UNIVERSITÉ DE RENNES 1

CENTRE DE MOBILITÉ INTERNATIONALE DE RENNES

Évaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française

Sébastien Le Maguer

Sous la co-direction de Nelly Barbot et Olivier Boëffard

IRISA/Université de Rennes 1, Lannion, France

2 Juillet 2013

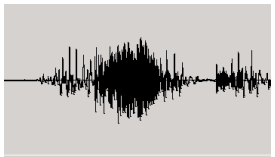
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTEMES ALÉATOIRES

UMR IRISA

Contexte - La synthèse de la parole

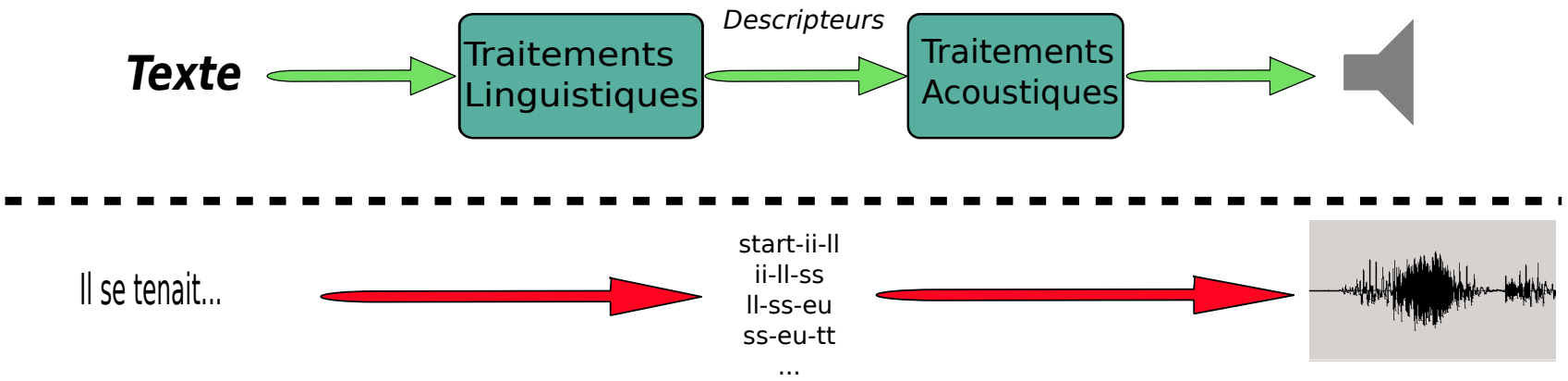
Contexte - La Synthèse TTS

Texte



Il se tenait...

Contexte - La Synthèse TTS



Contexte - La synthèse TTS

60 Synthèse par règles[?]

- Modélisation paramétrique basée sur les formants

70/80 Synthèse par diphones[?, ?]

90 Synthèse par corpus[?]

- Évolution de la synthèse par diphones
- Prise en compte du séquencement temporel

2000 Synthèse par règles statistiques[?]

- Modélisation paramétrique basée sur STRAIGHT[?]
- Évolution de la synthèse par règles
- HTS (HMM-Based Speech Synthesis System)

Contexte - Objectifs de la thèse

Objectif 1

Mise au point du système HTS pour la synthèse du français.

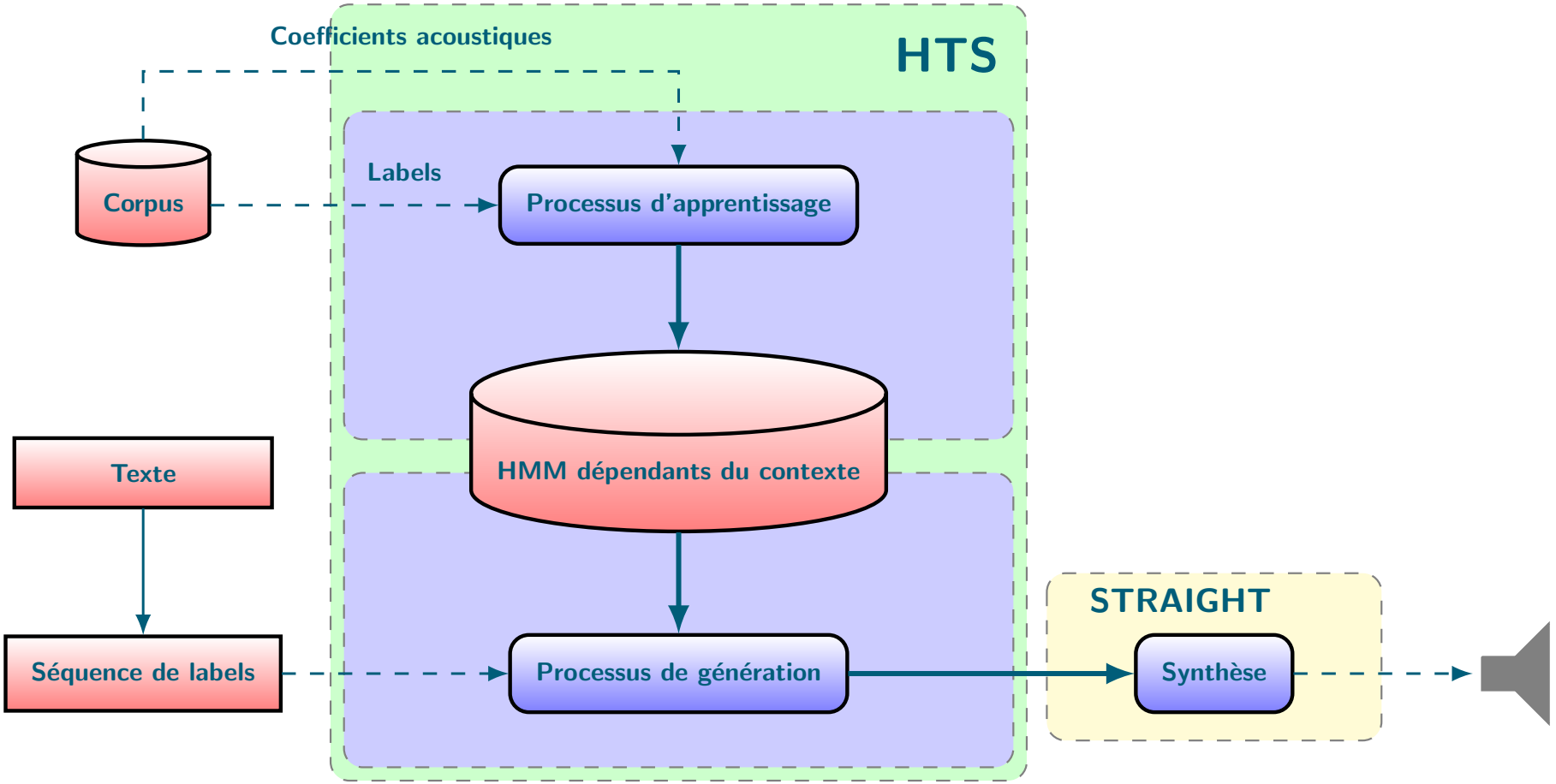
Objectif 2

Analyse de l'influence des descripteurs sur la qualité de la synthèse.

Déroulement

Overview

HTS - Architecture du système



HTS - Descripteurs linguistiques

- Anglais[**?**]
  - Jeu standard : 53 descripteurs
  - 5 horizons : phone, syllabe, mot, phrase, énoncé
  - Information : position, accentuation, étiquettes grammaticales et syntaxiques, étiquettes ToBI[**?**]

- Autres langues :
  - Une quinzaine de langues recensées dans le cadre HTS[**?**],
  - Allemand : 70 descripteurs[**?**],
  - Croate : 3 descripteurs (contexte phonétique direct)[**?**],
  - Descripteurs dérivés du jeu standard.

HTS - Problématique

- HTS = Système de synthèse statistique
- Une combinatoire élevée des descripteurs
- **Mise au point du système HTS pour la synthèse du français**

Problématique

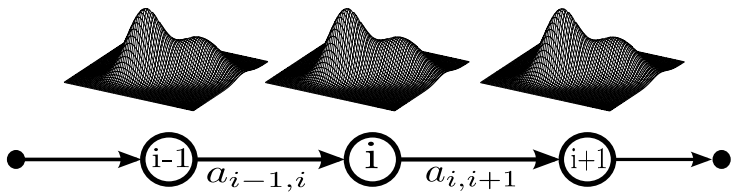
Quel est l'impact du choix des descripteurs sur la qualité du signal produit ?

HTS - Modèle

- Un segment phonétique = un HMM

Un énoncé = une phrase-HMM Λ obtenue par concaténation des HMM correspondants aux segments phonétiques

- Topologie de Bakis
- Durée :
  - 1 trame dure 5 ms
    5 états émetteurs
    $\Rightarrow$ durée min. d'un segment = $25ms$
  - Utilisation de HSMM[**?**]



HTS - Coefficients acoustiques

Objectif du système HTS

Produire une séquence de coefficients acoustiques C qui maximise $P(C|\Lambda)$

- Utilisation du modèle STRAIGHT[**?**]
- 3 types de coefficients :
  - F0
  - Coefficients MGC[**?**] (spectre sur une échelle logarithmique)
  - Bandes d'apériodicité BAP

HTS - Coefficients acoustiques

MGC
ΔMGC
$\Delta^2 MGC$
$f0$
$\Delta f0$
$\Delta^2 f0$
BAP
ΔBAP
$\Delta^2 BAP$

$o_t \in O$

- Apprentissage uniquement à partir des coefficients statiques[?]

  $\Rightarrow$ *faible* qualité du signal

- Introduction de la dynamique[?]
  - Matrice de fenêtrage W imposée
  - Dérivées d'ordre 1 et 2

