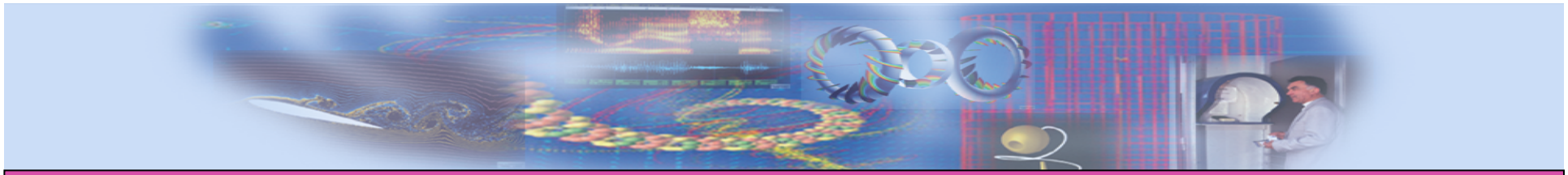


Laboratoire de Mécanique et d'Informatique pour les Sciences de l'Ingénieur

Interactions Voix Parole

Rôle et estimation quantitative de la force de voix

Jean-Sylvain Liénard

Limsi-Cnrs, Orsay

Atelier Sciences et Voix, Grenoble, 22-10-14

1. Voix et Parole

2. Effort Vocal et Force de Voix

3. Objectifs

4. Données voyelles

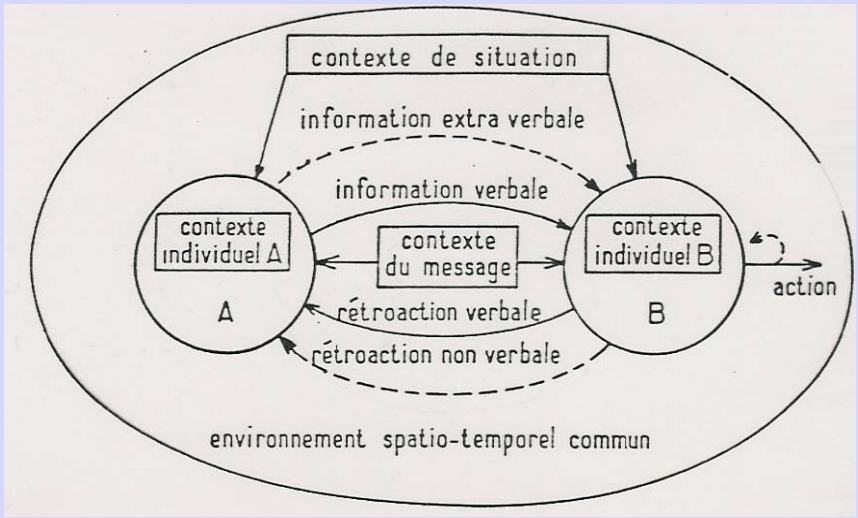
5. Analyse spectrale

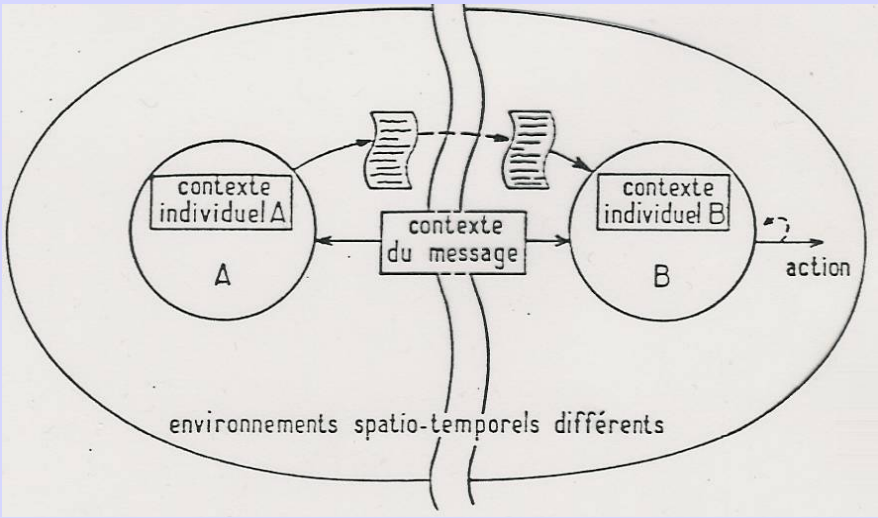
6. Analyse discriminante

7. Interactions

1. Voix et Parole

Communication parlée, communication écrite





- La parole est une activité située
- la parole n'est pas de l'écrit oralisé

Variabilité de la parole

Pour un même contenu linguistique, le signal varie dans toutes ses dimensions

- selon le genre vocal
- selon l'accent (variations régionales, sociologiques, individuelles)
- selon l'expression
- selon l'état physique (fatigue) ou psychologique (tonicité), etc.

mais aussi:

- selon la situation dans laquelle se trouvent les interlocuteurs
  ---> rôle (sous-évalué) de l'Effort Vocal EV

Dans le signal vocal toutes les sources de variabilité sont mélangées:

---> la recherche d'invariants absolus est illusoire
---> étudier les interactions entre tous les aspects du signal

Effets de la variation d'Effort Vocal

sur la fatigue vocale, le forçage vocal

sur l'identification du locuteur

- difficulté à reconnaître l'identité d'un locuteur à partir d'échantillons de sa voix émis selon divers degrés d'EV
- conséquences en identification automatique, authentification judiciaire...

sur les systèmes de reconnaissance automatique

- augmentation exponentielle du volume des données d'pprentissage

2. Effort Vocal et Force de Voix

Problématique de l'Effort Vocal

- reflète la prise en compte de l'interlocuteur et de la situation
  - le parleur ajuste son effort pour assurer intelligibilité et expression

- varier EV --> déformations du signal
  - variabilité pour la parole
  - information pertinente pour la voix

- classiquement: 3 nuances, + vx chuchotée et vx criée

- vx faible amplifiée ≠ vx forte

Effort Vocal et situation de communication

Situations naturelles vs situations artificielles (micro, hp, tél)

Le parleur ajuste son EV de façon à faire parvenir son message à l'auditeur auquel il s'adresse:

- le message n'est pas forcément de nature linguistique
- le parleur tient compte
  - de la distance (connue, supposée) de l'auditeur
  - des conditions acoustiques: bruit, réverbération
  - des capacités auditives de l'auditeur
- cette adaptation est largement inconsciente

Portée de la voix, modes de voix

Niveau sonore: de 30 à 120 dB spl

Portée: de qq cm à qq centaines de m

Atténuation (distance x 2): de -6 dB à -3 dB selon sol

Modes

- chuchoté (ch faible, ch fort), proximité (d<1 m), dégradé
- conversationnel (faible, neutre, fort) 0,5<d<15 m
  dynamique usuelle: 20 à 30 dB
- crié d>15 m, dégradé

Comment définir la voix neutre (normale, modale) ?

Les systèmes électroacoustiques modifient l'usage des modes de voix

Une mesure objective de l'EV: la Force de Voix

- ne pas confondre intensité émise par le parleur, et intensité reçue par l'auditeur

- EV notion qualitative: pas de méthode de mesure

- intensité acoustique (sonore niveau en dB à 1m) ok mais préciser sur quelle durée et à quel instant

- loudness (sonie) tient compte de la perception, mais varie avec la distance à la source

---> Force de Voix = intensité physique, sur fenêtre gaussienne de durée 50ms, prélevée au maximum observé sur le noyau vocalique. Pondération A

3. Objectifs et méthode

Objectifs à long terme

- démêler les divers aspects (descripteurs, dimensions) de la voix parlée ordinaire

- expliciter leurs relations et lois de variation

- étayer un modèle de la perception auditive

Actuellement

- Voyelles du français émises isolément:
- extraire du signal des indices liés à ses principaux aspects:
  - phonétiques: indices, traits, identité vocalique
  - force de voix fdv
  - échelle formantique, genre vocal
  - caractérisation du locuteur
  - conditions d'enregistrement

Méthode

- Partir de bases de données de voyelles isolées, dont la fdv soit connue ainsi que les autres aspects

- Etudier les variations spectrales obtenues en faisant varier un seul descripteur, les autres étant maintenus constants

Difficultés

- bd étalonnée
- analyse / représentation du signal
- élaboration des indices de chaque descripteur
- validation

4. Bases de données

Nécessité d'un calibrage de la FDV

Pour chaque séquence (syllabe) prononcée il faudrait connaître la fdv, pour la mettre en relation avec le signal

Ou au moins connaître: distance microphone-locuteur et gain de enregistrement

Habitudes de prise de son: maximiser rapport S/B. L'information de FDV est perdue

De plus: pas de variation systématique de l'EV,
hormis vx (modale, + fort, - fort), ou vx Lombard

---> pas de BD "reconnue", utilisable pour ce pb

BD Corenc CRC

voix "conversationnelle" en situation

le locuteur, assis, dans une pièce meublée
répète interactivement une phrase, puis les 12 voyelles proposées par l'opérateur,
celui-ci étant placé successivement à 0,4 1,5 et 6 m

