Extraction de paramètres phonétiques sur grands corpus

C. Gendrot

N. Audibert

Laboratoire de Phonétique et Phonologie

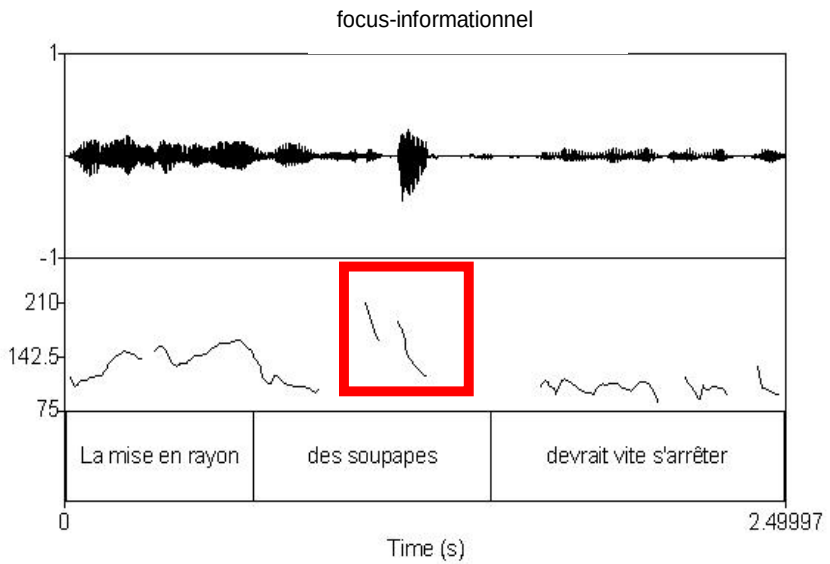
Université Sorbonne-Nouvelle, CNRS UMR7018

Préambule

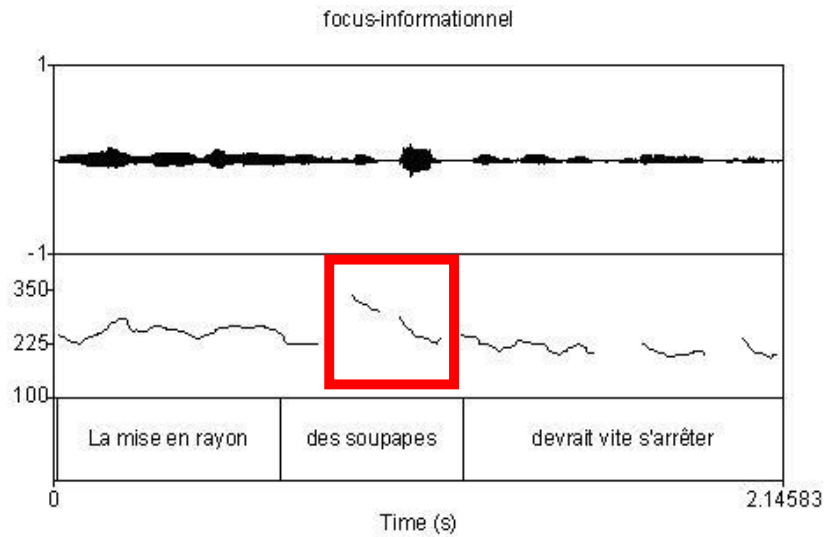
- Mesure locales (paradigmatiques) vs. Globales (syntagmatiques)

- Mesures statiques vs. Dynamiques

f0



f0



Plan

- 1- Pourquoi des mesures phonétiques ?
- 2- Des corpus ? (*corpora*)
- 3- Quels paramètres phonétiques ?
  - Paramètres classiques (durée, f0, intensité, formants)
  - Autres paramètres (qualité vocale, nasalité, etc.)

1- pourquoi des mesures phonétiques ?

- Les technologies vocales utilisent plus volontiers des MFCC (Mel Frequency Cepstral Coefficients) et leurs dérivées …

- MFCC => transformée *en cosinus discrète* appliquée au spectre d'amplitude d'un signal acoustique par bandes de fréquence MEL.

- Ces mesures sont peu utilisées en phonétique, car difficilement interprétables en termes physiologiques. Elles sont essentiellement utilisées pour des discriminations.
- Les mesures phonétiques ont souvent été déterminées grâce à des modélisations, elles permettent de mieux comprendre la production de la parole. Elles ont certaines contraintes malgré tout …

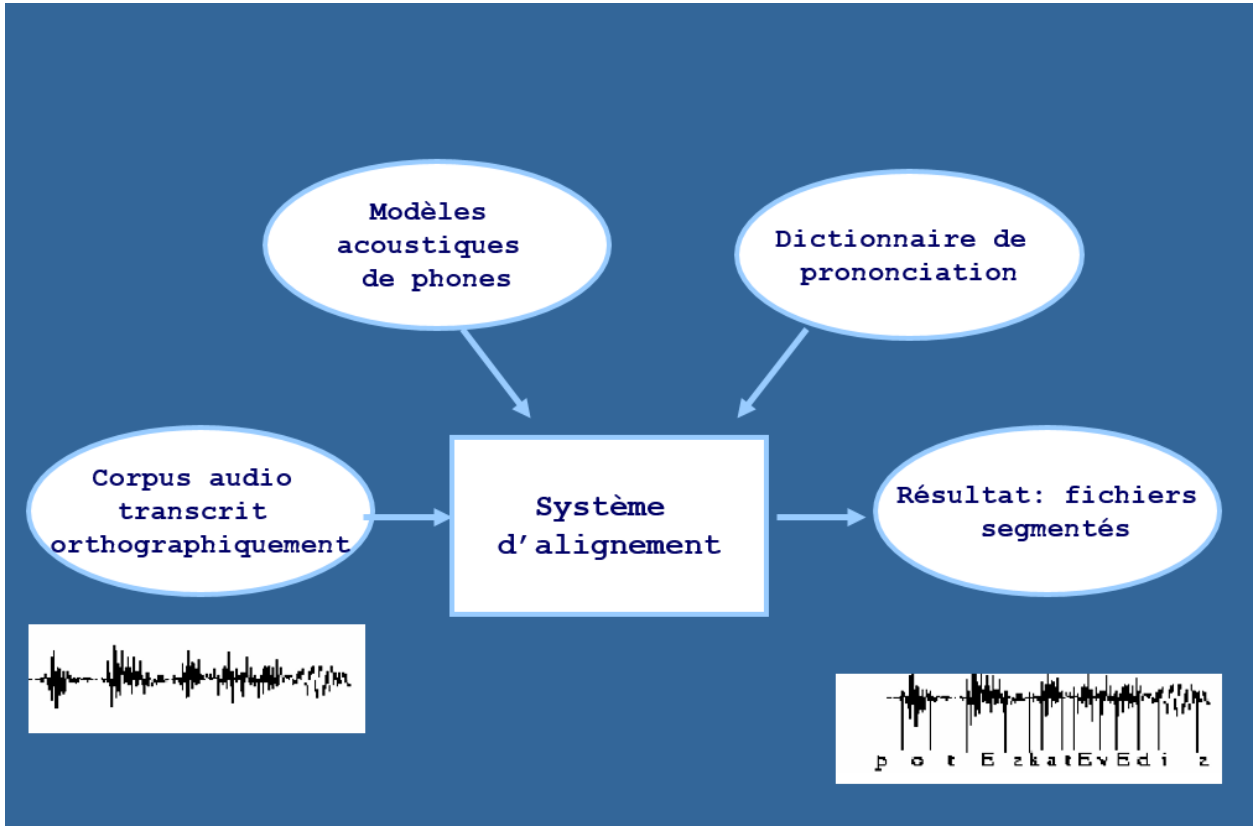
2- Des corpus ?

- Avant l'utilisation de grands corpus transcrits et alignés obtenus par et pour les technologies vocales, utilisation de corpus ad-hoc
  - Corpus ad-hoc : corpus lus, contrôlés d'un point de vue segmental et prosodiques, souvent petits (quelques minutes …)
  - « il a dit rire cinq fois », « il a dit rare cinq fois », « il a dit roure cinq fois »
  - « Paul et Tata-**N**adia arriveront demain matin », « Tonton, Tata, **N**adia et Paul arriveront demain »

- Les corpus ad-hoc ont l'avantage de pouvoir observer à coup sûr le phénomène recherché (par définition) …
- Conditions peu « écologiques » : leur défaut est leur manque de naturel évident, représentation artificielle de la réalité (fréquence, sémantique, etc.).

2- Des corpus ?

- L’utilisation de grands corpus (plusieurs heures, dizaines/centaines/milliers d’heures) obtenus par et pour les technologies vocales



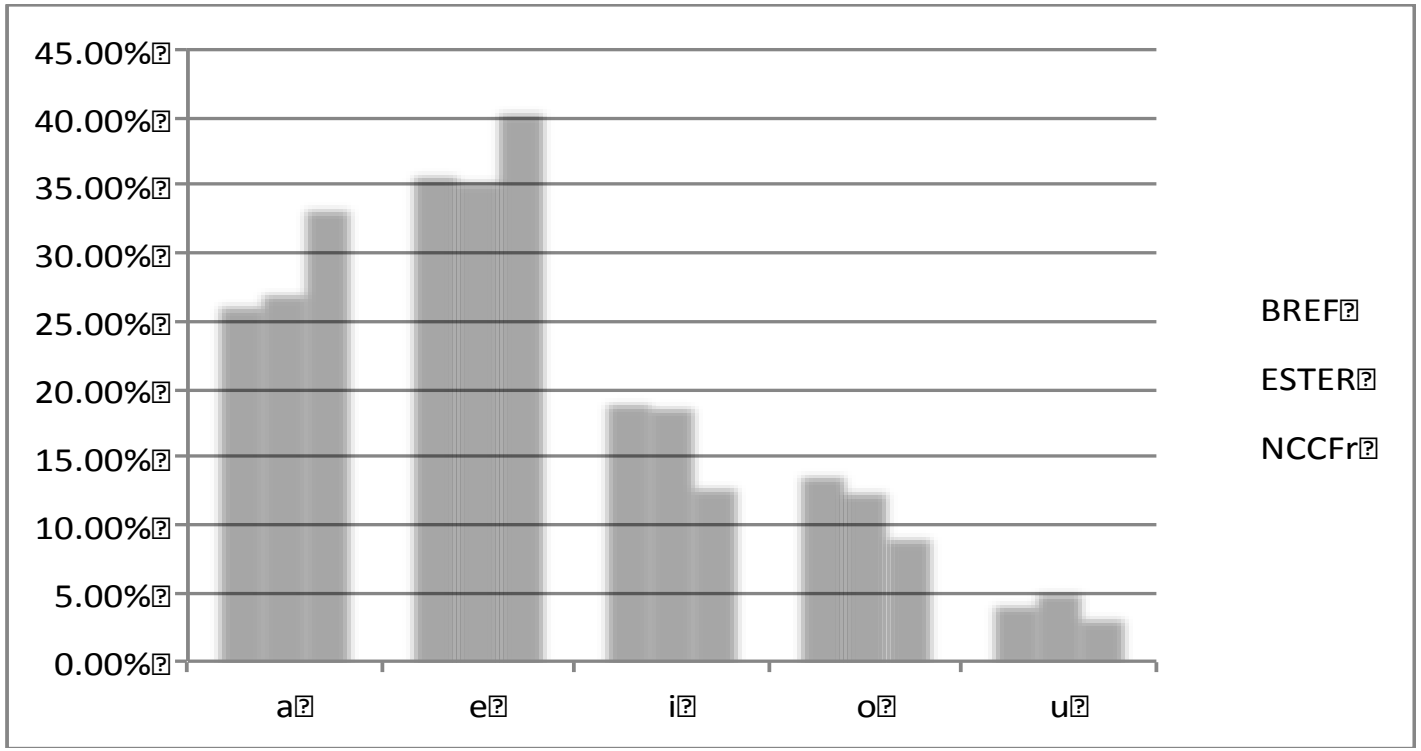
- La transcription orthographique peut être obtenue manuellement ou automatiquement

- Possibilité de jouer sur les modèles acoustiques, le dictionnaire de prononciation

2- Des corpus ?

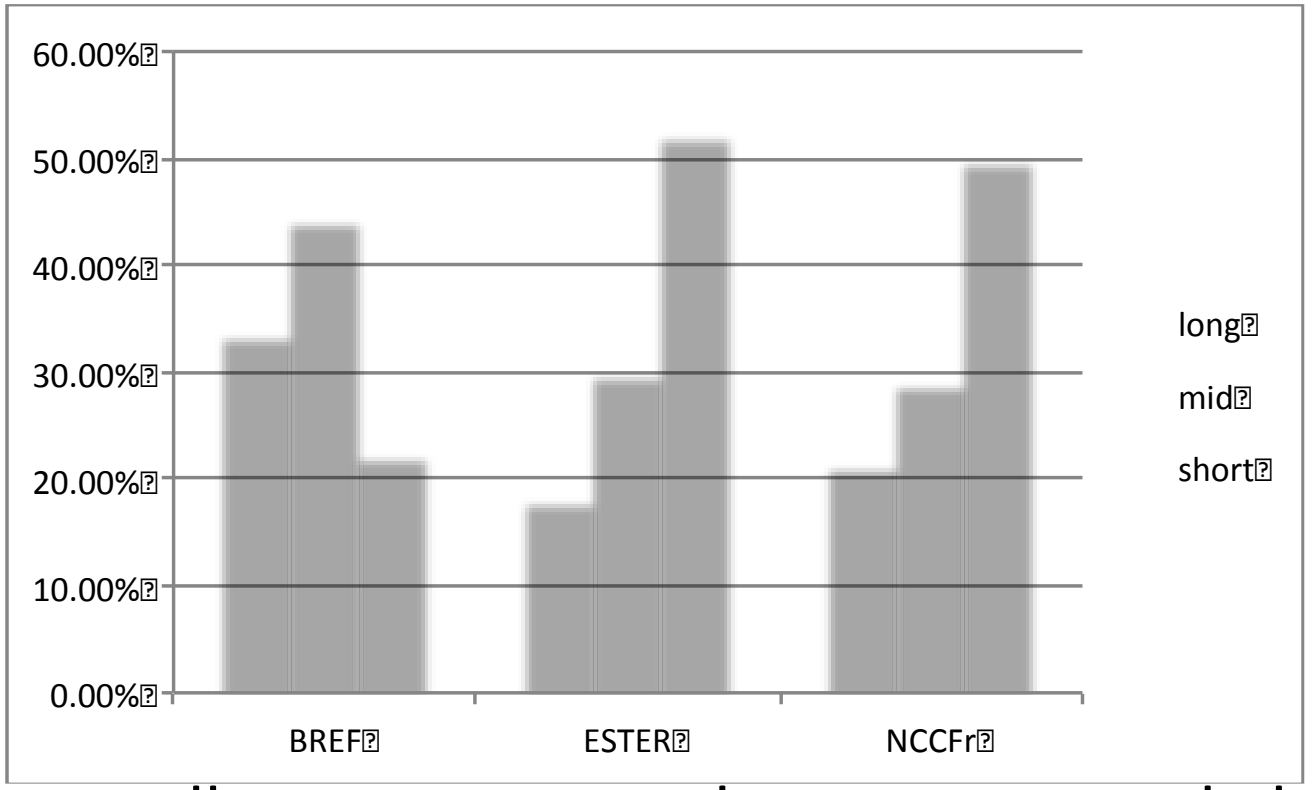
- L’utilisation de grands corpus (plusieurs heures, dizaines/centaines/milliers d’heures) a de nombreux avantages :
  - Contextes naturels (fréquence segmentale et lexicale, parole spontanée ou préparée, interactions)
  - Couplés à une analyse automatique, ils permettent d’obtenir très rapidement des milliers de résultats. Par exemple, 5000 /a/ par locuteur, etc.

Distribution des voyelles par corpus



- Dans les trois corpus, sur-représentation de /a/ et /e/, sous-représentation de /u/

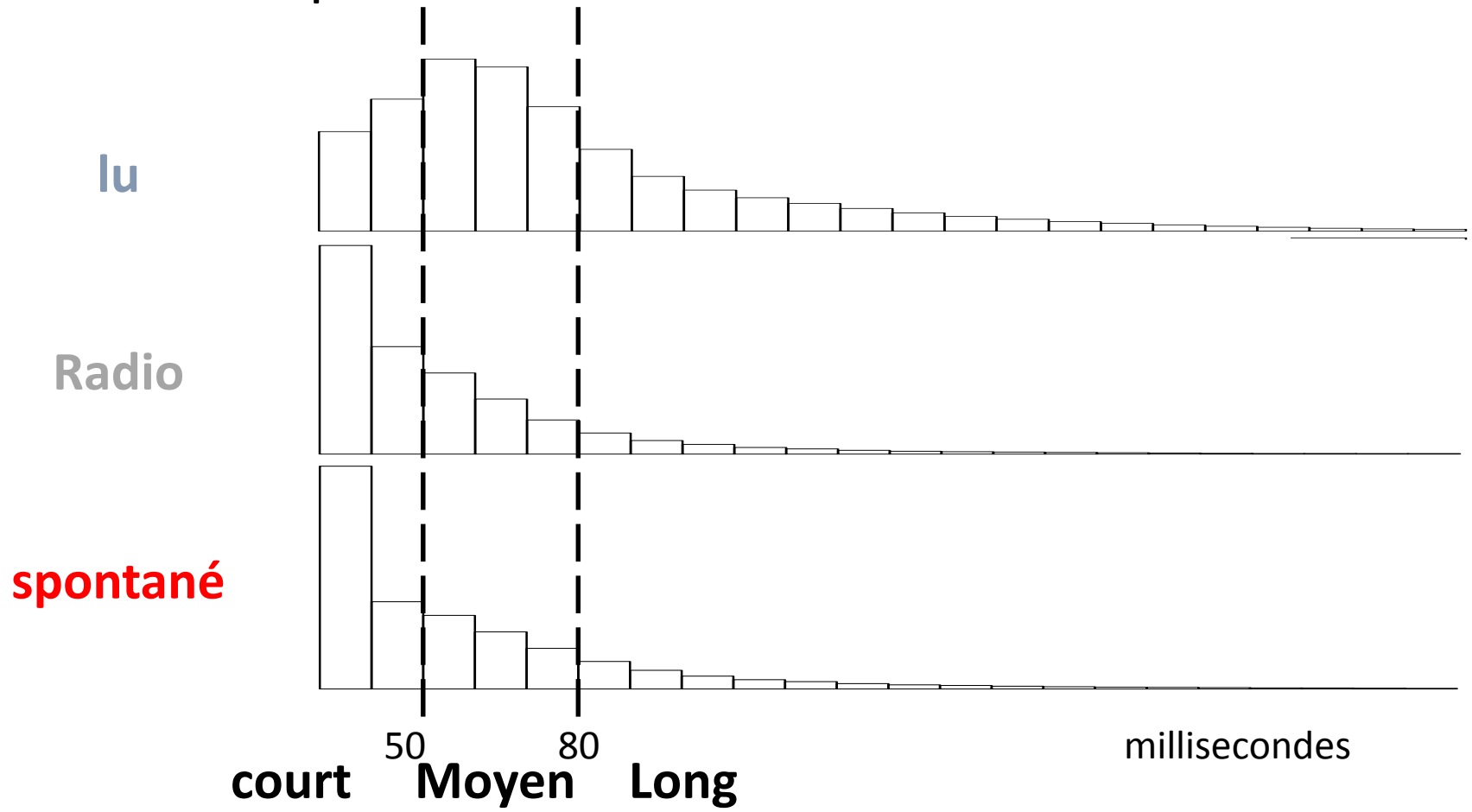
Distribution des classes de durée par corpus



- Plus de voyelles moyennes et longues en parole lue

Speech material and methods: DURATION factor

- 3 classes de durée pour nos mesures :



2- Des corpus ?