Objectif de la phase d'apprentissage

Obtenir les HMM qui maximisent $P(O|\Lambda)$ où $O = W.C$

HTS - Modélisation des paramètres

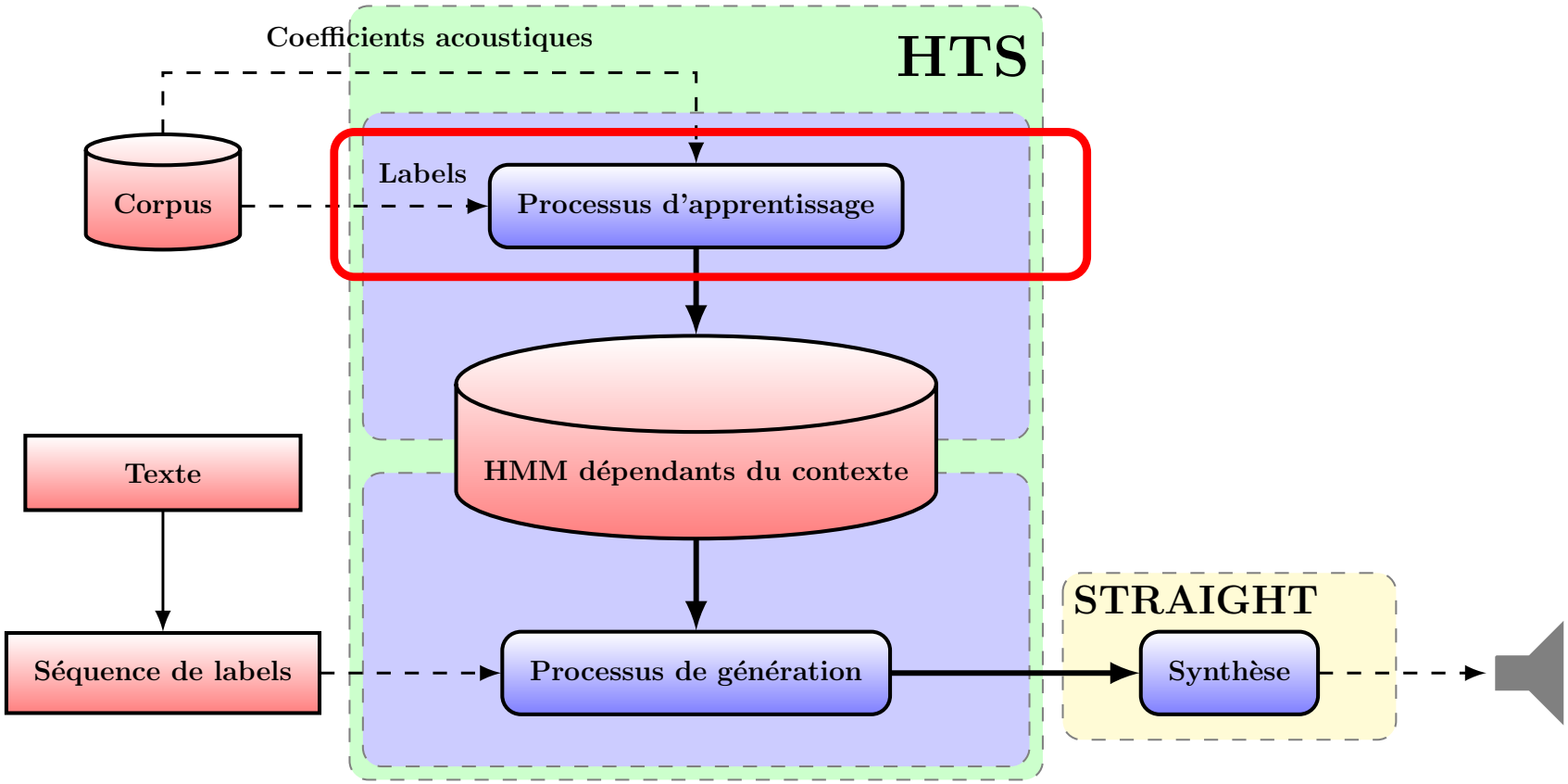
- MGC, BAP : modélisation Gaussienne

$$b(o_t) = \mathcal{N}(o_t; \mu, \Sigma)$$

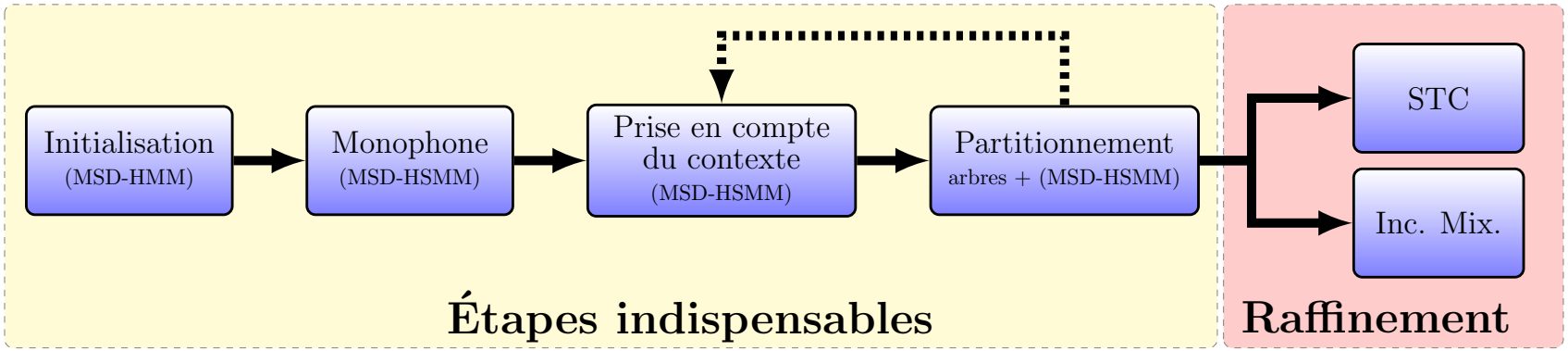
- F0
  - Problématique : 2 classes de valeurs (voisé/non voisé)
  - Utilisation de MSD[**?**]

$$b(o_t) = \begin{cases} w_1 \ \mathcal{N}(V(o_t); \mu, \Sigma), & S(o_t) = \{1\} \\ w_2 \ \delta_0(V(o_t)), & S(o_t) = \{0\} \end{cases}$$

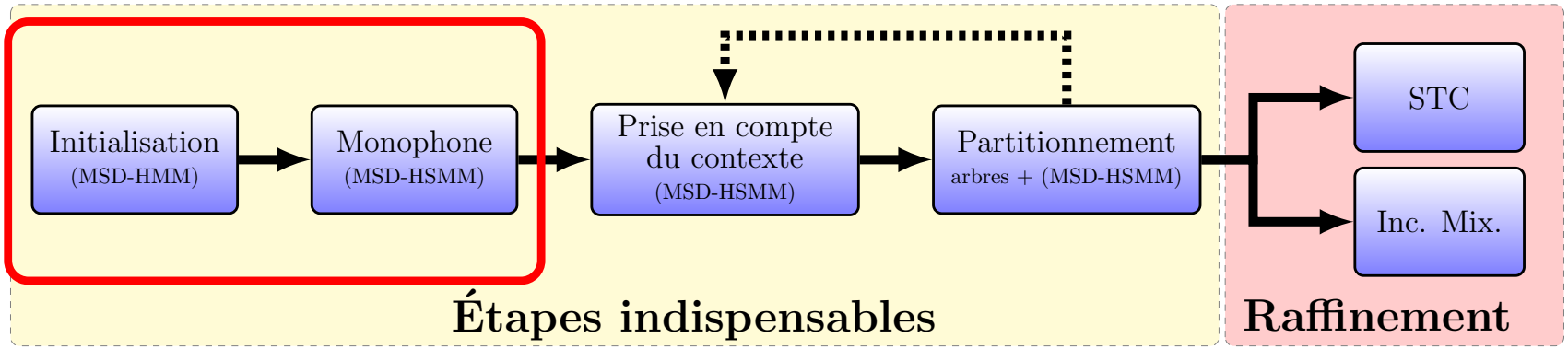
HTS - Architecture du système



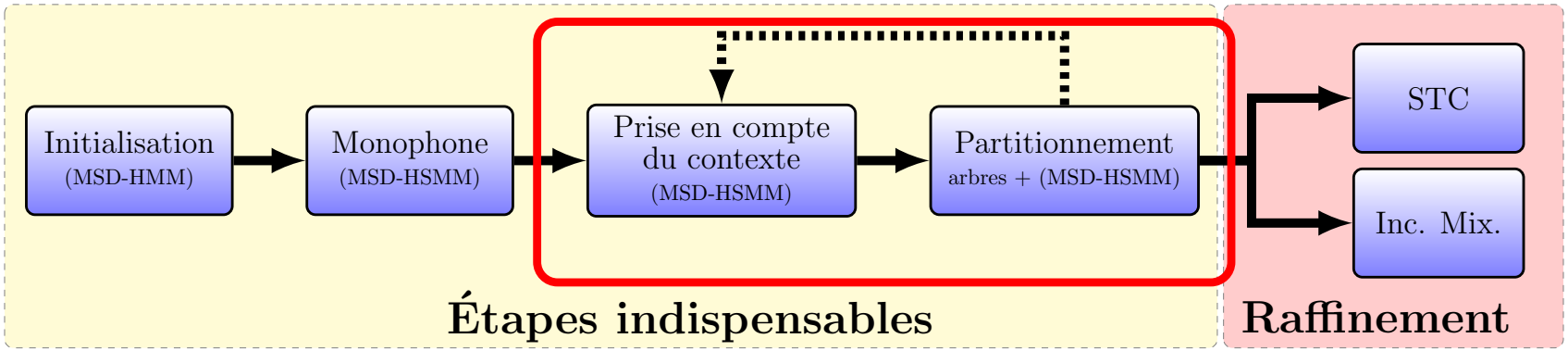
HTS - Processus d'apprentissage



HTS - Processus d'apprentissage



HTS - Processus d'apprentissage

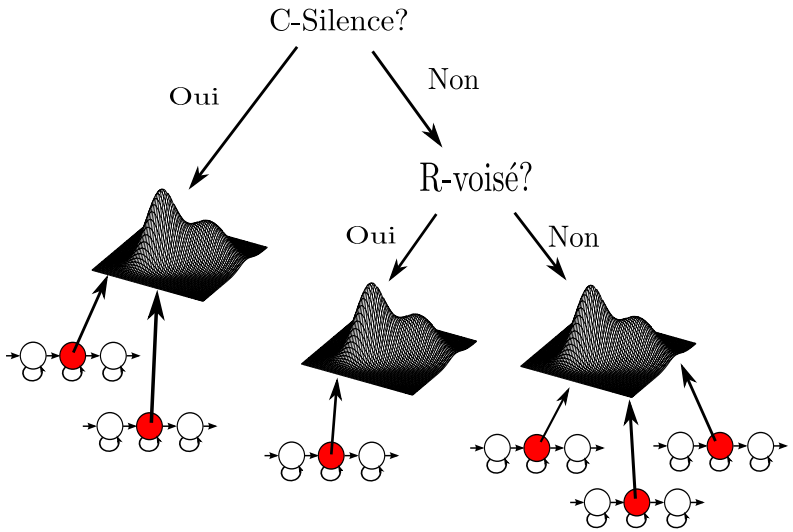


HTS - Partitionnement

Problématique

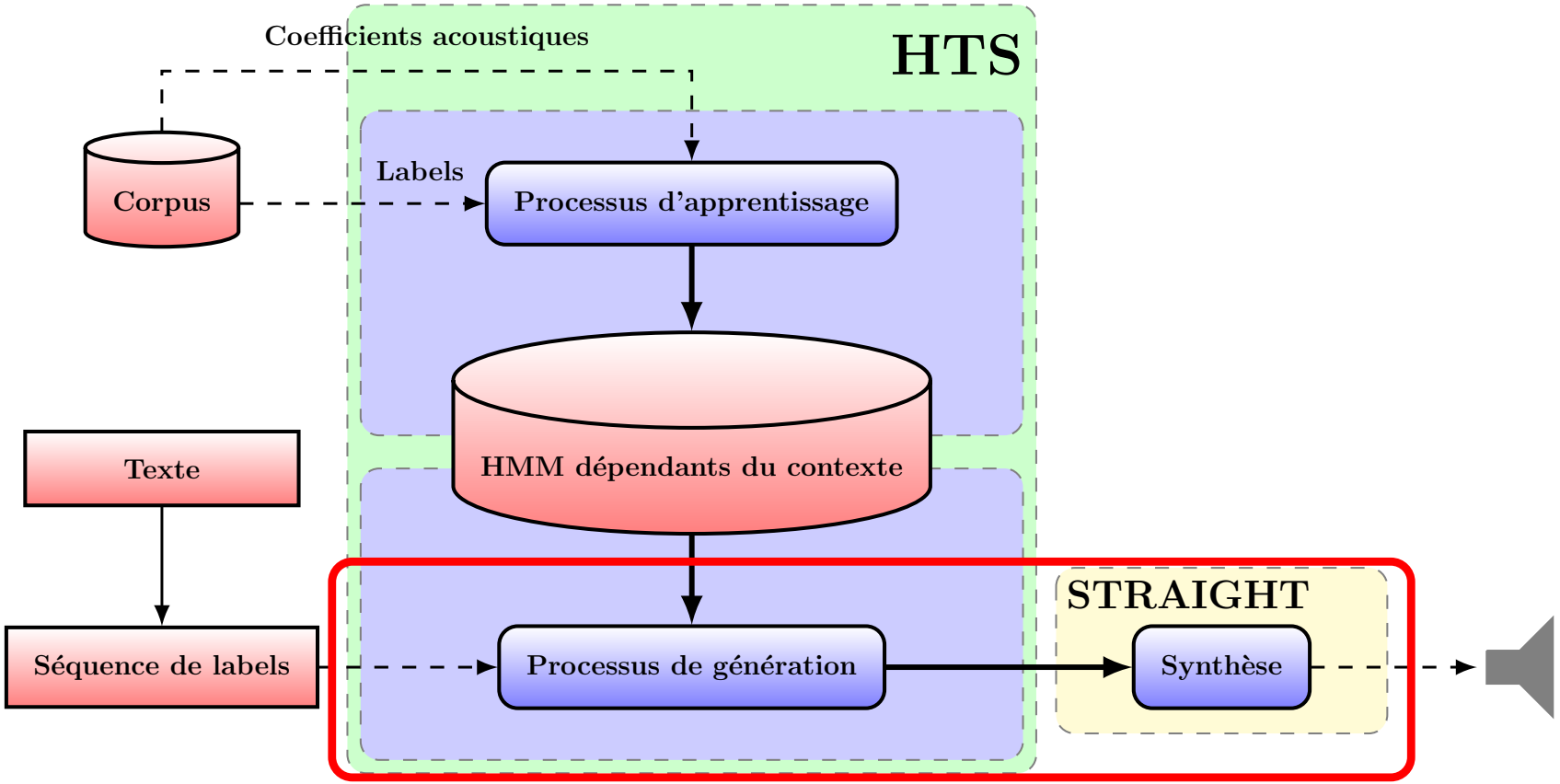
- Dispersion des valeurs des descripteurs,
- Robustesse de la modélisation.

- Arbres de décision[?] :



- Procédure de construction
  - Apprentissage supervisé (descripteurs et questions)
  - Critère MDL[?]

