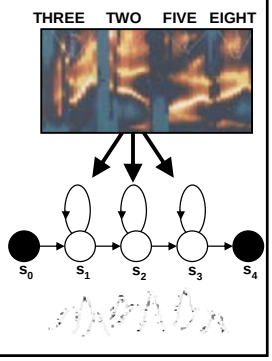
HMM pour parole (1)

- Nike



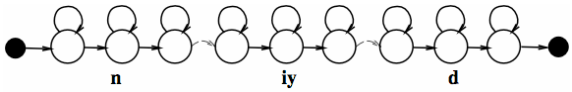
HMM pour parole (2)

- Décomposition de chaque phoneme en sous-modèles (états s)
  - Probabilités d'émission p(O/s) et de transition $p(s_i/s_j)$
  - Variations temporelles dans les probabilités de transition
  - Variations de locuteur, accent & prononciation dans les distributions des mixtures de Gaussiennes

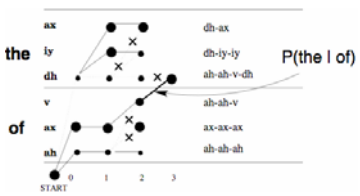


HMM pour parole (3)

- Un mot a donc cette allure

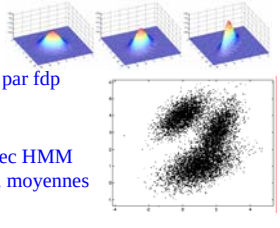


- Avec pb de transitions entre mots



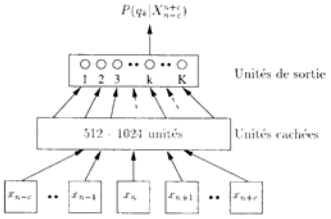
Observations

- Solution discrète
  - Observations: alphabet fini $\{V = b_n, n=1..N\}$ de symboles
  - Quantification vectorielle préliminaire
    - Métrique (moindre carré, mahalanobis, etc)
    - Algorithme de répartition (K-moyennes, LBG, etc)
- Solution continue
  - Observations paramétrées par fdp
    - Gausiennes
    - Multi-gaussiennes
  - Apprentissage conjoint avec HMM des paramètres (mélanges, moyennes et covariances)



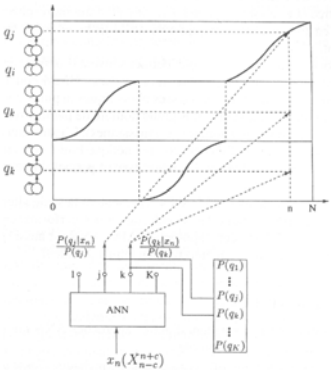
Réseaux de neurones

- P(O/s) estimé directement
  - Inférence statistique
  - Empan temporel centré sur la trame
- Couplage avec HMM pour contraintes séquentielles



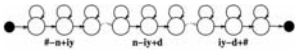
HMM/ANN

- Le réseau ANN fournit les vraisemblances normalisées
  - p(O/s)/p(s)
- En pratique:
  - HMM de phonèmes à un état
  - 40 à 60 classes phonétiques

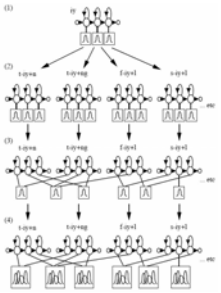


Contraintes linguistiques (1)

- Modèles phonologiques
  - Phonèmes en contextes

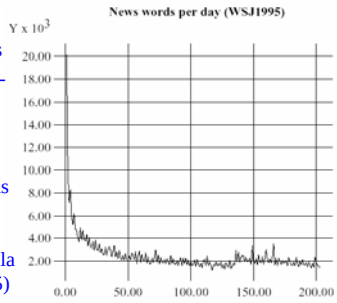


  - Modèles de syllabes, mots…
  - Treillis phonologique
- Mise en commun des états
  - State tying: monophones et simples gausiennes, clonage en triphones, regroupement, entraînement de multigausiennes
- Elagage



Contraintes linguistiques (2)

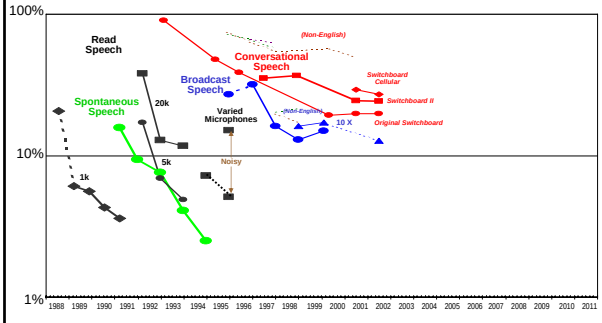
- Modèles de langage
  - Modèles d'états finis
  - HMM : bi-grams, tri-grams…
  - Treillis de mots, perplexité
  - Estimation sur corpus écrit: lissage (ex: « green table » n'apparaît jamais ds la totalité du WSJ 1995)



Applications

- Télécommunications
  - Automatisation des services de renseignement, composition vocale du numéro « mains libres », commande et contrôle de service d'accès à des bases de données, de services de réservation ou d'achat par téléphone
- Bureautique
  - Dictée vocale
- Contrôle de production
- Applications médicales
  - Création de rapports ou remplissage de formulaires
- Applications financières
  - Vérification du locuteur
- Applications Militaires
  - Contrôle vocal de commandes
- Applications éducatives
  - Apprentissage des langues, programmes éducatifs

Performances (1)



DARPA 1986-1998

Performances (2)

Problème	Tâche	Mode	Vocabulaire	%
Mots isolés	Mots équiprobables	DL	10 chiffres	0
			39 alpha-num	4.5
			1109 anglais de base	4.3
		IL	10 chiffres	0.1
			39 alpha-num	7.0
			1218 noms	4.7
Mots enchaînés	Séquence de chiffres (longueur connue)	DL	10 chiffres	0.1
		IL	11 chiffres	0.2
	Réservation avion (perp. 4)	DL	129 mots	0.1
Parole continue	RM (perp. 60)	IL	991 mots	3
	ATIS (perp. 25)	IL	1800 mots	3
	WSJ (perp. 145)	IL	20000 mots	12

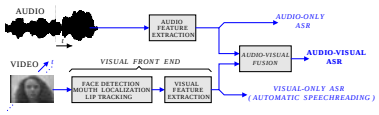
D'après Rabiner et al, 1996

Outils/ressources

- HTK ("HMM Tool Kit") de Cambridge, UK
- CMU Language Modeling Toolkit
  http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html
- SRI Language Modeling Toolkit
  http://www.speech.sri.com/projects/srilm

Reconnaissance audiovisuelle

- Extraction de primitives
- Reconnaissance audio robuste
  - Séparation de sources, réhaussement
- Fusion
  - Précoce/tardive
  - Modèles de fusion: conjointe, recodage, alignement…

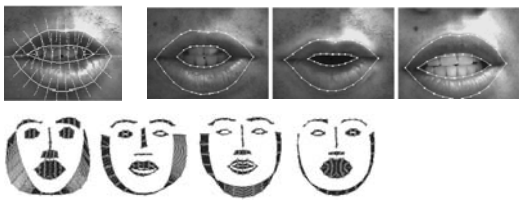


Primitives visuelles (1)

- Régions d'intérêt: visage puis lèvres
  - Teinte peau, modèle générique de visage

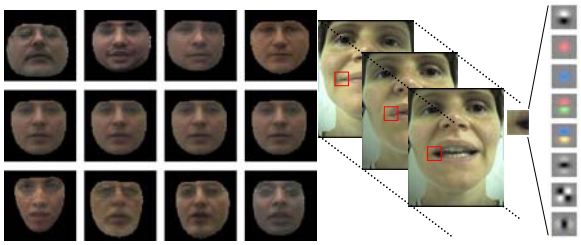
Primitives visuelles (2)

- Analyse de lèvres
  - Suivi de points d'intérêt
  - Modèles de forme



Primitives visuelles (3a)

- Modèles d'apparence
  - Analyse statistique d'images « shape-free » (Cootes 97)
  - Analyse statistique de primites au voisinage des points d'intérêt (ex: champs réceptifs)



Primitives visuelles (3b)

- Principe des modèles actifs d'apparence (AAM)

PCA/ Shape: $\mathbf{s} = \mathbf{s}_m + \boldsymbol{\phi}_s \mathbf{b}_s$

PCA/ Texture: $\mathbf{g} = \mathbf{g}_m + \boldsymbol{\phi}_g \mathbf{b}_g$

Coupling the two models: $\mathbf{b} = \begin{pmatrix} W_s \mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix}$

PCA / concatenated shape and texture parameters **b**:

$\mathbf{b} = \boldsymbol{\phi}_c \mathbf{c}$

Shape Instance: $\mathbf{s}_{model}(\mathbf{c}) = \mathbf{s}_m + \mathbf{Q}_s \mathbf{c}$

Texture Instance: $\mathbf{g}_{model}(\mathbf{c}) = \mathbf{g}_m + \mathbf{Q}_g \mathbf{c}$

Match a target face in a given image (iterative gradient search):

Minimize a texture residual: $r(\mathbf{c}, \mathbf{p}) = g_{model}(\mathbf{c}) - g_{image}(\mathbf{c}, \mathbf{p})$

Find the optimal correction ($\delta\mathbf{c}$, $\delta\mathbf{p}$) to apply in order to minimize $r(\mathbf{c}, \mathbf{p})$

$\delta\mathbf{c} = -R_c\, r(\mathbf{c}, \mathbf{p})$ $\quad$ $\delta\mathbf{c} = -R_p\, r(\mathbf{c}, \mathbf{p})$

R_p , R_p Matrices precomputed from training data

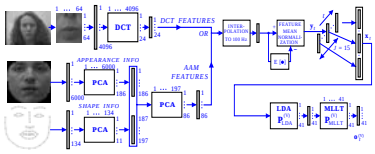
Crédit Hamlaoui & Davoine, UTC

Primitives visuelles (3c)

- Suivi de visage par AAM
  - Actualisation/spécialisation de l'apparence
    - Eclairage
    - Angle de vue
    - etc
  - Ex: Hamlaoui, S. and F. Davoine (2005). Facial action tracking using particle filters and active appearance models. International Conference on Smart Objects and Ambient Intelligence, Grenoble - France: 165-168.

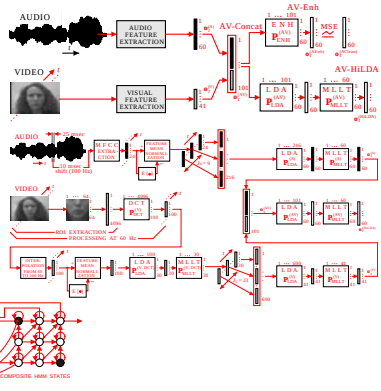
Primitives visuelles (4)

- Caractérisation intégrée forme/apparence
  - Ex: ViaVoice IBM



Modèles de fusion

- Concaténation simple
- Fenétrage
- Ou plus astucieux: HMM produit



Performances

Speech condition	Recognition task	Training set			Held-out set			Adaptation set			Test set		
		Utter	Dur	Sub	Utter	Dur	Sub	Utter	Dur	Sub	Utter	Dur	Sub
Normal	LVCSR	17111	34:55	239	2277	4:47	25	855	2:03	26	1038	2:29	26
	DIGITS	5490	8:01	50	670	0:58	50	670	0:58	50	529	0:46	50
Impaired	LVCSR	N / A			N / A			50	0:11	1	50	0:11	1
	DIGITS	N / A			N / A			80	0:08	1	60	0:06	1

Audio condition	Clean	Noisy	Audio Condition	Clean	Noisy
Audio-only	14.44	48.10	AV-MS-Joint (DF)	14.62	36.61
AV-Concat (FF)	16.00	40.00	AV-MS-Sep (DF)	14.92	38.38
AV-HiLDA (FF)	13.84	36.99	AV-MS-PROD (DF)	14.19	35.21
AV-DMC (DF)	13.65 →12.95	-	AV-MS-UTTER (DF)	13.47	35.27

D'après Potamianos et al, 2003. LVCSR vocabulaire de 10403 mots

Automates parlants

G. Bailly - ICP

Historique de la synthèse

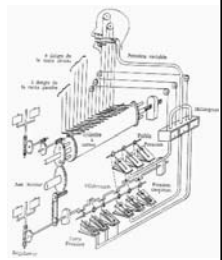
- Le faire-semblant
  - Des statues d'Alexandrie à la «fille invisible» du physicien Charles
- Le faire-comme
  - Vésale, De Vinci ...
  - Androïdes de Vausanson, Mical, Kempelen ...
- Triomphe du phonographe
- Développement instrumentation de mesure

Faire parler le non-vivant

- Vieille idée
  - Des statues d'Alexandrie à la «fille invisible» du physicien Charles (1780)
- Le faire-comme
  - Vésale, De Vinci ...
  - Androïdes de Vausanson, Mical, Kempelen ...
- Triomphe du phonographe
- Développement instrumentation de mesure

Vaucanson (1709-1782) et ses automates

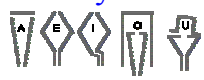
- Sons produits par mouvements



Mical (1730-1789) et les têtes parlantes

- Texte:
  - Le roi vient de donner la paix à l'Europe (traité de Versailles, Angleterre, 1783)
  - La paix couronne le roi de gloire
  - La paix fait le bonheur des peuples
  - O roi adorable! Père de vos peuples. Leur bonheur fait voir à l'Europe la gloire de votre trône
- Mémoires de Bachaumont…
  - "Dans les quatre phrases qu'elles articulent successivement, et en imitant à l'extérieur le mouvement des lèvres, il est des mots qu'elles mangent en entier; leur son de voix est rauque, leur articulation lente; et malgré tous les défauts, elles en disent assez pour qu'on ne puisse se refuser à leur accorder le don de parole…"

Kratzenstein (1723-1795) et les voyelles

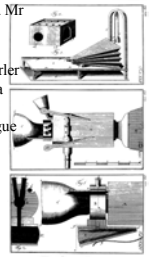


- Journal de Physique, 21, 358-380, 1782…
  - "Les voyelles sont des sons motivés par diverses ouvertures de la bouche et l'élévation de la langue … »
- Vox humana : « L'envoûtante et riche sonorité de l'instrument reste délicate à l'oreille. Ses principaux et autres jeux flûtés, dont les tuyaux sont construits de 97% d'étain, assurent une présence tout en étant des compagnons idéals pour les voix. Un large éventail de couleurs est disponible pour les solos et ce, principalement parmi les anches : un hautbois (Oboe) - réellement un Basson-Hautbois -, une trompette française (French Trompette), une trompette allemande (German Trumpet), un jeu d'anche de 4' à la pédale et une délicieuse Voix Humaine (Vox Humana) de type allemand. Enfin, une anche pleine longueur de 32' à la pédale, une rareté dans un instrument de cette taille.