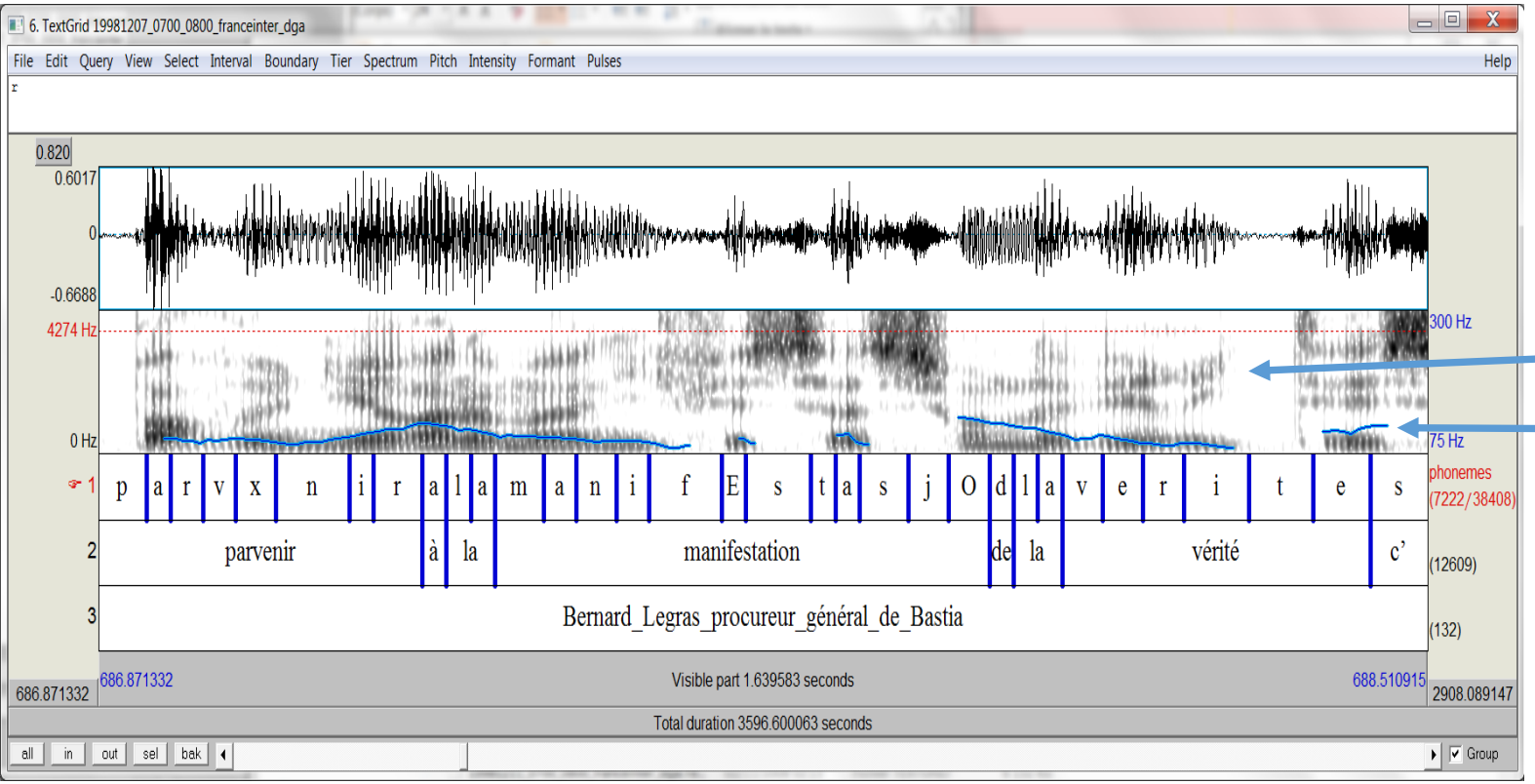
- De nombreux avantages aux grands corpus

  …

- Mais
  - Nécessite des précautions méthodologiques (mesures générales de la durée de /a/ vs. /i/, quel catégorie et fréquence de mot, syllabe, phrase ?)
  - Peut parfois rester limité dans l'analyse de contextes segmentaux/lexicaux peu fréquents (voyelles entre consonnes vélaires, analyse du / ɲ /)
  - Implique des erreurs de mesures de diverses origines (segmentation décalée et qu'on va pas toujours vérifier, mesures automatiques, etc.)
  - Statistiques : avec des ANOVAs classiques, tout est significatif (des ≠ de 5ms) …
  - Linguistique de corpus vs fauteuil !

3- quels paramètres phonétiques ?

- Paramètres « classiques » : <u>La durée</u>



Signal acoustique

spectrogramme

f0

3- quels paramètres phonétiques ?

- Paramètres « classiques » : La durée

  - Durée des phonèmes, des syllabes

  - Durée de portions de phonèmes (VOT, relâchement, transitions, etc.)

  - Cette mesure à priori simple est pleinement dépendante de la segmentation. Certains modèles acoustiques de phones (indépendants du contexte) peuvent être plus adaptés pour une meilleure précision

- Paramètres « classiques » : La durée
  - Les erreurs fines de segmentation peuvent être difficiles à traiter par l'humain également
  - La durée est dépendante du phonème, les voyelles nasales sont + longues que les orales, les voyelles arrondies sont + longues, les consonnes voisées + courtes que les sourdes, etc.
  - Pourquoi mesurer la durée ?
    - Pour mesurer le débit de parole, l'accentuation lexicale et/ou sémantique, les regroupements prosodiques/syntaxiques, mesures rythmiques.

Pour les études phonétiques … Exemple du schwa

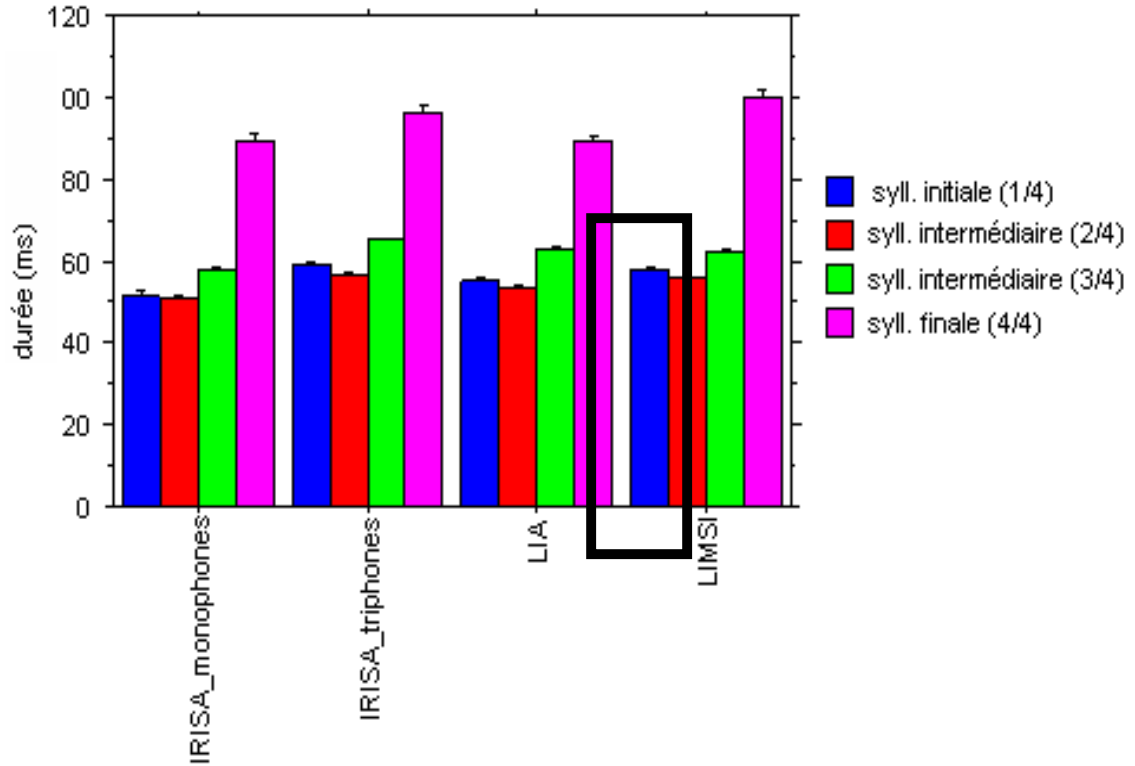
- La durée attribuée aux voyelles par un alignement automatique est généralement plus courte, avec une précision moins importante en fin de voyelle.

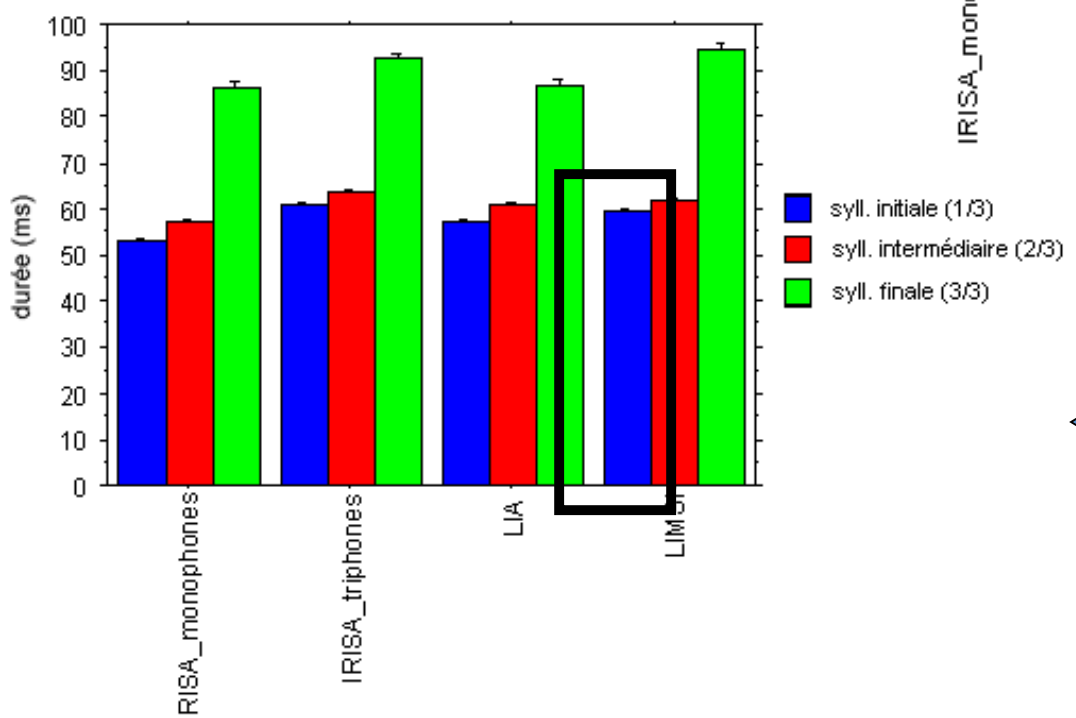
- Malgré tout, le milieu de la voyelle est correctement localisé dans près de 80% des cas.

- Dans les alignements automatiques évalués ici, la voyelle ne se voit jamais attribuer une durée inférieure à 30 ms, or, la durée minimale attribuée au schwa lors d'un alignement manuel est de 8 ms.

- (le schwa est une des voyelles, sinon la voyelle, qui pose le plus de problèmes de détection et d'alignement)

Durées des syllabes
dans des mots de 4 syll.





Durées des syllabes
dans des mots de 3 syll.

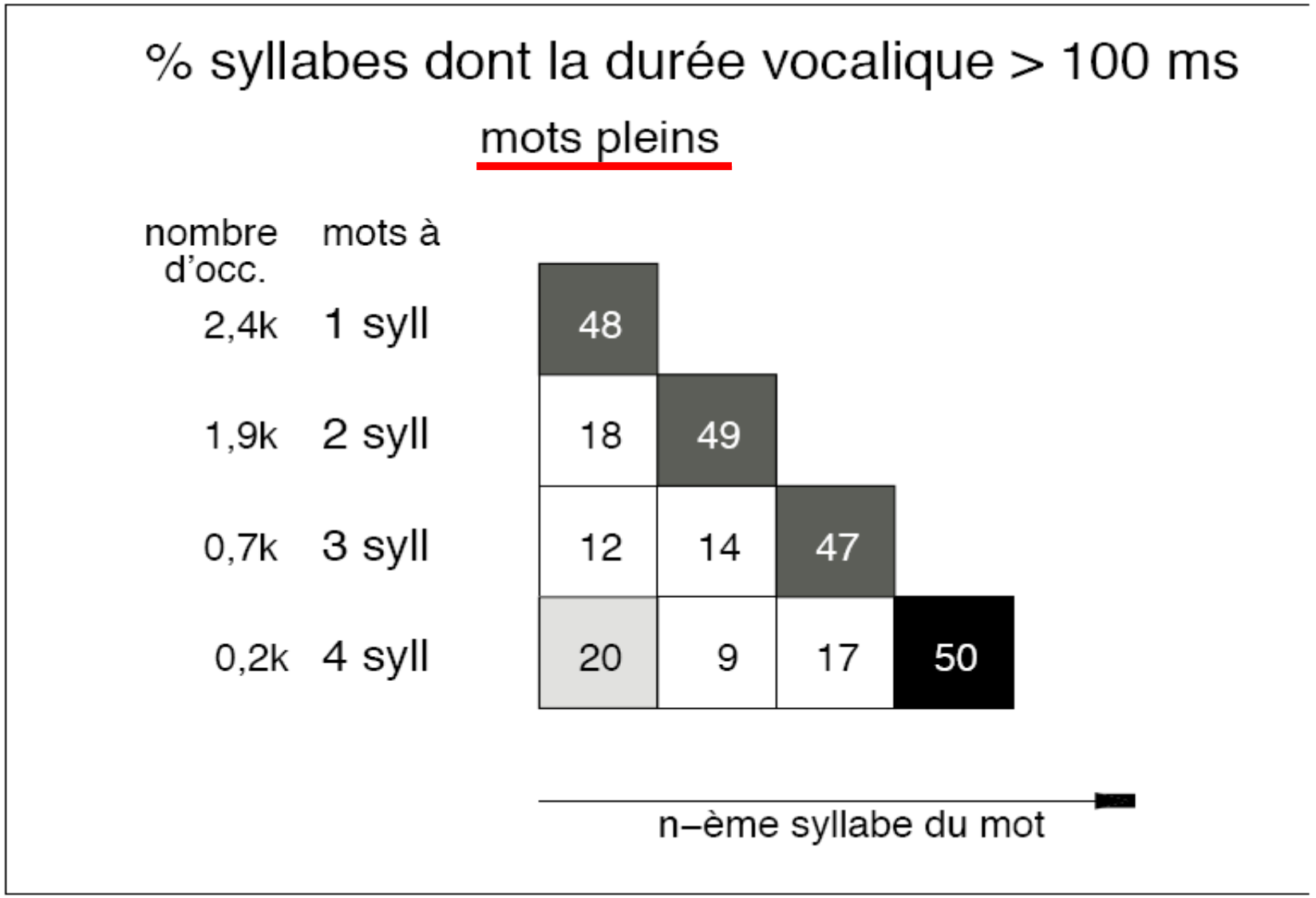


Fig : pourcentage de syllabes longues en fonction du nombre de syllabes

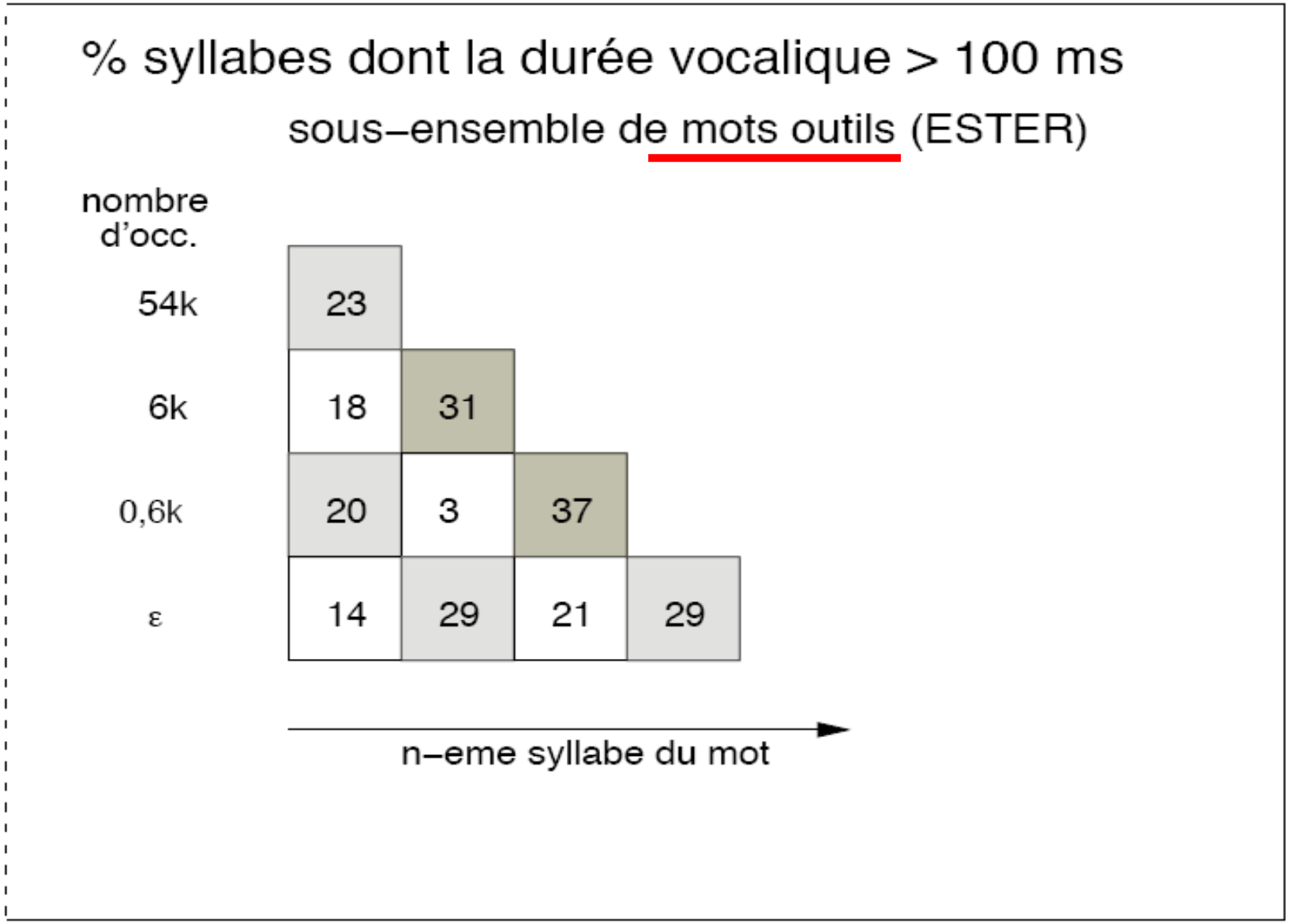
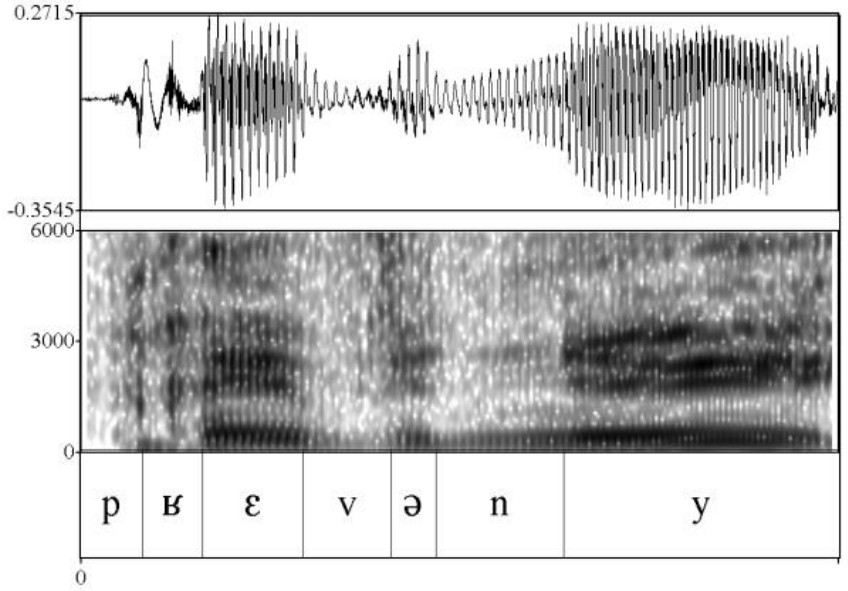


Fig : pourcentage de syllabes longues en fonction du nombre de syllabes

Apports pour la linguistique

phonologie



prédicteurs de l'élision du schwa ([pʁɛvəny] vs. [pʁɛvny]