HTS - Architecture du système



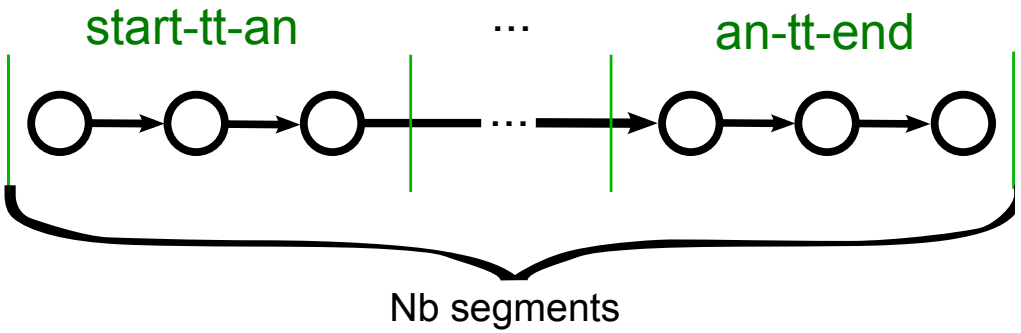
Processus de génération

0. Séquence de labels

start-tt-an ... an-tt-end

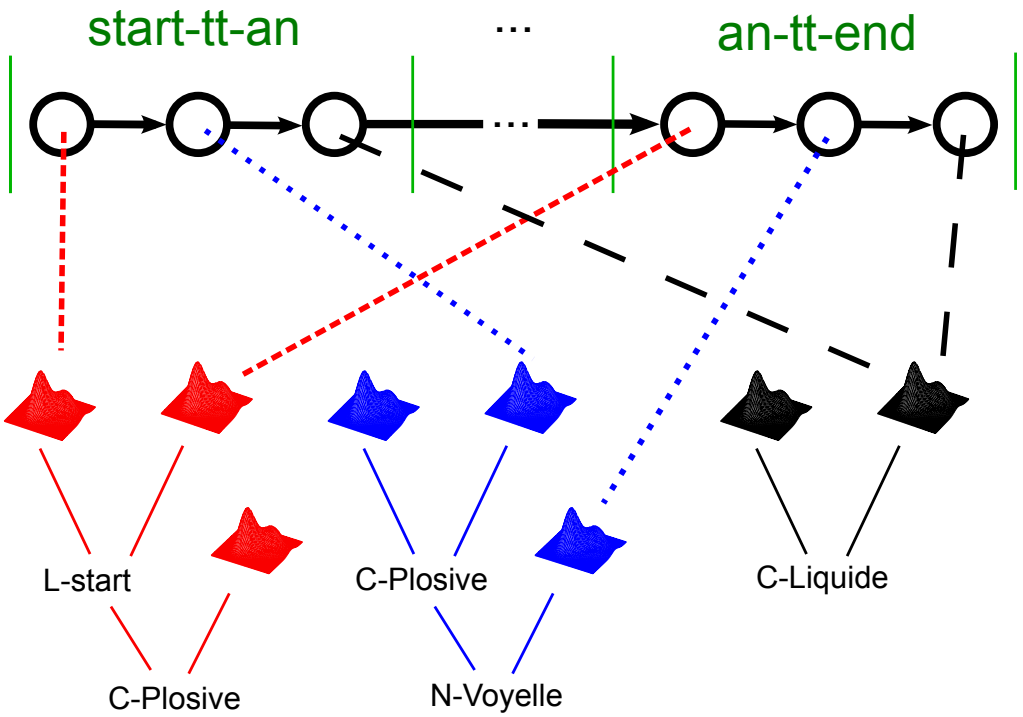
Processus de génération

1. Construction de la phrase HMM



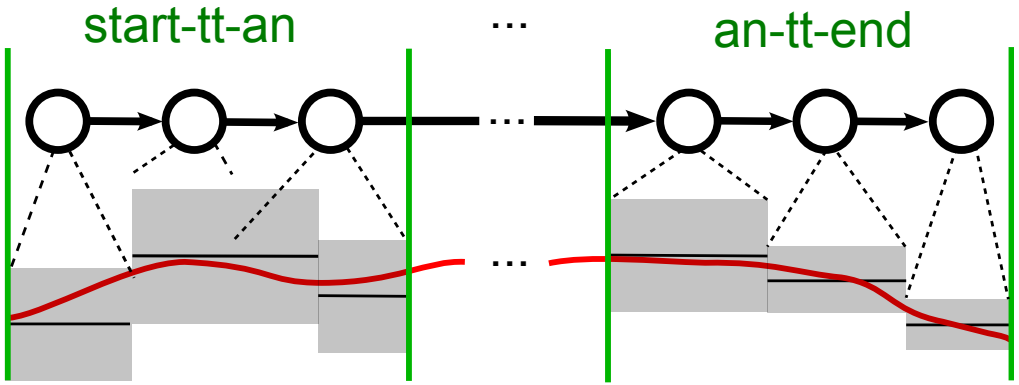
Processus de génération

2. Association des distributions[?]



Processus de génération

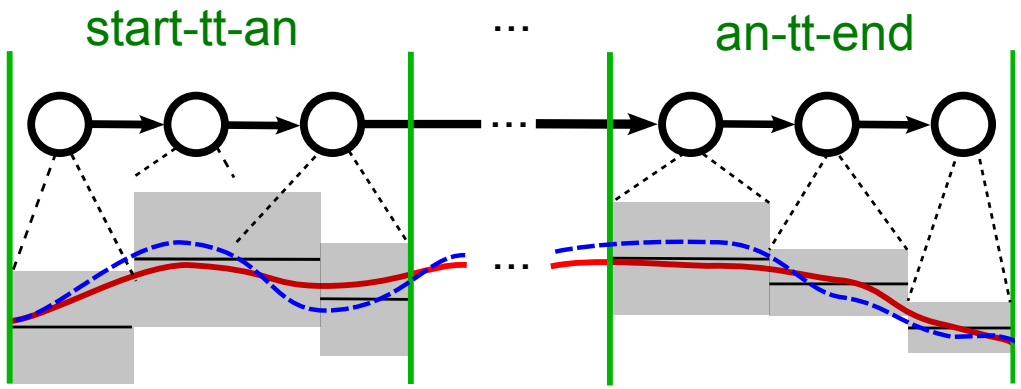
3a. Génération des paramètres[?]



- Production du vecteur C qui maximise $P(O|\Lambda) = P(WC|\Lambda)$
- Algorithme de type E.M.
  1. Etape E : Déterminer $\gamma_t(q) = P(q_t = q|O, \Lambda)$
  2. Etape M : Utiliser $\gamma_t(q)$ pour déterminer $P(O|Q, \Lambda)$ et obtenir $\overline{\Sigma^{-1}}$ ainsi que $\overline{\Sigma^{-1}\mu}$
  3. Estimation de $\overline{C}$ et $\overline{O}$: résoudre $(W^\top \overline{\Sigma^{-1}} W)\overline{C} = (W^\top \overline{\Sigma^{-1}\mu})$, puis $\overline{O} = W\overline{C}$

Processus de génération

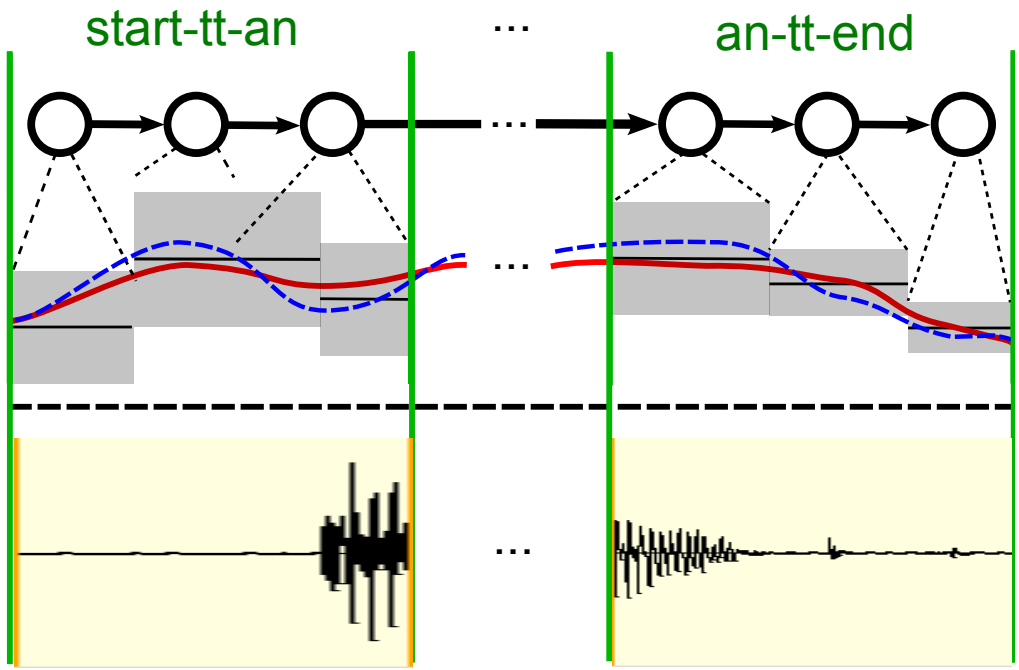
3b. Résolution du problème de surlissage[?]



Variance globale : $L = log\{P(O|\Lambda)^{\omega}.p(v(C)|\Lambda)\}$

Processus de génération

5. Génération du signal[?]



Utilisation de STRAIGHT[?]

Overview

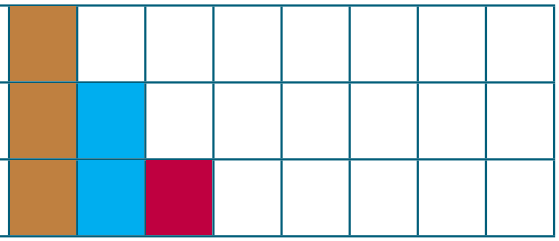
Jeu de descripteurs proposé pour le français

	Horizon	Description
Pho.		Label phonétique du segment courant
		Labels phonétiques des segments précédent/suivant
		Labels phonétiques des segments précédent-précédent/suivant-suivant
Syllabe	P/C/S	Nb. phones + position du phone courant dans la syl.
	C	Position de la syllabe dans le mot
	C	Position de la syllabe dans le Groupe de Souffle (GS)
	P/C/S	Information d'accentuation
	C	Nb. de syl. depuis la dernière accent./syl. cour. jusque la syl. cour./prochaine acc.
	C	Nb. de syllabes acc. avant/après la syllabe courante dans le GS
	C	Voyelle de la syllabe
Mot	P/C/S	Nb. de syllabes dans le mot
	C	Position du mot dans le GS
	P/C/S	Tag grammatical du mot
	C	Nb. de mots depuis le dernier sign./mot cour. jusqu'au mot cour./prochain sign.
	C	Nb. de mots signifiants avant/après le mot cour. dans le GS
GS	P/C/S	Nb. de syl. dans le GS
	P/C/S	Nb. de mots dans le GS
	C	Position du GS dans l'énoncé

Catégorisation des descripteurs

	Caté.	Description
Pho.		Label phonétique du segment courant
		Labels phonétiques des segments précédent/suivant
		Labels phonétiques des segments précédent-précédent/suivant-suivant
Syllabe		Nb. phones + pos. du phone courant dans la syl.
		Position de la syllabe dans le mot
		Position de la syllabe dans le Groupe de Souffle (GS)
		Information d'accentuation
		Nb. de syl. depuis la dernière accent./syl. cour. jusque la syl. cour./prochaine acc.
		Nb. de syllabes acc. avant/après la syllabe courante dans le GS
		Voyelle de la syllabe
Mot		Nb. de syllabes dans le mot
		Position du mot dans le GS
		Tag grammatical du mot
		Nb. de mots depuis le dernier sign./mot cour. jusqu'au mot cour./prochain sign.
		Nb. de mots signifiant avant/après le mot cour. dans le GS
GS		Nb. de syl. dans le GS
		Nb. de mots dans le GS
		Pos. du GS dans l'énoncé

Jeux de descripteurs soumis à évaluation

Identifiant	Description	Descripteurs
p1	Ph. courant	
p3	Contexte ph. direct (3 ph.)	
p5	Contexte ph. large (5 ph.)	
p5-sy_pos	Info. position de la syl.	
p5-sy_accent	Info. prosodique de la syl.	
p5-sy_full	Info. complète de la syl.	
p5-w_pos	Info. position du mot	
p5-w_content	Info. prosodique du mot	
p5-w_full	Info. complète du mot	
p5-s_pos	Info. position du GS	

Overview

Évaluation objective de l'influence d'un jeu de descripteurs

- [?]
  - Objectif : déterminer un jeu minimal de descripteurs.
  - Analyse ≪ globale ≫ basée sur des distances (RMSE et dist. cepstrale).
  - Evaluation appliquée à l'anglais et au japonais.
- [?]
  - Objectif : annotations manuelles vs annotations automatiques.
  - Analyse de l'arbre de décision.
  - Evaluation appliquée à l'anglais.

Nos objectifs

- Analyse ≪ extérieure ≫ au système HTS.
- Analyse des descripteurs par modèle.
- Application au français.

Corpus d'évaluation

- Caractéristiques du locuteur
  - Voix masculine.
  - Voix expressive.
- Caractéristiques acoustiques
  - Signal échantillonné à 16kHz
  - Paramétrisation standard (40 MGC + 1 log(F0) + 5 BAP)
- Trois corpus définis aléatoirement :
  - A = 520 000 trames (environ 60 min) pour 316 énoncés
  - V = 85 000 trames (environ 10 min) pour 152 énoncés
  - T = 85 000 trames (environ 10 min) pour 152 énoncés

Deux méthodes complémentaires

- Évaluation par GMM
  - Évaluation globale ⇒ pas de dépendance vis-à-vis de la position de la trame.
  - Nécessité d'une importante masse de données pour l'apprentissage.
- Évaluation par écarts entre trames appairées
  - Masse de données "faible".
  - Évaluation locale ⇒ alignement.

Évaluation par GMM - Description globale

- Objectif :
  - Évaluation adaptée à HTS
  - Évaluation globale
- Proposition :
  - Identification du locuteur[**?**] et transformation de voix[**?**]

Hypothèse principale

L'espace acoustique est correctement représenté par un GMM.