Van Kempelen (1734-1804)

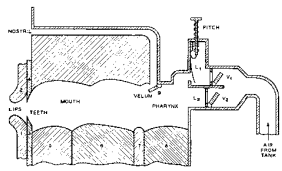
- Rivarol, 1808:
  - "M. Kempelen (!) avait aussi un coffret d'où il s'échappait quelques mots, à ce qu'on dit; mais cet honnête voyageur a rendu un véritable hommage à Mr l'abbé Mical; dès qu'il a eu connaissance des têtes parlantes…"
  - Avec un peu d'habitude et d'habileté, on pourra parler avec les doigts comme avec la langue, et on pourra donner au langage des têtes, la rapidité, le repos et toute la physionomie enfin que peut avoir une langue qui n'est point animée par les passions."





De Faber à Riez

- Perfectionnement de la synthèse mécanique



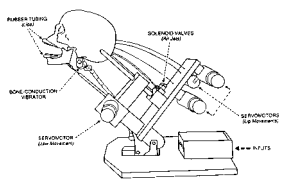
Riesz (1937)


Euphonia" (Faber, 1835-1846)

Jusqu'à nos jours...

- TADOMA

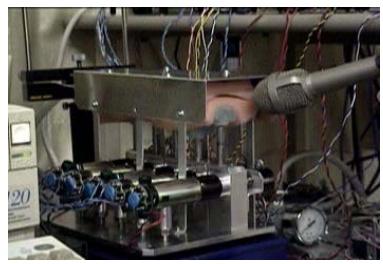

Reed et al, MIT, JASA 85, 77-1



Hong et al., MIT, JASA 89, 86-3

Jusqu'à nos jours...

- Sawada (Kagawa University)

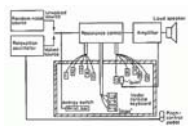


Et puis Edison...

- Du Moncel, 1880:
  - « On s 'est étonné que la machine parlante qui nous est venue, il y a quelques années d 'Amérique (Faber, Barnum, 1975) et qui a été exhibée au Grand Hotel fût d 'une extrême complication, alors que le phonographe résolvait le problème d 'une manière si simple: c 'est que l 'une de ces machines ne faisait que reproduire la parole, tandis que l'autre l'émettait… »

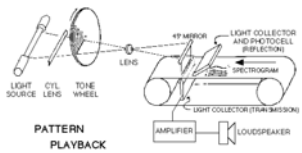
Dudley… la synthèse électrique

- VODER (Dudley) présenté à l'expo universelle (New York,1939)
- Pattern playback: enregistrement




Cooper: analyse et synthèse

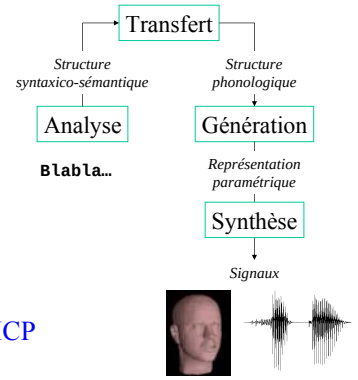
- Pattern playback (1950)
- Ove (1960)



Conclusions

- Le faire-semblant
  - Reproduction des effets
  - Traitement du signal et des images
    - Apprentissage automatique, fusion de données
- Le faire-comme
  - Reproduction des causes
  - Synthèse physique
    - Acoustique, aérodynamique, biomécanique, contrôle moteur … robotique anthropomorphique
  - Réalité virtuelle

Synthèse de la parole



- G. Bailly - ICP

Organisation générale

- Traitements linguistiques
  - Analyse morpho-syntaxique
  - Transcription orthographique-phonétique
  - Transfert phonologique
- Synthèse
  - Méthode de synthèse
  - Représentation du signal: technique de synthèse

Pourquoi? (1)

- Téléphonie
  - Services téléphoniques
  - Lecture e-mail
  - Accès bases de données (listes de prix, évènements culturels, météo…): 70% des appels ne nécessitent que très peu d'interaction
- Multimedia
  - Livres parlants, jeux interactifs
- Communication homme-machine

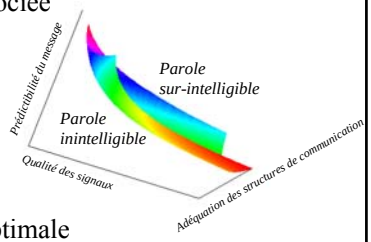
Pourquoi? (2)

- Aide aux handicaps
  - Voix artificielle
  - Aveugles: lecture automatique de documents

Handicap	Millions
Cannot use fingers	1
Wheelchair users	3
Dyslexia	25
Hard of hearing	80
Low vision	11

Négociation locuteur-auditeur

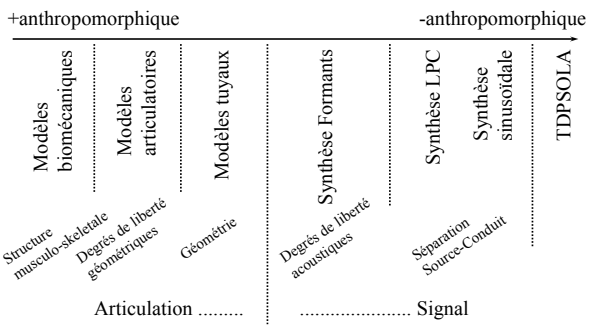
- Gestion "optimale" de l'articulation en fonction de l'espace de croyance mutuel
- Variabilité négociée
- Contenu
- Réalisation
- Nombreuses dimensions...
- *Enjeu*: restituer cette gestion optimale



Synthétiseurs

Traitement des signaux

Synthétiseurs

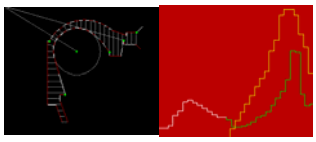


Synthèse articulatoire

- Instrument
  - Modèles géométriques
  - Modèles statistiques
  - Modèles biomécaniques
- Contrôle
  - Inversion
  - Contrôle musculaire: point d'équilibre

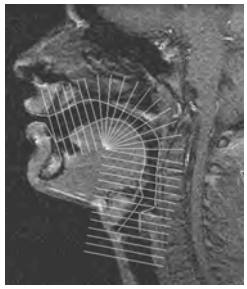
Modèles géométriques

- Mermelstein (1972)
  - angle of jaw opening (1)
  - tongue center position (2)
  - tongue tip position (3)
  - lip rounding (2)
  - hight of the hyoid (1)
  - state of the velum (1)
- Coker (1968)



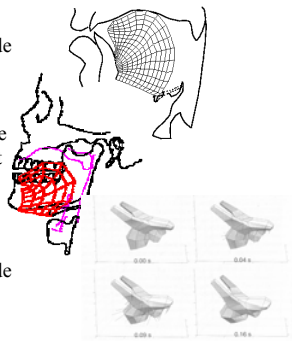
Modèles statistiques

- Maeda (1972)
  - Cinéradiographies
  - Données: intersection des contours du CV avec grille de lecture
  - Analyse en composantes principales par parties
  - 7 degrés de liberté
    - Mâchoire
    - Corps, dos, pointe de la langue
    - Protrusion & écartement des lèvres
    - Larynx



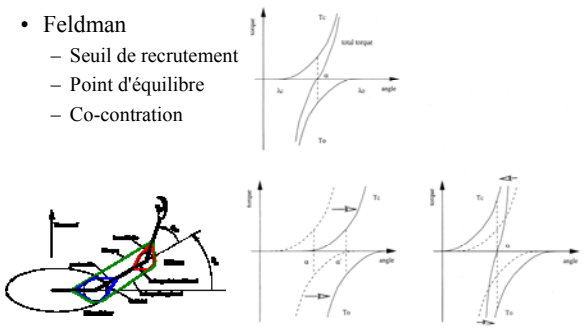
Modèles biomécaniques

- Payan & al (1996)
  - Modèle élément finis 2D de la langue
- Laboissière et al (1997)
  - Intégration dans un modèle de contrôle du mouvement de la mâchoire et de l'os hyoïde
- Wilhems (1995)
  - Modèle élément finis 3D de la langue



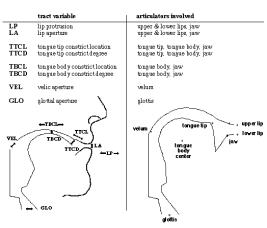
Modèles de contrôle musculaire

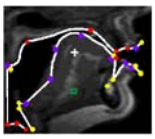
- Feldman
  - Seuil de recrutement
  - Point d'équilibre
  - Co-contration



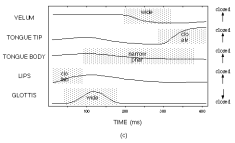
Modèles de contrôle de l'articulation

- Phonologie articulatoire
  - Spécification de pavés temporels de constrictions



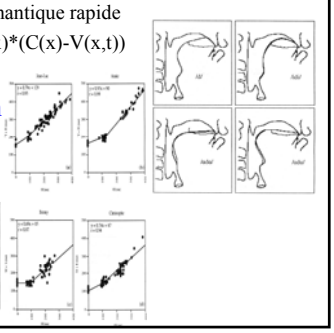


pan

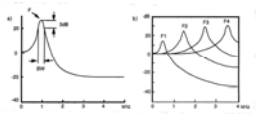


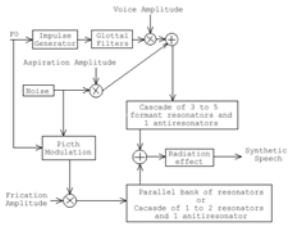
Modèles de coarticulation

- Modèle of coarticulation
  - Ohman [1967]: superposition d'un geste vocalique porteur (lent) et d'un geste consonantique rapide
  - F(x,t) = V(x,t) + k(t)*w(x)*(C(x)-V(x,t))
  - C(x) & w(x) estimés
- Modèle d'anticipation
  - MEM: Abry et al [1991]



Synthèse à formants

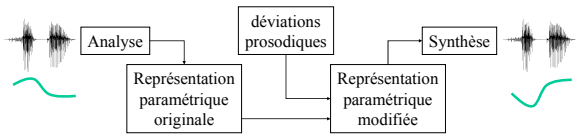




- Copie de naturel (Gobl, 87)
  - hommes
  - femmes
  - enfant

Systèmes d'analyse-modification-synthèse

- Traitement du signal
  - Représentations du signal temporel
  - Représentations fréquentielles
  - Transparence
  - Réalisme des covariations induites par déviations



TDPSOLA

- Signaux à court-terme (SCT)
- Durée = duplication/effacement des SCT
- F0 = décalage des SCT

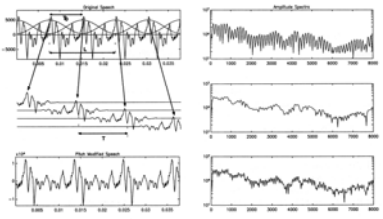
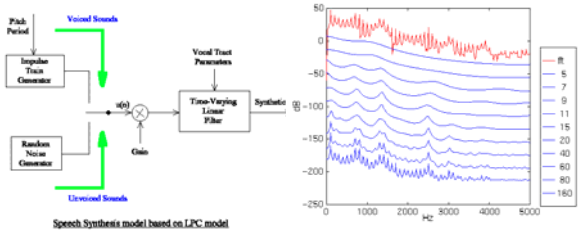


Fig. 10.1. The TD-PSOLA reharmonization process. The pitch-modified waveform (bottom plot) has the same spectral envelope as the original waveform (top plot).