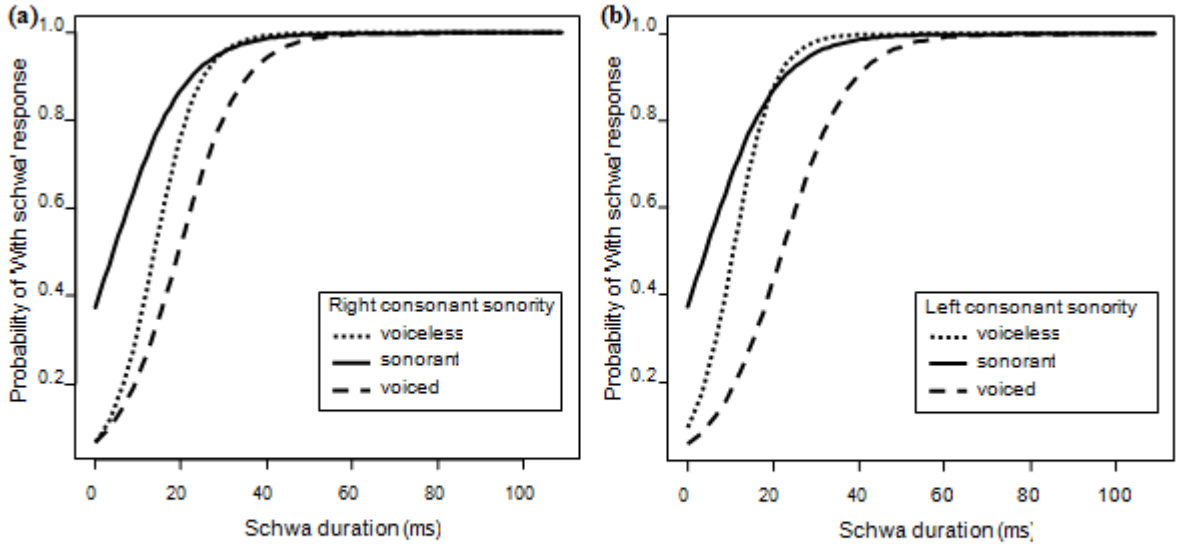
Débit de parole

Position du schwa dans le mot

Position du mot dans la phrase

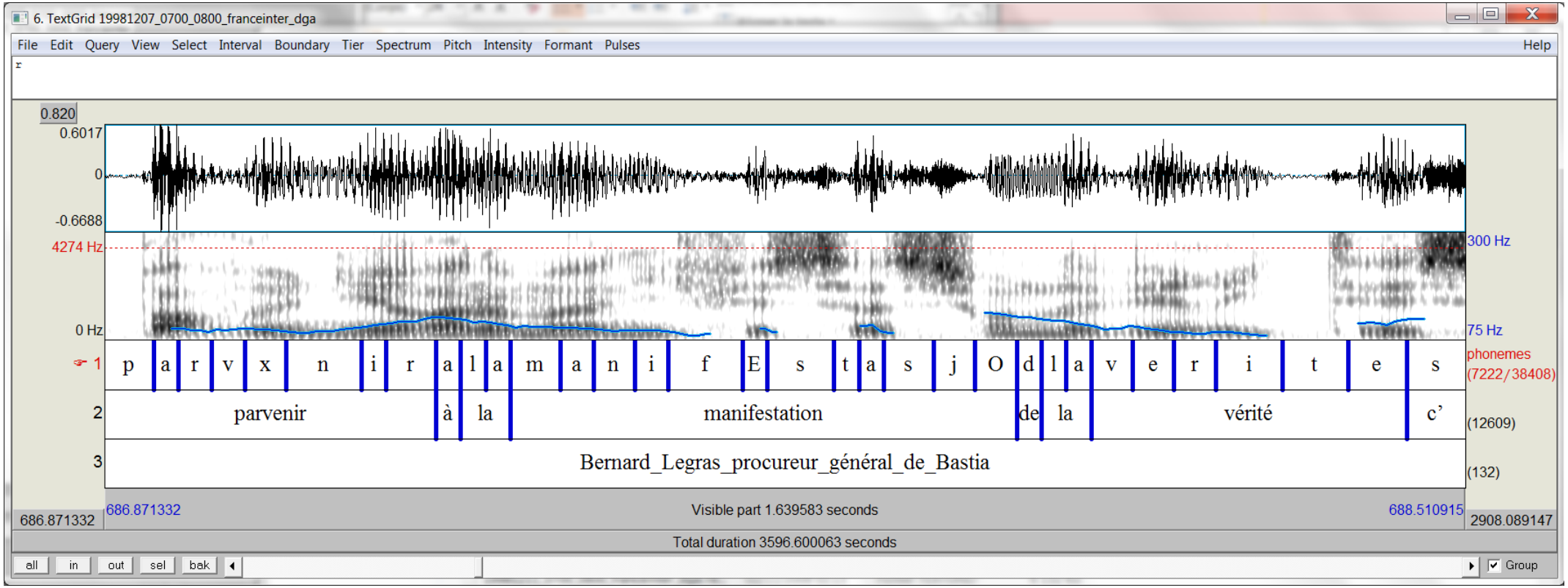
Nombre de consonnes dans le cluster

Respect du principe de sonorité



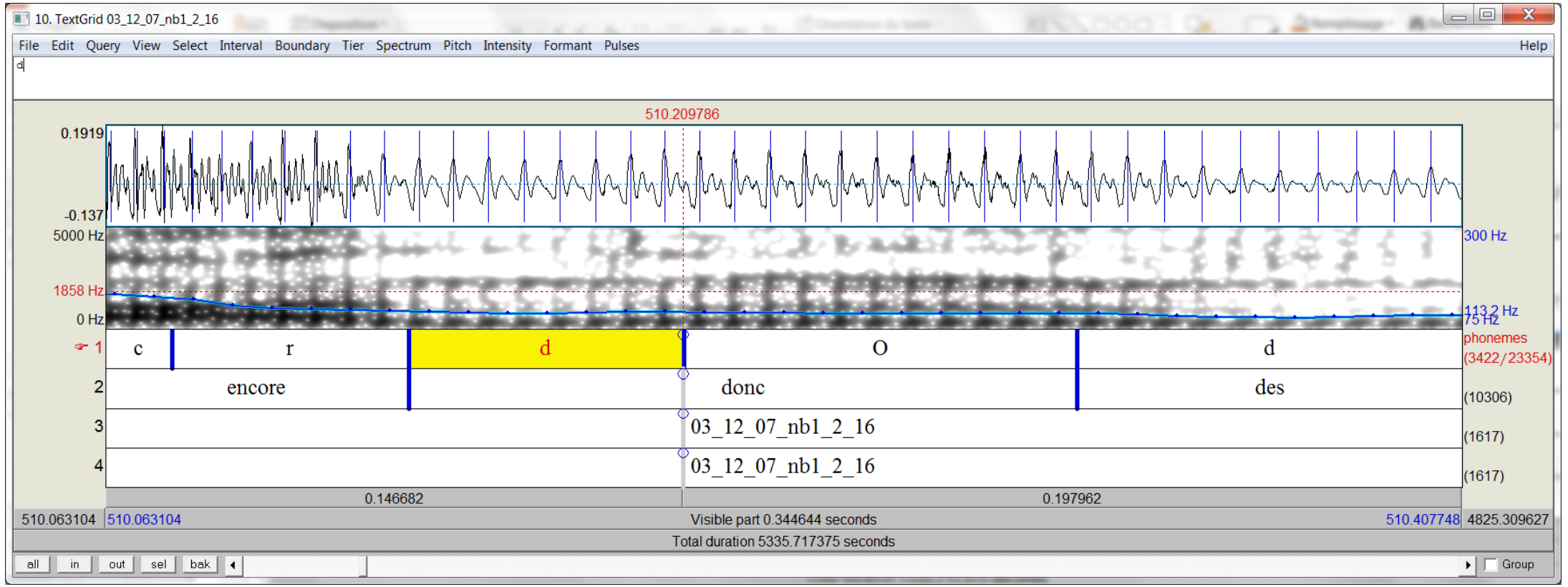
3- quels paramètres phonétiques ?

- Paramètres « classiques » : La f0



3- quels paramètres phonétiques ?

- Paramètres « classiques » : <u>La f0</u>



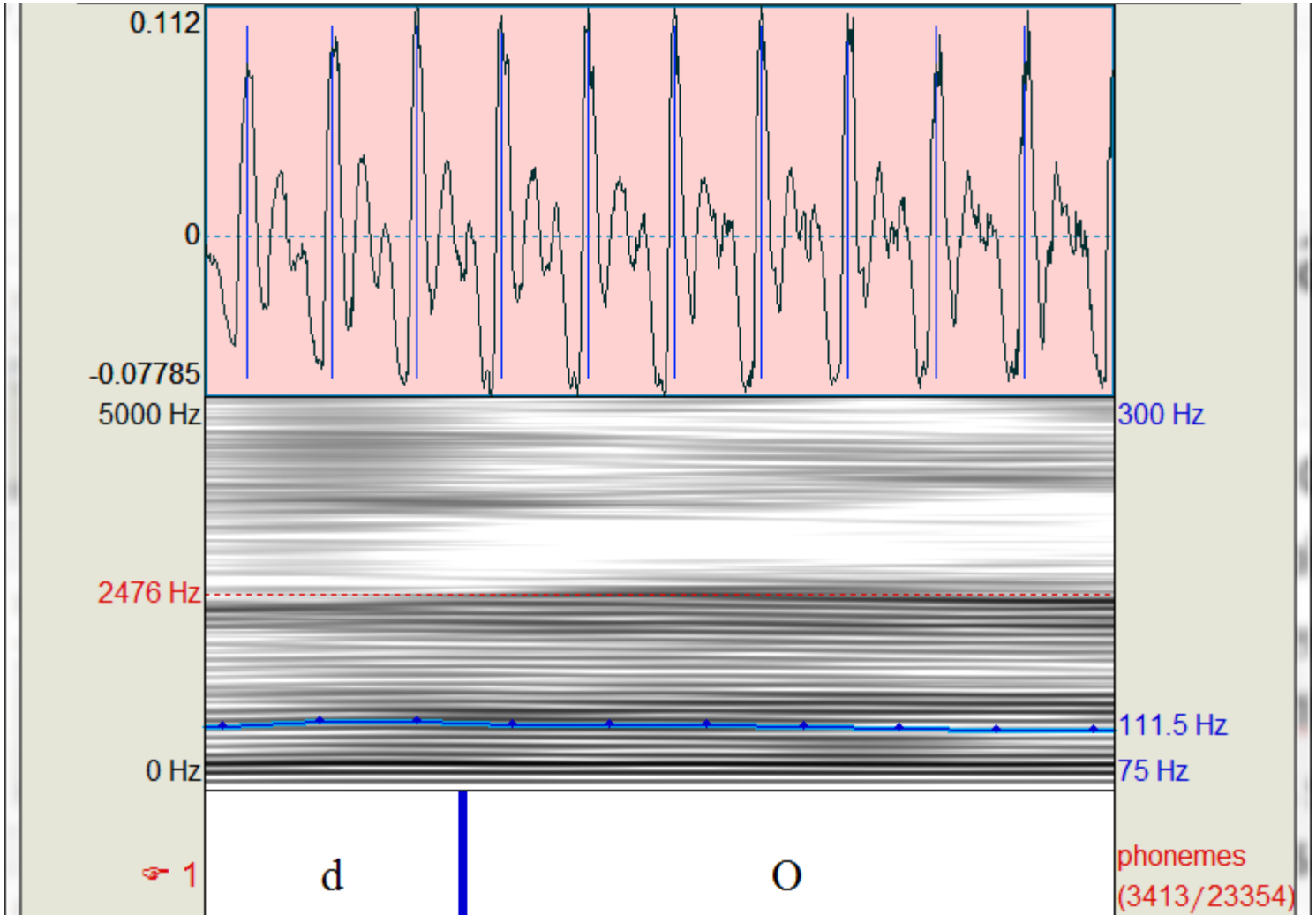
3- quels paramètres phonétiques ?

- Paramètres « classiques » :

  <u>La f0</u>

<u>Récurrence d'une période (par seconde)</u>

<u>L'harmonique fondamentale sur un spectre</u>

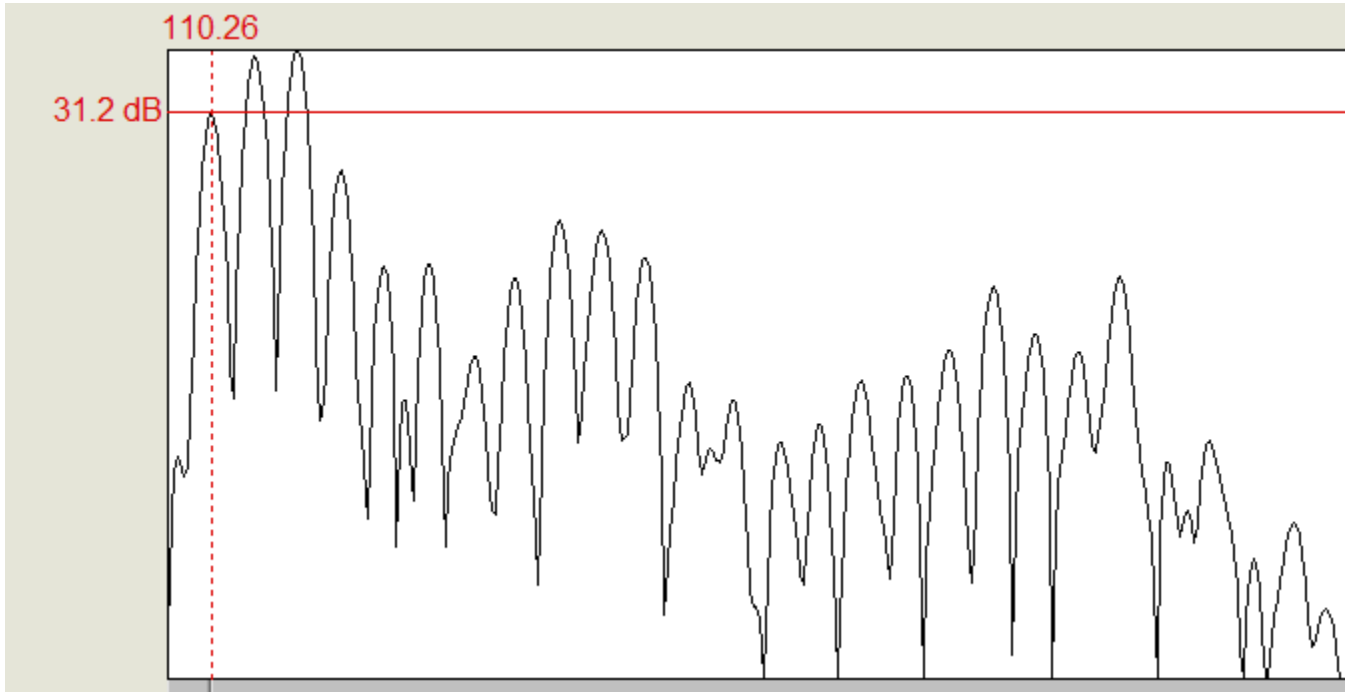


- Paramètres « classiques » :

  La f0

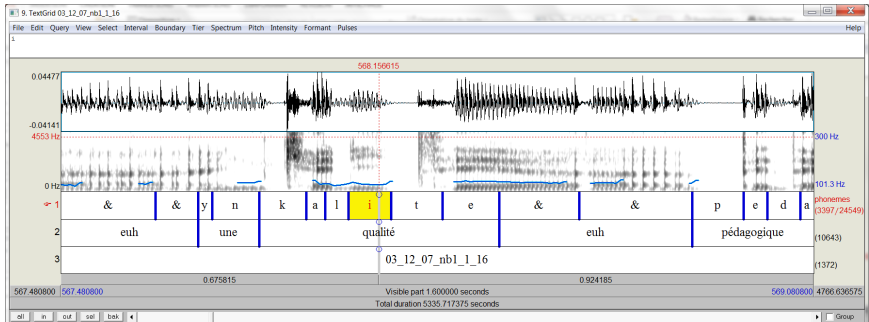
Récurrence d’une période (par seconde)

L’harmonique fondamentale sur un spectre



3- quels paramètres phonétiques ?

- Paramètres « classiques » : <u>La f0 (fréquence fondamentale)</u>
  - Automatisable facilement (fiable),
  - Les valeurs erratiques de f0 sont également difficilement mesurables à la main (dévoisement, sauts d'octave, craquements)
  - Les mesures sont paramétrables (rapport signal sur bruit, seuils, fourchettes de valeurs, mais pour un signal idéalement stable)



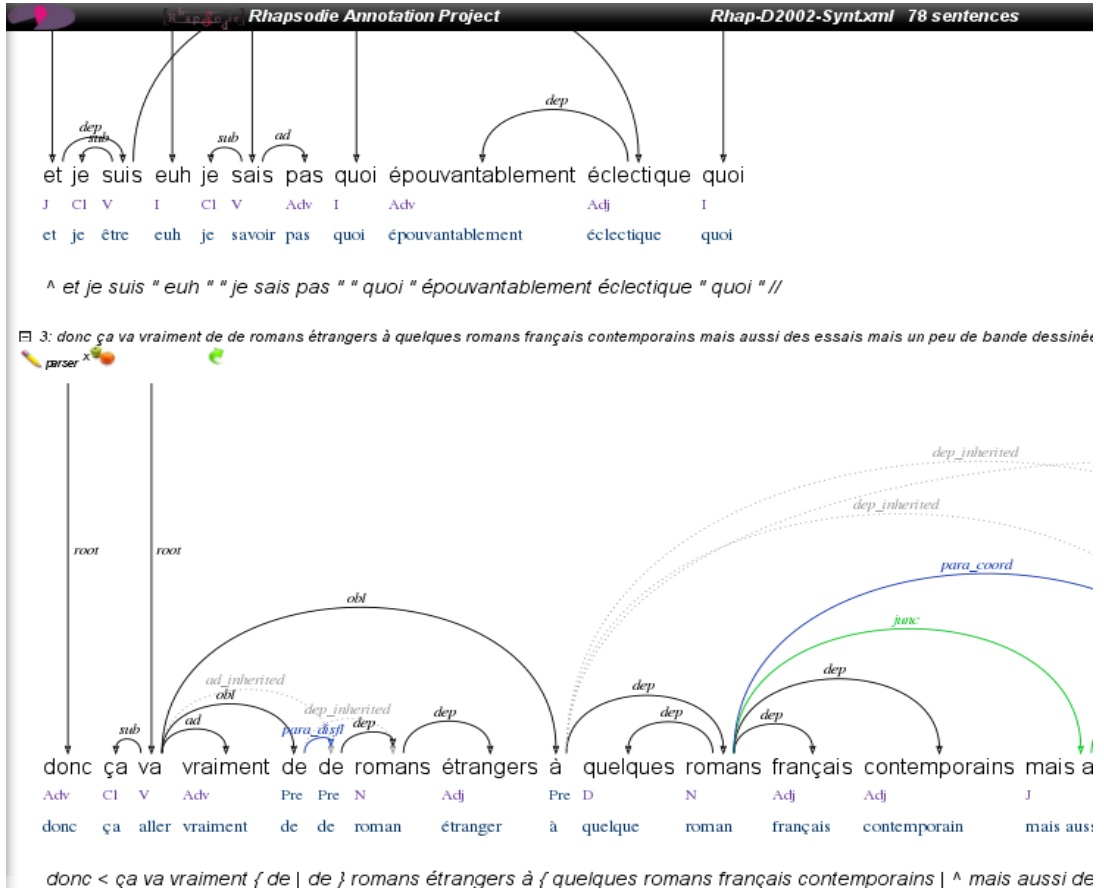
3- quels paramètres phonétiques ?