12 voyelles 13 locuteurs (6 h, 7 f), 3 conditions de distance
---> 720 "tokens", 20 séries en 3 sessions

microphone à 30 cm (LEM DO 21), enregistreur K7

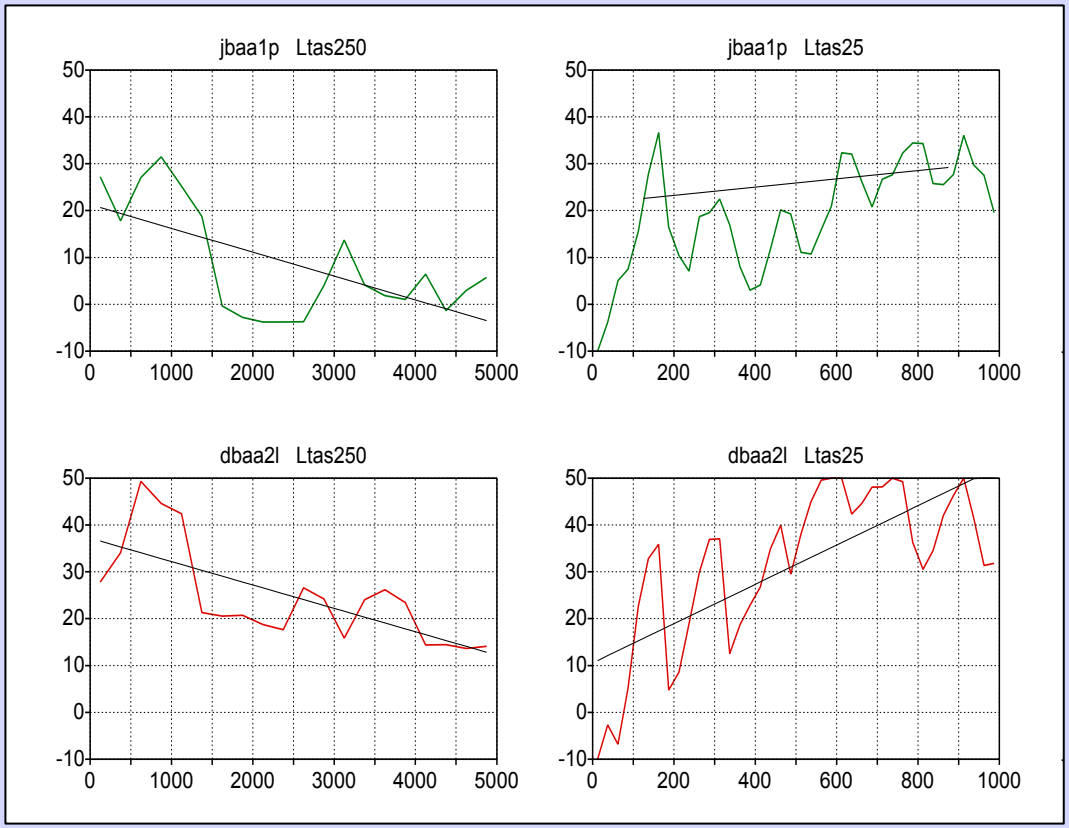
pas de calibration du niveau sonore, mais conditions d'enregistrement constantes d'une session à l'autre

- Exemple 1: sona d'un extrait de Corenc, distance l: phrase opérateur à 6 m + répétition sujet, puis 5 voyelles

- Exemple 2: même chose, distance p

2 tokens du corpus CRC

- même voyelle [a], même F_0 (152 Hz), loc et EV différents
  1. présentation au même niveau sonore --->
  2. présentation et analyse à prise de son constante (ci-dessous)



- en haut: voix féminine, EV faible 37 dBA
- en bas: voix masculine, EV moyen 55 dBA

- à gauche: Ltas à bande large (250 Hz), éch lin 0-5 kHz
- à droite: Ltas à bande étroite (25 Hz), éch lin 0-1 kHz

Analyse et représentation du signal

Banc de filtres Bark

fenêtre gaussienne 50 ms

intérêt d'une échelle Bark, ni lin ni log

Fréq Hz	Att dB
≥ 1000	0
500	-3
250	-8
125	-15
63	-23

Mesure de l'intensité

flat dB, sonie, ou pondérée dBA

Sélection de la trame la plus intense dans le noyau vocalique

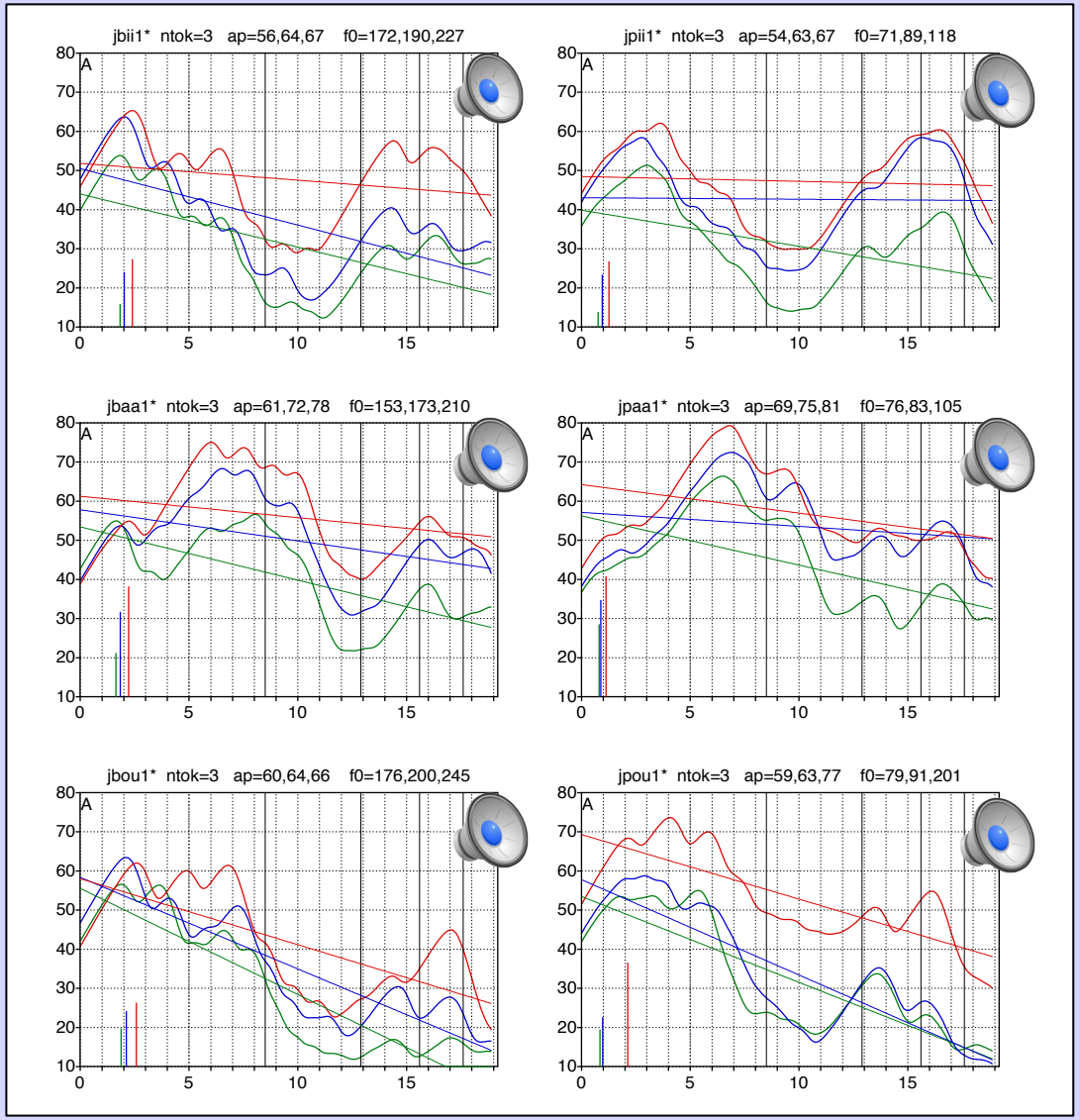
mesure de F_0 et de la FDV (DB ou dBA)

Indices acoustiques

pour visu: spectre Bark

pour analyses: F_0 et 18 coefficients Bark {b1, b18} normalisés p.rapp. au maxi 50 dB ou dBA

Spectres individuels, données CRC



Indices de la FDV ?

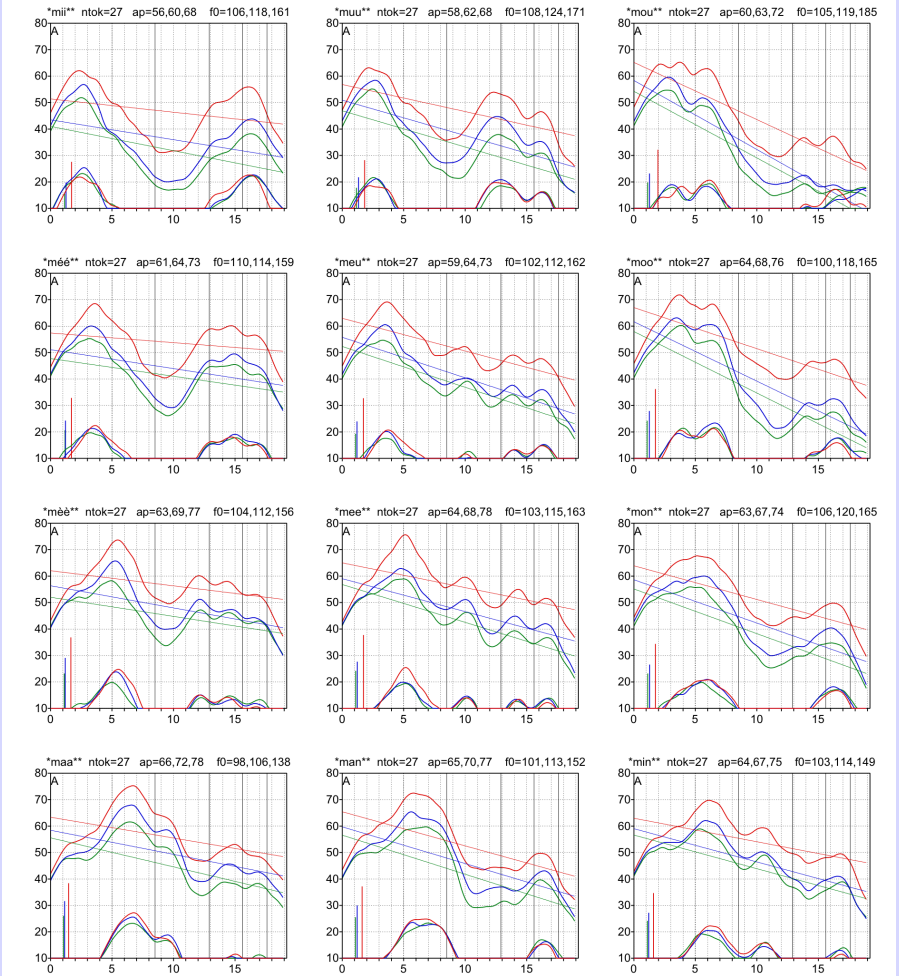
- pente globale
- pente dans les basses fréquences
- F_0