Prédiction linéaire



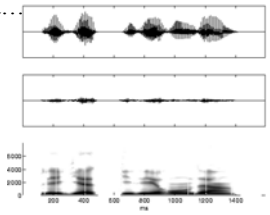
Modèle harmonique + bruit (HNM)

- Musique/parole
- Séparation **harmonique**/bruit
- Modèle sinusoïdal, LPC…FOF
- Analyse: FFT/**ABS**
- McAulay & Quatieri, Serra…

Original

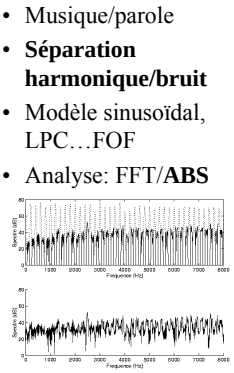
PS-ABS

$$s(t) = \sum_{l}^{L(t)} A_l(t)\exp\left(j\psi_l(t)\right)$$



Modèle harmonique + bruit (HNM)

- Musique/parole
- **Séparation harmonique/bruit**
- Modèle sinusoïdal, LPC…FOF
- Analyse: FFT/**ABS**



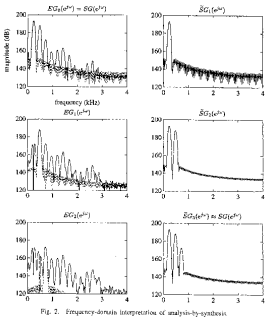


Fig. 2. Frequency-domain interpretation of analysis-by-synthesis

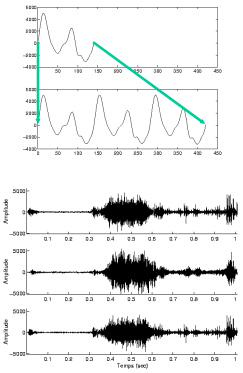
Modèle harmonique + bruit (HNM)

- Techniques d'interpolation
  - Modèle d'enveloppe (LPC discret, DCT)
  - Rééchantillonnage de l'enveloppe
  - Interpolation polynomiale

$$\psi_l(t) = a + bt + ct^2 + dt^3 \text{ with}$$

$$\begin{cases} a = \varphi_l^n \\ b = \omega_l^n \\ \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} \frac{3}{\Delta T^2} & \frac{-1}{\Delta T} \\ \frac{-2}{\Delta T^3} & \frac{1}{\Delta T^2} \end{bmatrix} \begin{bmatrix} \varphi_l^{n+1} - \varphi_l^n - \omega_l^n \Delta T + 2\pi M \\ \omega_l^{n+1} - \omega_l^n \end{bmatrix} \\ M = E\left[\frac{1}{2\pi}(\varphi_l^n + \omega_l^n \Delta T - \varphi_l^{n+1}) + \frac{\Delta T}{2}(\omega_l^{n+1} - \omega_l^n)\right] \end{cases}$$

- Modèle bruit
  - LPC modulé, FOF,



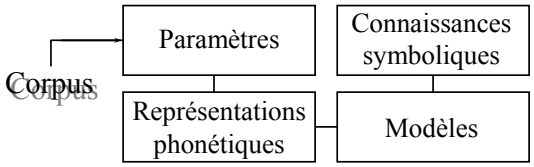
Méthodes de synthèse

Génération segmentale

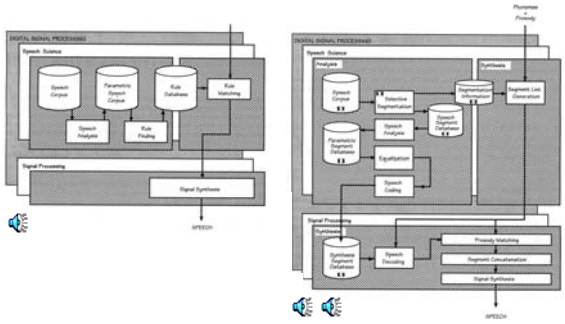
Méthodes de synthèse

Synthèse par concaténation ←

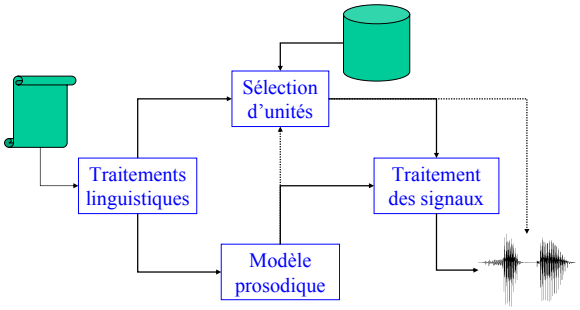
→ *Synthèse par règles*



Règles vs concaténation



Synthèse par concaténation



Unités infra-phonémiques

- MULTIVOX: entre règles et concaténation
- ABU

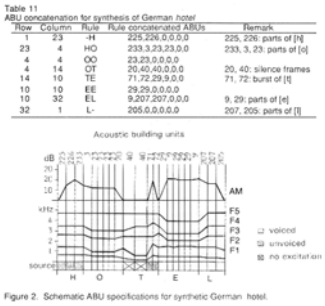


Table 11
ABU concatenation for synthesis of German *hotel*

Row	Column	Rule	Rule concatenated ABUs	Remark
1	23	-H	225,226,0,0,0,0	225, 226: parts of [h]
23	4	HO	233,3,23,23,0,0	233, 3, 23: parts of [o]
4	4	OO	23,23,0,0,0,0	
4	14	OT	20,40,40,0,0,0	20, 40: silence frames
14	10	TE	71,72,29,3,0,0	71, 72: burst of [t]
10	10	EE	29,29,0,0,0,0	
10	32	EL	9,207,207,0,0,0	9, 29: parts of [e]
32	1	L-	205,0,0,0,0,0	207, 205: parts of [l]

Figure 2. Schematic ABU specifications for synthetic German hotel.

Unités utilisées en synthèse (1)

- Diphones [Küpfmüller & Wrans, 57],[Estes et al, 64],[Dixon & Maxey, 68][Emerard 77]
- Di-syllabes VCV [Saito & Hashimoto, 68]
- Demi-syllabes [Fujimura, 76]
- Polyphones [Olive, 77], [Emerard, 86], [Bimbot, 88]
- Unités mixtes [Portele 94: 2182 unités dt 1086 demi-syllabes initiales, 577 finales, 88 suffixes, 234 diphones et 197 syllabes]
- Unités multi-représentées NTT [Nakajima & Hamada, 88] ATR [Sagisaka, 88: 5240 mots courants] AcuVoice [5 à 7 variantes prosodiques par syllabe: 150Mo]

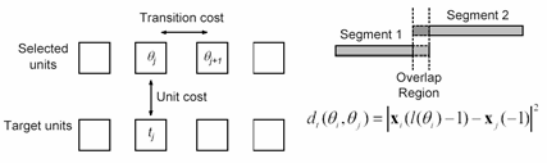
Unités utilisées en synthèse (2)

- Concaténation de mots
- Concaténation brute de diphones
- Synthèse par diphones standard
- Synthèse par unités multi-représentées
- Synthèse par treillis phonologique

Sélection/concaténation (1)

- Critères de sélection
  - Contenu phonétique, contexte phonologique, frontières prosodiques…
- Sélection des unités
  - Programmation dynamique

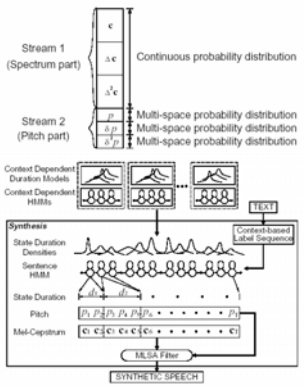
$$d(\Theta,T)=\sum_{j=1}^{N} d_u(\theta_j,T)+\sum_{j=1}^{N-1} d_t(\theta_j,\theta_{j+1})$$



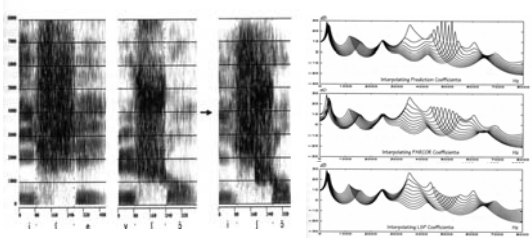
$$d_t(\theta_j,\theta_j)=\left|\mathbf{x}_s(l(\theta_j)-1)-\mathbf{x}_j(-1)\right|^2$$

Sélection/concaténation (2)

- Synthèse par HMM
  - Apprentissage de modèles HMM pour unités en contexte
  - Modèles de durée
- Comparaison
  - Formants
  - Diphones
  - Sélection
  - HMM



Lissage aux points de concaténation



- Fonction de la représentation
  - ex: LSP>LAR>AK

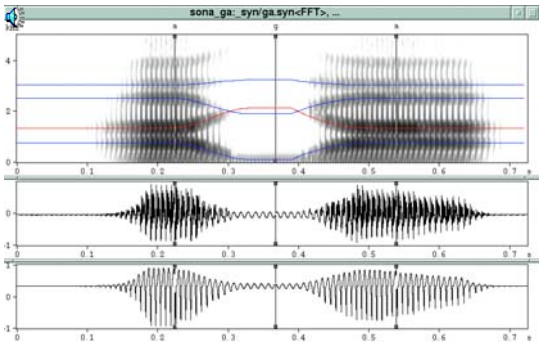
Synthétiseurs à partir du texte

- Bell Labs (diphones + LPC)
- Université de Caen (diphones + RELP)
- LAIP Lausanne (diphones + MBROLA)
- Elan/CNET (diphones + TDPOLA)
- ICP (polysons + TDPSOLA)

Synthèse par règles

- Principes
  - Cibles + coarticulation + transitions
- Stylisation
- Apprentissage
- Intelligibilité des représentations
- De la prosodie

Synthèse par règles...



Transformation de voix (règles)

- Exemples par C. Gobl (KTH, 87)
  - Original homme
  - F0 doublé
  - Renforcement des HF
  - Bandes passantes doublées
  - 1,115*F1, 1,15*F2, 1,17*F3 (Fant, 76)
  - 1,5*coefficient d'ouverture
  - 1,5*temps de retour à la fermeture
  - age: 5 - 12 - 14 - 21 ans
  - voisement

Transformation de voix (formants vs QV)

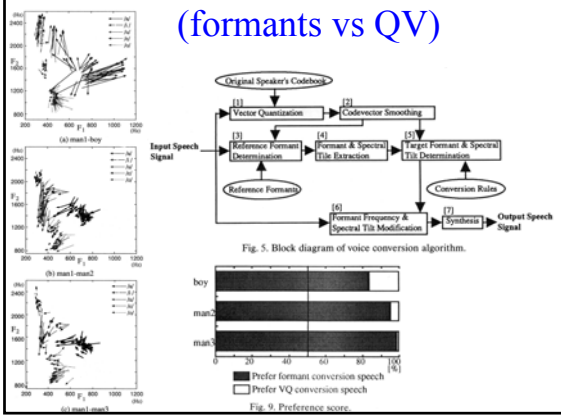
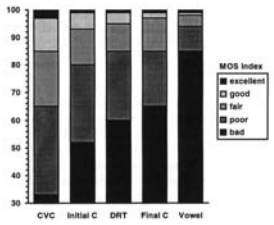


Fig. 5. Block diagram of voice conversion algorithm.

Fig. 9. Preference score.

Évaluation

- Quels tests?
  - Matériau
    - ex: bonne distribution CVC pour MOS
  - Intelligibilité
  - Acceptabilité
  - Adéquation
  - Charge cognitive



Traitements linguistiques

Traitement de la langue

Découpage du texte (1)

- Unité de traitement: la phrase
  - Relativement simple si la phrase se termine par ?!
  - Mais si elle se termine par “.”
  - Ambiguïté entre fin de phrase et fin d’abréviation
    - My place on Forest Ave. is around the corner.
    - I live at 360 Forest Ave.
    - (Not “I live at 360 Forest Ave..”)
  - Solutions ?