- Paramètres « classiques » : La f0 (fréquence fondamentale)
  - Les mesures de f0 pertinentes sont souvent dynamiques (pente, écart)
  - Mesurables sur les sons voisés (+micromélodie), dépendent du phonème, mais raisonnablement …
- Pourquoi mesurer la f0 ?
  - l'accentuation lexicale et/ou sémantique, les regroupements prosodiques/syntaxiques (corrélés ou non à la durée),
  - ligne de déclinaison, modélisation de l'intonation
  - Caractéristiques du locuteur

Apports pour la linguistique

syntaxe

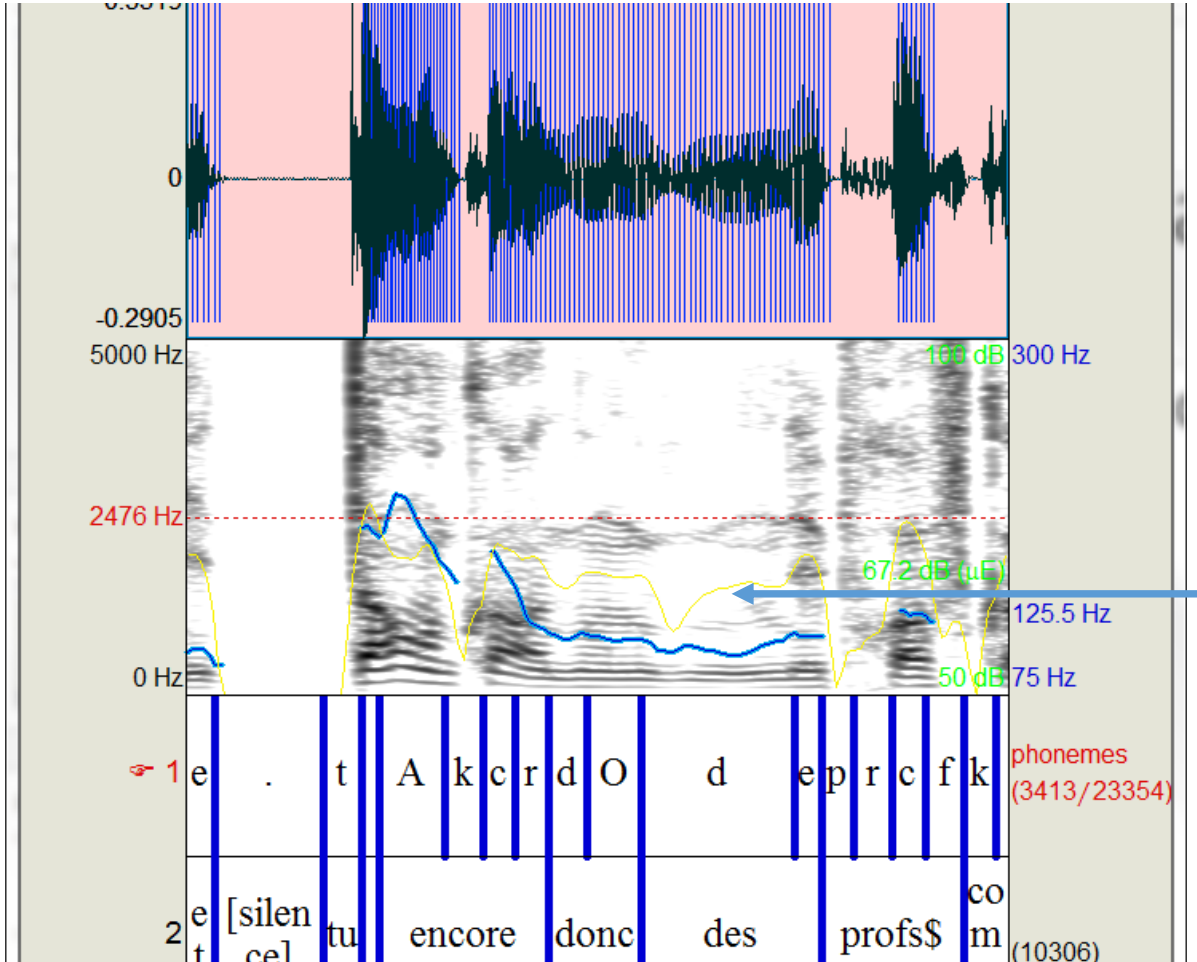
Connexions
entre syntaxe
et phonétique



Ou en bas un outil d'Annotation collaborative en ligne en graphes de dépendance

3- quels paramètres phonétiques ?

- Paramètres « classiques » : <u>L'intensité</u>



Intensité

3- quels paramètres phonétiques ?

- Paramètres « classiques » : L'intensité
  - En RMS ou dB
  - Automatisable facilement mais …
  - Les valeurs sont très variables d'un phonème à l'autre
  - Nécessité d'un micro casque (au minimum) pour l'enregistrement
  - Etalonnage avec un sonomètre indispensable pour une comparaison inter-locuteurs

- Pourquoi mesurer l'intensité ?
  - l'accentuation lexicale et/ou sémantique
  - Caractéristiques du locuteur

3- quels paramètres phonétiques ?

- Paramètres « classiques » : <u>les formants</u>

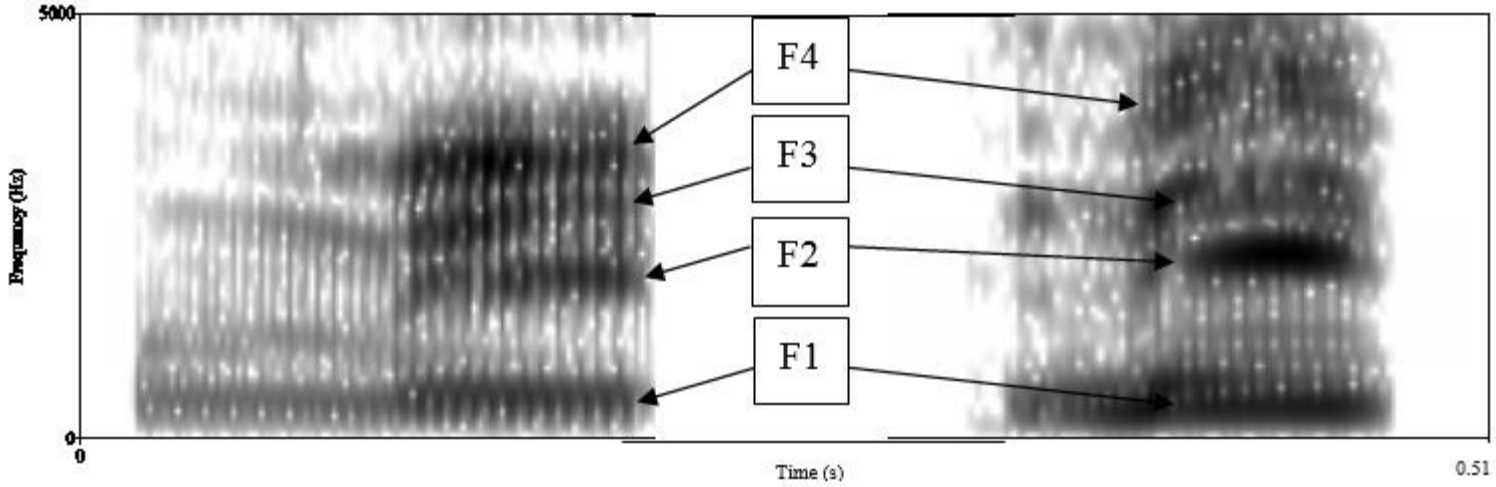
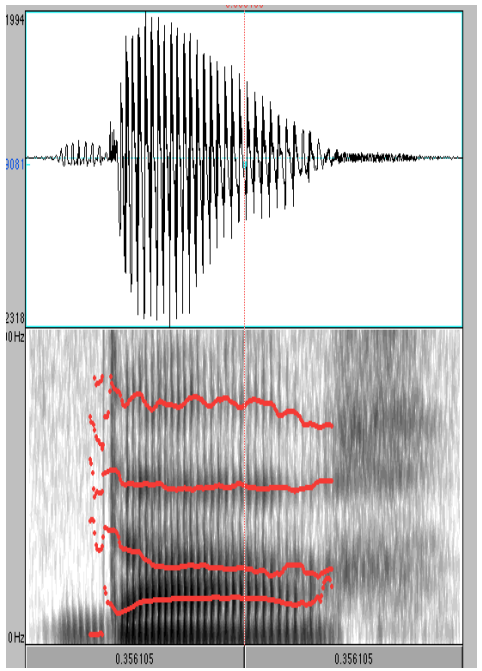


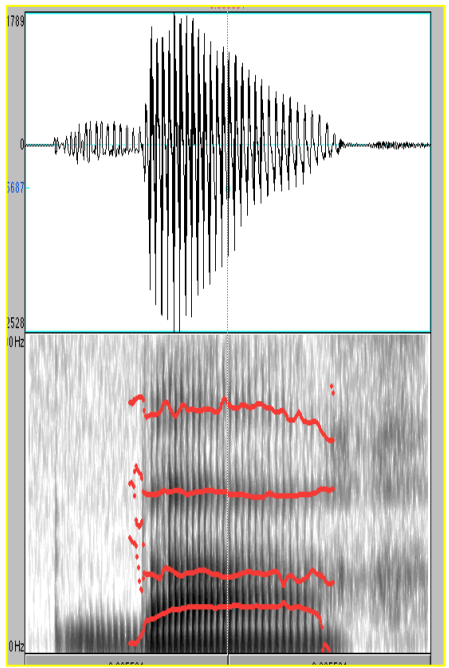
figure 1 : exemples de spectrogramme pour le mot « mais » [mɛ] prononcés par un homme (Alain Passerel) et une femme (Pascale Clark). Données extraites du corpus ESTER.

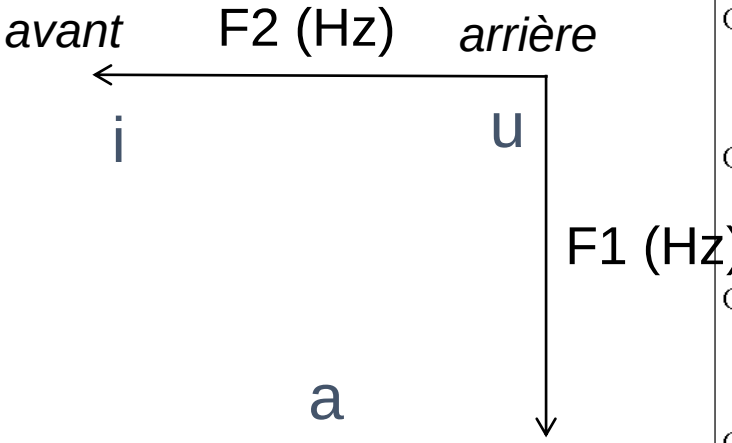
Paramètres « classiques » : les formants

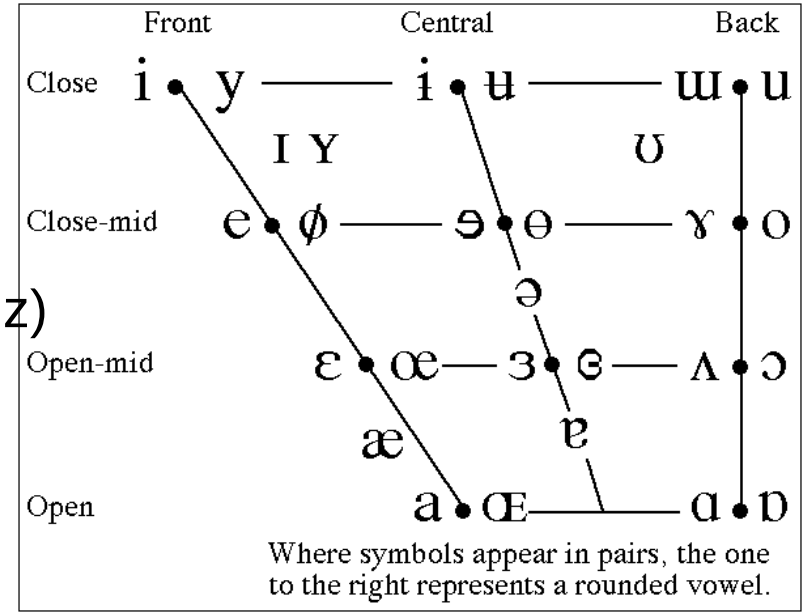
⇒ Pics de résonance qui permettent de distinguer acoustiquement les voyelles entre-elles

- 1er formant (F1) entre 250 et 800 Hz : correspond principalement à l'aperture de la voyelle
- 2ème formant (F2) entre 800 et 2400 Hz : correspond principalement à l'antériorité
- 3ème formant (F3) entre 2500 et 3500 Hz : correspond principalement à l'arrondissement









3- quels paramètres phonétiques ?

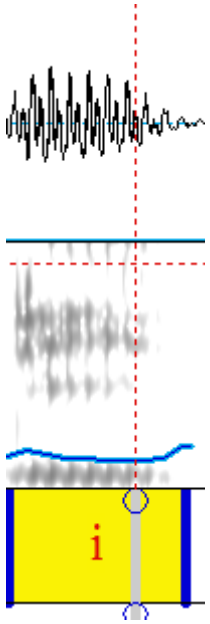
- Paramètres « classiques » : les formants
  - Automatisable avec une certaine prudence …
  - Principalement F1, F2 et F3 (F3 pour le fr notamment)

- Pourquoi mesurer les formants ?
  - Permet de caractériser l'articulation, reliés à l'abaissement (F1), antériorité (F2) de la langue, et à l'arrondissement (notamment F3)

  - l'accentuation lexicale et/ou sémantique
  - Permet de caractériser le locuteur

Fiabilité des mesures acoustiques automatiques

- Les mesures acoustiques automatiques de formants effectuées sont-elles fiables ?
  - Vérification manuelle sur un petit pourcentage des données
  - Filtrages pour les valeurs de formants ou de f0 établis sur la base de connaissances acoustiques ou bien de vérifications visuelles
  - Les mesures erronées ne fournissent-elles pas des indications utiles ? Pourquoi y a-t-il des erreurs ?

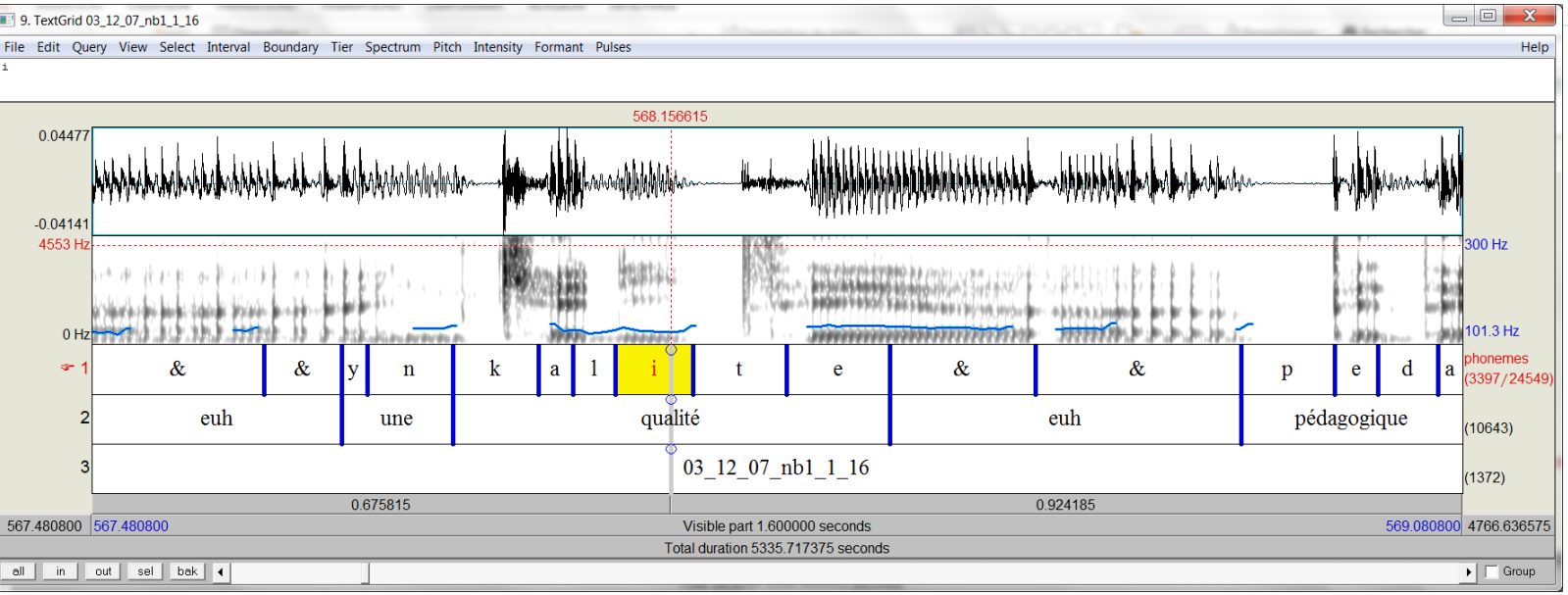


- Les erreurs de mesures automatiques ne sont pas dues au hasard, peuvent être justifiées et révèlent des phénomènes intéressants

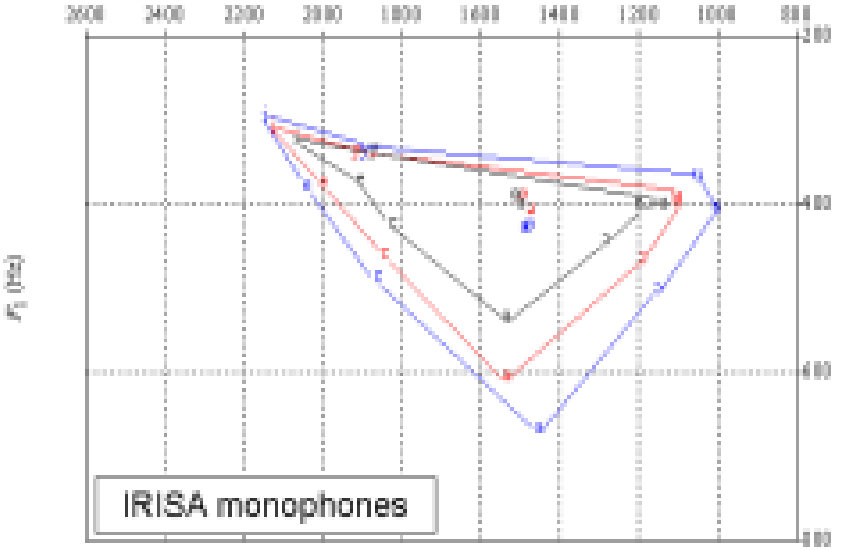
- **Formants :**
  - non détection du 2ème formant de /i/, voyelle reconnue comme mieux perçue par le rapprochement des 3ème et 4ème formants

  - Il en va de même pour les deux premiers formants de /u/ qui se rapprochent Dans ce cas, une baisse du seuil max de détection des formants ➔réduction des taux d'erreurs de détection de 45% à 19%