Principe fondamental

- Données fixes,
- Évaluation de la vraisemblance du GMM.

Évaluation par GMM - Méthodologie

A

4600000 10900000 ww^ii-insp+spause=ss
11000226 12900226 ii^insp-spause+ss=ai
12900226 13700226 insp^spause-ss+ai=ss
13700226 14000226 spause^ss-ai+ss=
14000226 15200226 ss^ai-ss+ll=aa
15200226 15800226 ai^ss-ll+aa=kk

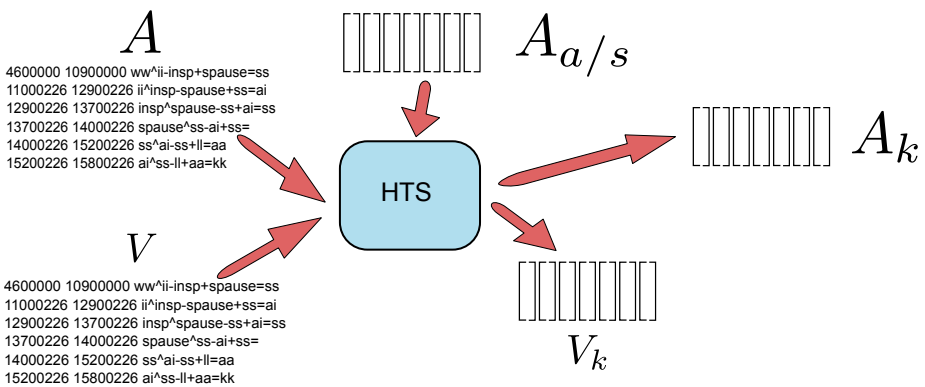
V

4600000 10900000 ww^ii-insp+spause=ss
11000226 12900226 ii^insp-spause+ss=ai
12900226 13700226 insp^spause-ss+ai=ss
13700226 14000226 spause^ss-ai+ss=
14000226 15200226 ss^ai-ss+ll=aa
15200226 15800226 ai^ss-ll+aa=kk

$A_{a/s}$

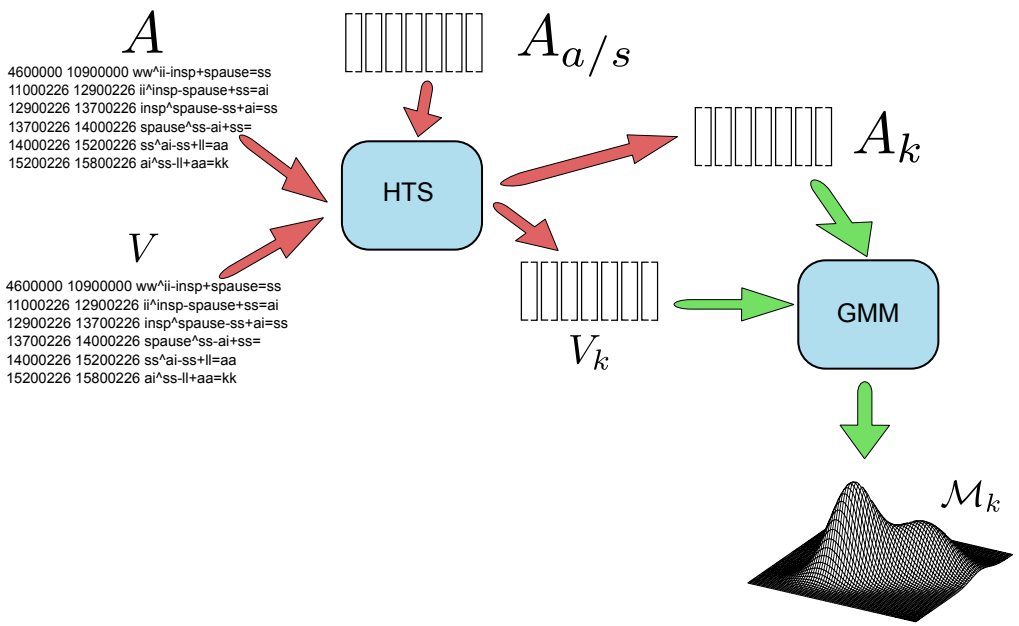
- Corpus :
  - A = apprent.
  - V = validation
  - T = test
- Identifiant :
  - a/s = ana.-synth.
  - k = jeu de desc.

Évaluation par GMM - Méthodologie



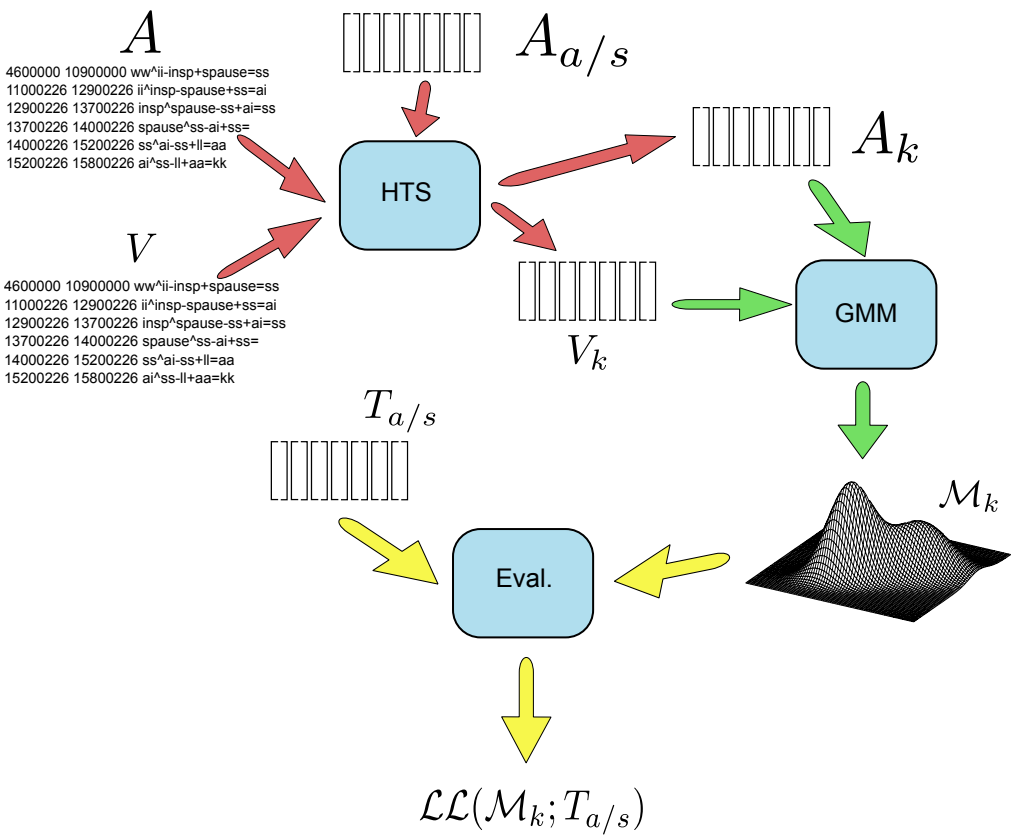
- Corpus :
  - A = apprent.
  - V = validation
  - T = test
- Identifiant :
  - a/s = ana.-synth.
  - k = jeu de desc.

Évaluation par GMM - Méthodologie



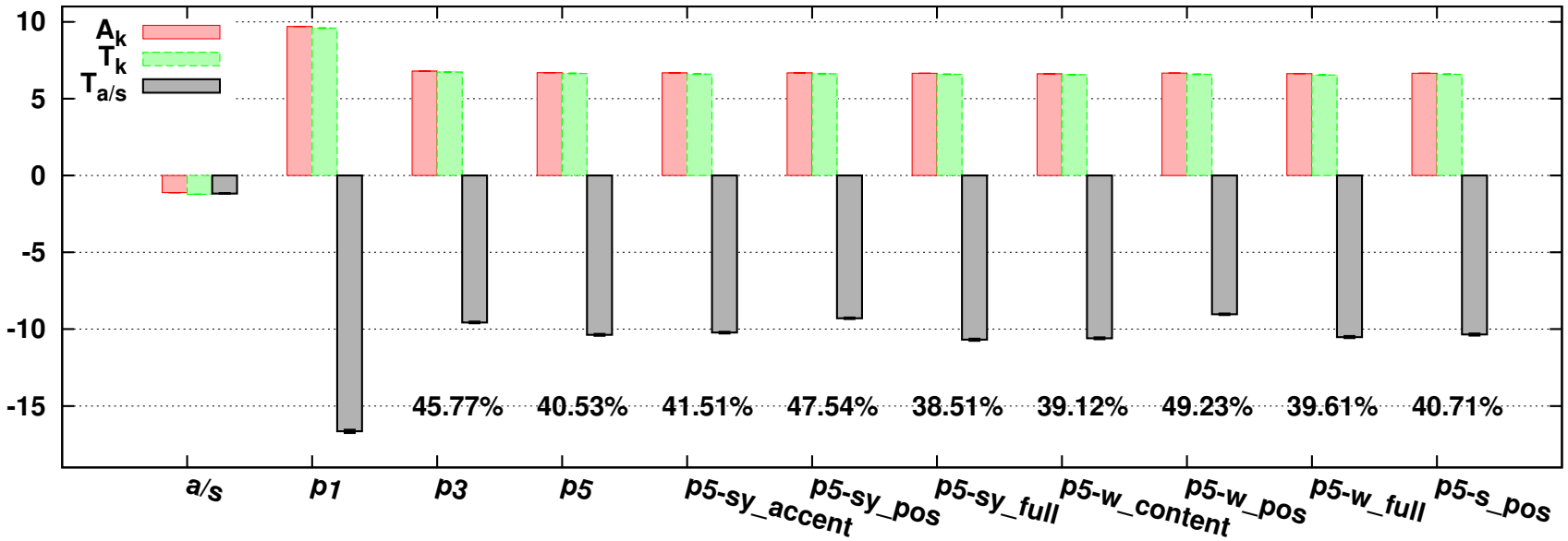
- Corpus :
  - A = apprent.
  - V = validation
  - T = test
- Identifiant :
  - a/s = ana.-synth.
  - k = jeu de desc.

Évaluation par GMM - Méthodologie

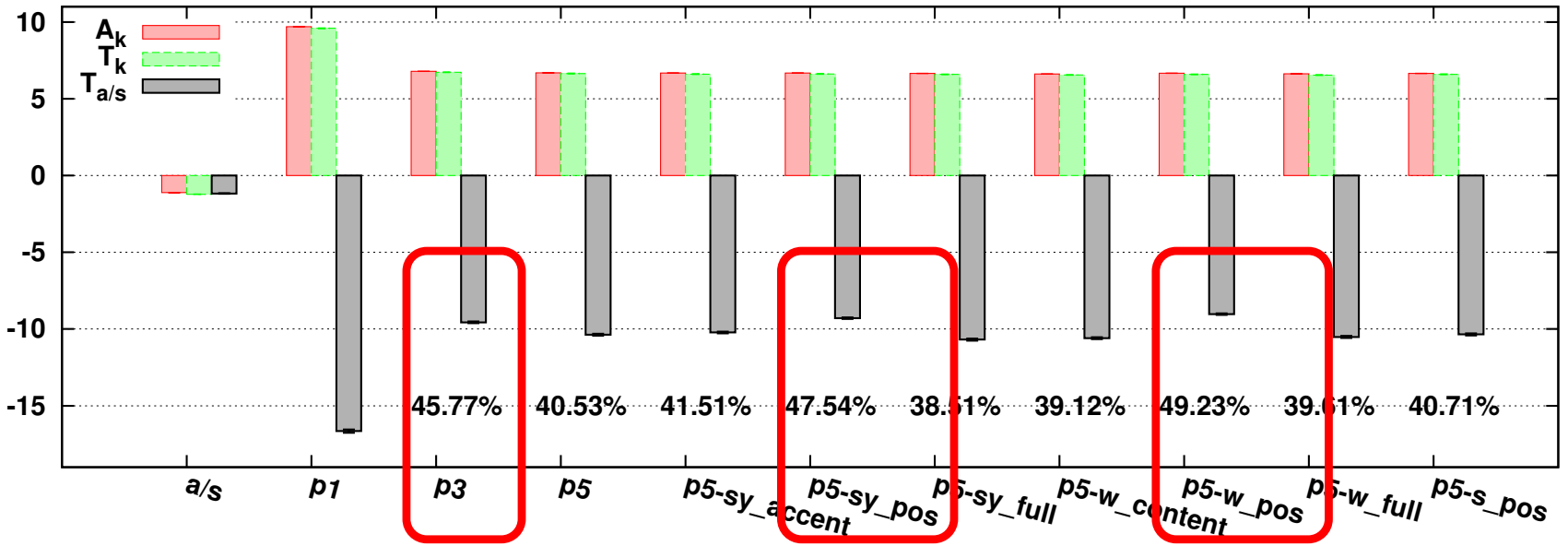


- Corpus :
  - A = apprent.
  - V = validation
  - T = test
- Identifiant :
  - a/s = ana.-synth.
  - k = jeu de desc.