Arbre de décision pour fin de phrase

```
((n.whitespace matches ".*\n.*\n[ \n]*") ;; A significant break in text
 ((1))
 ((punc in ("?" ":" "!"))
  ((1))
  ((punc is ".")
   ;; This is to distinguish abbreviations vs periods
   ;; These are heuristics
   ((name matches "\\(.*\\..*\\|[A-Z][A-Za-z]?[A-Za-z]?\\|etc\\)")
    ((n.whitespace is " ")
     ((0))          ;; if abbrev, single space enough for break
     ((n.name matches "[A-Z].*")
      ((1))
      ((0))))
    ((n.whitespace is " ")  ;; if it doesn't look like an abbreviation
     ((n.name matches "[A-Z].*")  ;; single sp. + non-cap is no break
      ((1))
      ((0)))
     ((1))))
  ((0)))))
```

Arbre de décision précédent

- Un point avec une ou deux lettres est une abbrev
- Un point avec 3 lettres capitales est une abbrev.
- Une abbrev suivie de 2 espaces et une lettre capitale est une fin de phrase
- Des non-abbrevs suivies par un mot commençant par une capitale sont des fins de phrase
- Ca ne marche pas pour
  - Cog. Sci. Newsletter
  - Plein de cas à la fin d’une ligne
  - Phrases avec mots mal espacées/capitalisés

Indices plus sophistiqués

- Prob(mot avec “.” arrive à la fin-de-p)
- Prob(mot après “.” arrive au deb-de-p)
- Longueur du mot avec “.”
- Longueur du mot après “.”
- Cas du mot avec “.”: Upper, Lower, Cap, Number
- Case du mot après “.”: Upper, Lower, Cap, Number
- Ponctuation après “.” (le cas échéant)
- Classe d’abréviation du mot avec “.” (nom de mois, unité de mesure, titre, adresse, etc)

Analyse lexicale (1)

- « POS tagging » Affecter une fonction lexicale à chaque « mot »
  - Définition du « mot »: e.g locutions « à la fois », etc.
  - Inventaire de fonctions
    - Fondamentales: nom, verbe, adjectif, préposition, adverbe, article, interjection, pronom, conjonction, etc
    - Appelées: parts-of-speech (POS), lexical category, word classes, morphological classes, lexical tags
  - Classe ouverte vs. fermée
    - fermée: prépositions (of, in, by, …), auxiliaires (may, can, will had, …), pronoms (I, she, mine, his, them, …). Usuellement mots de fonction (mots courts et communs qui jouent un rôle central dans la grammaire)
    - ouverte: mots créés en permanence. La plupart des langues en ont 4 (noms, verbes, adjectifs, adverbes)… mais pas toutes!

Analyse lexicale (2)

- Classes fermées
  - Idiosyncrasique
  - Exemples:
    - prépositions: on, under, over, …
    - particules: up, down, on, off, …
    - déterminants: a, an, the, …
    - pronoms: she, who, I, ..
    - conjonctions: and, but, or, …
    - verbes auxiliaires: can, may should, …
    - numéraux: one, two, three, third, …
- Jeu d'étiquettes
  - Minimum 10
  - "UPenn TreeBank tagset" 45

Analyse lexicale (3)

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WPS	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	"	Left quote	(‘ or “)
POS	Possessive ending	*'s*	"	Right quote	(’ or ”)
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PPS	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Analyse morpho-syntaxique (1)

- Un mot a souvent de nombreuses étiquettes
  - Ex: the *back* door, on my *back*, get money *back*, to *back* the bill
  - WSJ: non ambigus (38,857), ambigus (8,844)

2 tags	6,731	
3 tags	1621	
4 tags	357	
5 tags	90	
6 tags	32	
7 tags	6	well, set, round, open, fit, down
8 tags	4	's, half, back, a
9 tags	3	that, more, in

Analyse morpho-syntaxique (2)

- Solutions
  - Règles
    - Eliminer suites d'étiquettes impossibles dans listes d'options. Fastidieux
    - Pb. de déterminisme
  - HMM
    - Affecter l'étiquette la plus probable
    - Ex: la belle ferme le voile… Sujet Verbe Complt
    - On limite la dépendance ds le temps. Ex: bigrams

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(t_1^n \mid w_1^n) \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} P(w_i \mid t_i) P(t_i \mid t_{i-1})$$

Analyse syntaxique (1)

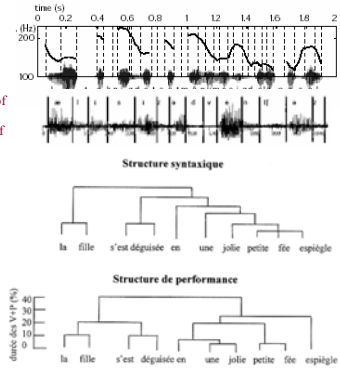
- But: segmentation en unités minimales de sens (constituants immédiats) et hiérarchiser ces constituants
  - Segmentation Chunks
    - Noun phrase (NP) (a draught beer)
    - Verb cluster (would be)
    - Adverb phrase (ADVP) (as rapidly as possible)
    - Adjective phrase (AP) (very nice)
    - Prepositional phrase (after this talk)
    - Une dizaine de « chunks » couramment utilisés

Modèles prosodiques

Génération supra-segmentale

Prosodie

- Fonctions
  - Segmentation/hiérarchisation
    French [bread and cheese]
    [French bread] and [cheese]
  - Mise en relief
    LEGUMES are a good source of vitamins
    Legumes are a GOOD source of vitamins
    Legumes are a good source of VITAMINS.
  - Position du locuteur
  - Locuteur...
- Paramètres
  - Mélodie
  - Durées
  - Intensité/timbre
- Organisation
  - Accents
  - Tons



Accents

- Accent lexical (stress) vs accent de phrase (accent)
- ***Stress*** **is a structural property of a word — it marks a potential (arbitrary) location for an accent to occur, if there is one.**
- ***Accent*** **is a property of a word in <u>context</u> — it is a way to mark intonational prominence in order to 'highlight' important words in the discourse.**

(x)					(x)		(accented syll)
x					x		stressed syll
x			x		x		full vowels
x	x	x	x	x	x	x	syllables
vi	ta	mins	Ca	li	for	nia	

Prosodie: nécessaire

- Synthèse
  - Avec prosodie plate et durées constantes
  - Avec prosodie déclinante
  - Avec des accents sur les syllabes accentuées
  - Avec des accents sur les syllabes accentuées et tons finaux
  - Avec prosodie apprise par un modèle statistique (f0 et durées)

Modèles prosodiques

- Modèle phonétique
  - Relation à la chaîne segmentale
  - Paramétrage
- Modèle phonologique
  - Relation aux fonctions discursives
  - Apprentissage automatique
  - Capacité à accroître sa compétence

Fonctions discursives

Modèle phonologique profond

Prototypes, évènements…

Modèle phonologique de surface

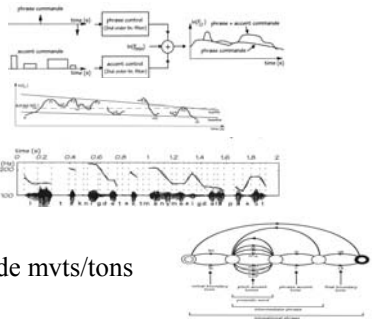
Contours, commandes, points-clés…

Modèle phonétique

Paramètres prosodiques

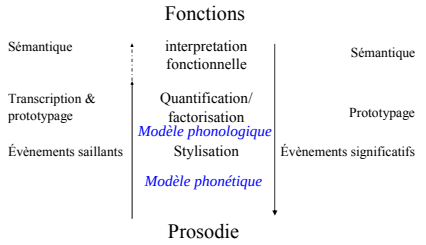
Modèles phonétiques de la mélodie

- Commandes
- Mouvements
- Tons
- Grammaires de mvts/tons



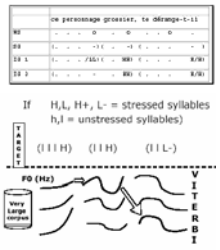
Approches montantes/descendantes

- Phonologie prosodique: de la prosodie aux fonctions
- Morphophonologie: des fonctions à la prosodie



Modèles de prosodie

- Plus de théories que de chercheurs!
- Mise à l'échelle d'un modèle
  - Modèle versus concaténation de contours préstockés
  - Nécessité d'un modèle
    - REOF
  - Apprentissage automatique
  - Maîtrise du corpus d'apprentissage
    - Linguistique
    - Paralinguistique
    - Signification statistique



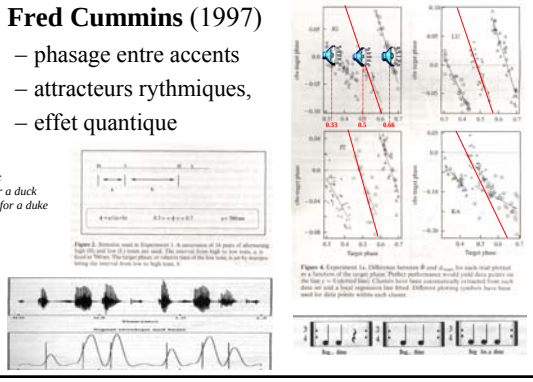
Gestalts

- Théorie de la perception de formes
- Perception holistique
- Schémas préférentiels/attendus
  - Production du rythme (Cummins 97)
  - Perception de la mélodie (Jones 89) et de l'intonation (Aubergé, Grépillat 97)
- Apprentissage des formes
  - Formes pré-établies « imposées » par le système de production (ex: Fujisaki) ou système de perception (ex: lignes ds t'Hart)
  - Formes émergentes négociées entre production, perception et cognition (ex: Holm)

Production du rythme

- **Fred Cummins** (1997)
  - phasage entre accents
  - attracteurs rythmiques,
  - effet quantique

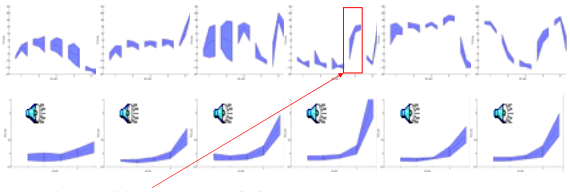
CVCs:
big for a duck
geese for a duke
...



Modèle morphophonologique de l'intonation: formes globales

- Attitudes: contours globaux indépendants du substrat morphosyntaxique

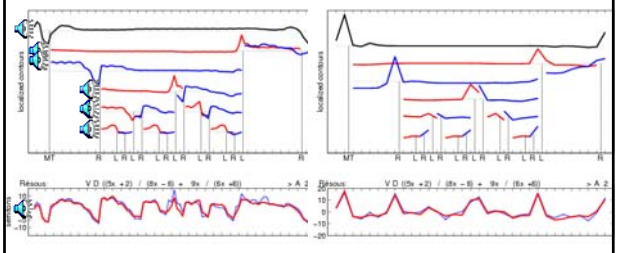
déclaration, question, exclamation, incrédulité, ironie de soupçon, évidence



Forme fait sens: événement saillant intégré dans le contour global

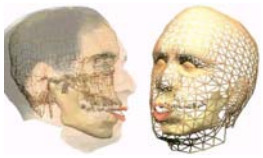
Modèle morphophonologique de l'intonation: émergence de contours

- Formules mathématiques: fort enchâssement de relations de dépendance



1

Synthèse audiovisuelle



G. Bailly - ICP

2

Plan de l'exposé

- Instruments
  - Synthèse basée-image
  - Synthèse basée-modèle
- Modèles de coarticulation
- Évaluation
- Modèles basés-données
  - Mouvements visibles
  - Mouvements invisibles/partiellement visibles

3

Synthèse basée-image

Placage/mélange de régions

VideoRewrite

ATT AnimatedHead

Déplacements de pixels

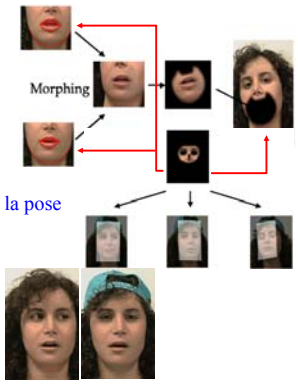
MikeTalk

4

VideoRewrite

C. Bregler, M. Covell & M. Slaney … Stanford Univ & IBM

- Basé-images
  - Mélange de régions
    - Bouche/arc mandibulaire
    - Face
  - Caractérisation
    - "Eigenpoints"
  - Référentiel: estimation de la pose
- Triphones
- Vidéo de fond
  - Longueur de la phrase

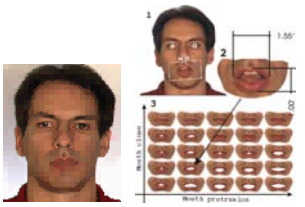


5

AnimatedHead

J. Ostermann, E. Cosatto & H. Peter … ATT Research Lab

- Basé-image
  - Superposition de régions
- Control paramétrique
  - Exemplaires
    - hypercube de formes de bouche possibles
    - triphone
- Mouvements de tête
  - Aléatoires

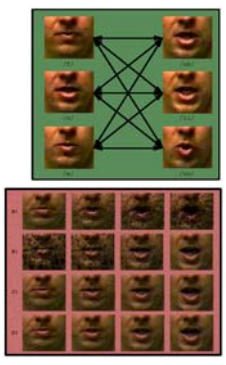


6

MikeTalk

T. Ezzat & T. Poggio … AI Lab MIT

- Basé-image
  - Interpolation d'images
  - Flux optique
- Contrôle paramétrique
  - Interpolation entre visèmes
- Mouvements de tête corrélés aux mouvements de parole



7

Synthèse basée-modèle

Modèles géométriques

Descendants de Parke

Modèles biomécaniques

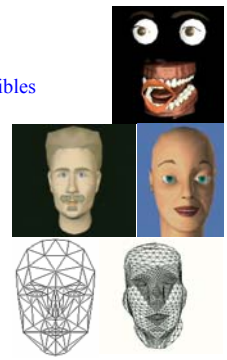
Terzopoulos & Waters...