Problème

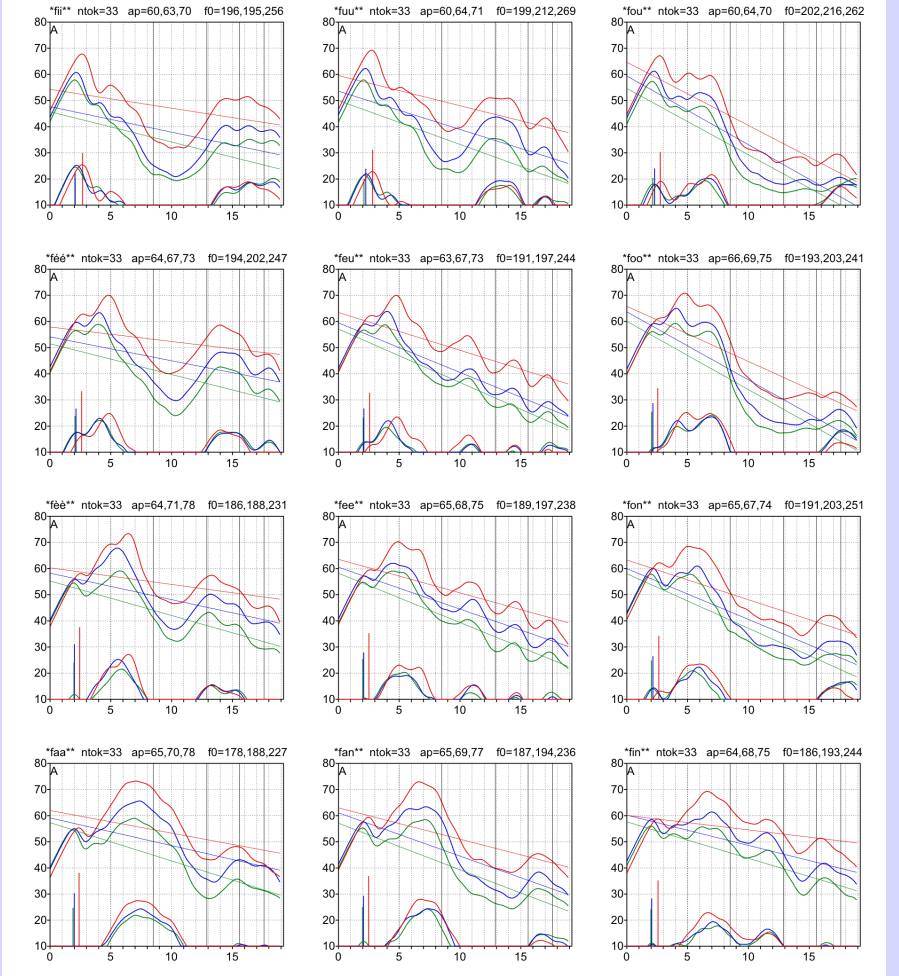
- les indices dépendent du locuteur et de la voyelle

- # Exemples spectres Bk moyens toutes voyelles vx h et vx f

Système moyen locuteurs m, AVEC pondA



Système moyen locutrices f, AVEC pondA



BD José JAE

pièce insonorisée
microphones de qualité (Schoeps, B&K) à 20 cm du locuteur
enregistrement numérique 24 bits (dynamique > 90 dB)

Chaque locuteur, partant d'un niveau qu'il estime "neutre"
répète une voyelle à intensité croissante jusqu'à vx criée
puis redescend au niveau "neutre"
puis recommence vers vx faible, voire chuchotée
et remonte au niveau "neutre"

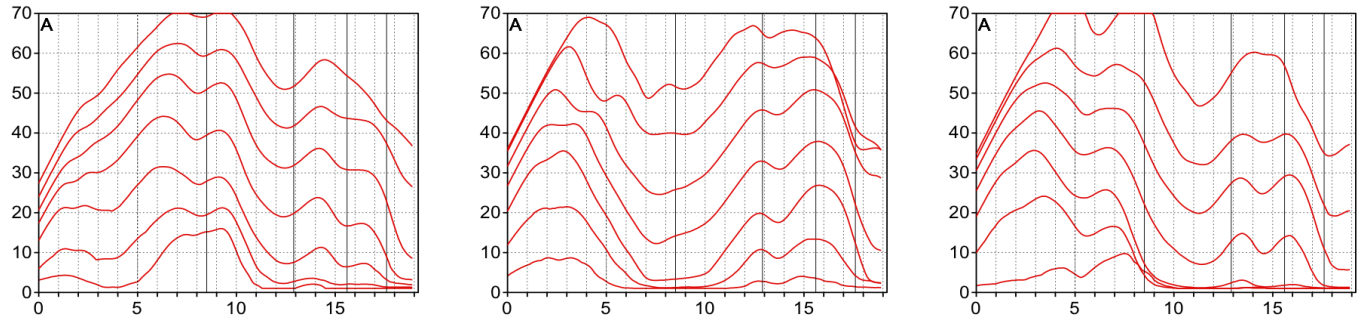
3 voyelles [a, i, u], 17 locuteurs (7 h, 7 f, 3 e)
signal d'étalonnage du microphone, mais quelques erreurs

---> grandes différences avec Corenc:
on est bien au delà de la vx conversationnelle
répétitions de la même voyelle
nb de tokens variant selon le locuteur

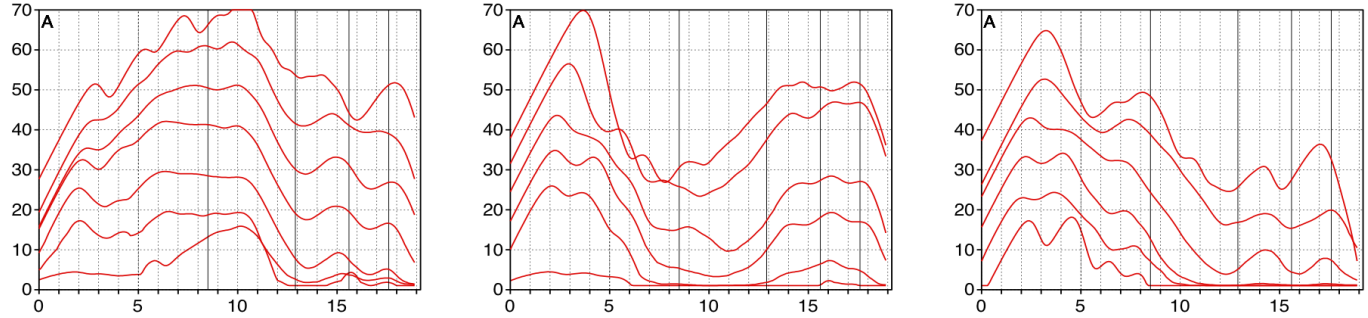
BD José: locuteur ch, voyelle [a], série montante puis descendante

BD José: locuteur ch, voyelle [a], série descendante puis montante

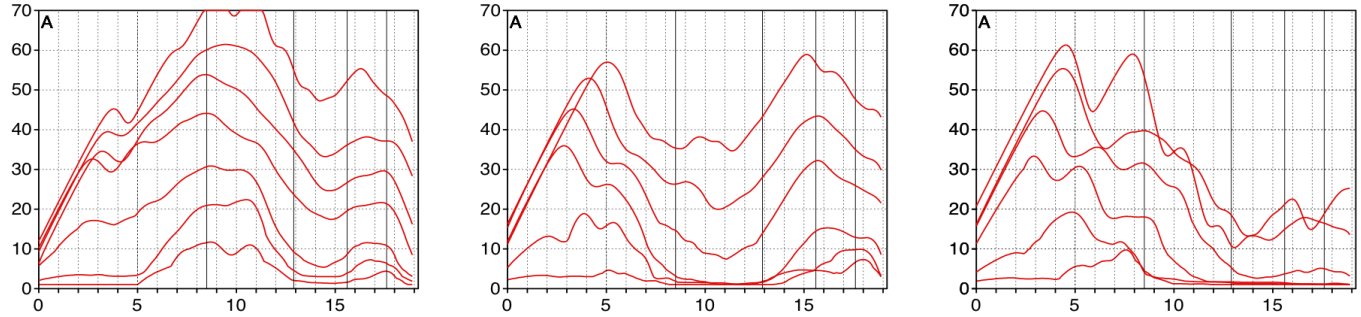
Toutes les voix homme de José



Toutes les voix femme de José



Toutes les voix enfant de José



5. Analyse Discriminante

- ensemble de tokens, chacun caractérisé par
  - sa catégorie vocalique (parmi 12 pour CRC, 3 pour JAE)
  - 19 indices acoustiques (F_0+18 Bk)