Évaluation par GMM - Modélisation spectrale (256 comp.)

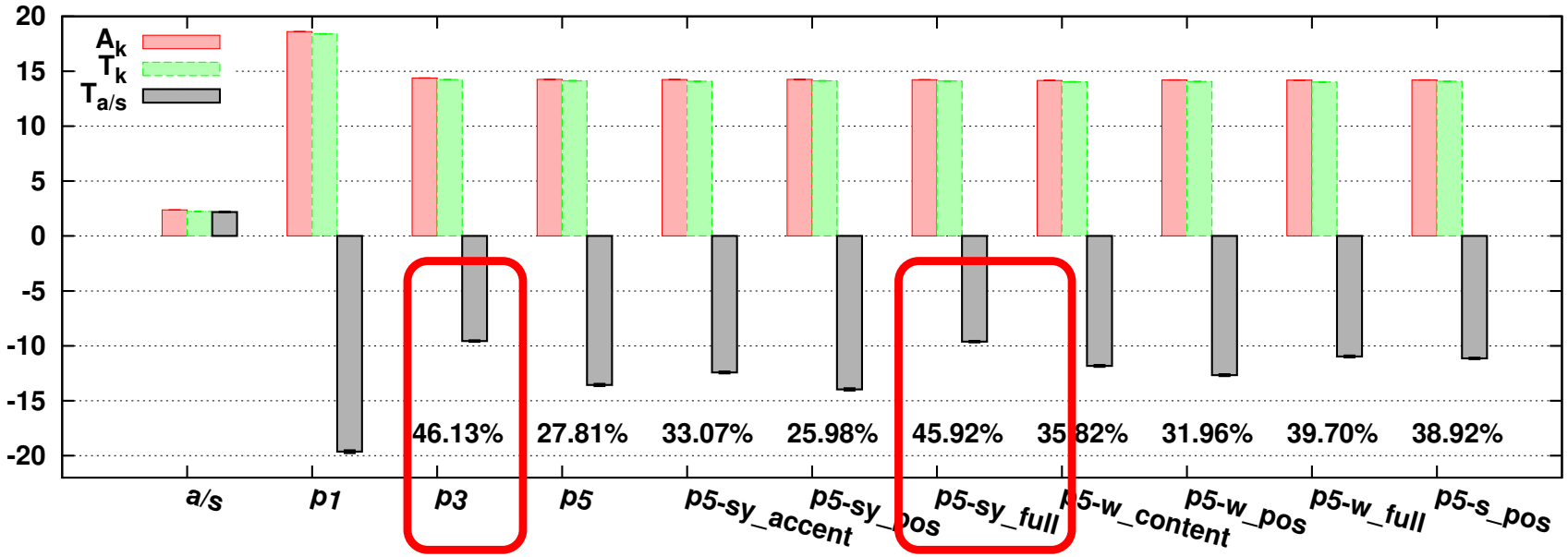


Évaluation par GMM - Modélisation spectrale (256 comp.)



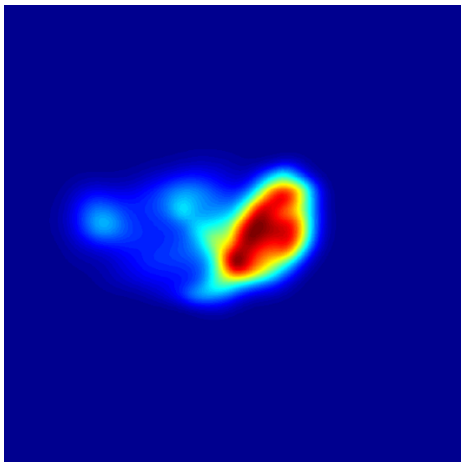
- Les jeux obtenant la vraisemblance la plus forte sont p3, p5-sy_pos et p5-w_pos
- L'introduction de nouveaux descripteurs n'aboutit pas à une amélioration (p5/p3)

Évaluation par GMM - Modélisation spectrale (Sans l'énergie - 512 comp.)

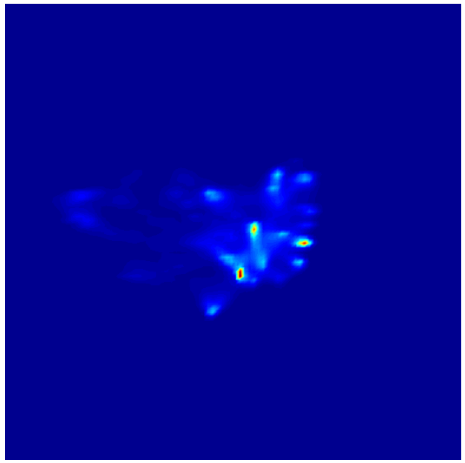


- Accentuation des écarts
- p5-sy_pos et p5-w_pos ne se distinguent plus
- Jeu de descripteurs optimal = p3

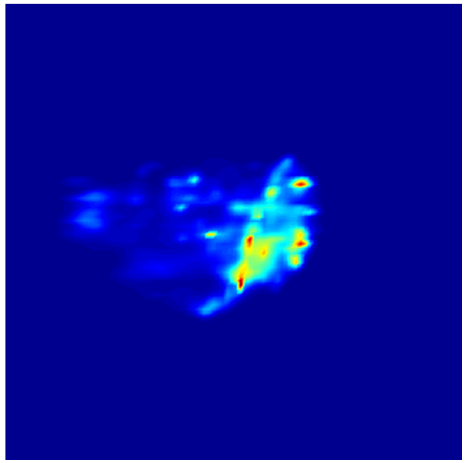
Évaluation par GMM - Modélisation spectrale - Visualisation de l'espace acoustique



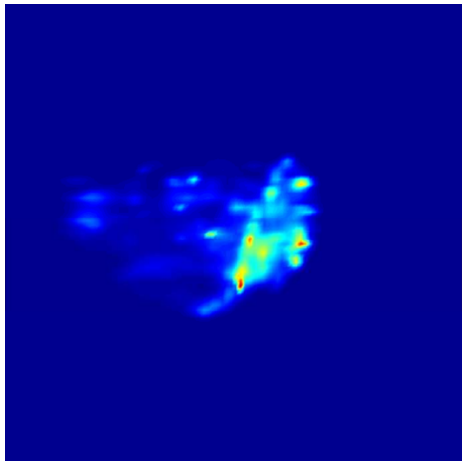
(a) a/s



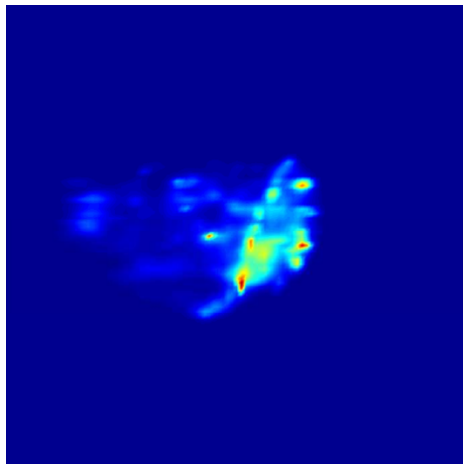
(b) p1



(c) p3

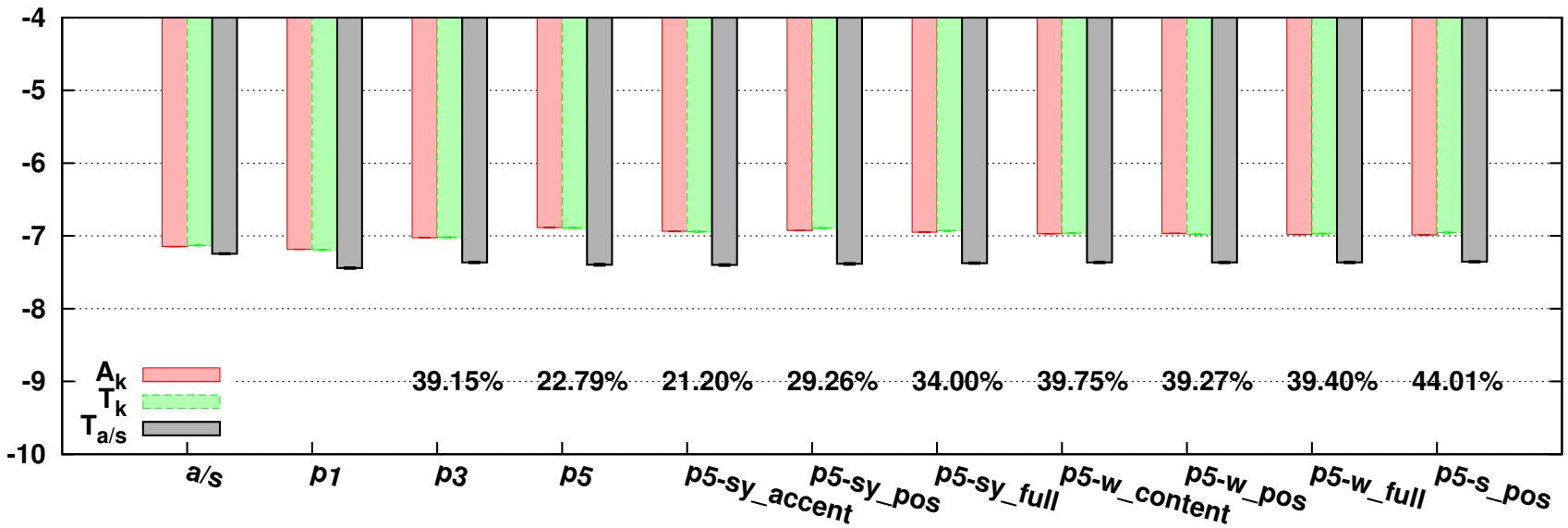


(d) p5

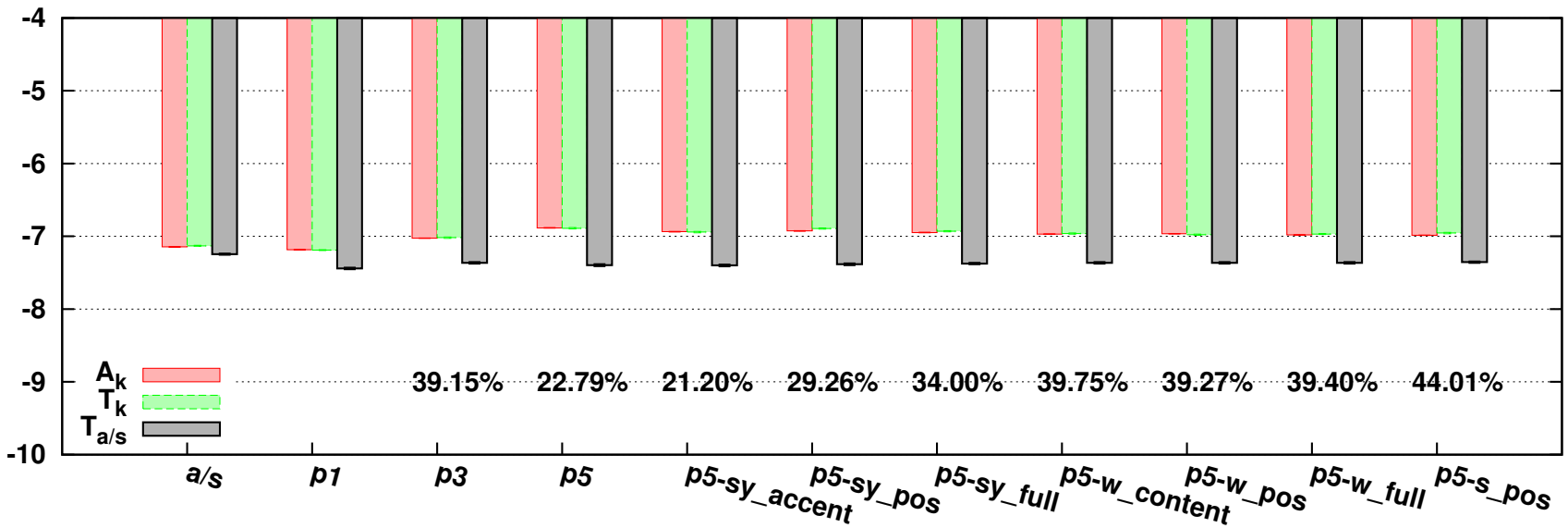


(e) p5-s_pos

Évaluation par GMM - F0 (cent - 128 comp.)