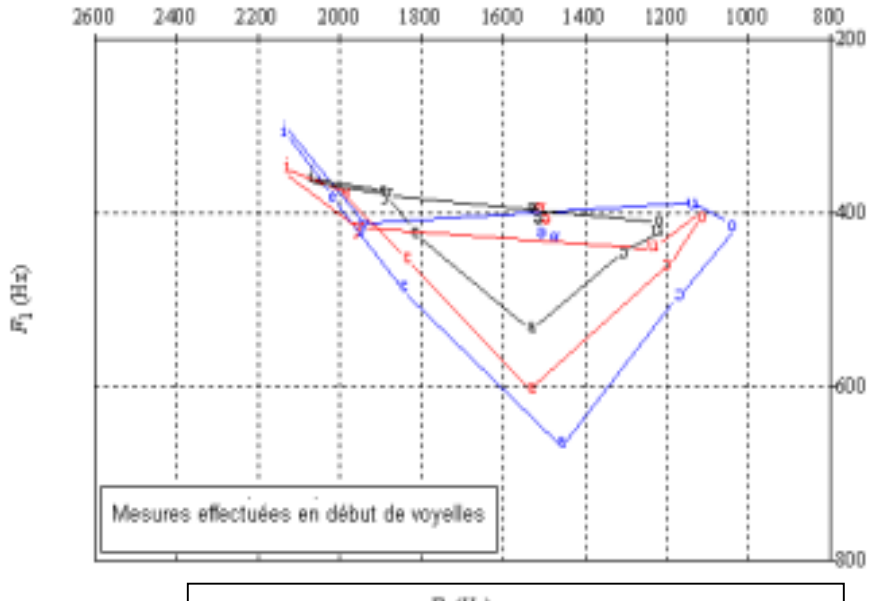
Précision de l'alignement



Segmentation et analyse automatique : Fiabilité des outils et des mesures

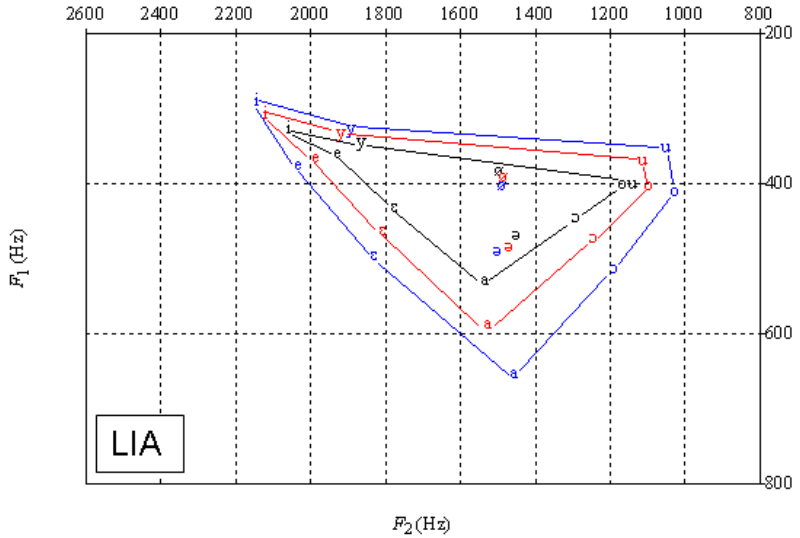


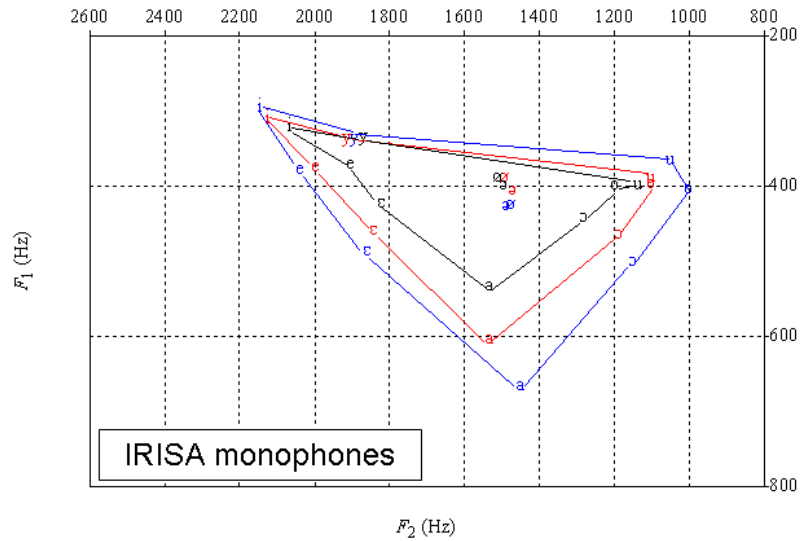
Mesures en partie médiane des phonèmes

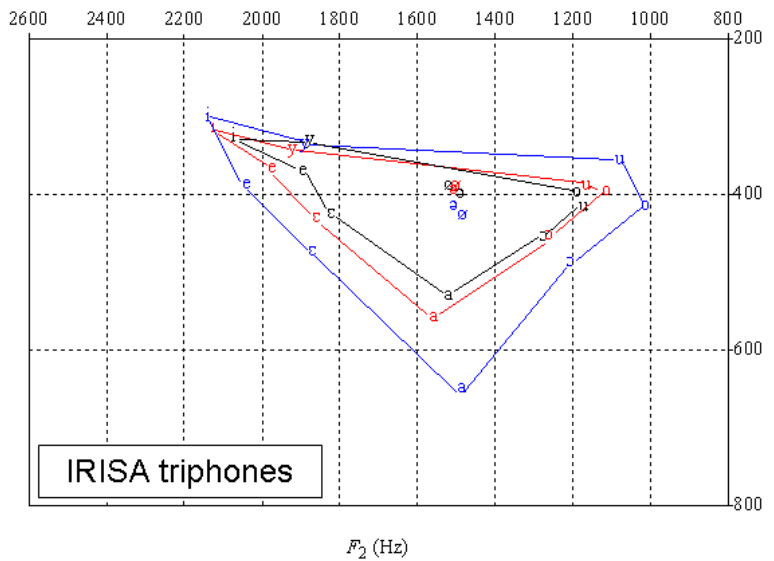


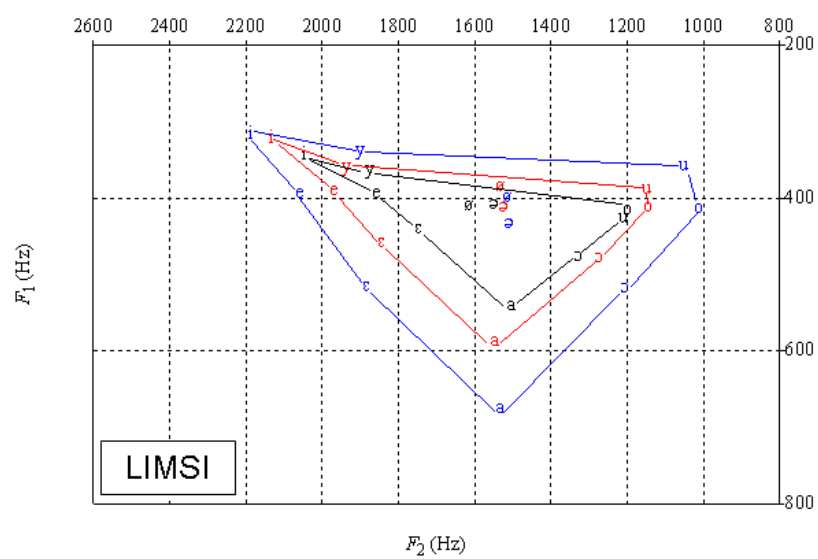
Mesures en début de phonèmes

Mesures de formants sur plusieurs systèmes d'alignement

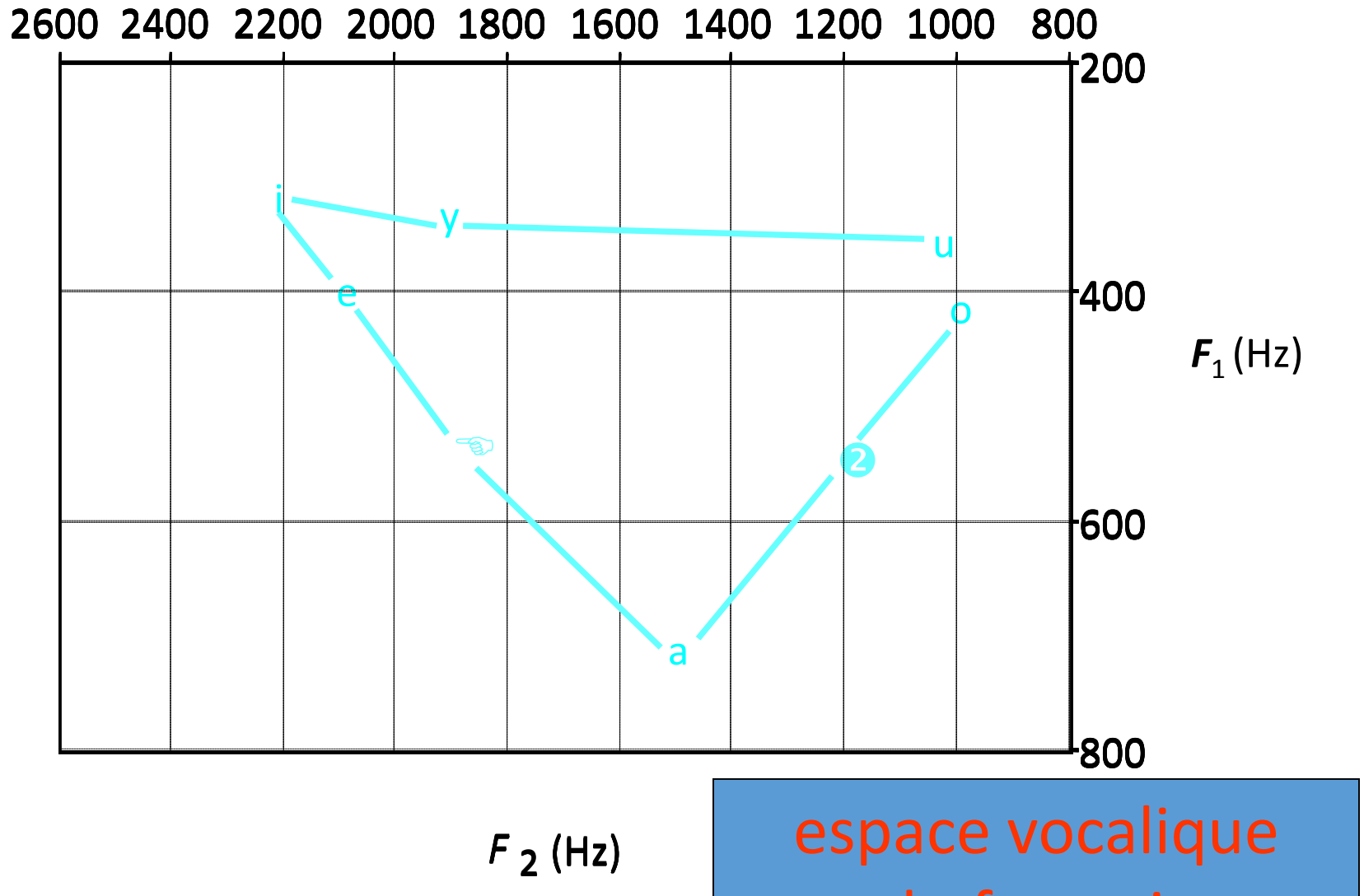








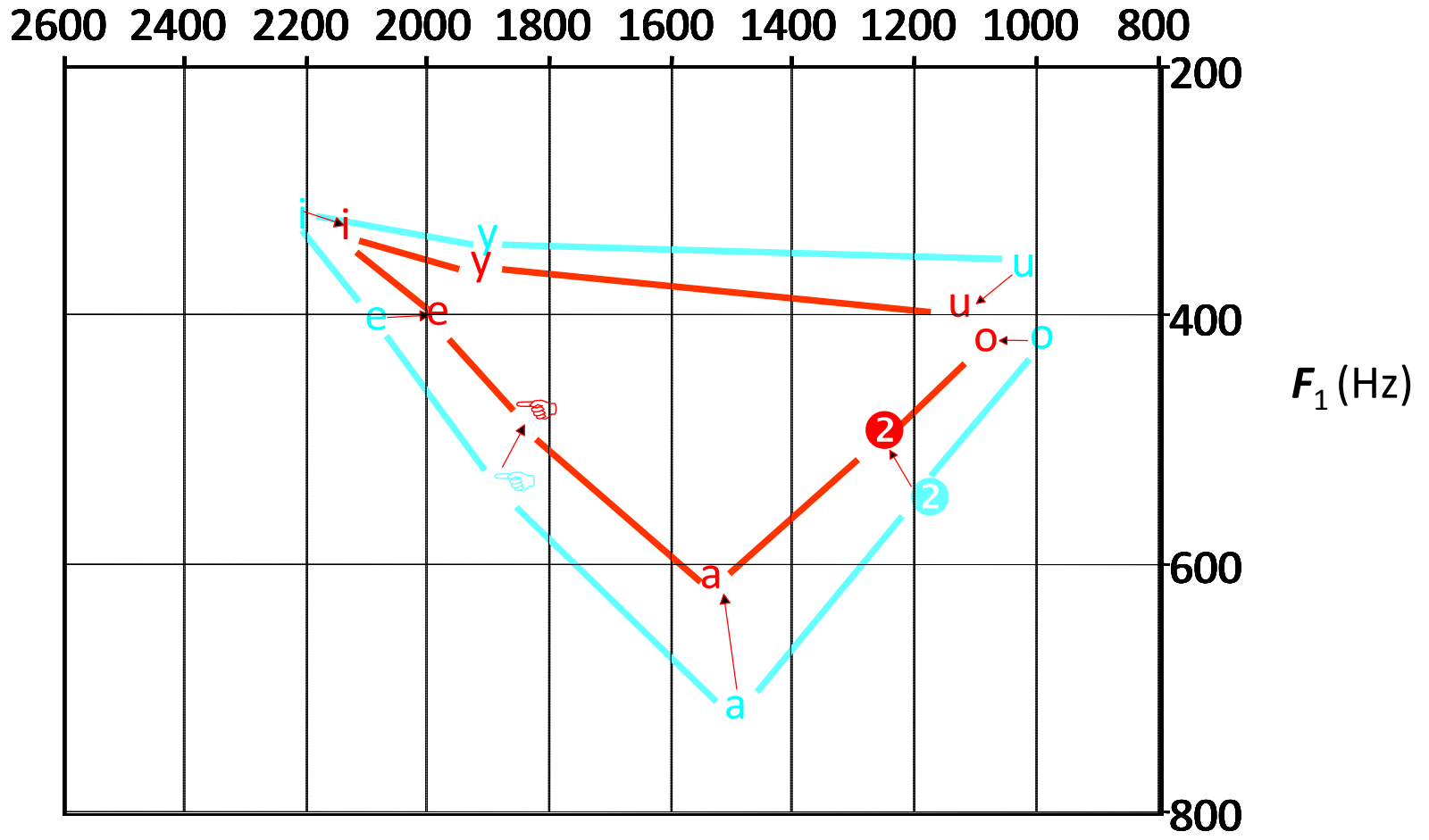
- Interaction entre durée et formants



(ici durée ≥ 90 ms)