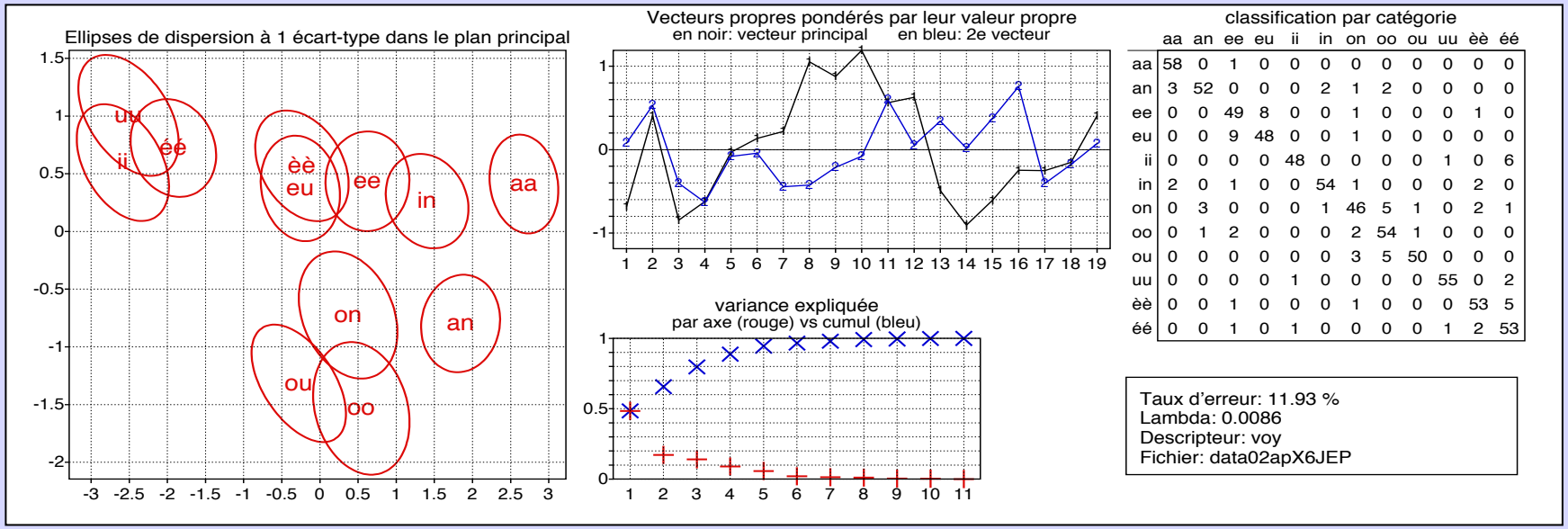
- l'AD apprend un classifieur (discriminant)
  - variance intra catégories minimale
  - variance inter catégories maximale

- Test
  - nouveaux tokens classifiés à partir des seuls indices acoustiques
  - erreur: taux de tokens mal classés

    ou critère statistique (lambda de Wilks)

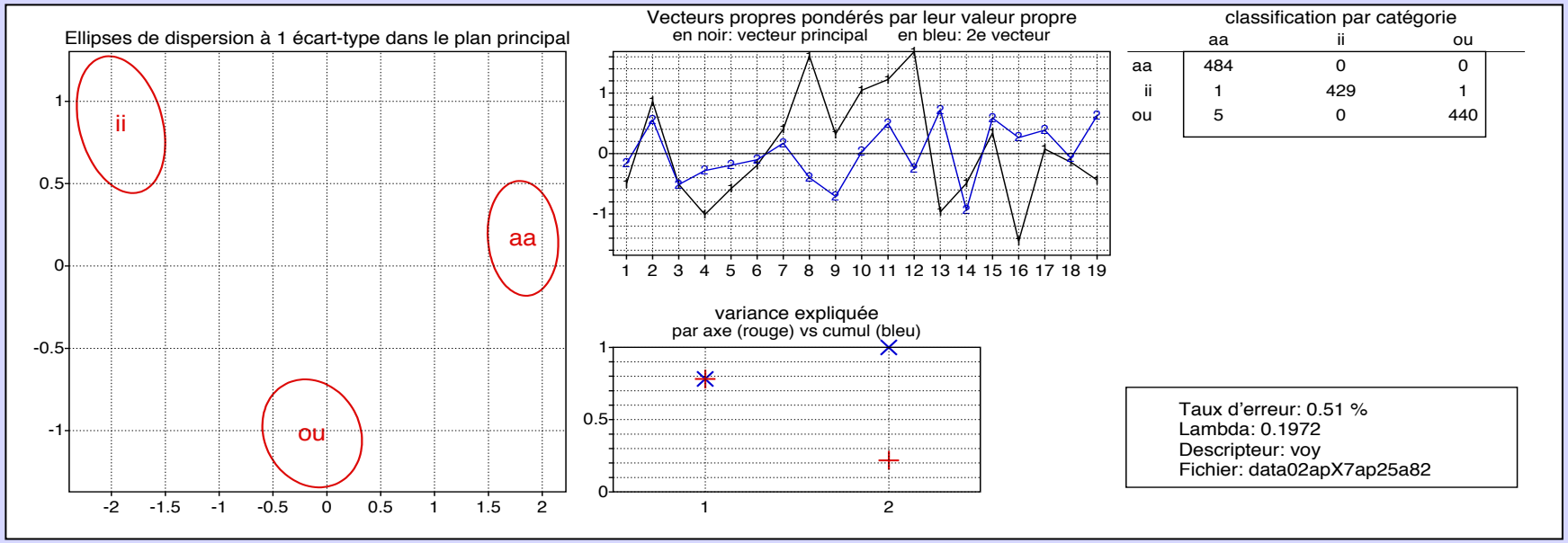
  - autocohérence (classification) vs validation croisée (généralisation)

Classification des voyelles, données Corenc



- peu d'erreurs (12%), voy isolées, nasales
- 6 axes suffisent (sur 11)
- "triangle vocalique" ds plan principal

Classification des voyelles, données José



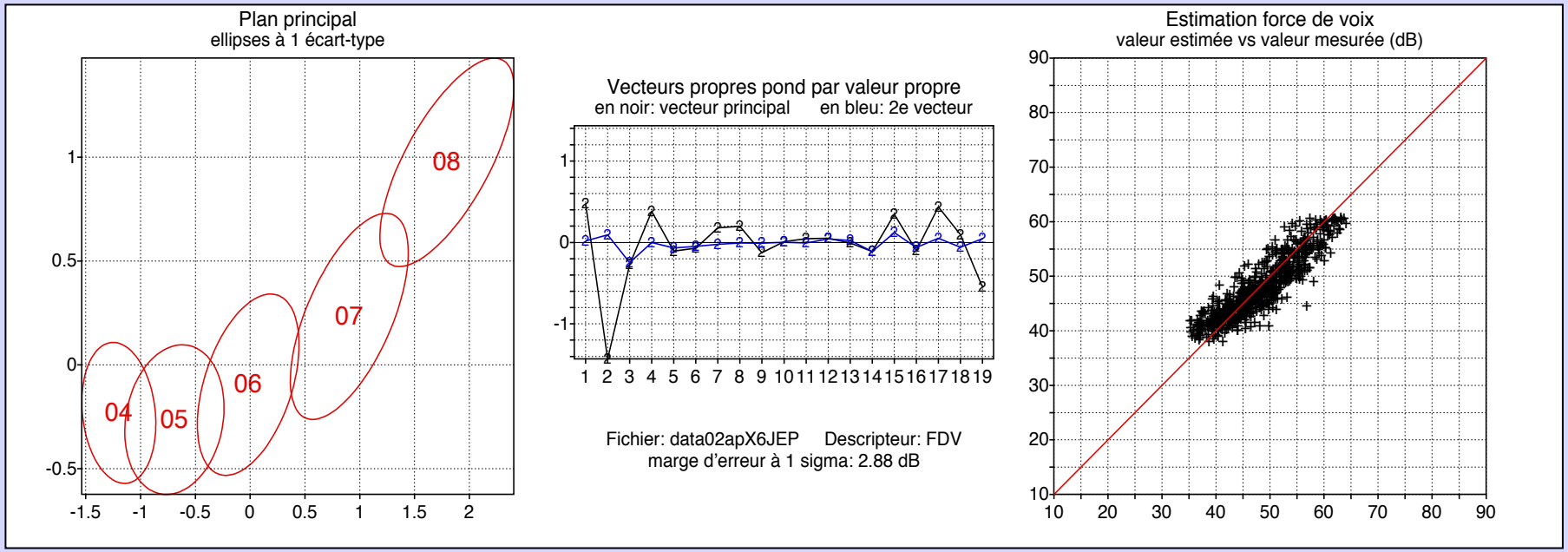
- erreurs quasi-nulles:
- 3 voy, grande dynamique
- profil des vecteurs propres

Analyse Discriminante pour la FDV

- Même processus, mêmes indices acoustiques F_0+18Bk
  catégories "voyelles" remplacées par "degrés de FDV" (échelons de 6 dBA)