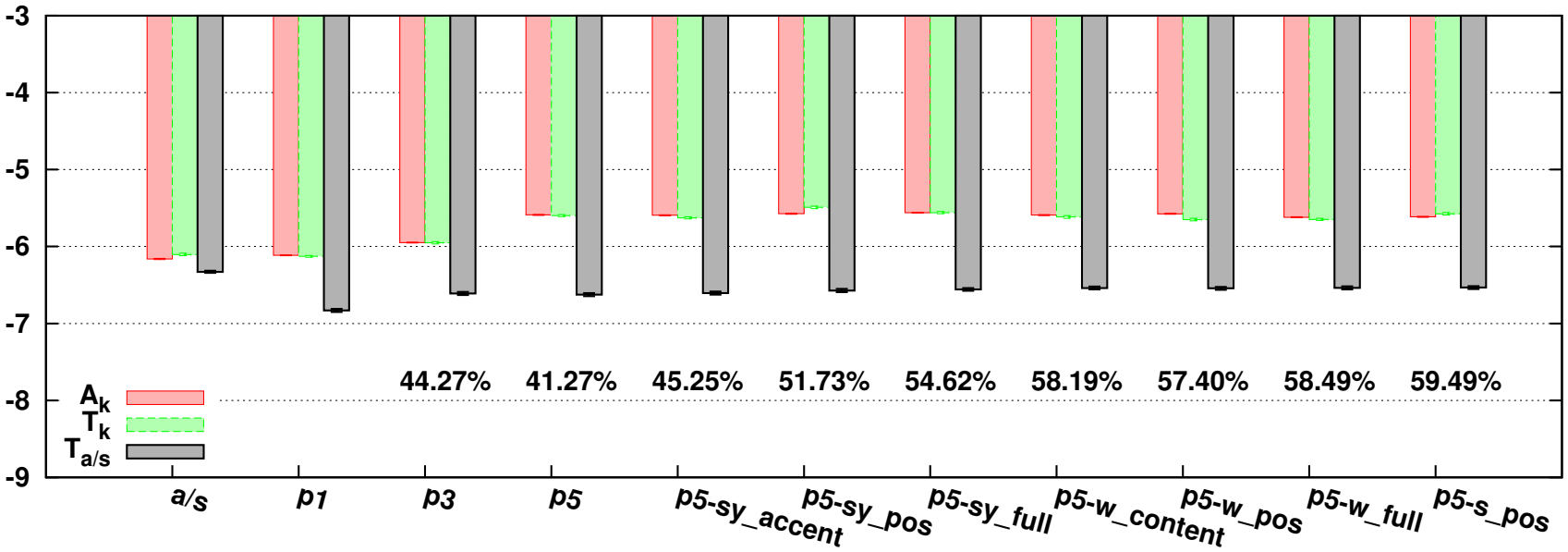


Évaluation par GMM - F0 (cent - 128 comp.)



- Faibles différences de vraisemblance
- Aucune différence significative à partir de p3

Évaluation par GMM - F0 + Δ F0 (64 comp.)



- Amélioration significative par palier jusque `p5-sy_full`
- Jeu de descripteurs optimal = `p5-sy_full`

Évaluation par GMM - Bilan

- Conclusion
  - MGC : jeu optimal = `p3`
  - F0 : jeu optimal = `p5-sy_full`
- Remarques :
  - Utilisation de certains jeux $\Rightarrow$ dégradation (MGC)
  - Faible discrimination pour le F0
  - Évaluation de la partie voisée pour le F0

Évaluation non paramétrique - Motivation

- Objectif = analyse plus fine des modèles
  - Quels modèles posent problème ?
  - Est-ce que les descripteurs ont une influence sur ces modèles ?

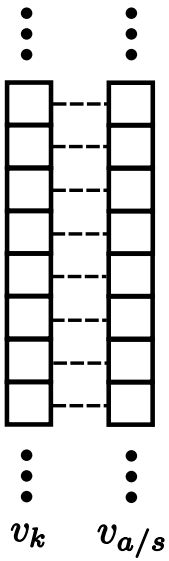
- Problème : masse de données insuffisante

  $\Rightarrow$ Évaluation par GMM non applicable

Proposition

Méthodologie basée sur des distances entre trames appairées

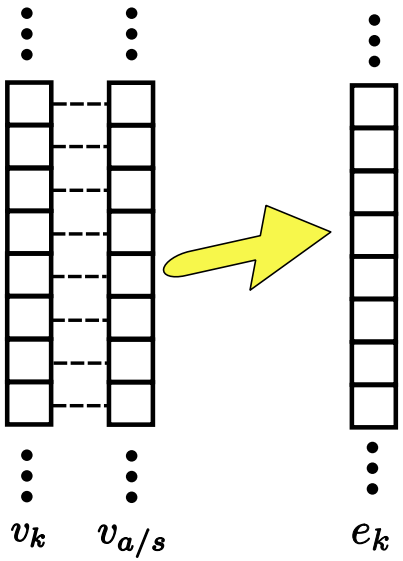
Évaluation non paramétrique - Méthodologie



14000226 15200226 ss
15200226 15800226 ll

0. Données nécessaires
   - Alignement des trames obtenu en imposant la durée lors de la phase de génération

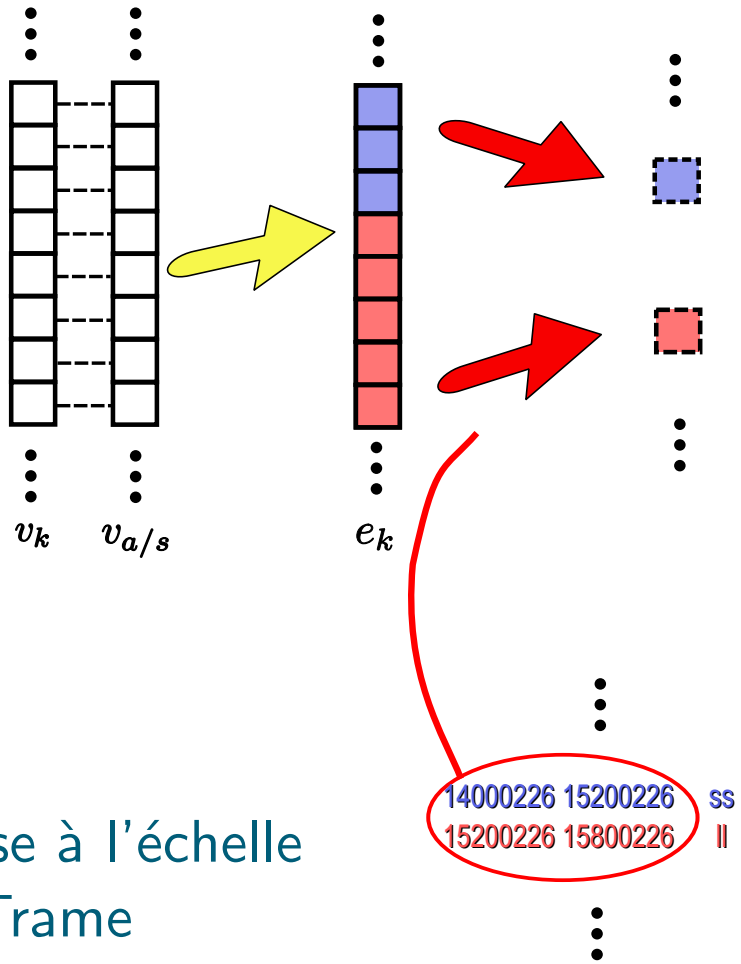
Évaluation non paramétrique - Méthodologie



1. Distance par trame
   - $e_k = v_k - v_{a/s}$
   - $e_k = |v_k - v_{a/s}|$
   - $e_k = 1200 * log_2(\frac{v_k}{v_{a/s}})$
   - $e_k =$ err. de voisement

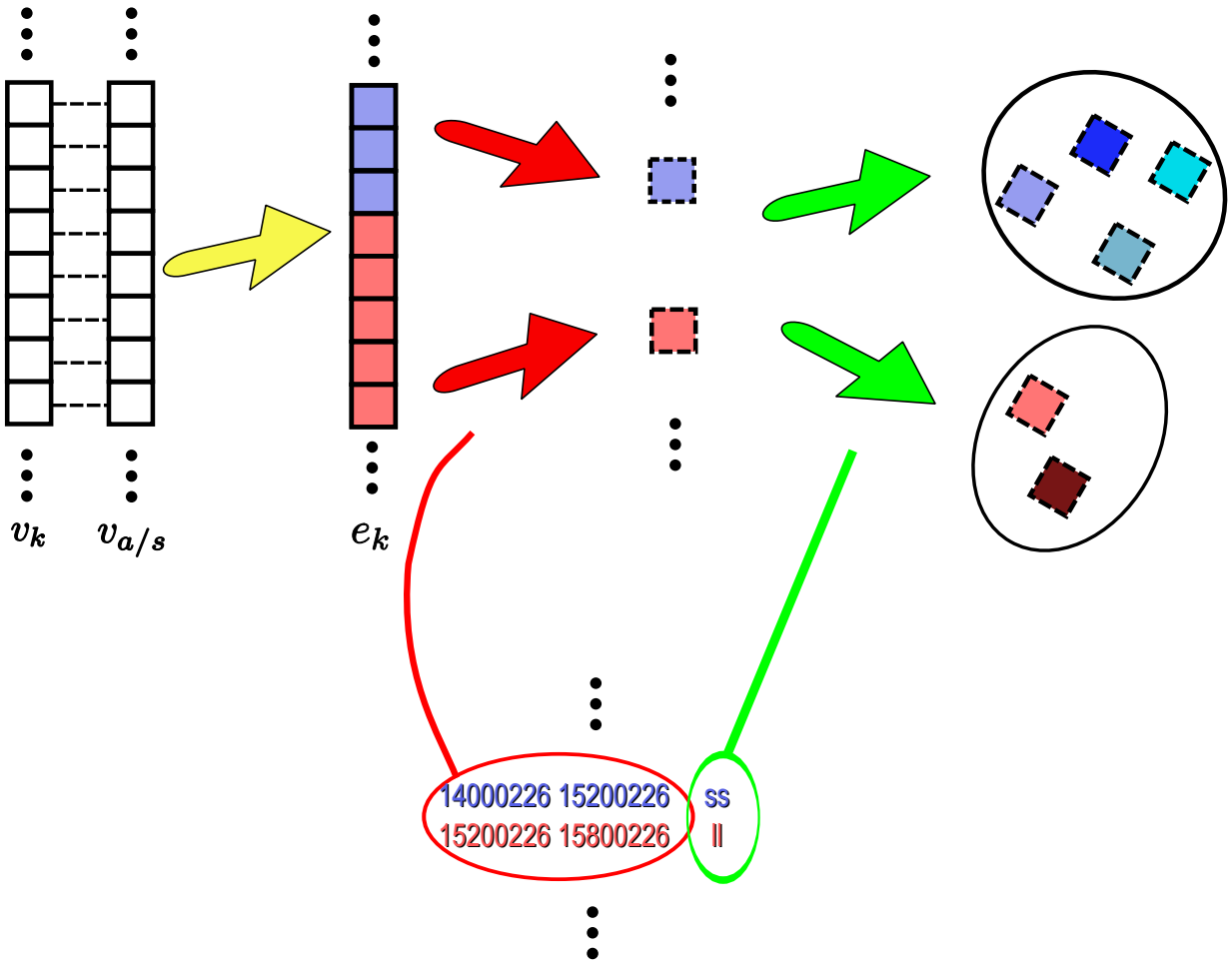
14000226 15200226 ss
15200226 15800226 ll

Évaluation non paramétrique - Méthodologie



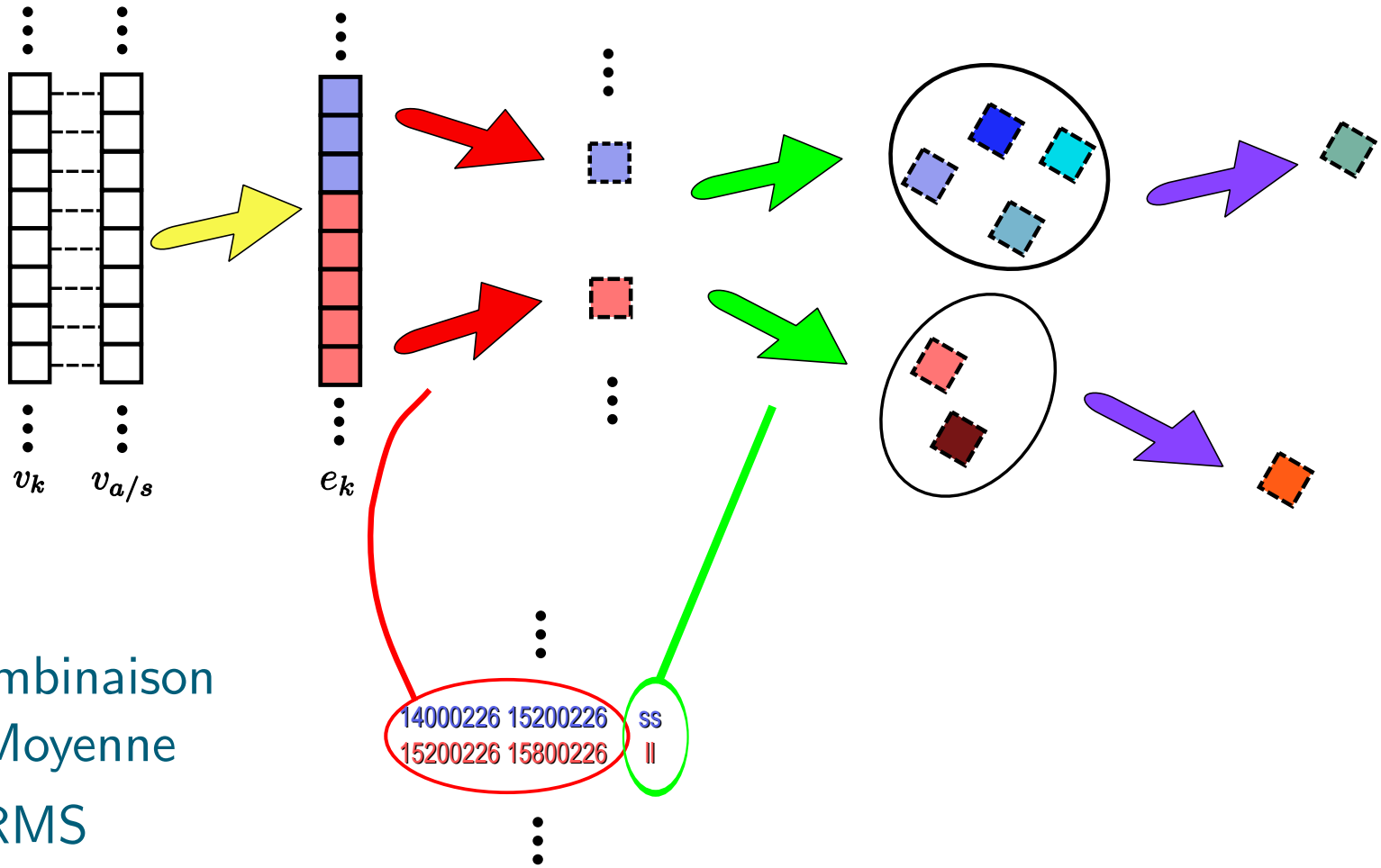
2. Mise à l'échelle
   - Trame
   - Segment