espace vocalique du français

- Interaction entre durée et formants

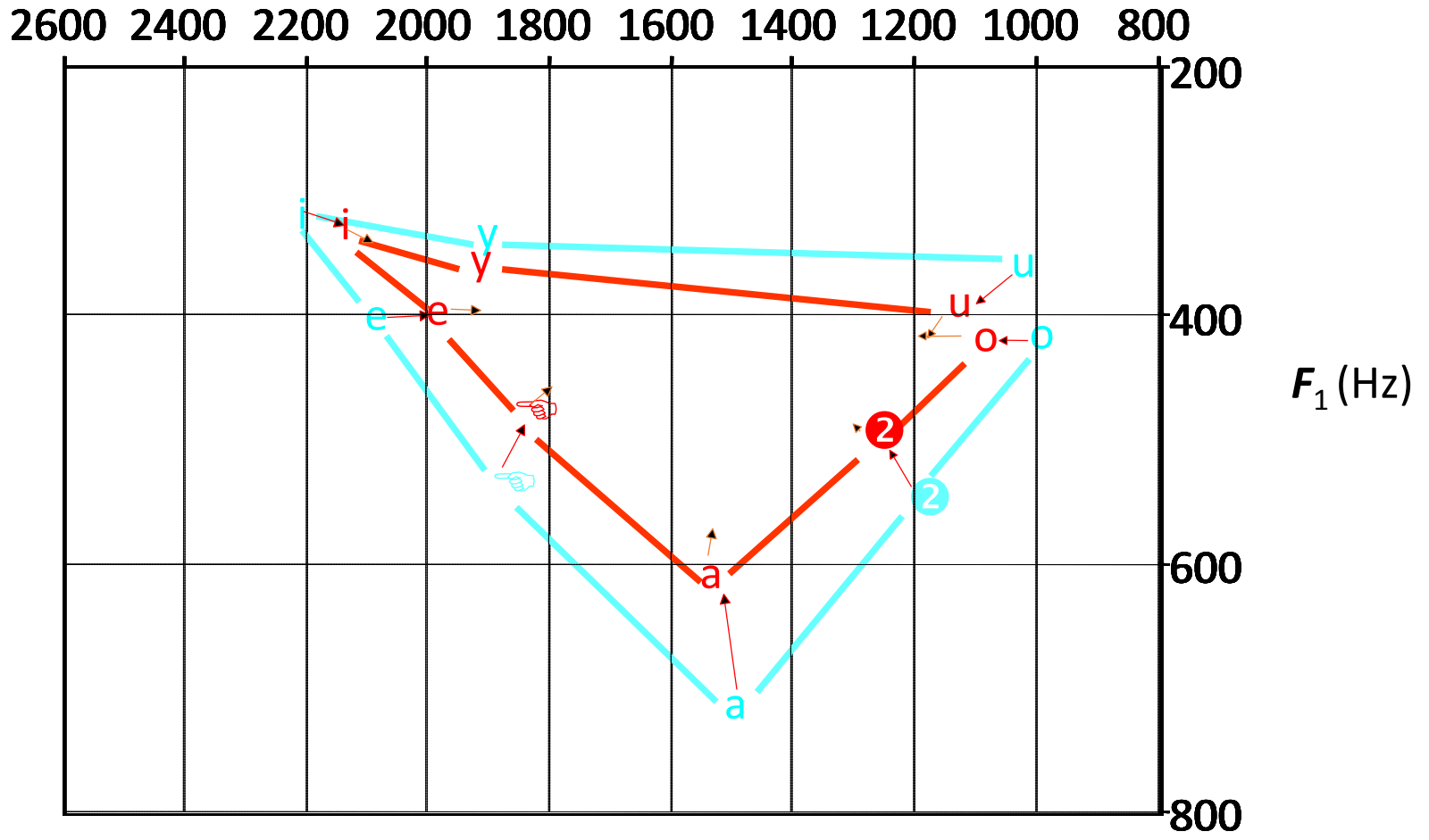


en bleu ... durée ≥ 90 ms

en rouge ... 90ms≥ durée ≥ 60 ms

espace vocalique du français

- Interaction entre durée et formants

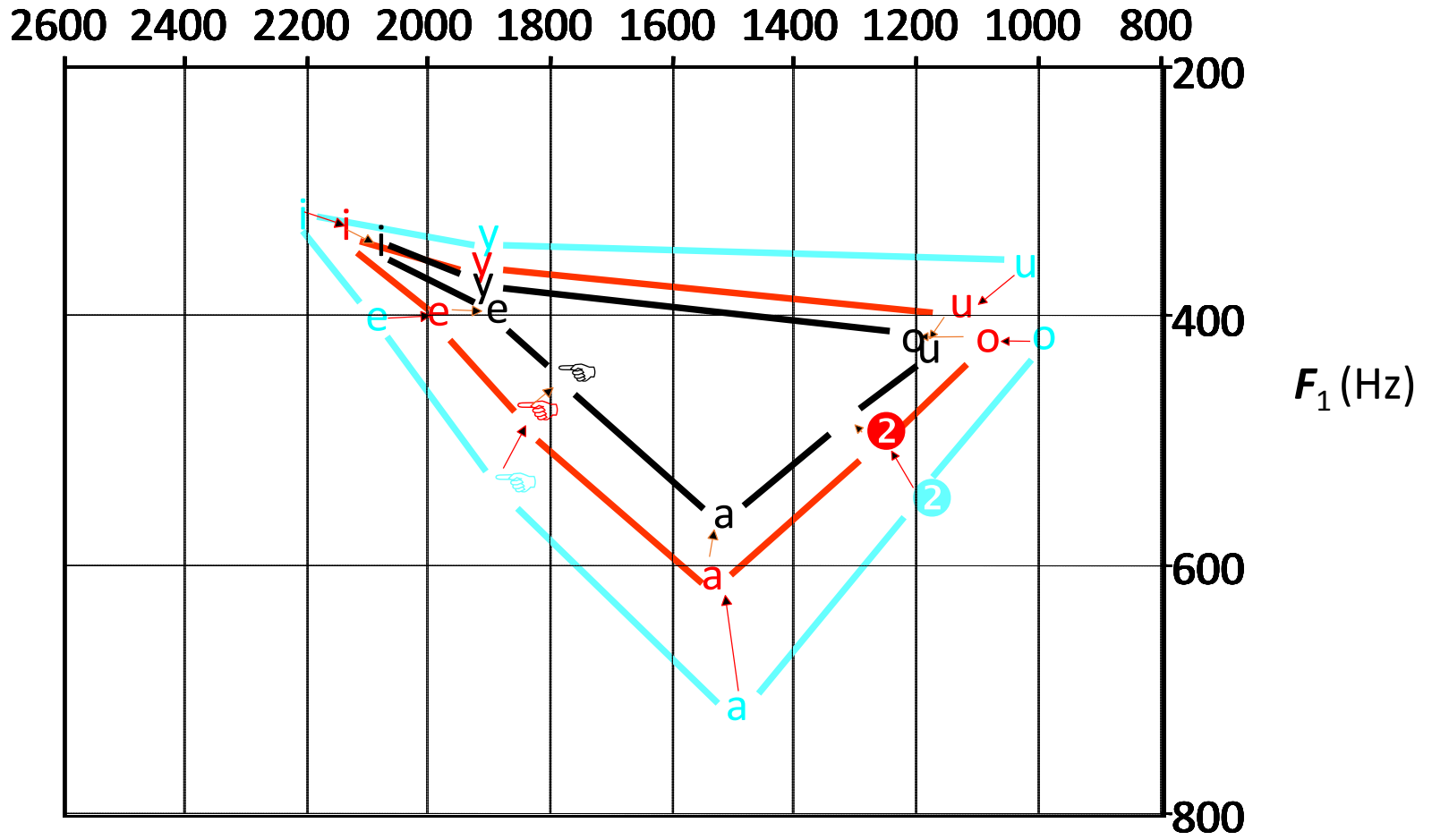


en bleu ... durée ≥ 90 ms

en rouge ... 90ms≥ durée ≥ 60 ms

espace vocalique du français

- Interaction entre durée et formants



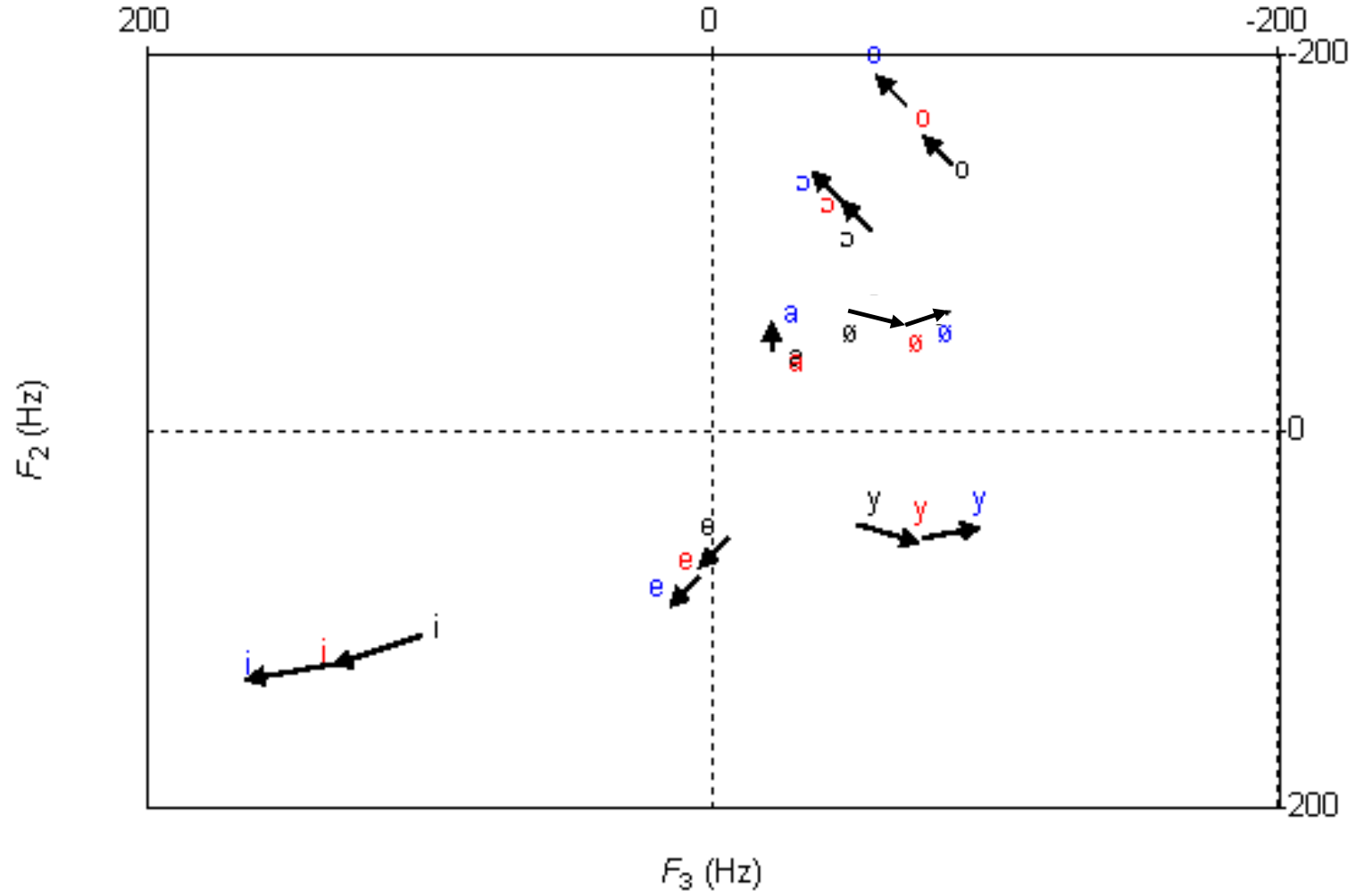
en bleu ... durée ≥ 90 ms

en rouge ... 90ms≥ durée ≥ 60 ms

en noir ... durée ≤ 50 ms

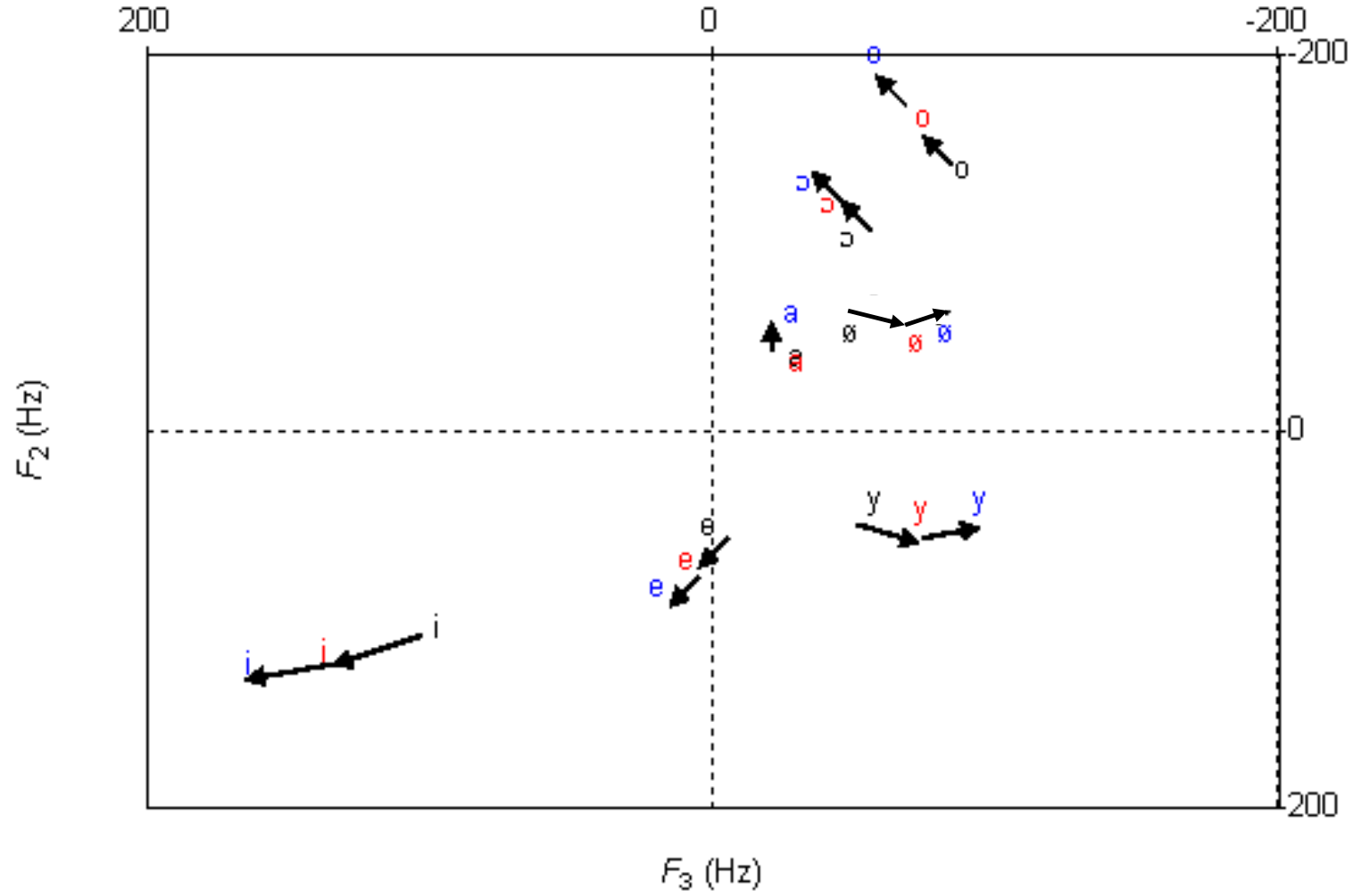
espace vocalique du français

- Interaction entre durée et formants



Lorsque la voyelle est plus longue, F3 augmente pour la voyelle /i/, et c'est **le mouvement le plus large observé parmi toutes les voyelles**.

- Interaction entre durée et formants



Les variations de /y/ : les valeurs baissent quand la durée augmente. Ces observations valent pour /ø/ et /œ/, (cf. arrondissement)

- Interaction entre position prosodique et formants

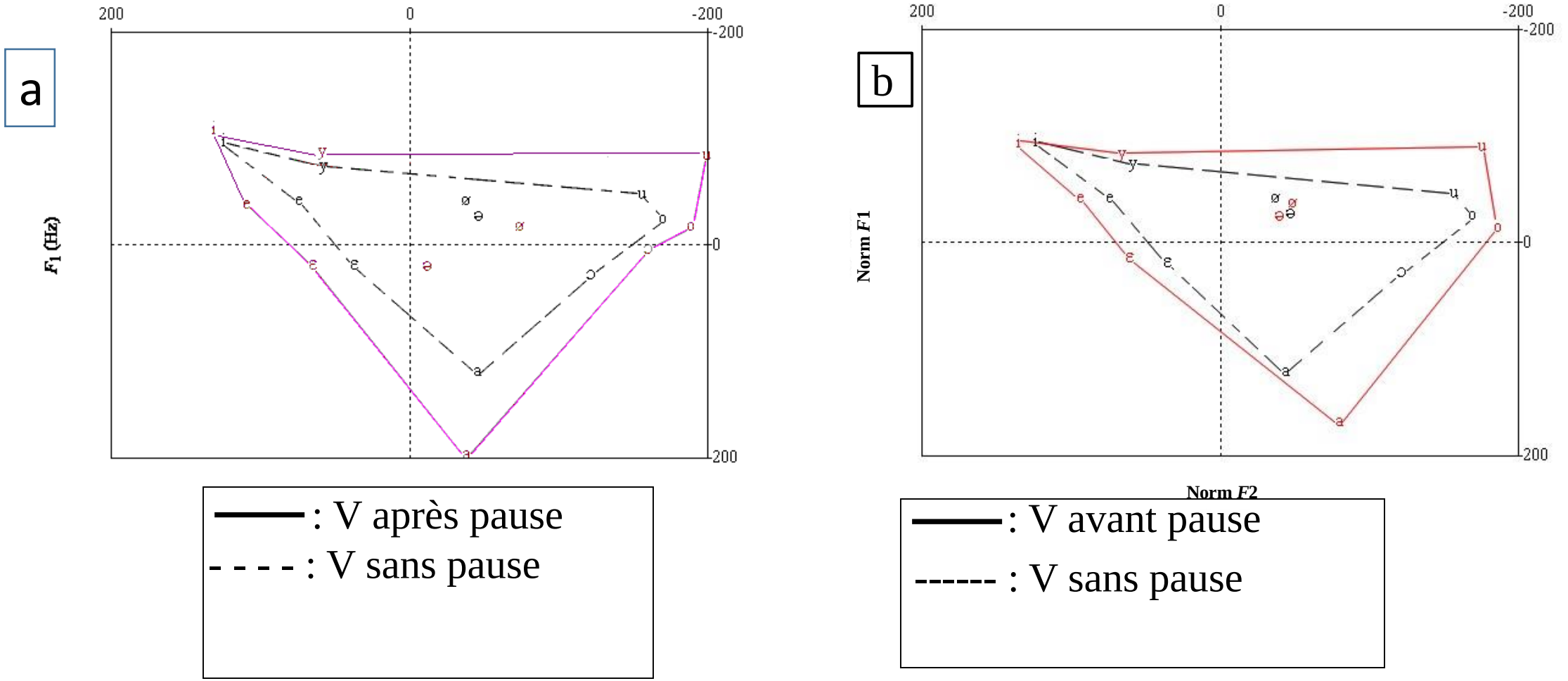
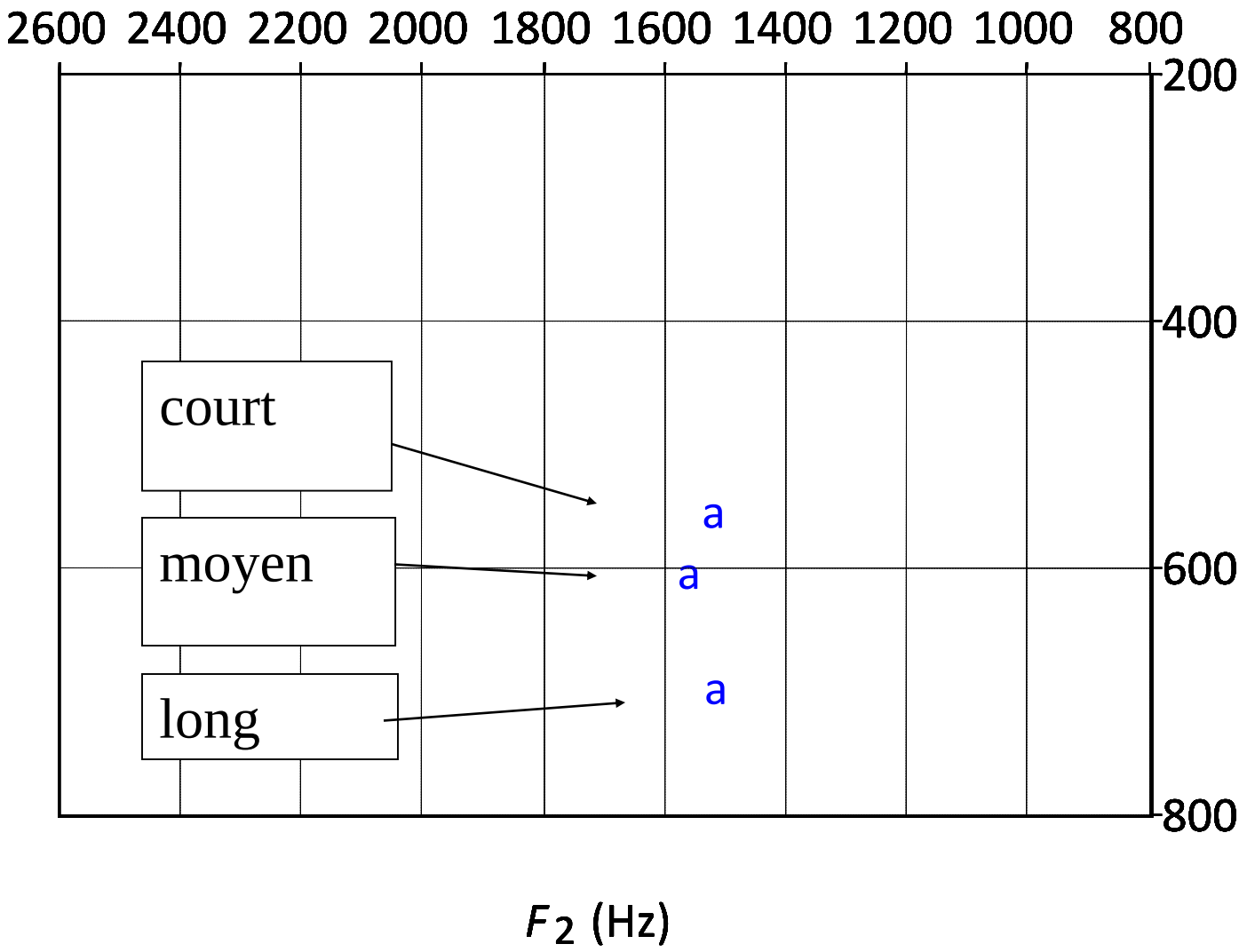


Figure ab: Valeurs moyennes de F1 et F2 pour les voyelles en fonction de l'absence/présence de la pause à proximité de la voyelle analysée.
a(gauche) : pause précédant la voyelle ; b(droite) : pause suivant la voyelle.

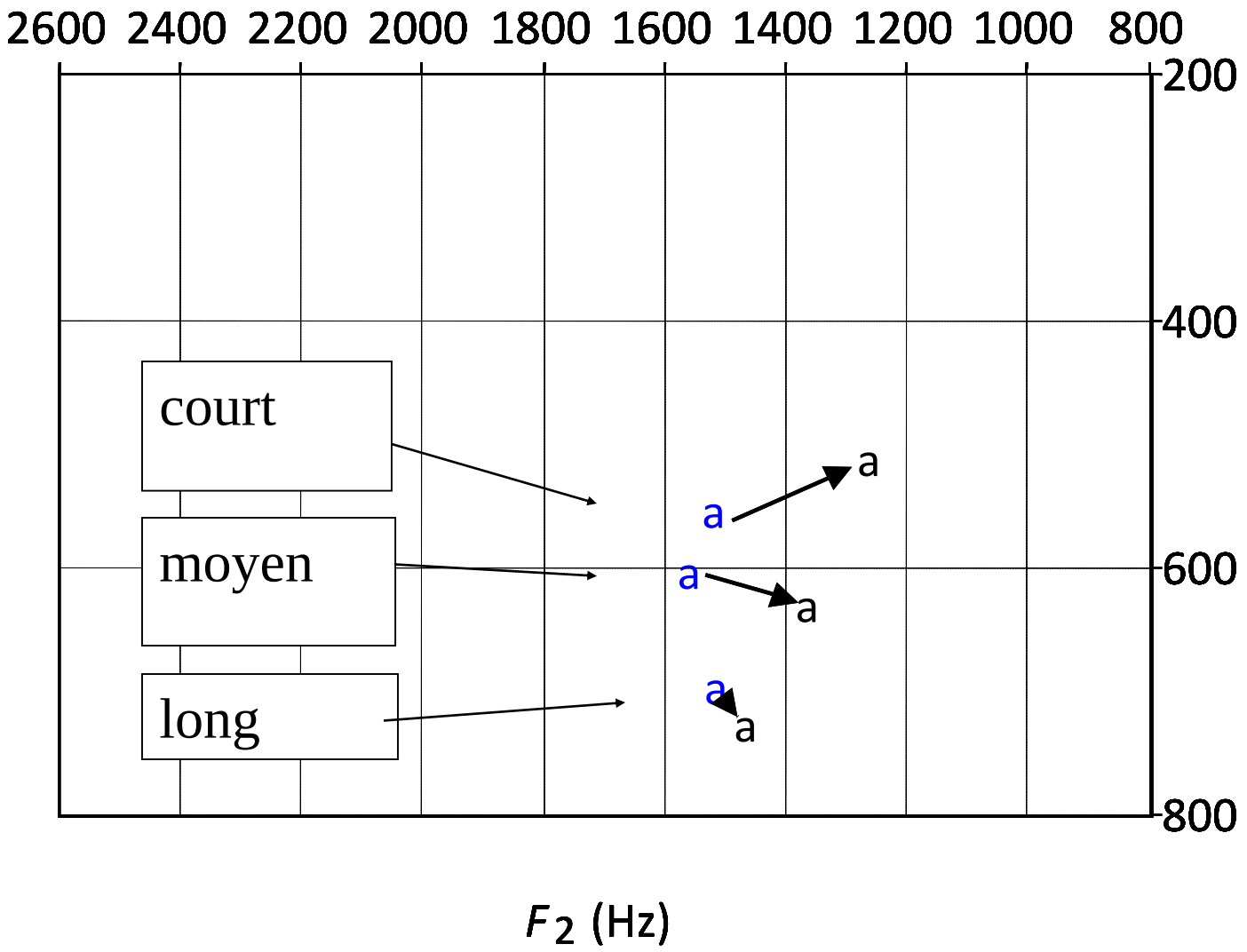
Interprétation

- Peut-on dire qu'on observe une centralisation pour les voyelles les plus courtes ?

➔ Une réduction de l'espace vocalique en tout cas ...