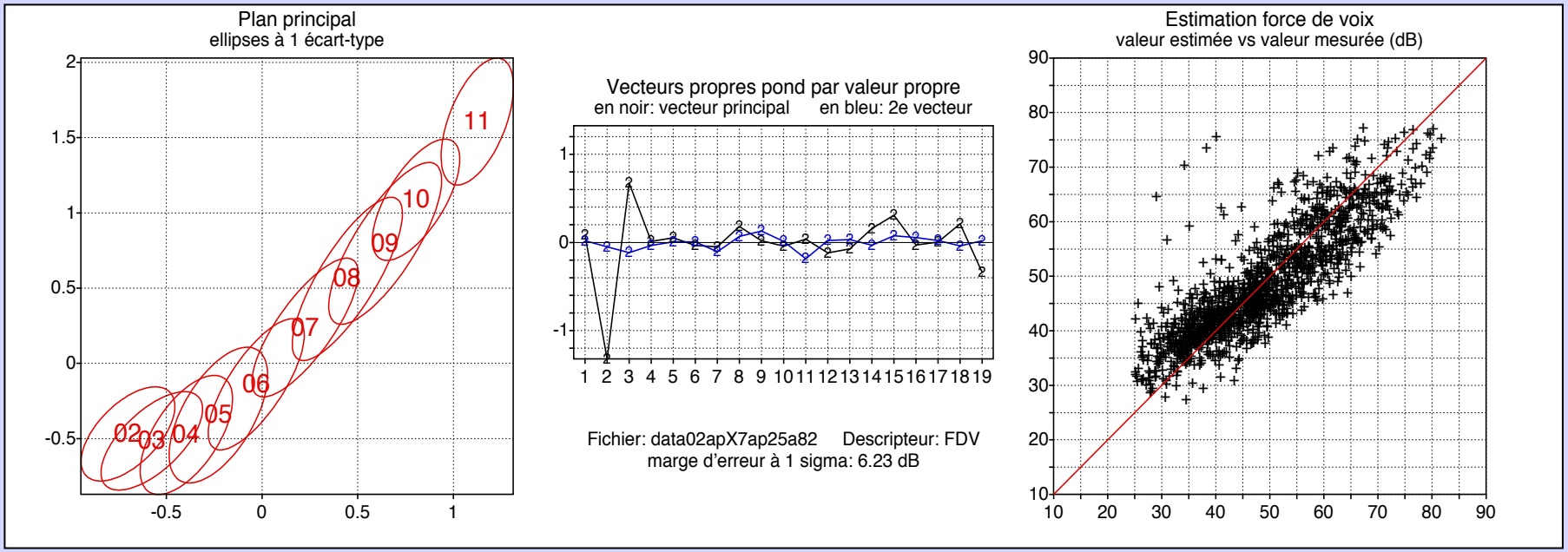
- Test
  indices produisent le degré de FDV le plus proche
  FDV continue --> interpolation
  erreur: différence |(FDV estimée)-(FDV mesurée)|
  sur l'ensemble de test: écart-type en dB ou dBA

Classification des degrés d'EV, données Corenc



- Non-linéarité pour voix faibles
- Corrélation (FDV estimée, FDV mesurée) > 90%
- marge d'erreur de l'ordre de 3 dBA pour dynamique 30 dBA

Classification des degrés d'EV, données José



- même allure générale, dispersion plus grande
- marge d'erreur de l'ordre de 6 dBA pour dynamique 60 dBA

Classification des degrés d'EV, à partir d'autres indices acoustiques

Au lieu de 19 indices (F0+18Bk), autres mesures

- sur le spectre: harmoniques h1-h10, pentes, centres de gravité
- sur le signal temporel: jitter, autocorrélation

---> résultats comparables (marge d'erreur env. 3 dB) avec moins d'indices

mais pb de sélection: choisir indices en rapport avec les études du signal glottique

6. Interactions Voix-Parole

(1) On suppose la voyelle connue, quid de la FDV ?

Marge d'erreur sur la FDV en DBA	voy inconnue	voy connue
CRC (12 voy, 720 tokens, dyn 30 dBA)	2,97	2,12
JAE (3 voy, 1395 tokens, dyn 60 dBA)	6,51	4,87

--> Connaître la voyelle facilite l'estimation de la FDV

(2) On suppose la FDV connue (3 à 4 grandes catégories), quid de la voyelle ?

Erreurs de classification voy% (lambda)	FDV inconnue	FDV connue
CRC (12 voy, moy consignes FDV: p, n, l)	12,5 (0,009)	9,0 (0,005)
JAE (3 voy, moy 4 tranches FDV de 12 dBA)	0,5 (0,202)	0,3 (0,131)

--> Connaître la FDV, même de manière grossière, facilite l'identification de la voyelle

7. Petit bilan

1. l'EV peut être indexé par une grandeur objective (FDV), calculable à partir d'indices spectro-temporels

2. le nombre de degrés d'EV encodés dans le signal oral est bien supérieur à 3

3. la catégorisation de la FDV et celle de la voyelle sont interdépendantes: connaître l'une facilite l'identification de l'autre

4. dissymétrie: "EV plus fort" entraîne "F_0 plus élevée"
   "F_0 plus élevée" n'entraîne pas "EV plus fort"

Quelques publications

AUGUSTE-ETIENNE, J. (1999). "Etude d'un protocole d'enregistrement pour l'analyse du timbre de la voix", mémoire de recherche, ENS Louis Lumière, Noisy le grand.

BOERSMA, P. and WEENINK, D. (2012). "Praat: doing phonetics by computer" [computer program], version 5.3.32, retrieved 17 October 2012 from http://www.praat.org/.

BRUNGART, D., SCOTT, K. and SIMPSON, B. (2001) "The influence of vocal effort on human speaker identification", *Eurospeech*, Scandinavia.

D'ALESSANDRO, C. (2006). "Voice source parameters and prosodic analysis", in S. Sudhoff et al [Eds] *Methods in Empirical Prosody Research*, 63-87, Walter de Gruyter.

DOVAL, B., D'ALESSANDRO, C. HENRICH, N. (2006). "The spectrum of glottal flow models", *Acustica united with Acta Acustica*, 92:1026-1046, 2006.

FANT, G., LILJENCRANTS, J. and LIN, Q. (1985). "A four parameter model of glottal flow", STL-QPRS, 26(4):1-13, 1985.

GARNIER, M. (2007). "Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal", thèse de doctorat, université Paris VI.

HANSEN, J. and VARADARAJAN, V. (2009). "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition", IEEE tr. on ASLP, 17 (2), 366-378.

HANSON, H. (1997). "Glottal characteristics of female speakers: acoustic correlates", *J. Acoust. Soc. Am.* 101 (1), 466-481, 1997.

HENRICH, N., D'ALESSANDRO, C., DOVAL, B. and CASTELLENGO, M. (2005). "Glottal open quotient in singing: measurement and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency", *J. Acoust. Soc. Am.* 117 (3), 1417-1430.

HUBER, J.E., STATHOPOULOS, E.T., CURIONE, G.M., ASH T.A. and JOHNSON, K. (1999). "Formants of children, women, and men: the effects of vocal intensity variation", *J. Acoust. Soc. Am.* 106 (3), 1532-1542.

JUNQUA, J.-C. (1992). "The Lombard reflex and its role on human listeners and automatic speech recognizers", *J. Acoust. Soc. Am.* 93, 510-524.

LIENARD, J.S. and DI BENEDETTO, M.G. (1999). "Effect of vocal effort on spectral properties of vowels", *J. Acoust. Soc. Am.* 106 (1), 411-422.

LIENARD J.S. and BARRAS C. (2013). "Fine-grain voice strength estimation from vowel spectral cues", *InterSpeech*, Lyon.

Remerciements

A

Christophe d'Alessandro, Jean-Jacques Gangolf,
Philippe Boula de Mareüil, Albert Rilliard (LIMSI)
Nicolas Audibert (ILPGA, Paris)
Pierre Divenyi (CCRMA, Stanford)

et

le comité scientifique du LIMSI
les concepteurs du logiciel Praat

Quelques points à discuter ou approfondir

1. FDV et timbre de la voix
2. FDV et prosodie
3. comment obtenir des données calibrées ?
4. des voyelles isolées aux séquences
5. contrôler la FDV en synthèse ?
6. transformation de voix
7. FDV et reconnaissance automatique
8. perception: 7 ± 2 nuances de FDV ?
9. multicatégorisation
10. analyse acoustique: formants ou enveloppe spectrale ?
11. organisation de la perception
12. sur l'intensité des voyelles

Q1. FDV et timbre de la voix

la notion de FDV peut simplifier la notion de timbre de la voix,
sans s'y substituer

une partie des variations de timbre peut être attribuée
simplement à l'évolution de la FDV, quantifiable

Q2. FDV et prosodie

l'intensité est svt considérée comme secondaire pour la prosodie

pb de mesure ? proéminence ?

différencier accent de groupe (final, syntaxique) d'un accent de FDV ou éclat de voix ?

caractéristique: F0 et FDV corrélés

Q3. comment obtenir des données calibrées ?