Modèles statistiques

8

Descendants de Parke

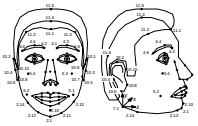
- Transformations géométriques
  - Translations, rotations …
  - Interpolation entre configurations-cibles
  - Constructions procédurales ad hoc
- Avatars
  - Baldi Massaro & Cohen … PSL USCL
  - OLGA et autres , Beskov & al… KTH
  - LCE talking head, Sams & al
- Très utilisé en suivi de visages
  - Candide

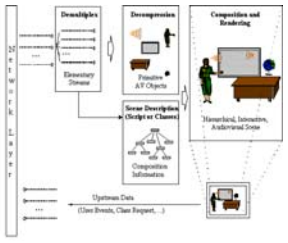


9

MPEG4

- Codage d'objets audiovisuels
  - décomposition/recomposition
- objets SNHC
  - maillage, points-clés, FDP et FAP





Lucent: face2face

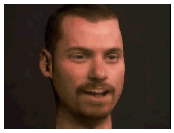
Eisert : conférence virtuelle

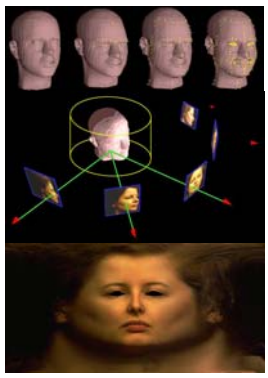
ATT: AVTTS

10

Édition de maillages

- RealFace
  - Model générique
  - Photos simultanées
  - Points-clés
  - Texture cylindrique
  - Mélange entre expressions faciales

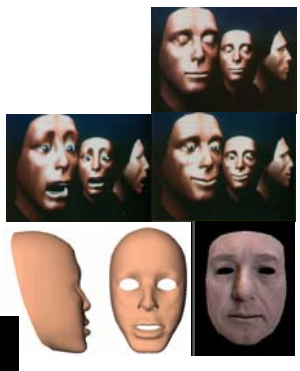




11

Descendants de Waters

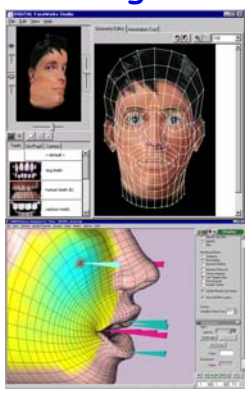
- Modélisation biomécanique
  - Couches de peau
  - Système musculo-squelettal
- De simples matelas de ressorts... aux modèles d'éléments finis
  - Waters
  - Munhal & Terzopoulos
  - Payan & al, TIMC
  - Basu et al, MIT



12

Édition de maillages

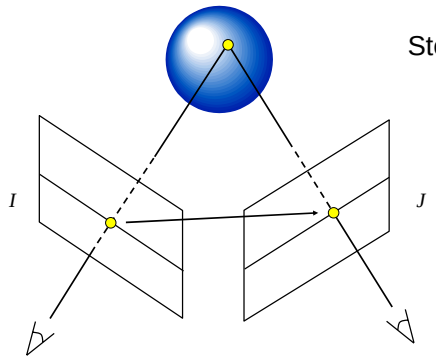
- FaceWorks
  - Ressorts
- Famous 3D animator
  - Points de contrôle
  - Nurbs
  - Régions d'action



13

Principe 1: triangulation

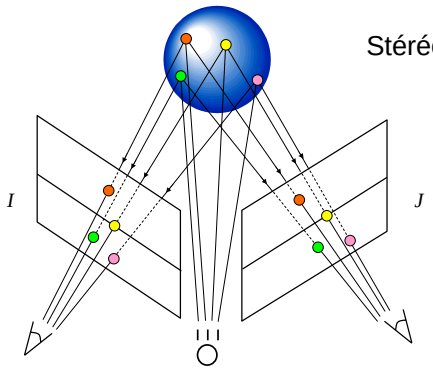
Stéréo



14

Principe 1: triangulation

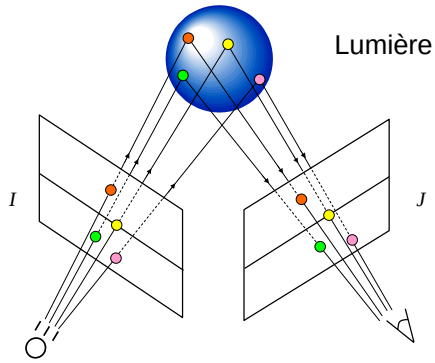
Stéréo active



15

Principe 1: triangulation

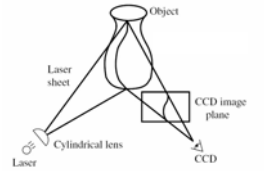
Lumière structurée



16

Collecter la géométrie (1)

- Surface
  - Cyberscan
    - + très précis <0.01mm
    - − 10sec par scan

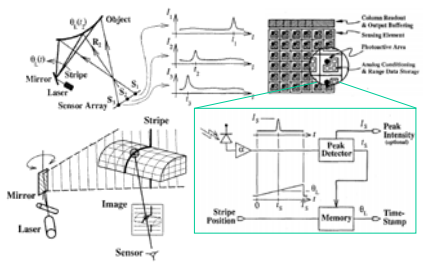


Cyberware® face and head scanner

17

Collecter la géométrie (2)

- Surface
  - Scanner haute vitesse
    - + vitesse d'acquisition 1000Hz
    - − custom, surface convexes

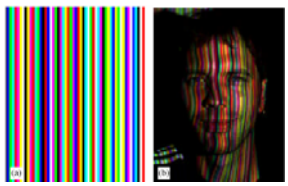


A. Gruss, S. Tada, and T. Kanade "A VLSI Smart Sensor for Fast Range Imaging," ICIRS 1992
Working Volume: 350-500mm - Accuracy: 0.1%
Spatial Resolution: 28x32 - Speed: 1000Hz

18

Collecter la géométrie

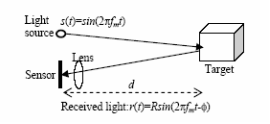
- Surface
  - Lumière structurée

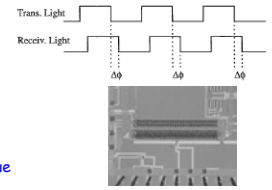


P. Huang, C. Zhang, F. Chiang, "High-speed 3-D shape measurement based on digital fringe projection", Journal of Optical Engineering, 2003
Working Volume: 10-2000mm - Accuracy: 0.025%
Spatial Resolution: 532x500 - Speed: 120Hz

19

Principe 2: temps de vol



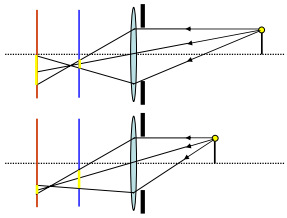


\+ Pas d'ombre
\+ alignement mécanique pas trop critique
− resolution

Miyagawa, R., Kanade, T., "CCD-Based Range Finding Sensor", IEEE Transactions on Electron Devices, 1997
Working Volume: 1500mm - Accuracy: 7%
Spatial Resolution: 1x32- Speed: ??

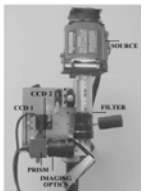
20

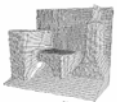
Principe 3: focales multiples



21

Principe 3: focales multiples





\+ Bonne resolution et précision, temps réel
− hardware ad hoc

Nayar, S.K., Watanabe, M., Noguchi, M., "Real-Time Focus Range Sensor", ICCV 1995
Working Volume: 300mm - Accuracy: 0.2%
Spatial Resolution: 512x480 - Speed: 30Hz

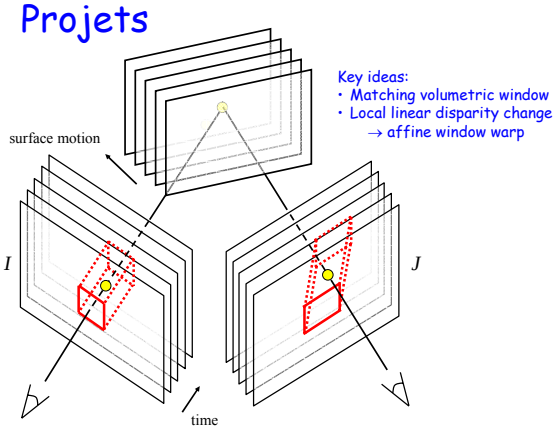
22

Produits commerciaux

Company	Working principle	XY resolution	Depth accuracy	Speed
Cyberware	Laser	>500x500	0.01mm	>10sec per scan
XYZRGB	Laser	Very high	0.01mm	>10sec per scan
Eyetronics	Structure light	High	<2mm	<0.1sec
3Q	Structure light	High	<2mm?	<0.1sec
3DV	Time of flight	720x486	1-2cm	30Hz
Canesta	Time of flight	64x64	1cm	30Hz

23

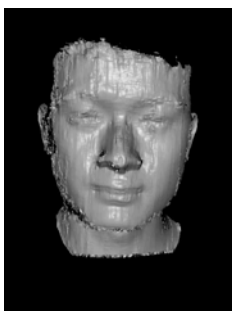
Projets



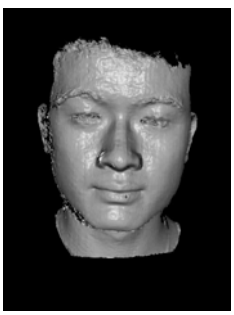
Key ideas:
• Matching volumetric window
• Local linear disparity change
→ affine window warp

24

Résultats



Frame-by-frame stereo
WxH=15x15 window



Spacetime stereo
WxHxT=9x5x5 window

25

Contrôler les clones

Approche paramétrique : modèle de coarticulation

De Cohen-Massaro, Öhman au contrôle moteur

Stockage/déformation/lissage de trajectoires naturelles

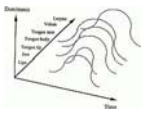
Visèmes

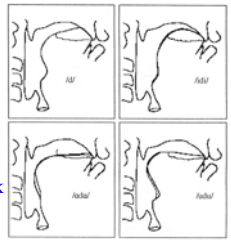
Modèles de n-phones

26

Modèles de coarticulation

- Cohen & Massaro
  - Mélange de gestes indépendants du contexte avec des poids indépendants du contexte
- Öhman
  - Superposition de gestes de constriction rapides sur un geste vocalique « lent »
- Jusqu'aux modèles de planification du mouvement…
  - Planification vs exécution (ex: Task dynamics…)

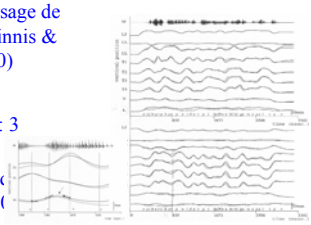




27

modèles de triphones

- Stockage
  - Sélection/superposition/lissage de triphones audiovisuels (Minnis & Breen, 98; Cosatto et al, 00)
- Modèles de trajectoires
  - HMM visuel (Brooke, 98): 3 états/phonème en contexte
  - Modèle de la cinématique (triphones (Okadome et al, (
    - Pilotage par l'acoustique
  - HMM audiovisuel (Tamura & al, 98; Bregler & al, 97; Yamamoto et al, 98; Brant, 99)



28

Évaluation

Intelligibilité

Intégration multimodale

Charge cognitive

29

Résultats impressionnants

- Mary
  - Test de Turing effectué par Ezzat (2002)
  - Le **test de Turing** est une proposition de test d'intelligence artificielle ayant la faculté d'imiter la conversation humaine.
    - Décrit par Alan Turing en 1950 dans sa publication "Computing machinery and intelligence", ce test consiste à mettre en confrontation verbale un humain avec un ordinateur et un autre humain à l'aveugle. Si l'homme qui engage les conversations n'est pas capable de dire qui est l'ordinateur et qui est l'autre homme, on peut considérer que le logiciel de l'ordinateur a passé avec succès le test. Cela sous-entend que l'ordinateur et l'homme essayeront d'avoir une apparence sémantique humaine. Mais pour conserver la simplicité et l'universalité du test, la conversation est souvent limitée à un échange textuel entre les protagonistes.