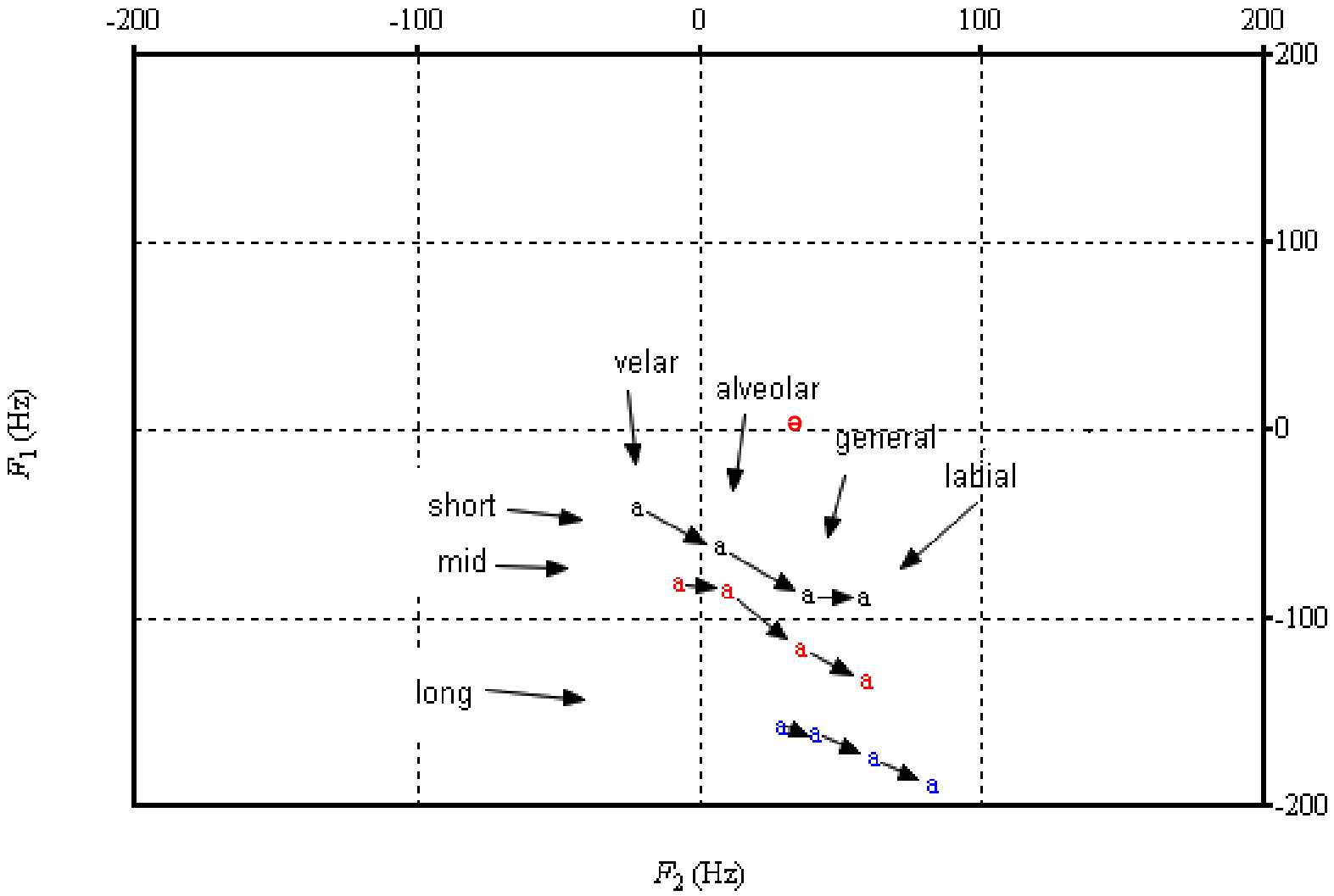


en bleu : tous contextes consonantiques



en bleu : tous contextes consonantiques

en noir : contexte labial seulement


-200
-100
0
100
200
200
100
0
-100
-200
velar
alveolar
general
labial
short
mid
long
ə
a
F_1 (Hz)
F_2 (Hz)

Représentativité

- La méthodologie est fondamentalement inversée en comparaison des corpus linguistiques construits ad hoc.

- La représentativité de certains phénomènes et/ou de certains contextes ne doit pas être négligée.
  - Grande quantité de mots outils en parole continue qui diffèrent considérablement des mots lexicaux dans leur réalisation (?)
  - Spécifiquement pour certains phonèmes : la réalisation sera fortement altérée par la fréquence d'utilisation (/y/ dans « tu »)
  - Style : en parole conversationnelle, « tu » est très fréquent mais rare en parole journalistique

Phonotactique

- /œ/ en français fréquemment suivi d'un /ʁ/ (bonheur, leur, heure), contexte qui modifie sa réalisation moyenne. Ces différences de distribution doivent être prises en compte dans les analyses.

- Notons également le contexte alvéolaire majoritaire en parole continue ...

- ... ou d'autres phénomènes comme la fréquence lexicale, le voisinage phonologique et l'ordre d'apparition.

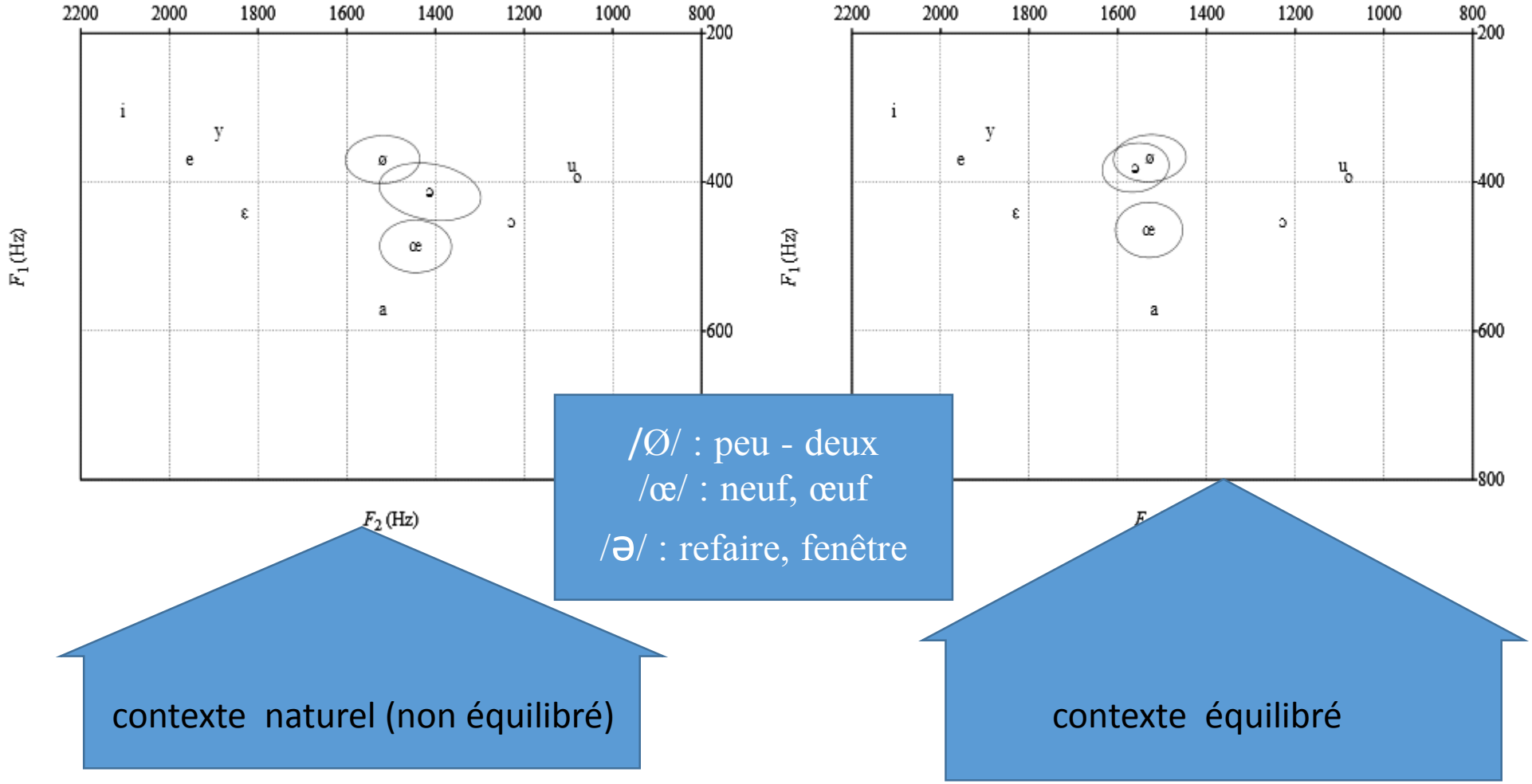
Phonotactique

- Exemple : pour des schwas internes de mots
- Importance de prendre en compte la distribution du contexte phonémique. Fréquence des mots commençant par <re> recommencer, refaire, etc.)

- Les résultats présentés ici indiquent des valeurs de formants très différentes de ceux des contextes équilibrés à cause de la phonotactique.

Précautions méthodologiques

distributions



Mesures de formants pour les voyelles centrales du français. Les résultats diffèrent nettement selon que les contextes sont rééquilibrés (à droite) ou non (à gauche). Les voyelles périphériques (sans ellipses) sont positionnées à titre indicatif

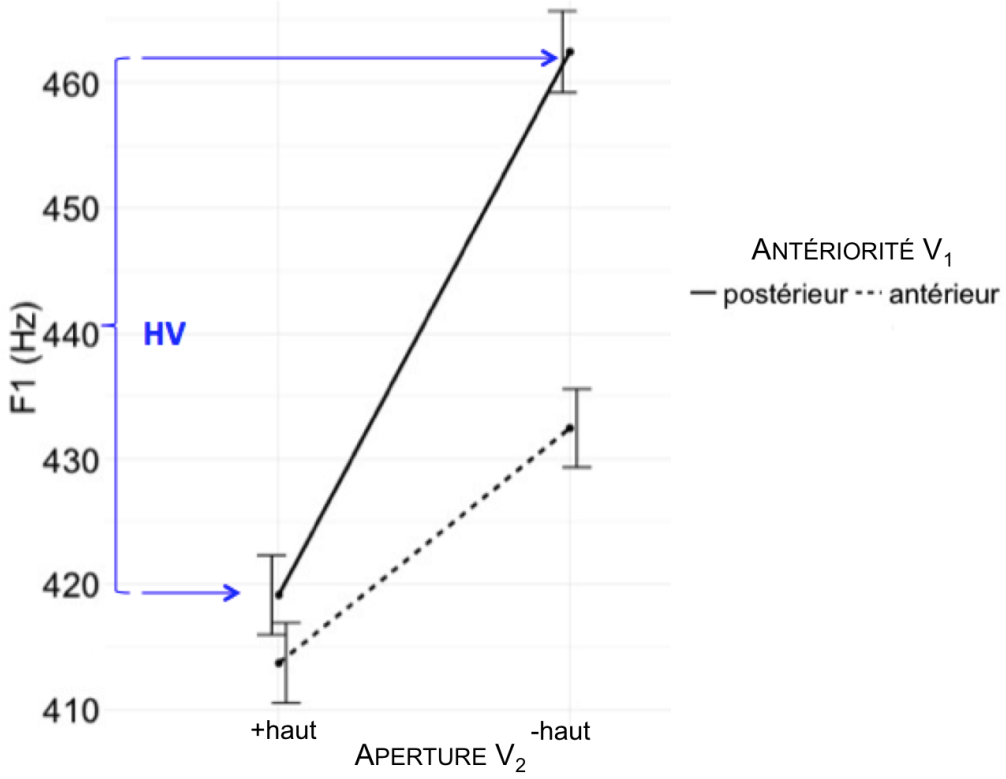
Autre exemple d'étude sur de grands corpus de parole

- Etude de l'harmonie vocalique en français : tendance à la modification du degré d'aperture des voyelles moyennes (mi-ouvertes et mi-fermées) en fonction de celui de la voyelle suivante
- Exemple : *aimait* [**ɛ**mɛ] / *aimer* [**e**me]
- Objectif : mieux comprendre où apparaît l'harmonie vocalique, et dans quelle mesure les facteurs décrits dans la littérature la conditionnent

(Turco, Fougeron, Audibert, 2016)

Autre exemple d'étude sur de grands corpus de parole

- Mesure retenue de l'harmonie vocalique : différence de F1 sur V1 entre une V2 ouverte et une V2 fermée
- Evaluation sur 33k mots (ESTER 19k – NCCFr 14k)



(Turco, Fougeron, Audibert, 2016)

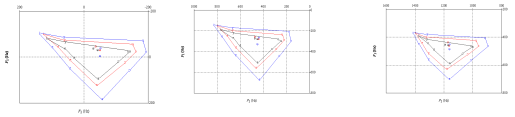
Autre exemple d'étude sur de grands corpus de parole

- Plus d'harmonie vocalique quand :
  - La 1e voyelle V1 est postérieure
  - Une graphie telle que « é » ou « au/eau » favorise une prononciation mi-fermée
  - V1 et V2 séparées par une consonne labiale (/p, b, m, f, v/) dont l'articulation n'implique pas la langue, plutôt que linguale
  - V1 et V2 ne sont pas séparées par un schwa sous-jacent (« e muet » non prononcé mais supposé appartenir à la représentation du mot)

(Turco, Fougeron, Audibert, 2016)

Normalisations ...

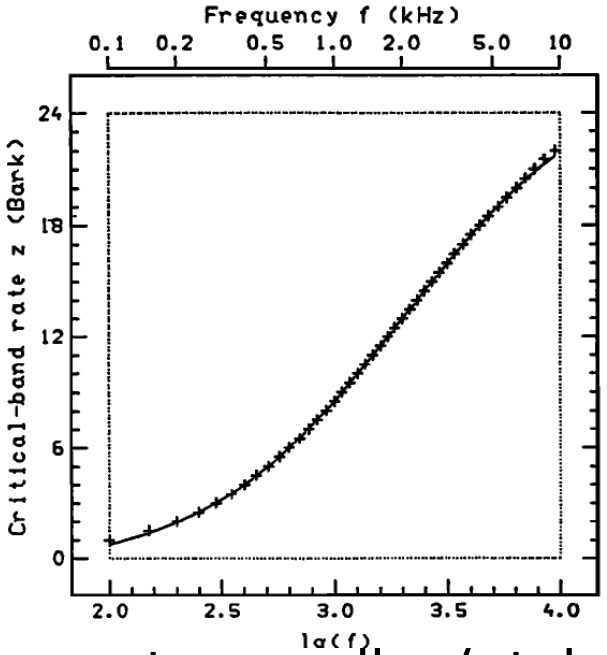
- De par le grand nombre de locuteurs présents, des procédures de normalisation sont souvent utilisées (Lobanov, Gerstman, Bark, F’2, etc.)

Cf. Adank 

- Des normalisations spécifiques au mesures spectrales (Mel, Bark, etc.)
- Ou aux mesures de f0 (demi-tons)
- Mais aussi des normalisations standards (z-score)

Corpus et méthodes : valeurs formantiques

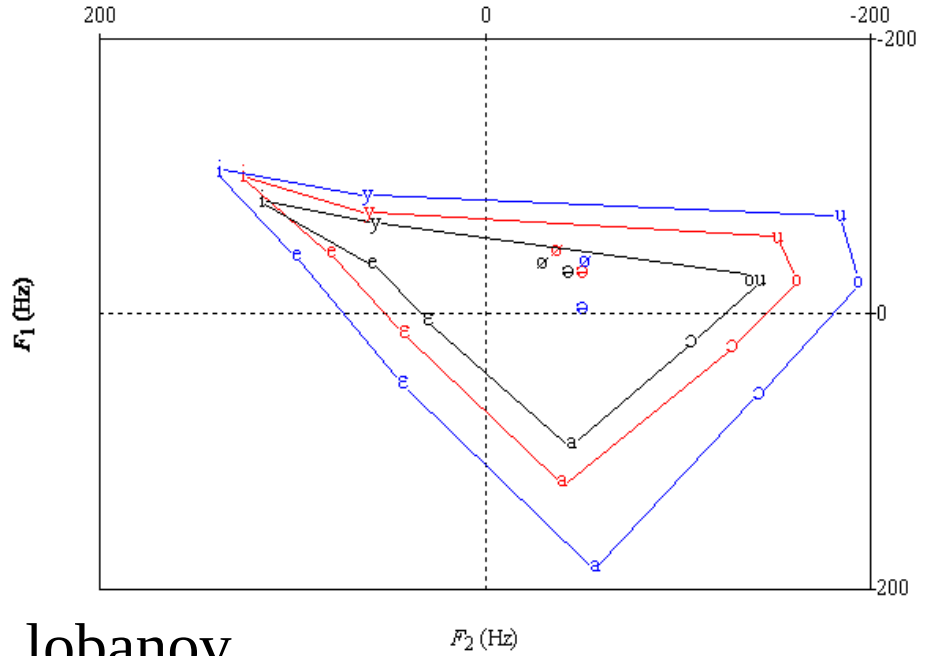
- Conversion des fréquences formantiques de Hertz en Bark (Traunmüller, 1990)



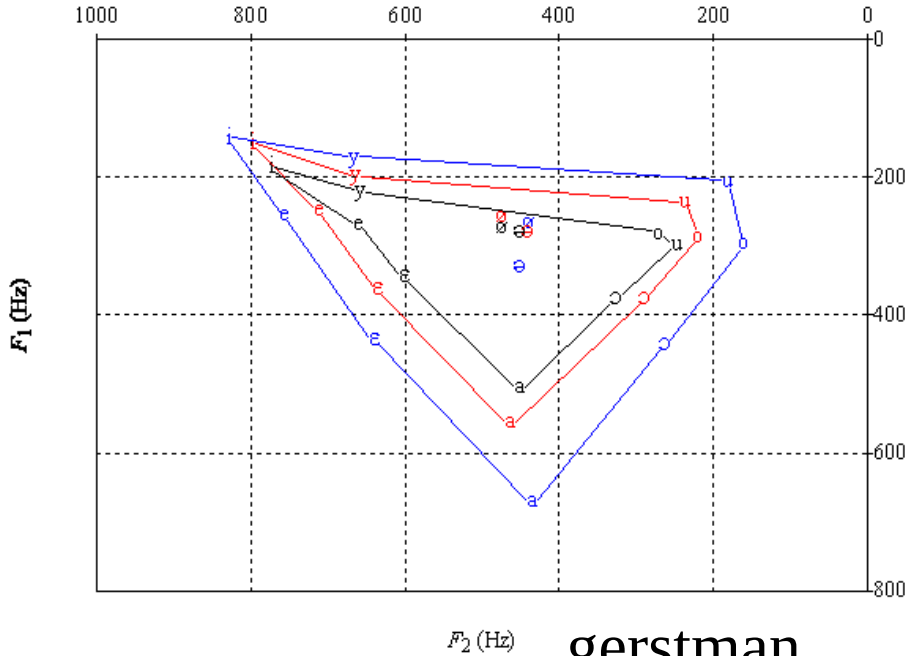
Tables de correspondances issues d'expériences de psychoacoustique, approximées par l'équation :

$z = [26.81/(1 + 1960/f)] - 0.53$

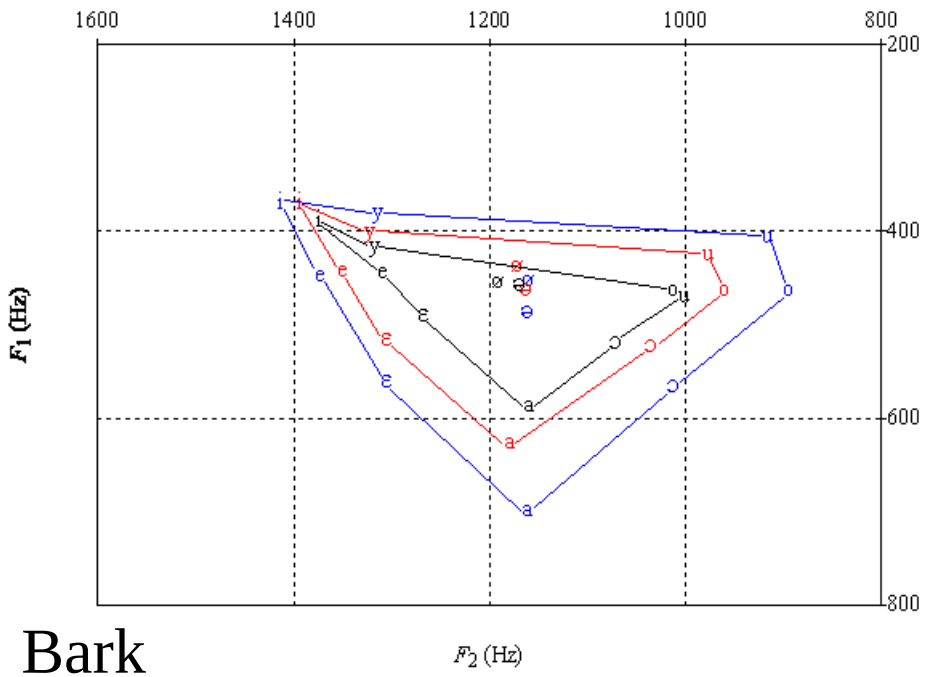
=> Mesures de distances entre voyelles (et donc métriques pour quantifier les variations de l'espace vocalique) plus conformes à la perception



lobanov



gerstman



Bark

Comment quantifier la réduction vocalique ?

- *Reduction or expansion of a vowel space is "neither uniform nor simple"* (Ferguson & Kewley-Port, 2002, 2007, voir aussi Harmegnies & Poch-Olivé, 1992)

⇒ Multiples dimensions relatives aux valeurs formantiques à prendre en compte

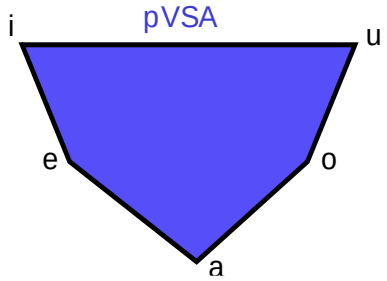
- Centralisation / réduction de l'espace vocalique
- Dispersion au sein de chaque catégorie de voyelle
- Neutralisation des contrastes entre catégories

Dimensions corrélées, mais seulement jusqu'à un certain point...