étalonnage distance et prise de son

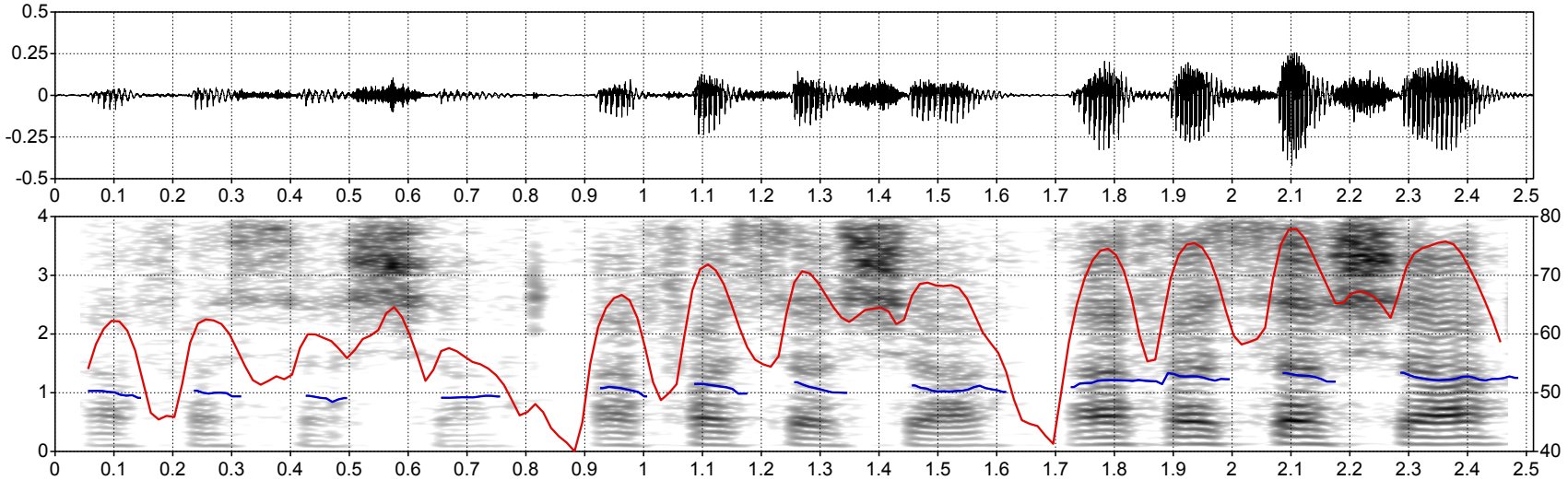
éviter parole de laboratoire (soutenue, contrainte) tout en
suggérant la variation souhaitée ?

système de régulation auditive:
rétroaction positive pour faire baisser la voix
ajout de bruit pour la faire monter ?

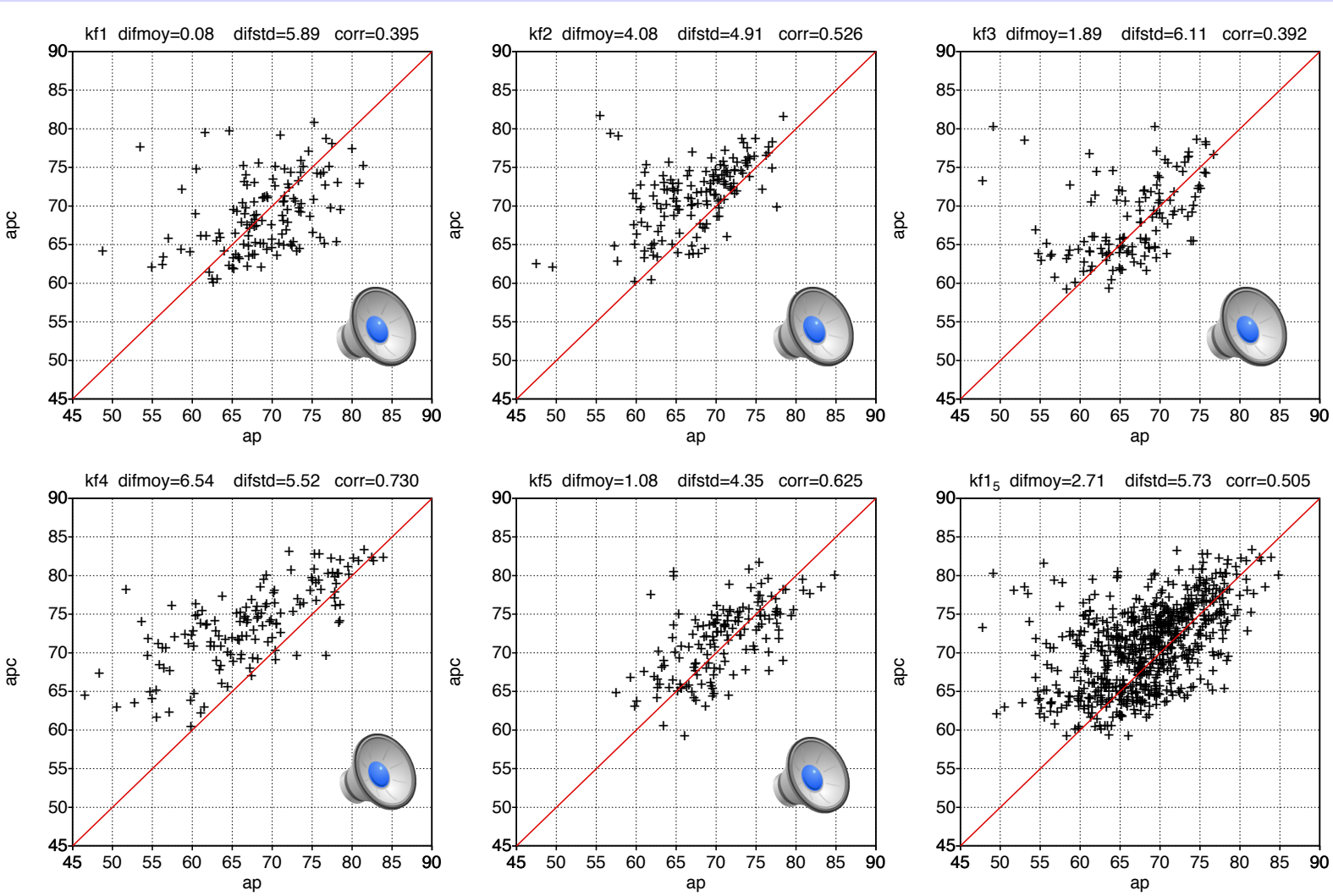
Q4. des voyelles isolées aux séquences

série de mesures sur
- les noyaux vocaliques
- les creux consonantiques (sil, fric, barre de voisement)
- identification des fricatives: intensité relative /voyelles

"afasacha", loc h



Essais sur voix Keele



Q5. contrôler la FDV en synthèse de parole ?

méthode:

il faut pouvoir commander la source de manière indépendante

pour augmenter la FDV:

- augmenter l'amplitude et la fréquence de l'onde glottique
- diminuer Oq
- question: Oq est-il le bon paramètre dans tous les cas ?
- voir aussi l'effet de la durée ouverte, qui commande Fg ?

manipulation impossible avec les systèmes de synthèse par concaténation

Q6. transformation de voix

Comme pour la synthèse, mais on part d'un enregistrement

il faut
- estimer la FDV
- extraire automatiquement les paramètres de source.
  - Déconvolution source/conduit vocal, pb difficile

potentiellement, intérêt pour
- reconstruire une voix avec la FDV adaptée à une nouvelle
  - situation de communication
- ajuster les FDV de divers enregistrements les uns par
  - rapport aux autres

Q7. FDV et reconnaissance automatique

Bénéfices envisageables en reconnaissance:

de la parole
- simplifier l'apprentissage
- fournir une info supplémentaire

du locuteur
- réduire la variabilité du locuteur enregistré dans des situations diverses

"diarisation" (quoi et qui)
- estimation de la distance de l'interlocuteur auquel le parleur s'adresse
- détection des tours de parole ?

Q8. perception: 7±2 nuances de FDV ?

cf Miller 1955

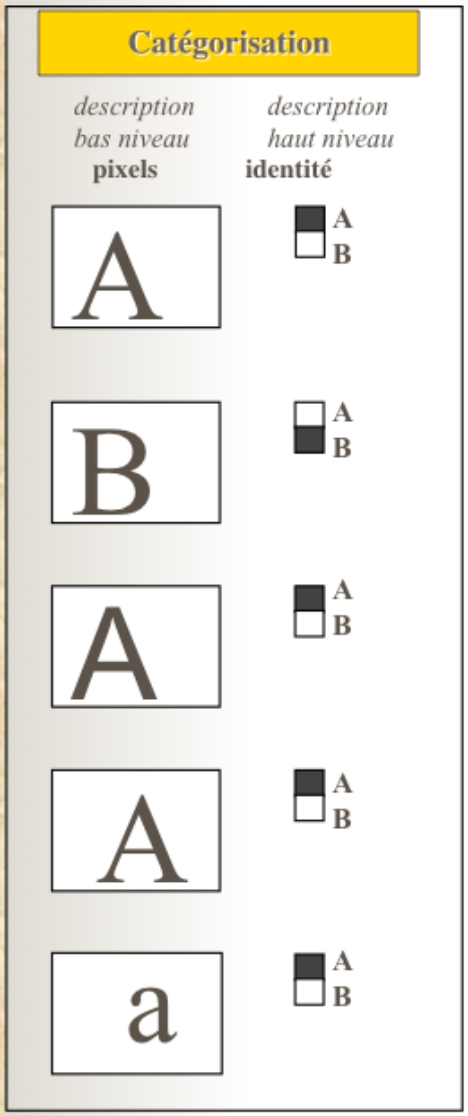
perception du degré de FDV dans l'absolu