Évaluation non paramétrique - Méthodologie



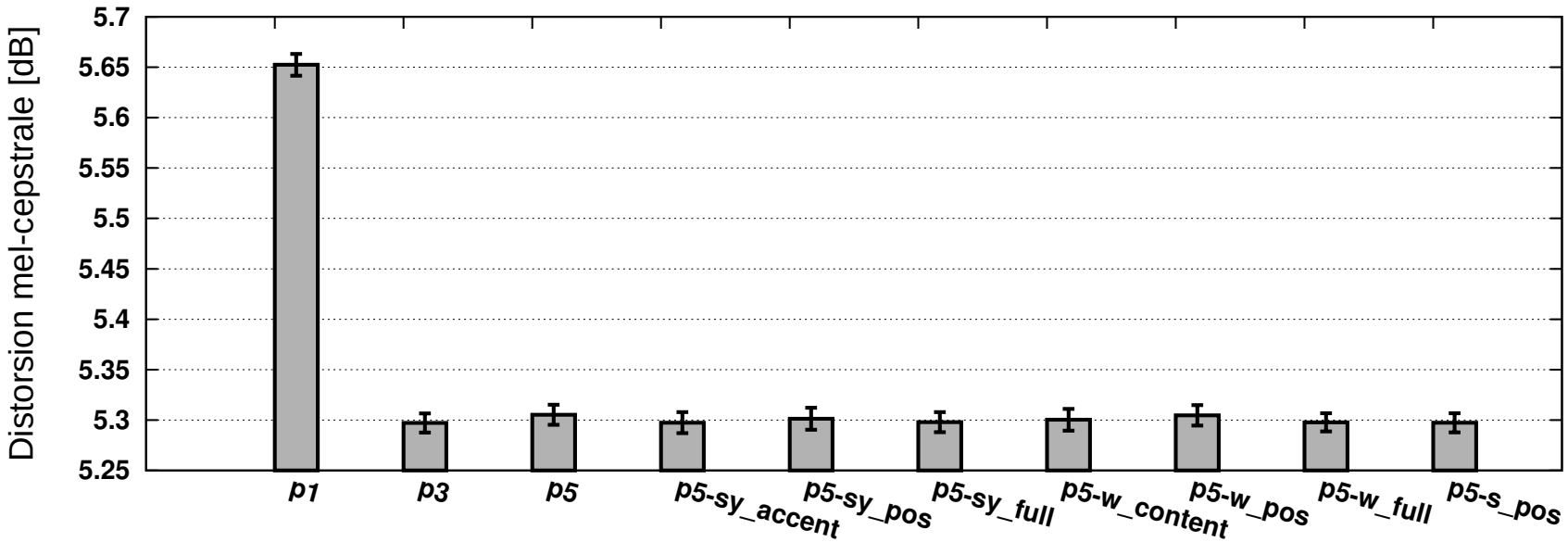
3. Partitionnement

Évaluation non paramétrique - Méthodologie



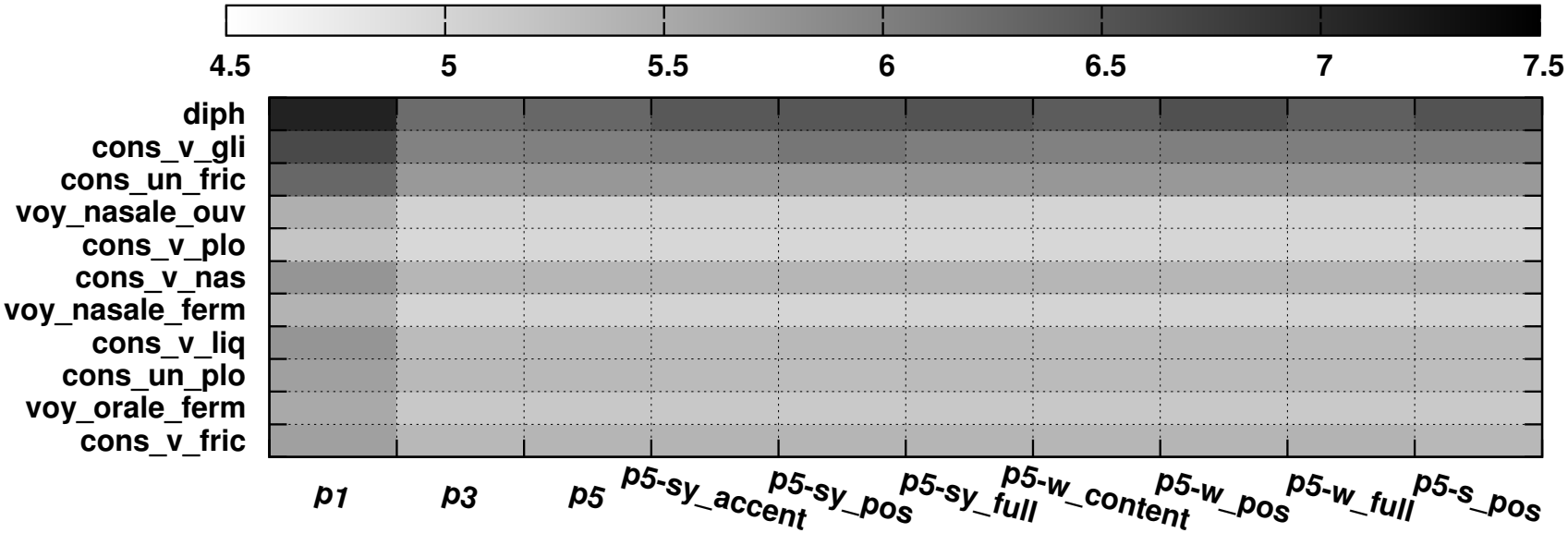
4. Combinaison
   - Moyenne
   - RMS
   - Distance cepstrale

Évaluation non paramétrique - Spectre - globale (distance cepstrale)



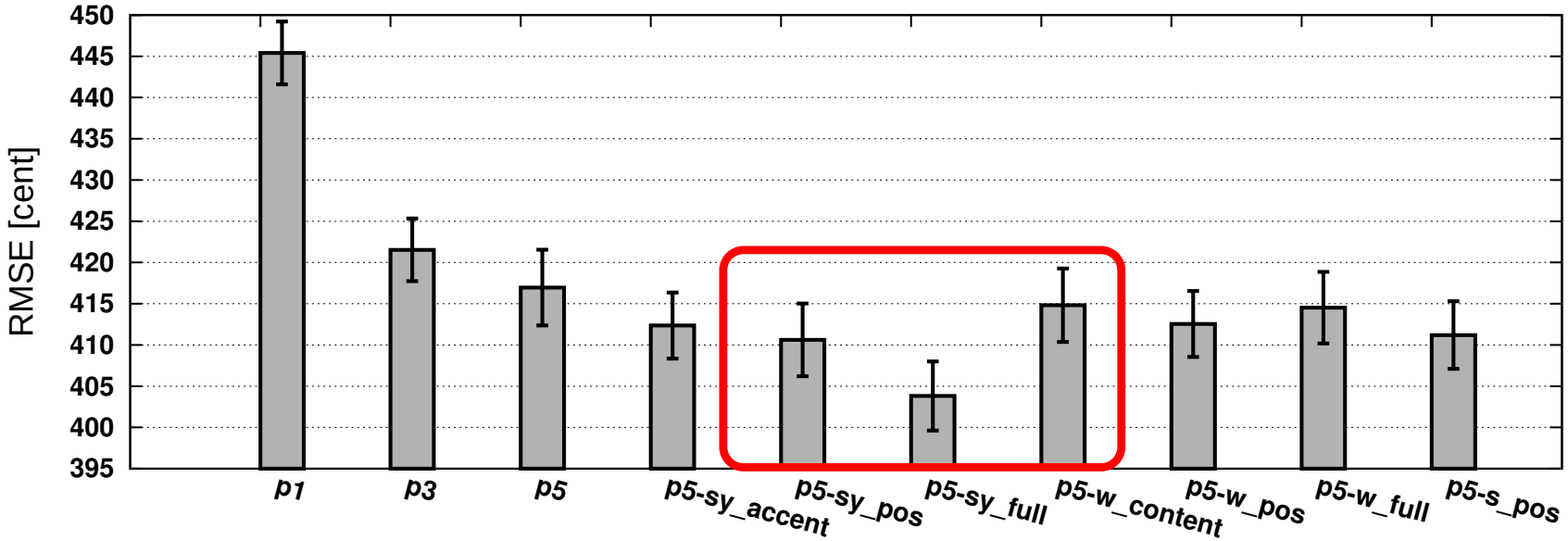
- Aucune différence significative hormis p3/p1
- Jeu optimal = p3

Évaluation non parametrique - Spectre - par catégorie phonétique (distance cepstrale)



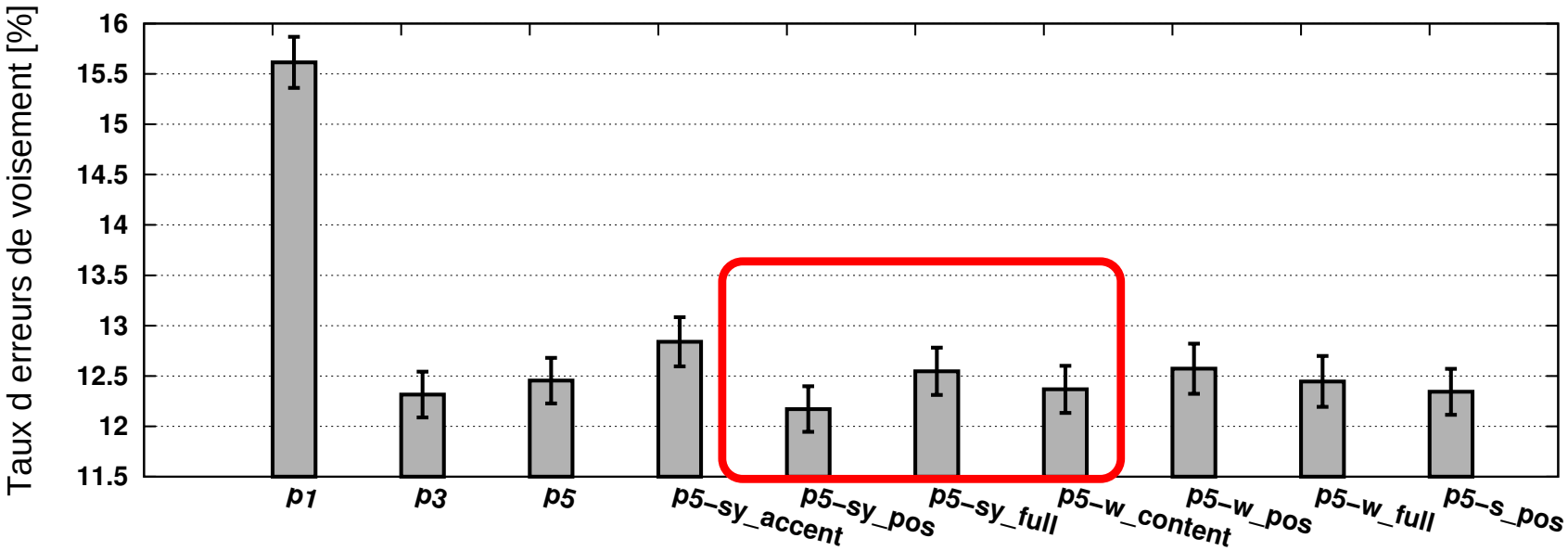
- L'amélioration apportée par p3 est globale à l'ensemble des modèles
- Quelques différences de qualité de modélisation entre les catégories phonétiques

Évaluation non paramétrique - F0 - globale (RMSE)