30

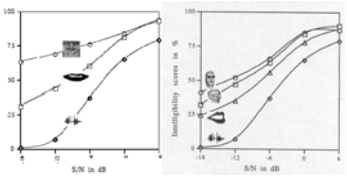
Résultats impressionnants

- Mary
  - Test de Turing effectué par Ezzat (2002)
  - Mais au niveau intelligibilité, ne trompe pas les sujets…

Banc de test standard

31

- Intelligibilité
  - Parole dans le bruit
  - AV>A & >V
  - Gain ≈ 11dB

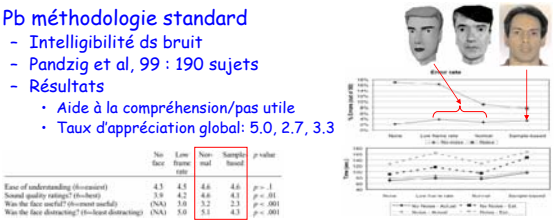


- Vérifier la fusion AV précoce
  - Mc Gurk: $[b]_A+[g]_V = [d]_{AV}$

Evaluation (1)

32

- Pb méthodologie standard
  - Intelligibilité ds bruit
  - Pandzig et al, 99 : 190 sujets
  - Résultats
    - Aide à la compréhension/pas utile
    - Taux d'appréciation global: 5.0, 2.7, 3.3

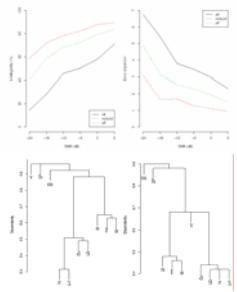


- Points lumineux
  - se débarrasser du modèle d'apparence
  - Peuvent accroître l'intelligibilité *(Rosenblum et al, 1996), (Bergeson, Pisoni & Reynolds, 2003), (Cohen, Walker & Massaro, 1996),*
  - Sont effectivement intégrés McGurk *(Rosenblum & Saldana, 1996),*
  - Activent les mêmes aires que parole normale *(Santi et al., 2003).*

Evaluation (2)

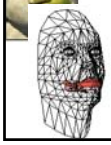
33

- Performances
  - Environ 70% de la vidéo complète
  - Même structure d'erreur



De l'avatar à l'agent conversationnel incarné

34



- Cog Project R. Brooks MIT:

« In studying human intelligence, … our alternative methodology is based on evidence from cognitive science and neuroscience which focus on four alternative attributes which we believe are critical attributes of human intelligence: *embodiment and physical coupling*, *multimodal integration*, developmental organization, and social interaction. We believe that not only are (these principles) critical to the understanding of human intelligence but also that *they actually simplify the problem of creating human-like intelligence.* »

- Être humble… observer (et copier) les comportements humains

« Nous avons créé les personnages en bâtissant d'abord un squelette, puis en rajoutant des muscles. Il était indispensable de respecter rigoureusement les règles de l'anatomie. *Notre idée est que le cerveau humain, même s'il méconnaît notre travail, le percevra néanmoins et les personnages gagnerons en réalisme.* »

Shrek's visual artist, PDI, Palo Alto CA. (le monde du 10 mai 01)

- Percevoir la parole audiovisuelle naturelle vs synthétique

« Much to our dismay, however, we failed to replicate the prototypical McGurk fusion effect... *Whatever reason*, the auditory information dominated the perceptual judgment. »

D. Massaro, AVSP'98, p.22

Modèles articulatoires (statistiques)

35

Capture de mouvement : collecter les mouvements de points de chair (fleshpoints)

Identifier les degrés de liberté

Cloner l'articulation visible

36

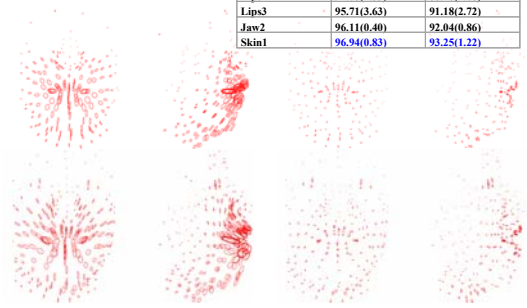
- Degrés de liberté des gestes visibles
  - Expérimenter
  - Modéliser

37

Réduction d'information

- Variance expliquée

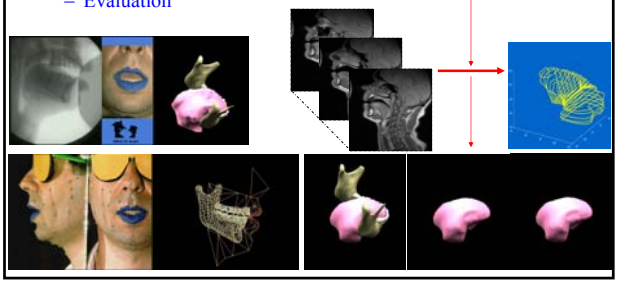
Paramètre	Locuteur français (197 points)	Locuteur arabe (230 points)
Jaw1	30.52(30.52)	14.80(14.80)
Lips1	87.55(57.03)	83.78(68.99)
Lips2	92.08(4.53)	88.46(4.68)
Lips3	95.71(3.63)	91.18(2.72)
Jaw2	96.11(0.40)	92.04(0.86)
Skin1	96.94(0.83)	93.25(1.22)



38

Articulation partiellement visible

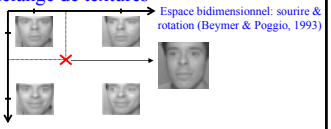
- Degrés de liberté des gestes articulatoires sous-jacents
  - Expérimention, modélisation
  - Évaluation



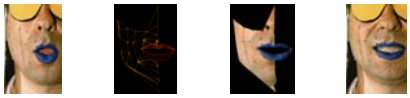
39

Video-réalisme (1)

- Mouvements de tête
  - Texture cylindrique vs mélange de textures



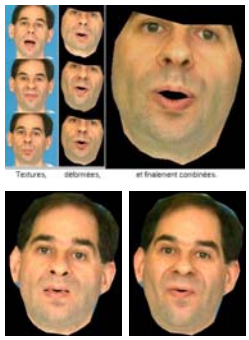
- Mouvements faciaux
  - Une texture unique n'est pas suffisante



40

Video-réalisme (2)

- Multi-références
  - Coefficients de mélange sont fonction de la distance entre le maillage cible et les maillages de référence
  - Illusion du 3D: voir nez



41

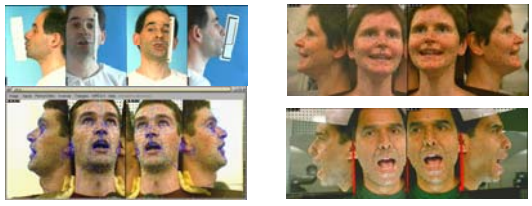
Modèle d'apparence

- Statistical Model of appearance (Cootes 1997)
  - Images « shape-free »
  - Analyse statistique

42

Galerie de locuteurs clonés

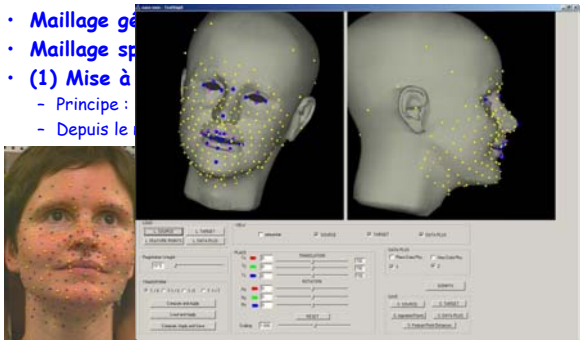
- **4 locuteurs**
  1. Un locuteur français
  2. Une locutrice française (bilingue Français/Anglais)
  3. Un locuteur allemand
  4. Un locuteur algérien
- **Tous articulés par les mêmes paramètres**



43

Tête parlante générique (1)

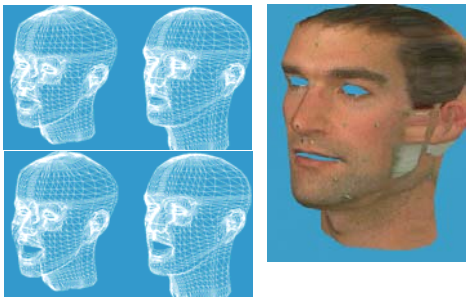
- Maillage g
- Maillage sp
- (1) Mise à
  - Principe :
  - Depuis le



44

Tête parlante générique (2)

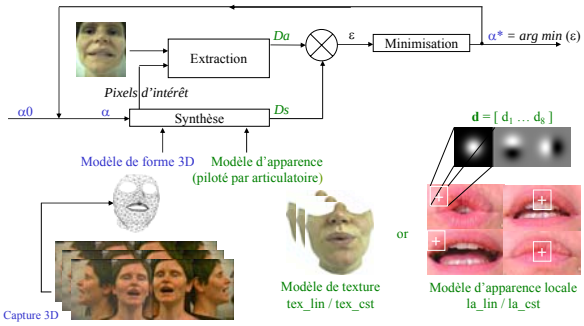
- (2) Articuler le maillage générique
  - Mise à l'échelle des visèmes avec tous les FPs, modèle de forme



45

Suivi de mouvements

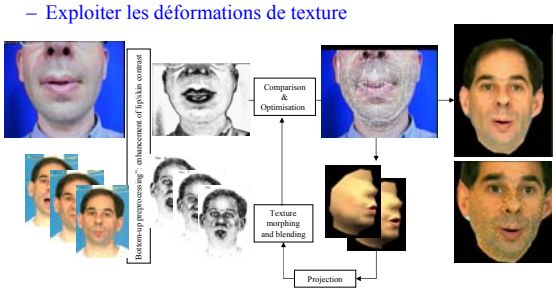
- Boucle de minimisation: analyse/synthèse



46

Capture de mouvement globale

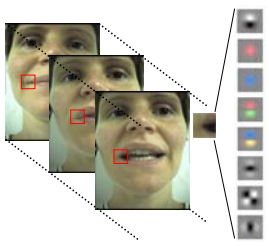
- Analyse par la synthèse
  - Cartes colorimétriques
  - Exploiter les déformations de texture



47

Capture de mouvement locale

- **De l'apparence à la forme**
  - Modèle d'interaction I (local) reliant variation de forme et d'apparence locale (LAM © Prima/GRAVIR)
  - Inversion many-to-few (© LASMEA)
- **Equations de base pour objets rigides**
  N points d'intérêt
  M champs réceptifs chromatiques
  P paramètres articulatoires
  Rfij, i=1..N, j=1..M
  I = Flin(Δ P, Δ RF)
- **Objets déformables**
  I(P) = Flin([P, Δ P], Δ RF(P))



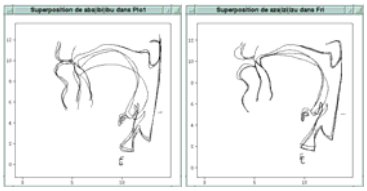
48

Résultats

49

Calculer l'articulation

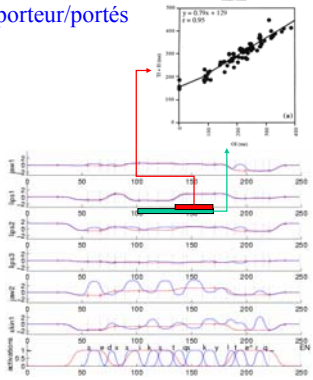
- Degrés de liberté prédits depuis une chaîne phonétique
  - Modèles de variabilité: coarticulation, anticipation et synchronisation inter-articulatoire
  - Ohman: $s(x,t) = v(x,t) + k(t)\, w_c(x)\, (C(x)-v(x,t))$



50

Calculer la (co)articulation

- Ohman 67 : Mouvements porteur/portés

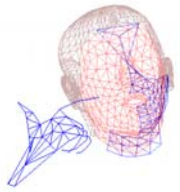


- MEM: Abry et al 91

51

Stocker l'articulation

- Segments multimodaux
  - Cibles non réduites
  - Asynchronie des gestes encapsulée
  - Coût de concaténation acoustique/gestuel
  - Lissage + facile a priori sur paramètres articulatoires
- Ex. ARTUS

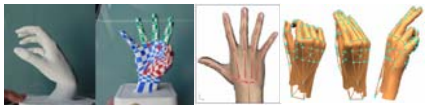


52

ARTUS

- Capture de mouvements sur codeuse

- Passage main BD à HD



53

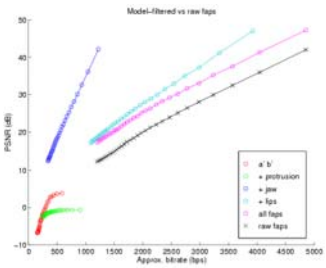
ARTUS

- Modèles de forme non linéaires
- Deux dictionnaires
- Face par photogrammétrie
- Main par modèle «skinné»

54

Transmettre l'articulation

- Couche articulatoire sur FAP/MPEG4
  - Encode avec une bande passante bien plus petite pratiquement la même information géométrique qui pilotera de toutes façons l'animation faciale



55

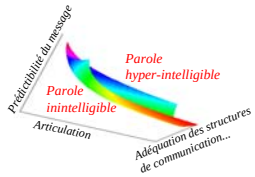
Vers des agents conversationels

- Immerger des clones virtuels dans des scènes naturelles
- Réalité augmentée
- Téléconference virtuelle
- Vie artificielle
- Apprentissage de la langue

56

Épilogue

- De la poupée vers l'agent autonome
- Modéliser l'interaction locuteur-interlocuteur
  - Quoi dire: intelligence
  - Comment(Quoi dire): squelette fonctionnel de l'acte de dialogue
  - Comment(Comment(Quoi dire)): planifier l'acte de parole
  - Comment(Comment(Comment(Quoi dire))): exécuter les gestes articulatoires
- Encodage conditionné
  - Négocié
  - Adapté
  - Augmenté



57

Evaluation (1)

- Modularité
  - Modèles de contrôle, de forme et d'apparence
- Exemple: modèles de contrôle
  - Modèle de forme unique
  - Modèle d'apparence simplifié
    - paradigme des points lumineux
    - Autorise la comparaison avec naturel
  - Modèles de contrôle paramétrés par données d'apprentissage
    - Modèle de coarticulation (Ohman), unités AV, modèle simple d'association son/articulation
  - MOS: Cohérence entre mouvements naturels/calculés et signal acoustique naturel

58

Evaluation (2)