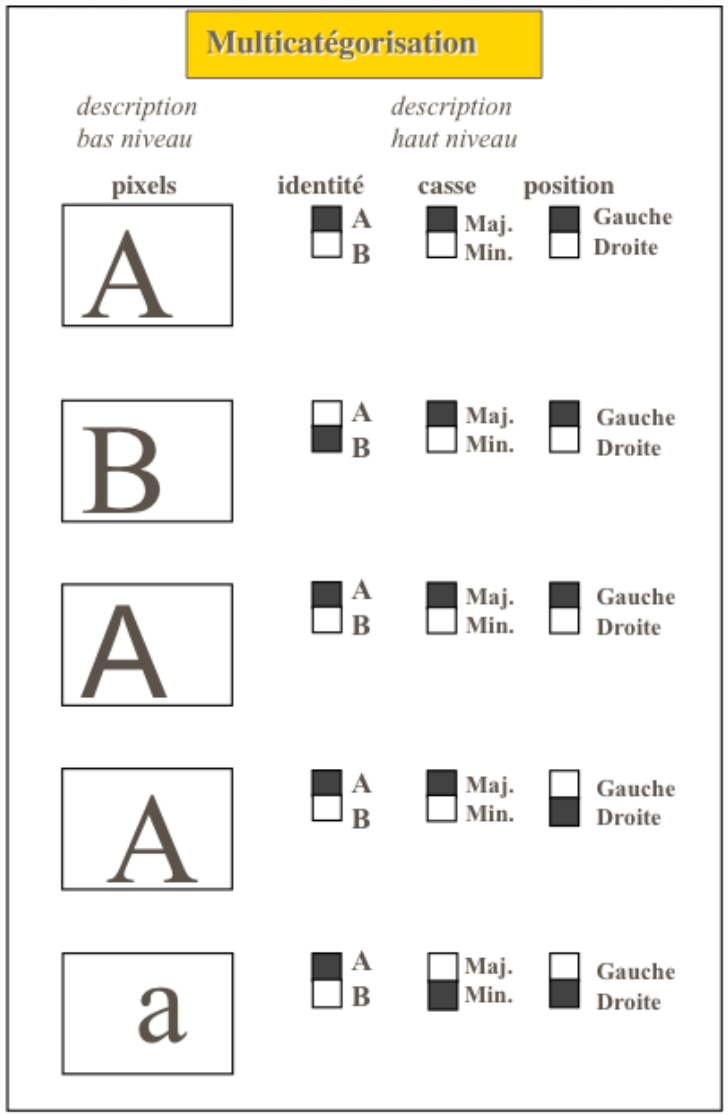
d'après Corenc: dynamique 25 à 30 dBA, résolution 3dBA
---> 7 à 10 nuances distinctes dans le signal

d'après José: dynamique env 60 dBA, résolution 6 dBA
---> env 10 nuances id

Comment monter la manip perceptive ? comparaison par paires de tokens de FDV différentes ? égalisation des niveaux ?

Q9. multicatégorisation





En Catégorisation, seul un descripteur haut niveau est défini (ici l'identité de l'objet).

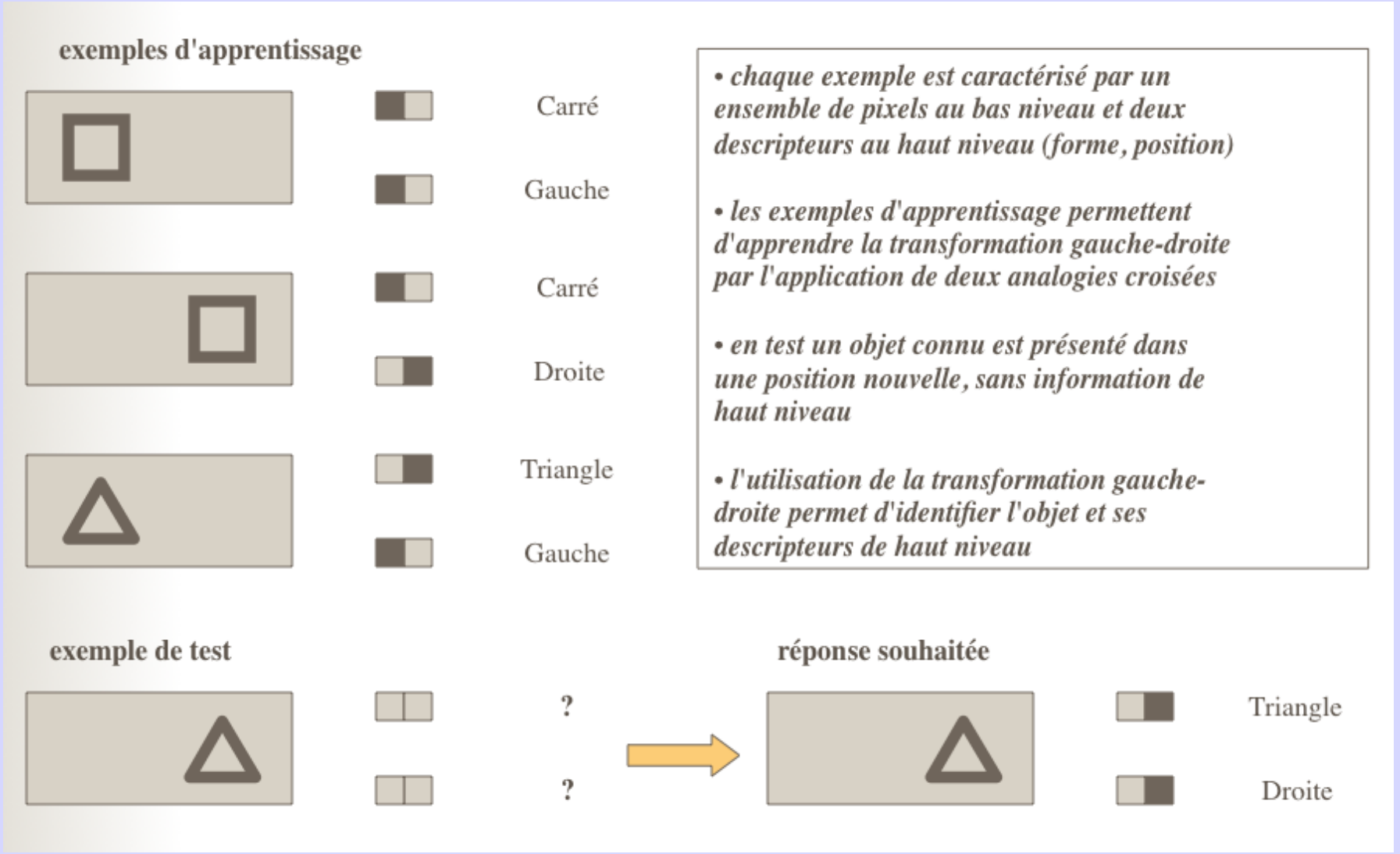
Il en résulte une grande variabilité (non-coïncidence des classes bas et haut niveau).

En définissant plusieurs descripteurs haut niveau (multicatégorisation = plusieurs points de vue sur l'objet) la variabilité est fortement réduite.

La catégorisation est une forme particulière de multicatégorisation.

En ingéniérie : Traitement des Formes vs Reconnaissance des Formes

Induction analogique



Q10. formants ou enveloppe spectrale ?

Formants

- décrivent les modes propres du conduit vocal
- robustes en synthèse et transmission
- impossible de les détecter et de les identifier à 100% sans connaissances préalables
- voix d'enfants, quand F0>F1
- F'2 et intégration à 3.5 Bark
- échec en transmission et reconnaissance

---> les formants ne décrivent pas la perception

Enveloppe spectrale

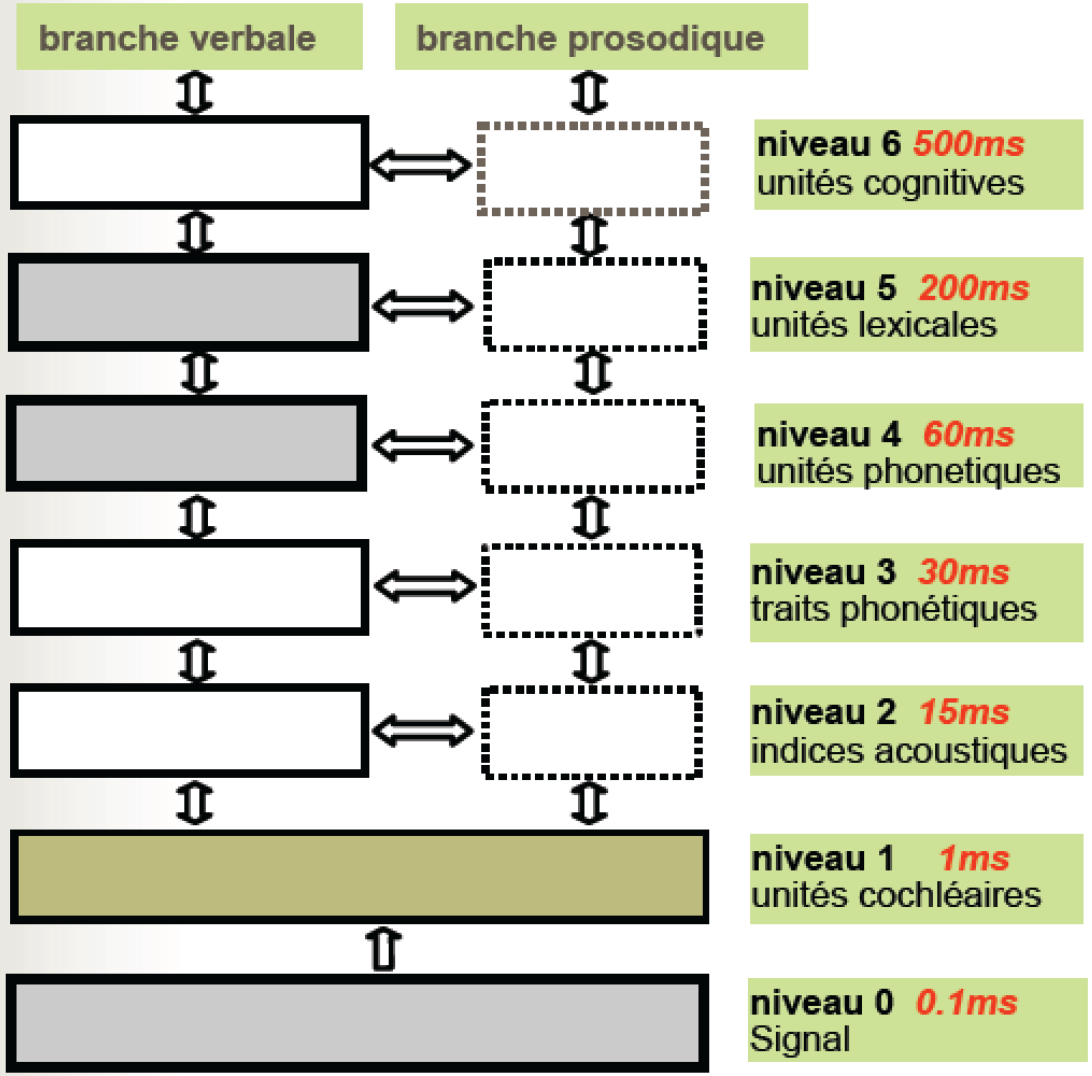
- comportent toute l'information, pas seulement phonétique
- sensibilité à distorsion spectrale