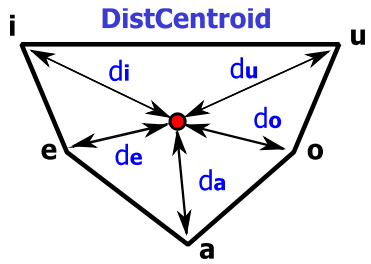
Mesure des variations vocaliques : Centralisation / réduction

- Points de référence : centroïdes des catégories

- Générale

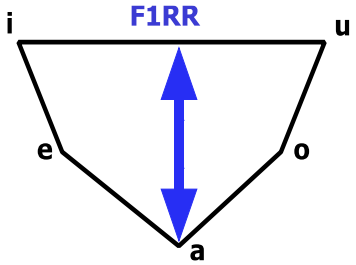


Aire du polygone

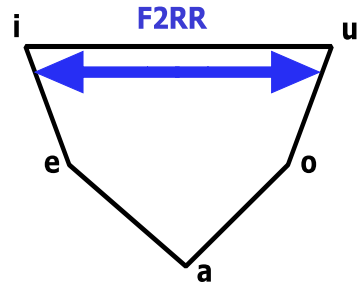


Distance moyenne au centroïde

- Spécifique à une dimension acoustique (Sapir et al. 2010)

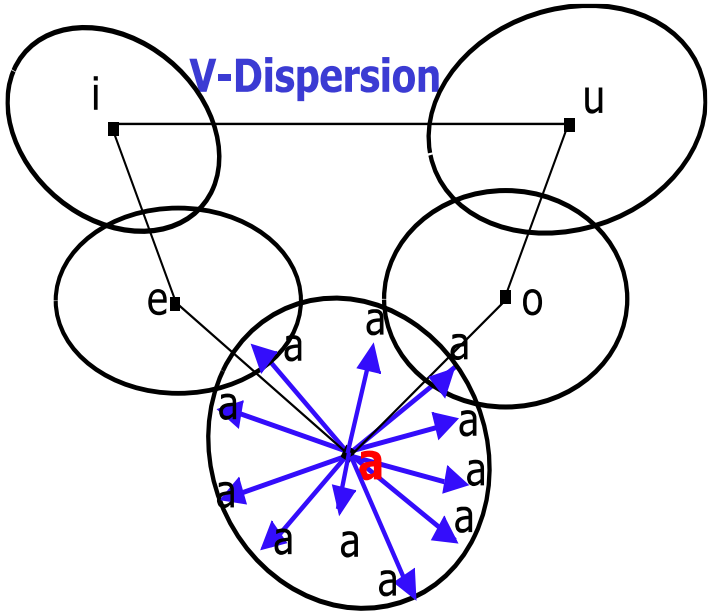


Amplitude sur F1



Amplitude sur F2

Mesure des variations vocaliques : Dispersion intra-catégories



Aire moyenne des ellipses de confiance à 95%

(ellipse qui inclut 95% des valeurs en considérant une distribution normale bivariée)

Alternative : distance moyenne au centroïde de chaque catégorie

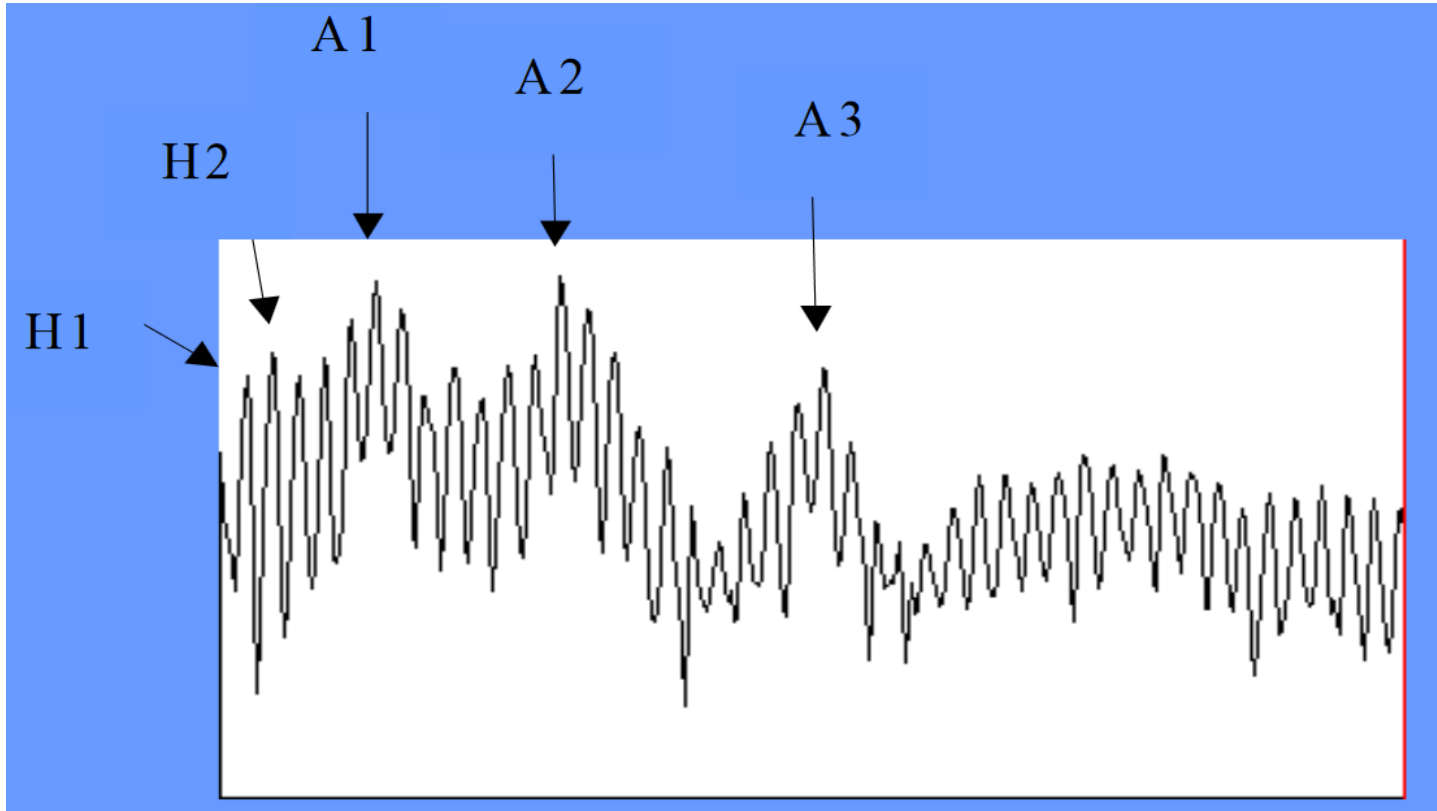
D’autres types de mesures

Nous avons vu les mesures phonétiques classiques, d’autres bien sûr ont été envisagées dans la littérature, celles-ci visent souvent à calquer des mesures physiologiques

- Mesures de qualité vocale

(logiciel VoiceSauce)

  - h1 – h2, (souffle)
  - h1 – A3 (pente spectrale, voix tendue)
  - Cepstral Peak prominence



D’autres types de mesures

- Mesures acoustiques de la nasalité (A1 – P0, Chen, 1997)

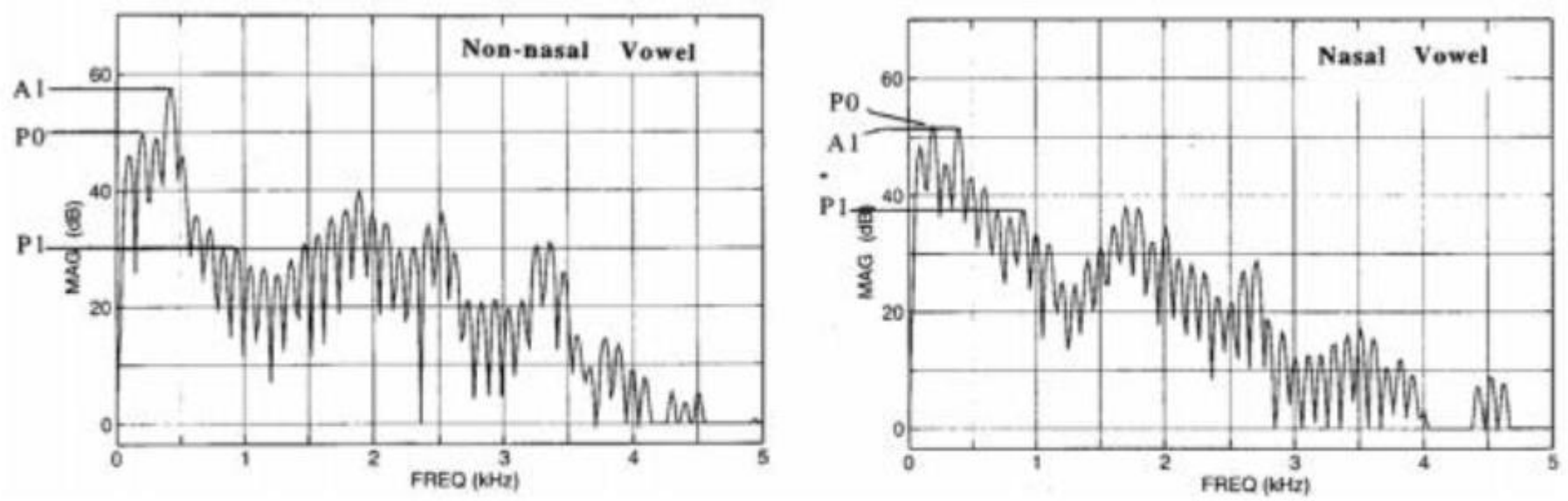


Figure 8: Spectra of nasal (right) and non-nasal (left) vowels with marled A1 and P0. It can be seen that the amplitude of P0 is boosted relative to A1 in the nasalized vowel. After Chen (1997)

D’autres types de mesures

- Rapport Harmoniques sur bruit (friction, raucité)

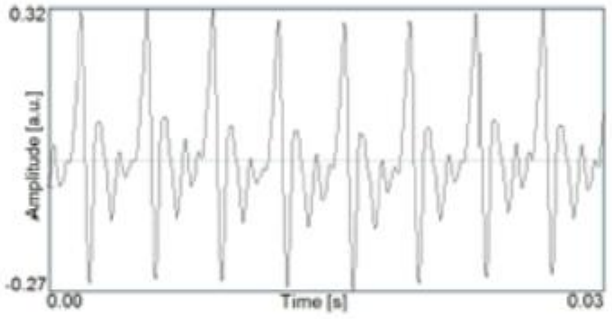


Figure 1: Wave shape of the /a/ sound.

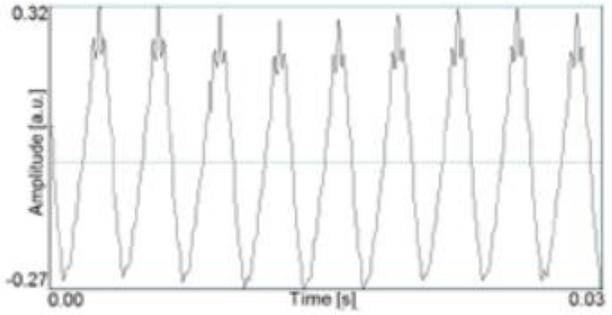


Figure 2: Wave shape of the /i/ sound.

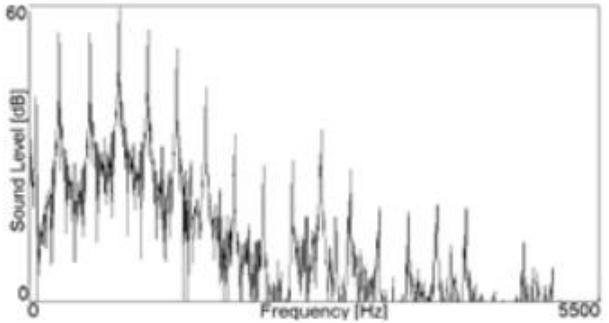


Figure 3: Harmonics of the /a/ vowel.

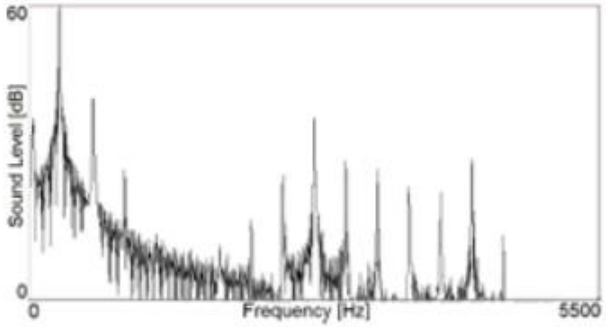
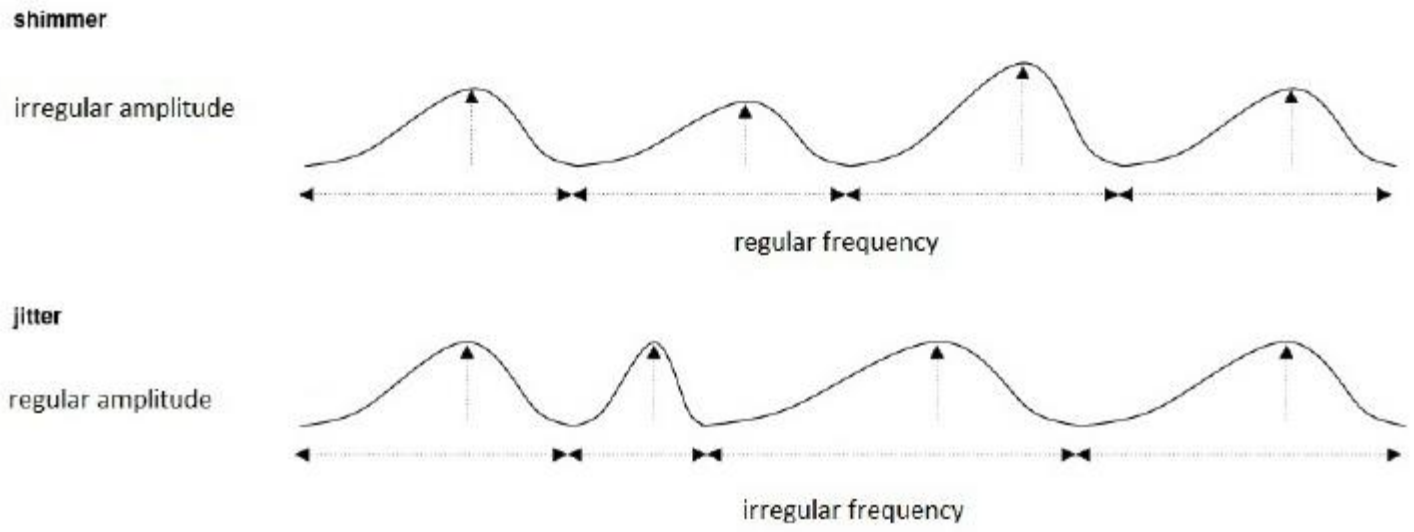


Figure 4: Harmonics of the /i/ vowel.

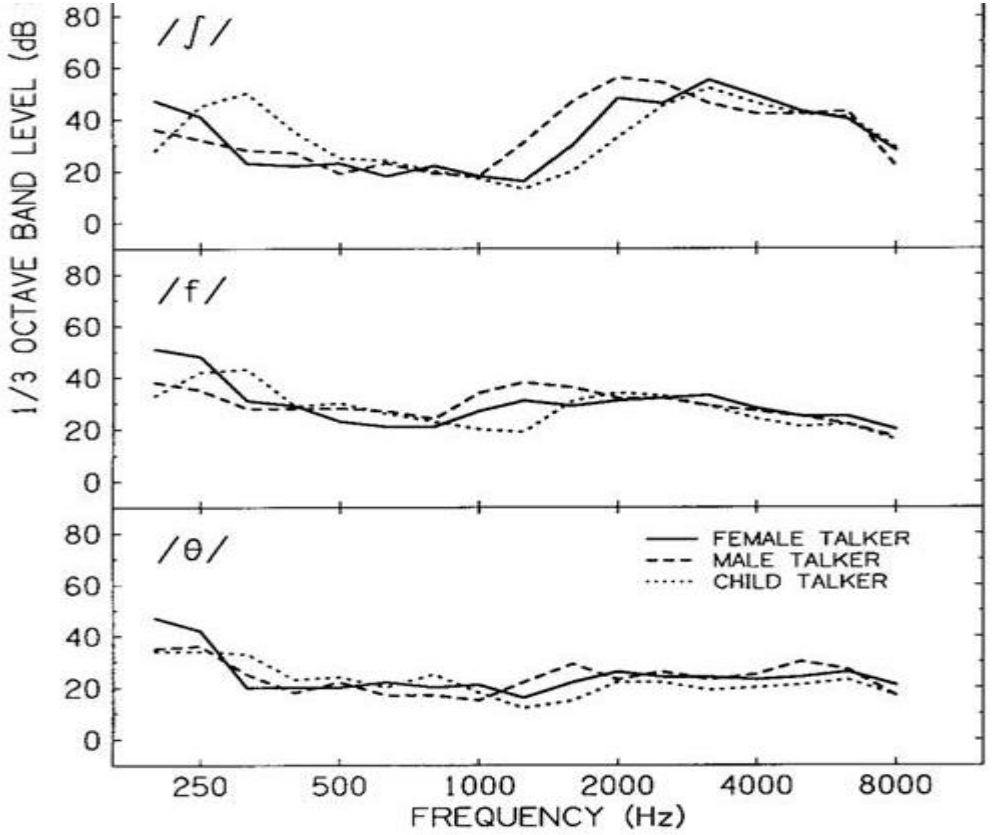
D’autres types de mesures

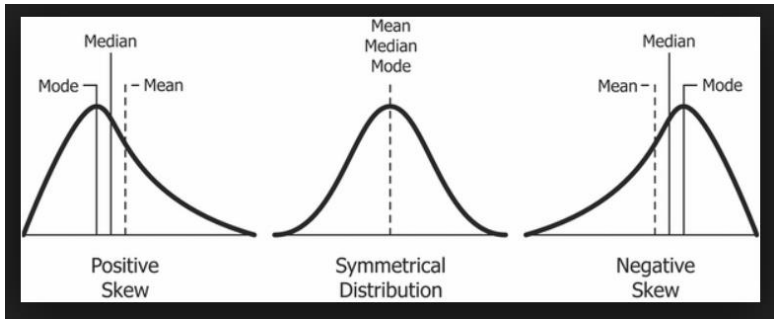
- jitter, shimmer (stabilité de la voix)



D’autres types de mesures

- Moments spectraux (stridence, pente, …)
  - Le spectre est considéré comme une distribution de données et on mesure sa moyenne, son asymétrie, son aplatissement, son écart-type





D’autres types de mesures

- Celles-ci sont dépendantes du phonème analysé, et sont parfois difficiles à mesurer, et/ou nécessitent une taille de fenêtre d’analyse incompatibles avec la parole continue.

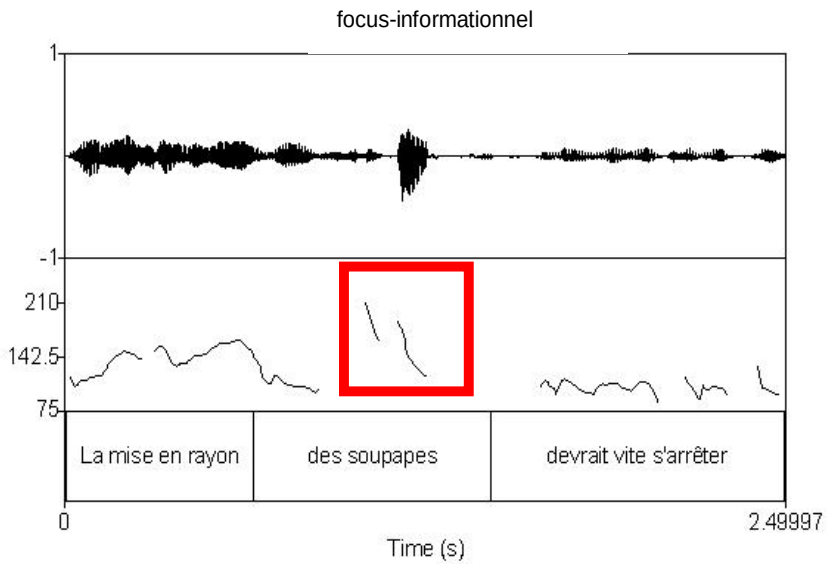
- Une alternative est de revenir aux MFCC pour discriminer des catégories pré-établies.

Préambule

- Mesure locales (paradigmatiques) vs. Globales (syntagmatiques)

- Mesures statiques vs. Dynamiques

f0



f0