- Base de données
  - 76 phrases + 96 stimuli VCV
  - Paramètres faciaux estimés par analyse/synthèse globale
  - Données d'apprentissage: 66 phrases + 96 stimuli VCV
  - Données de test: 10 phrases
- 6 modèles
  - *Syn* est un système de concaténation de diphones AV
  - *Synl* utilise une procédure de lissage par anticipation
  - *Reg* utilise le modèle d'Ohman. Le MEM décrit les transitions intervocaliques
  - *Mltst* lie les trajectoires LSP filtrées passe-bas (10Hz) avec les mouvements articulatoires
  - *Mlapp* utilise la phrase de test comme base d'apprentissage

59

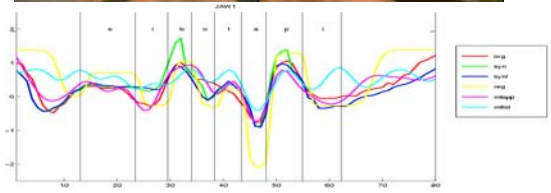
Résultats (1)

- Corrélations
  - Faibles corrélations pour Reg et Mltst
  - Bonnes corrélations pour Syn
  - Très bonnes corrélations pour Mlapp

	Jaw1	Jaw2	Lips1	Lips2	Lisp3	Skin1
Syn	0.85	0.62	0.66	0.70	0.65	0.65
Synl	0.84	0.70	0.73	0.65	0.64	0.63
Reg	0.32	0.66	0.27	0.32	0.37	0.46
Mlapp	0.85	0.88	0.90	0.89	0.87	0.89
Mltst	0.44	0.44	0.55	0.41	0.30	0.53
Inv	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00

60

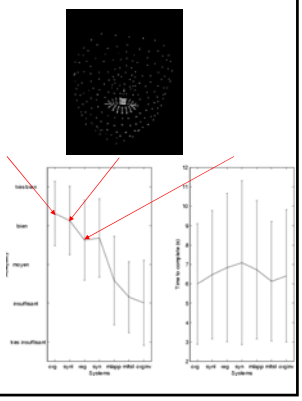
Résultats (2)



61

Perception de mouvements purs

- **Résultats du test**
  - Invalide estimation multilinéaire depuis l'acoustique
  - Valide concaténation AV
  - Sujets estiment qu'une série de stimuli était adéquats en termes d'indices phonétiques mais non « naturels » (trop hyperarticulés). *Reg* a en effet de plus longs temps de décision.



62

Pour en savoir plus…

- Sites WEB
  - http://www.haskins.yale.edu/haskins/heads.html
  - http://mambo.ucsc.edu
  - http://www.research.att.com/~osterman/AnimatedHead/
- Papiers, conférences
  - AVSP workshops 1995, 1996, 1997, 1998, 1999 & 2001
- Bouquins
  - Parke & Waters: *Computer Facial Animation* A.K. Peters 96
  - Cassell, Sullivan, Prevost & Churchill: *Embodied Conversational Agents* MIT Press 00

1

Interaction multimodale

Gérard Bailly

ICP, CNRS/INPG/U3
46, av. Félix Viallet - 38031 Grenoble France

2

Plan de l'exposé

- Systèmes d'interaction
  - Gestion du dialogue
  - Analyse de scène
  - Interaction
- Le cas de la gestion du regard
  - Regard et cognition
  - Regard et technologie
- Objets d'intérêt et conversation face-à-face
  - Atteinte de cible en deux bandes…
  - Agent conversationnel

3

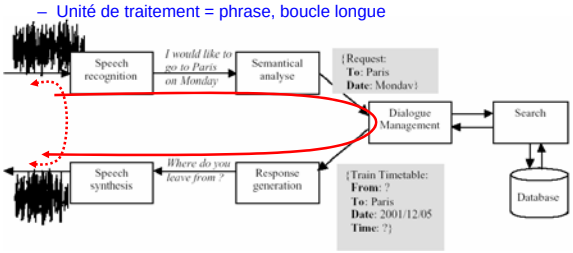
Plan de l'exposé

- Systèmes d'interaction
  - Gestion du dialogue
  - Analyse de scène
  - Interaction
- Le cas de la gestion du regard
  - Regard et cognition
  - Regard et technologie
- Objets d'intérêt et conversation face-à-face
  - Atteinte de cible en deux bandes…
  - Agent conversationnel

4

Les systèmes de dialogue (1)

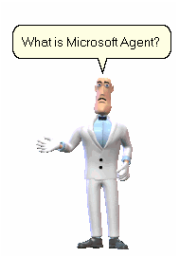
- Le système de reconnaissance est l'oreille du système de synthèse
  - Restitution d'une réponse
  - Esclave du système de dialogue et de génération de « phrases »
- Échange d'informations « symboliques »
  - Unité de traitement = phrase, boucle longue



5

Les systèmes de dialogue (2)

- Incarnation du système
  - Présence
  - Agents Conversationnels Animés (ACA, ex: Agent Microsoft)
  - Souvent limité à la présentation multimodale d'informations
- Nombreuses applications
  - Présentations multimédia
  - Assistant personnel
  - Formation
  - Jeux de rôle interactifs



6

Agents conversationnels animés

Agents conversationnels animés

7

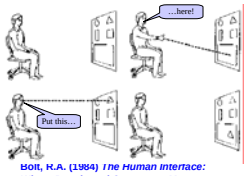
- Idéalement
- Agent individué par:
  - Apparence: conformation faciale, âge, race, gendre, …
  - Comportement: choix des comportements NV accompagnant la parole, répertoire of NV, redondance / seuil de recrutement des actes NV
  - Expressivité du mouvement: tempo, amplitude, tension
- Gestion du dialogue
  - Compréhension du discours, historique, etc
- Comportement influencé par:
  - Facteurs sociologiques: culture, rôle social, personnalité
  - Contexte conversationnel: interlocuteur (capacités cognitives et linguistiques), relation avec l'interlocuteur
  - Contexte situationnel: environnement (bruit, public/privé), contexte d'interaction (formel, amical)

Catherine Pelachaud
LINC, IUT de Montreuil
University of Paris 8

Commentaires

8

- Gestion du dialogue… mais pas de l'interaction
  - Interface du système d'information conscient de l'usager et de l'environnement d'interaction
    - Gages de présence et d'attention
  - Dimensions informationnelles, expressives… et émotionnelles
    - Communication multimodale et multi-segments (voix, face, mains, regard, posture…) en entrée… et en sortie
  - … Piocher l'«intelligence» de l'interaction (et ses symptômes) du côté usager
- Une multimodalité opportuniste et mobilisatrice



Bolt, R.A. (1984) *The Human Interface: Where People and Computers Meet.*

Idem avec ACAs

9

- Systèmes interactifs
  - Couplage Réel/virtuel
  - Gage de présence par communication

Bases cognitives de l'interaction

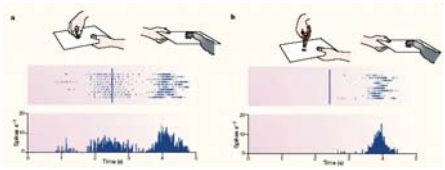
10

- Neurones miroirs
- Théorie de l'esprit
- Les modèles d'interaction
- Constructions d'ontologies

Les neurones de l'empathie

11

- Neurones miroirs
  - Neurones qui déchargent à la perception et à l'exécution d'une classes de gestes (Rizzolati, Parme): Aire 5. Attention dépend des acteurs de l'action

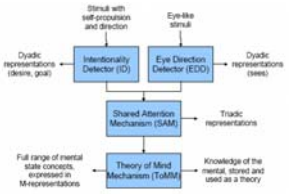


  - Encodent l'action et son contexte: base de la compréhension du la signification des gestes (répertoire)
  - De l'imitation à l'empathie

Les théories de l'esprit

12

- TOM: capacité à pouvoir faire des hypothèses sur ce que se représentent les autres, prédire leur comportement, leurs intentions, imaginer qu'ils ont telles ou telles préoccupations, croyances. etc.
- Avoir une théorie de l'esprit c'est pouvoir se représenter ce que se représentent les autres
- Bases de la théorie de l'esprit
  - ToBY (Theory of body motion) lois physiques de causalité distinction objet/sujet
  - EDD
    Guidage de l'attention par tiers
  - ID
    Répertoires



Baron-Cohen, 1995

13

Les ontologies

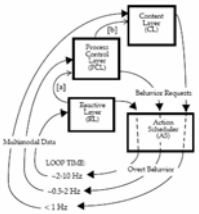
- Définitions
  - une ontologie est une spécification explicite d'une conceptualisation [Gruber, 1993]
  - une ontologie est une spécification formelle d'une conceptualisation partagée [Borst, 1997]
- Apprentissage de la signification par l'usage
  - Généralisation des thésaurus du TALN
  - Ingénierie des connaissances et intégration d'information: WEB sémantique, indexation, résumé de documents multimédia
- Usages
  - Interface d'interrogation de serveurs d'information :
    - vocabulaire structuré servant de support à l'expression de requêtes
  - Interface d'intégration de sources de données hétérogènes :
    - connexion entre différentes sources accessibles

CEA - DIST, 2005

14

Les modèles d'interaction (1)

- Ymir Turn Taking Model (YTTM, Thorisson)
  - Gestion des tours de parole
  - Écho de regulation: « back channeling »
  - Modèle en 3 niveaux: Contenu, Contrôle de processus, Réactif
  - Appliqué à Gandalf, MIT



- `http://alumni.media.mit.edu/~kris/gandalf.html`
- Thórisson, K.R. (1999) *A mind model for multimodal communicative creatures and humanoids.* International Journal of Applied Artificial Intelligence, **13**(4-5): p.449-486.
- Thórisson, K. (2002) *Natural turn-taking needs no manual: computational theory and model from perception to action*, in *Multimodality in language and speech systems*, B. Granström, D. House, and I. Karlsson, Editors. Kluwer Academic: Dordrecht, The Netherlands. p. 173–207.

15

Les modèles d'interaction (2)

- Ymir Turn Taking Model (YTTM, Thorisson)
  - Distinction entre
    - États
    - Percepts, actions
    - Conditionnement mutuel
  - Lexique de séquences motrices
    - Ex: détourner le regard pour réfléchir, regarder dans les yeux pour affirmer, etc
  - .

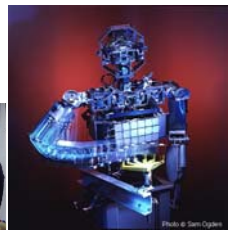
Table 6. Overt Decision Modules used in Gandalf's Reactive Layer. These modules control Gandalf's reactive behavior. For discussion see Section 3.3.4.

Show-Im-taking-turn 38 EL 5000 msec BehaviorRequest: Show-im-taking-turn PRE-CONDS: (Intake-turn = T) RESTORE-CONDS: (Intake-turn = F)	**Show-Im-giving-turn** 43 EL 2000 msec BehaviorRequest: Show-Im-giving-turn PRE-CONDS: (Log-w-turn = T) RESTORE-CONDS: (Log-w-turn = F)
Show-I-know-other-is-addressing-me-1 39 EL 200 msec BehaviorRequest: Smile PRE-CONDS: (Im-attending-speech-act = T) (Other-is-turned-to-me = T) RESTORE-CONDS: (Other-is-turned-to-me = F)	**Show-Im-giving-turn-2** 44 EL 2000 msec BehaviorRequest: Show-Im-giving-turn PRE-CONDS: (Log-w-turn = T) RESTORE-CONDS: (Log-w-turn = F)
Show-I-know-other-is-addressing-me-2 40 EL 200 msec BehaviorRequest: Eyebrow-quirt PRE-CONDS: (AND (Other-is-saying-any-name = T) (Other-is-turned-to-me = T) (Other-is-facing-me = T)) RESTORE-CONDS: (Other-is-turned-to-me = F)	**Show-Im-listening** 45 EL 200 msec BehaviorRequest: Brows-in-passive-shape PRE-CONDS: (AND (Other-is-saying-any-name = T) (Other-is-turned-to-me = T) (Other-is-facing-me = T)) RESTORE-CONDS: (Other-is-turned-to-me = F)
Initialize-dialogue 41 EL 1000 msec BehaviorRequest: Face-neutral PRE-CONDS: (Dialog-On = F) RESTORE-CONDS: (Dialog-On = T)	**Show-I-know-other-is-not-addressing-me** 46 EL 1000 msec BehaviorRequest: (Turn-At-Work-space) PRE-CONDS: (AND (Dialog-On = T) (Other-is-turned-to-me = F)) RESTORE-CONDS: (Other-is-taking-turn = T)
Look-puzzled-during-awkward-pause 42 EL 1000 msec BehaviorRequest: Look-puzzled PRE-CONDS: (AND (other-is-turned-to-me = T) (other-is-facing-me = T) (*time-since* (Other-is-turned-to-me = 400)) RESTORE-CONDS: (Other-is-turned-to-me = F)	**Look-aloof** 47 EL 1000 msec BehaviorRequest: Look-aloof PRE-CONDS: (AND (Other-is-turned-to-me = T) (Other-is-facing-me = T) (*time-since* (Other-is-facing-me = 200) (Dialog-On = T) (Other-is-speaking = F) (*topic-knowledge-system* Strange-Speech-Data = F)) RESTORE-CONDS: (Other-is-turned-to-me = F)

16

Les modèles d'interaction (3)

- ETOM (Brooks, Scassellati)
  - Implémentation d'une TOM pour robot
  - COG / Kismet
  - Apprentissage par imitation
  - Gestion de l'attention & empathie



- `http://www.ai.mit.edu/projects/humanoid-robotics-group/`
- Scassellati, B. (2001) *Foundations for a theory of mind for a humanoid robot*, in *Department of Computer Science and Electrical Engineering*. MIT: Boston - MA. 174 pages.