un compromis

- intégration 3.5 Bark: spectre à bosses ?

Q11. organisation de la perception

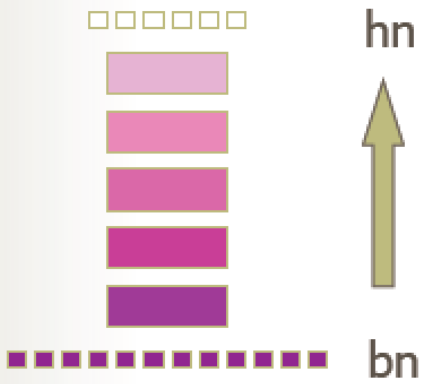
Perception de la parole et de la voix



- *les niveaux d'abstraction sont liés à la résolution temporelle*
- *traitement conjoint des informations*
- *à chaque niveau la description du contenu perceptif est complète (linguistique et non-linguistique)*
- *descripteurs de plus en plus indépendants*
- *deux flux d'information: ascendant et descendant*

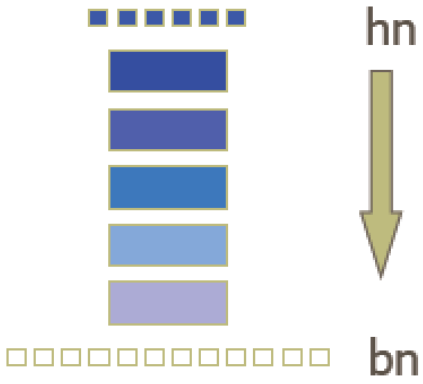
Divers modes de fonctionnement

ascendant



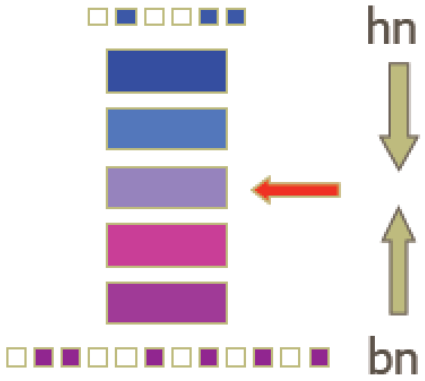
- l'information bn est prédominante
- prévisibilité nulle
- streaming, pop-up, descripteurs intrinsèques (bn), Gestalt, émergence

descendant



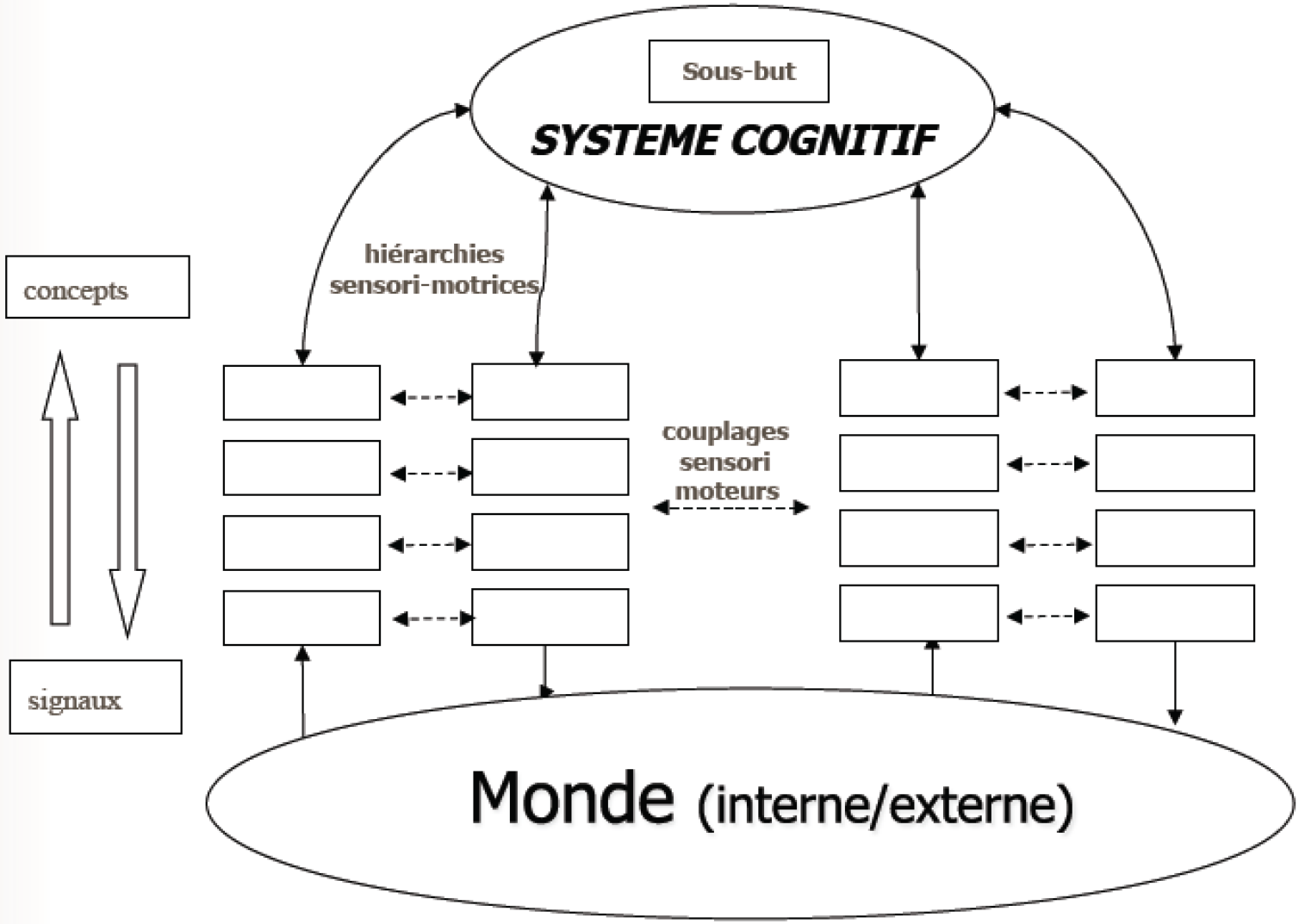
- l'information hn est prédominante
- prévisibilité totale
- attention et connaissances attachées aux niveaux supérieurs

double flux



- les informations hn et bn sont partielles
- prédominance d'un niveau
- conflit possible

Modèle comportemental



Q12. sur l'intensité des voyelles

José	aa	ii	ou	ensemble
nb tokens	490	443	452	1385
moy_ap dBA	52.92	45.41	46.69	48.48
sigma_ap dBA	*12.93*	*11.55*	*11.69*	*12.54*
ap-moy_ap	4.44	-3.07	-1.79	-
moy_ax dB	54.69	54.03	54.62	54.69
sigma_ax dB	*10.30*	*9.66*	*10.13*	*10.30*

Corenc	aa	ii	ou	ensemble
nb tokens	60	60	60	180
moy_ap dBA	71.49	62.75	64.96	66.40
sigma_ap dBA	*6.79*	*6.25*	*5.93*	*7.31*
ap-moy_ap	5.09	-3.65	-1.44	-
moy_ax dB	74.42	73.55	74.50	74.16
sigma_ax dB	*4.85*	*5.10*	*4.69*	*4.87*

---> ap dépend de la voyelle alors que ax est constant

--->remarquable constance des niveaux ap-moy_ap voyelles d'une bd à l'autre

---> Corenc enregistré plus fort que José, d'environ 20 dB ou 18 dBA. Attention, la dynamique n'est pas la même, on ne peut pas appliquer cette correction telle quelle, il faut voir comment se calent les fdv respectives

--->la dynamique de José est env. 2 fois plus forte que celle de Corenc (voir les écarts-types)