17

Les modèles d'interaction (4)

- Analyse/synthèse multimodale de scènes; identification/suivi des intervenants, modèles de comportement: rôle, relations sociales, etc
- Gestion des tours de parole, deixis, etc.

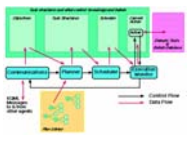
`http://www.plyojump.com/cant_jump_yet.html`

ROBITA, Waseda University

18

Les langages

- Beaucoup de développement ad hoc
- Peu de langages génériques
- SOAR (U. Michigan)
  - Sur Tcl/Tk
  - `http://sitemaker.umich.edu/soar`
- URBI
  - Interface Universelle pour Systèmes Interactifs
  - Utilisé pour AIBO
  - `http://www.urbiforge.com/index.php`
- Et évidemment les systèmes multi-agents
  - RETSINA (CMU)
  - `http://www.cs.cmu.edu/%7Esoftagents/retsina_agent_arch.html`
- Pour la parole, extensions de XML
  - IBM, Motorola and Opera Software (VoiceXML), Microsoft SALT



Les lieux de recherche

19

- Projets européens
  - COMIC COnversational Multimodal Interaction with Computers
    `http://www.hcrc.ed.ac.uk/comic`
- Réseaux d'excellence
  - SIMLILAR Interaction multimodale
    `http://www.similar.cc`
  - HUMAINE Emotions & interaction multimodale
    `http://emotion-research.net/`
- Groupes structurés
  - Affective computing
    `http://www.bartneck.de/link/affective_portal.html`

Plan de l'exposé

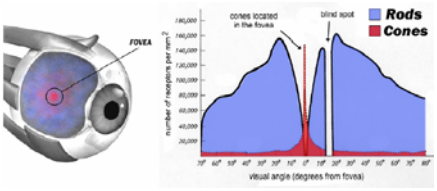
20

- Systèmes d'interaction
  - Gestion du dialogue
  - Analyse de scène
  - Interaction
- Le cas de la gestion du regard
  - Regard et cognition
  - Regard et technologie
- Objets d'intérêt et conversation face-à-face
  - Atteinte de cible en deux bandes…
  - Agent conversationnel

Vision fovéale & périphérique

21

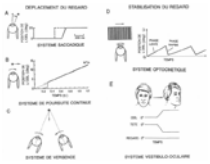
- Bâtonnets & cônes
  - Cônes: analyse colorimétrique
  - Bâtonnets: analyse cinématique



Mouvements occulaires

22

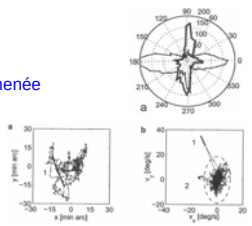
- Deux systèmes visuels
  - le système rétino-géniculo-strié (RGS) dédié aux fonctions perceptives (apprentissage visuel discrimination, reconnaissance des objets : formes, couleurs, estimation des distances, etc.)
  - le système rétino-géniculo-tectal (RGT) ayant la capacité de détecter les déplacements afin de déclencher rapidement les mouvements orientés vers la cible détectée… qualifié par certains de « détecteur de mouche » chez les amphibiens (Barlow, 1953; Mach, 1959). Le RGT est donc impliqué dans le contrôle des habiletés motrices (par exemple : préhension, saisie, attraper etc.).
- Temps de réaction
  - Simple (réaction à spot) = 190ms
    - De l'apparition, à l'appui touche
    - Influencé par intensité (Fischer and Boch 1983; Fischer and Breitmeyer 1984)
  - Choix multiple = 350ms



Fixations et saccades

23

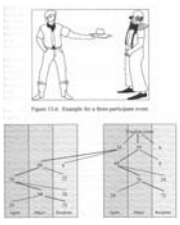
- Saccades
  - Attention: compétition entre exogènes (événements saillants de la scène, apparition abrupte..) et endogènes (buts fixés par l'observateur)
  - Vitesse max: 500°/s
  - 30 à 120 ms, 1 à 40°
  - 150,000 par jour [Abrams, 92]
  - Exploration écran: 7 m/mn
  - Coordination binoculaire: vergence
- Fixations
  - Analyse de la cible de l'attention amenée dans la fovéa de la rétine
  - 200-300 ms, 3/5 par seconde
  - Mouvements miniatures: trémolo, dérive et micro-saccades (compensation à l'adaptation rétinienne)

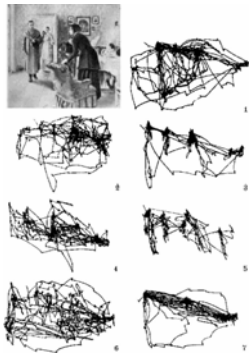


Demande cognitive

24

- Saccades exploratoires
  - Dépendent de la demande cognitive (Yarbus, 1967: général, environnement, âge, activité, vêtements, positions, durée de l'absence)
  - Coordination avec parole

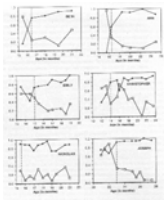


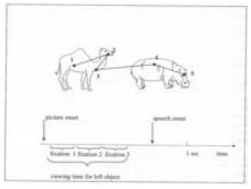


Regard et production de parole

25

- Pointage
  - Aspects développementaux
  - Redondance/complémentarité
  - Multimodalité: degrés de liberté des segments
- Regard et parole
  - Lecture
  - Analyse de scènes

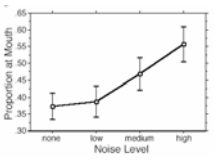


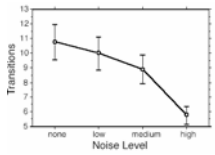


Regard et perception de parole

26

- Redondance et complémentarité audiovisuelle
  - Scrutation du visage
  - Yeux – bouche
- Fonction
  - Tâche
  - Attention
- Références
  - Lee, S.P., Badler, J.B., and Badler, N. (2002) *Eyes alive.* ACM Transaction on Graphics, **21**(3): p. 637-644.
  - Langton, S., Watt, J., and Bruce, V. (2000) *Do the eyes have it ? Cues to the direction of social attention.* Trends in Cognitive Sciences, **4**(2): p. 50-59.

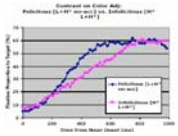




Regard et perception de parole

27

- Deixis prosodique
  - Séquences d'actions
  - Gestion du rhème/thème
    - Felicitous L+H* (e.g. First, hang the green drum. Next, hang the ORANGE drum.) compared to infelicitous L+H* (e.g. Next, hang the orange DRUM)
  - Bénéfice de 200ms pour emphase correcte
  - Ito & Speer (2005) Ohio U.

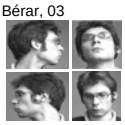


Regard de l'autre

28

- Stimulus exogène particulier
  - Module de base de la Théorie de l'espr (ToM)
    - Détecteurs d'intentionnalité et de la directio de l'œil / Mécanisme d'attention partagée Module de théorie de l'esprit (Baron-Coher Leslie et al. 1985; Baron-Cohen 1995)
    - Signal d'alerte (me regarde)
    - Objet d'intérêt (champ vision commun)
    - Attire attention (signal tierce)
  - Base de l'apprentissage du monde et du langage
    - Where/what multimodal
- Biaise stimulus exogène
  - Langton & Bruce
    - avantage de 20ms si lieux cohérents
    - + important en haut/bas vs. gauche/droite
  - Compétition

Langton & Bruce 99; Bérar, 03



Regard et perception de parole

29

- Deixis multimodale
  - Jeu de cartes
  - Indications congruentes vs. Incongruentes
  - Bénéfice de 200ms pour indications congruentes; moins de cartes scrutées
  - Bénéfice de 200ms supplémentaire en multimodal
  - Raidt et al (2006) ICP

Plan de l'exposé

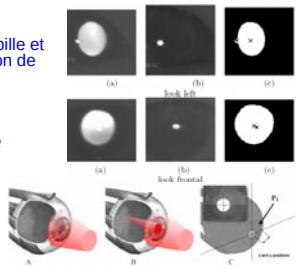
30

- Regard et cognition
  - Analyse de scène
  - Langage
- Regard et technologie
  - IHM
- Objets d'intérêt et conversation face-à-face
  - Atteinte de cible en deux bandes…
  - Agent conversationnel

Suivi du regard

31

- Lumière réfléchie
  - Écart entre centre de la pupille et celui de la première réflexion de Purkinje
  - Calibration
    - Entre écart et [x,y] écran
- Champ électrostatique
- Capture
  - Video
  - Lentilles de contact
    - miroirs
    - aimants
- EMG muscles



Occulomètre (1)

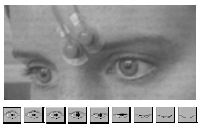
32

- Invasif ou non-invasif

Regard et souris (1)

33

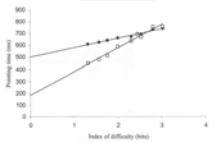
- Regard et point d'intérêt
  - « Eye-mind hypothesis »: regard correspond au sommet de la pile des opérations mentales
- Problèmes
  - « Midas Touch Problem»: résoudre l'opération de sélection
  - « One-way Zoom Problem »: rendre réversible l'opération de zoom automatique sur le point d'intérêt
- Solutions
  - Clignement, fronçage [Surakka 2001: EMG surface, corrugator supercilii], temps de fixation (150-250ms)
  - EyeCons icônes attachés à tout objet sélectionnable que l'on doit regarder un certain temps
  - MAGIC [Zhai 1999] regard redéfinit la position du curseur, IGO [Salvucci 2000] accroissement contextuel de la zone de sélection



Regard et souris (2)

34

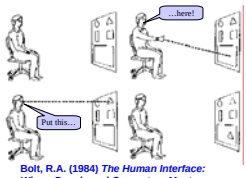
- Retours
  - Test signature: impossible de « tracer » sans retour (Humphrey, Johnx2, Carol)!
  - curseur regard (peut affecter attention: tremblements...)
  - objet sélectionné, beep…
- Evaluations
  - Loi de Fitts: MT=a+b.log$_2$(A/W+1), A distai & W taille, IP=1/b indice de performance ou log$_2$(A/W+1)/I



Objets d'intérêt

35

- Objet et interlocuteur
  - Un ou deux?
- Agent conversationnel
  - Sélection d'un « espace d'intérêt »
  - Référence, ange gardien…
- Dynamique de la conversation
  - Objets virtuels/réels
  - Locuteur(s)/agent(s)
- Cognition spatiale
  - Gestion de la vergence
  - Multimodalité

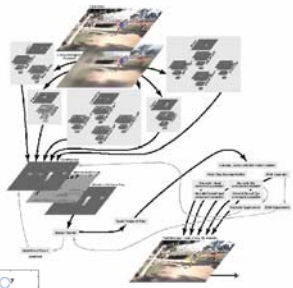


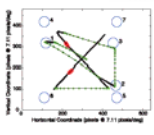
Bolt, R.A. (1984) *The Human Interface: Where People and Computers Meet.*

Couplage analyse/synthèse scène

36

- Scrutation de scènes
  - Analyse/synthèse multimodale de scènes
  - Couplage perception/action
  - Perméabilité cognitive
    - Saillance
    - Pertinence
    - Attention



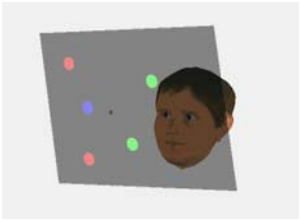


Itti, L., N. Dhavale and F. Pighin (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. SPIE 48th Annual International Symposium on Optical Science and Technology, San Diego, CA

Couplage analyse/synthèse scène

37

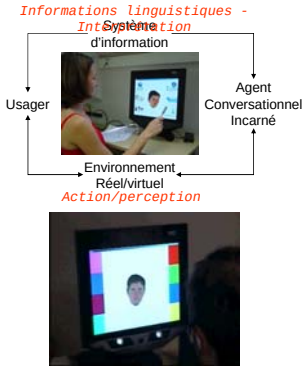
- Scrutation de scènes
  - Analyse/synthèse multimodale de scènes
  - Couplage perception/action
  - Perméabilité cognitive
    - Saillance
    - Pertinence
    - Attention
  - Couplage action/perception
    - Regard pilote l'analyse



Projets

38

- Communication face-à-face
- Intelligence ambiante
  - Deixis multimodale
    - Tâches simples
    - Objets virtuels
    - Regard, parole, souris
  - Analyse de l'attention mutuelle
    - Conversation humaine face-à-face
    - Deux oculomètres
  - Magicien d'Oz
    - Deixis multimodale + complexe
    - Regard, parole sur complice



Conclusions & perspectives

39

- Termes de la gestion dynamique du regard
  - Endogène: dynamique du dialogue
  - Exogène: interlocuteur/environnement
  - ? Négociation
- Attention mutuelle
  - Conséquences
    - Robustesse de la communication
    - Croyance en l'information
- Multimodalité
  - Vers une communication expressive…