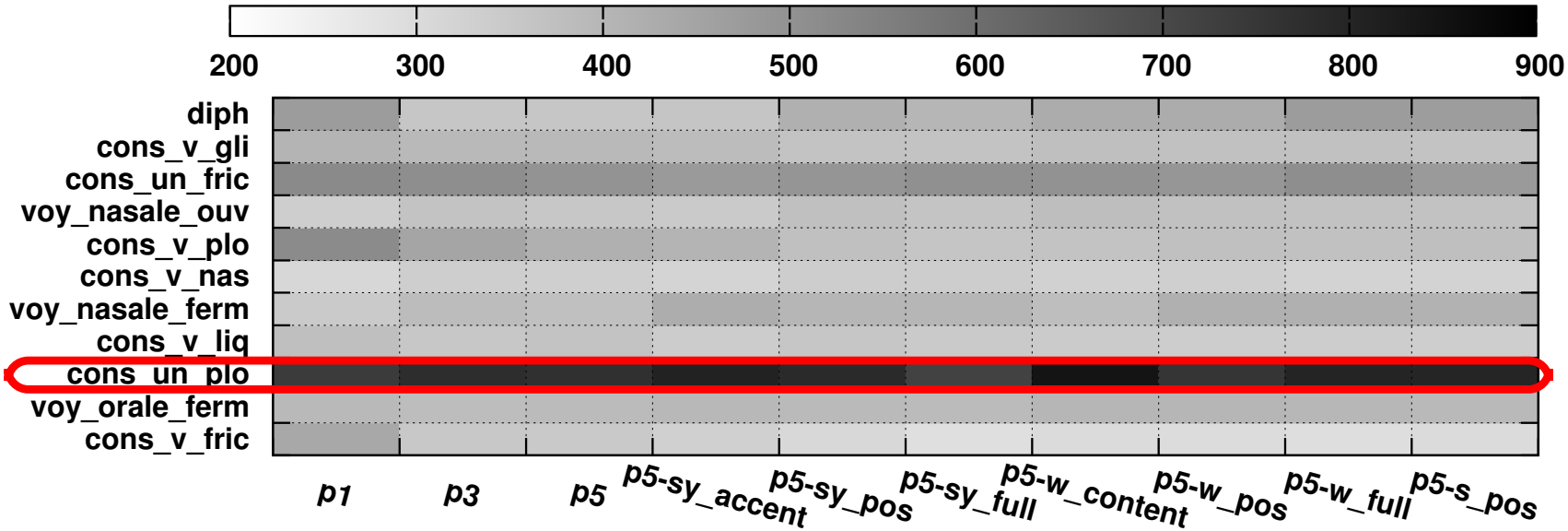
- Amélioration par palier jusqu'à `p5-sy_full`
- Jeu optimal = `p5-sy_full`

Évaluation non paramétrique - F0 - globale (taux d'erreurs de voisement)



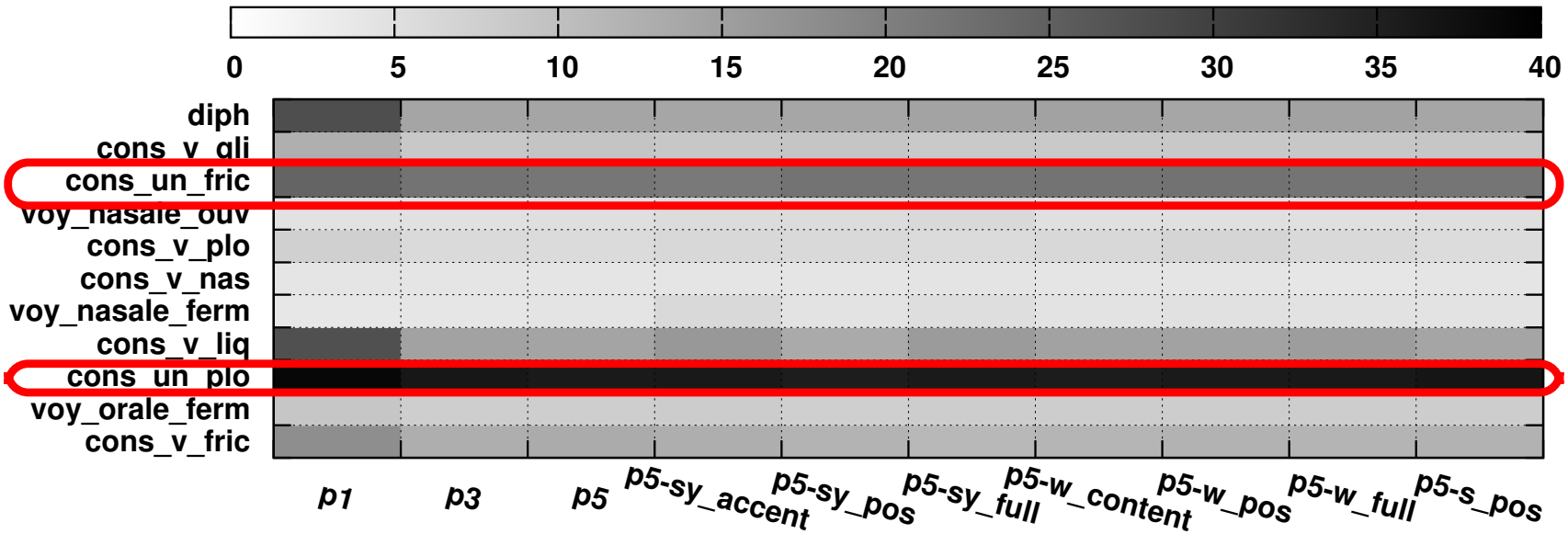
- Amélioration importante apportée par p3
- Faibles différences entre les autres jeux de descripteurs
- Jeu optimal = p5-sy_pos mais différence non significative avec p5-sy_full

Évaluation non parametrique - F0 - par categorie phonetique (RMSE)



- Les plosives non-voisées se distinguent pour l'ensemble des jeux de descripteurs
- Tendance pour les fricatives non voisées

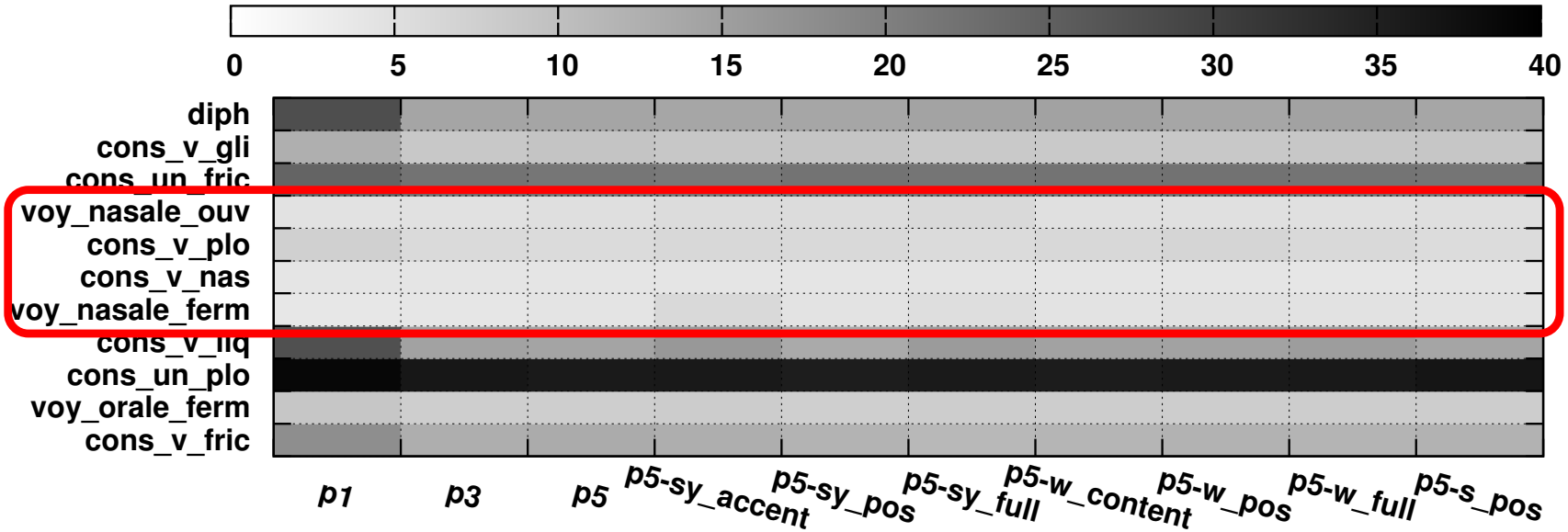
Évaluation non parametrique - F0 - par categorie phonetique (taux d'erreurs de voisement)



- Confirmation de la tendance indiquée par la RMS

$\Rightarrow$ Mauvaise modélisation du F0 pour les phonèmes contenant en majorité des frontières de voisement

Évaluation non paramétrique - FU - par catégorie phonétique (taux d'erreurs de voisement)



- Erreur de voisement quasi-inexistante pour une majeure partie des phonèmes considérés comme voisés

Évaluation non paramétrique - Bilan

- Bilan
  - MGC : contexte phonétique direct suffisant
  - F0 : ensemble des descripteurs associés au phone et à la syllabe suffisant

- Jeux de descripteurs et modèles
  - Apport global
  - Différences importantes dues à la modélisation (MSD)

$\Rightarrow$ Jeu optimal `p5-sy_full`

Bilan des évaluation objectives

	GMM	non-paramétrique
MGC	- Jeu optimal = p3 - Dégradation p3 → p5	- Jeu optimal = p3 - Dégradation p3 → p5 non significative
F0	- Jeu optimal = p5-sy_full - Partie non-voisée non évaluée	- Jeu optimal = p5-sy_full - Prise en compte des erreurs de voisement - Différences importantes entre catégories phonétiques

⇒ Jeu optimal p5-sy_full

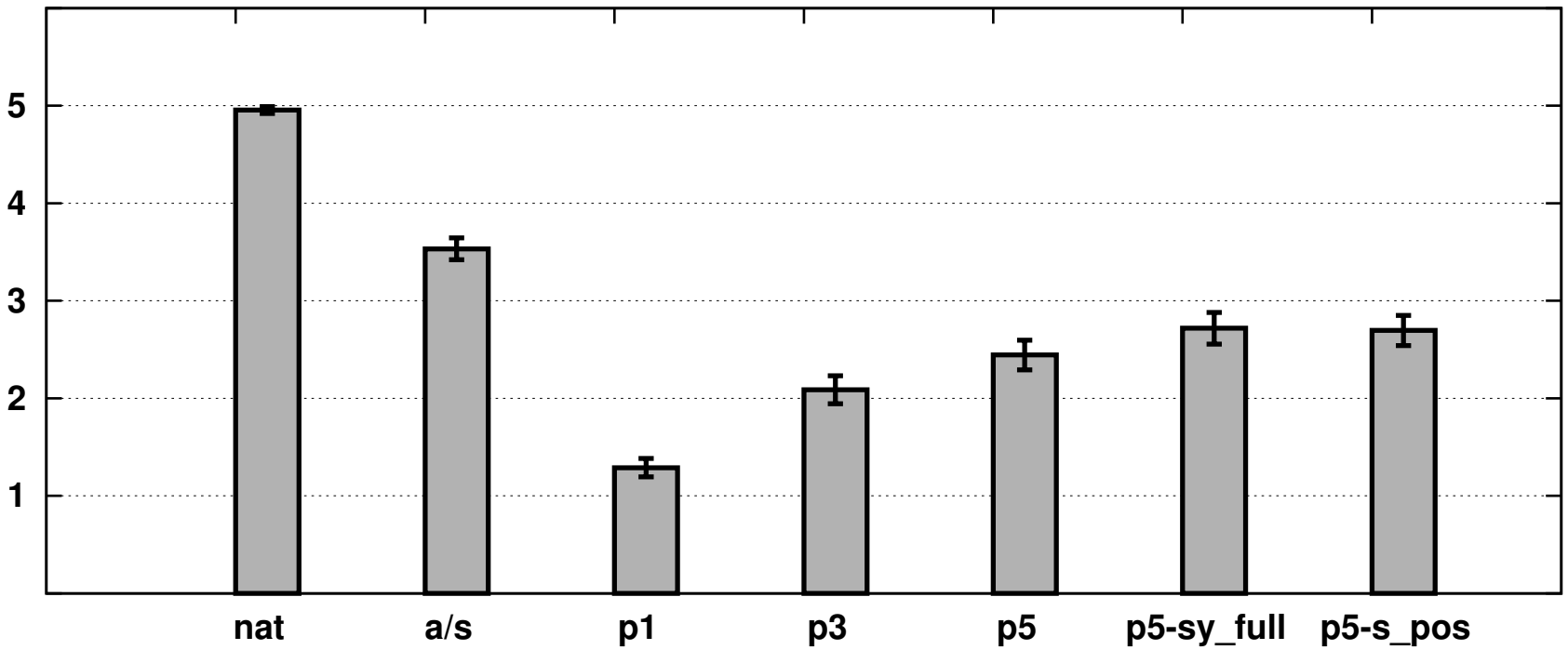
Overview

Description

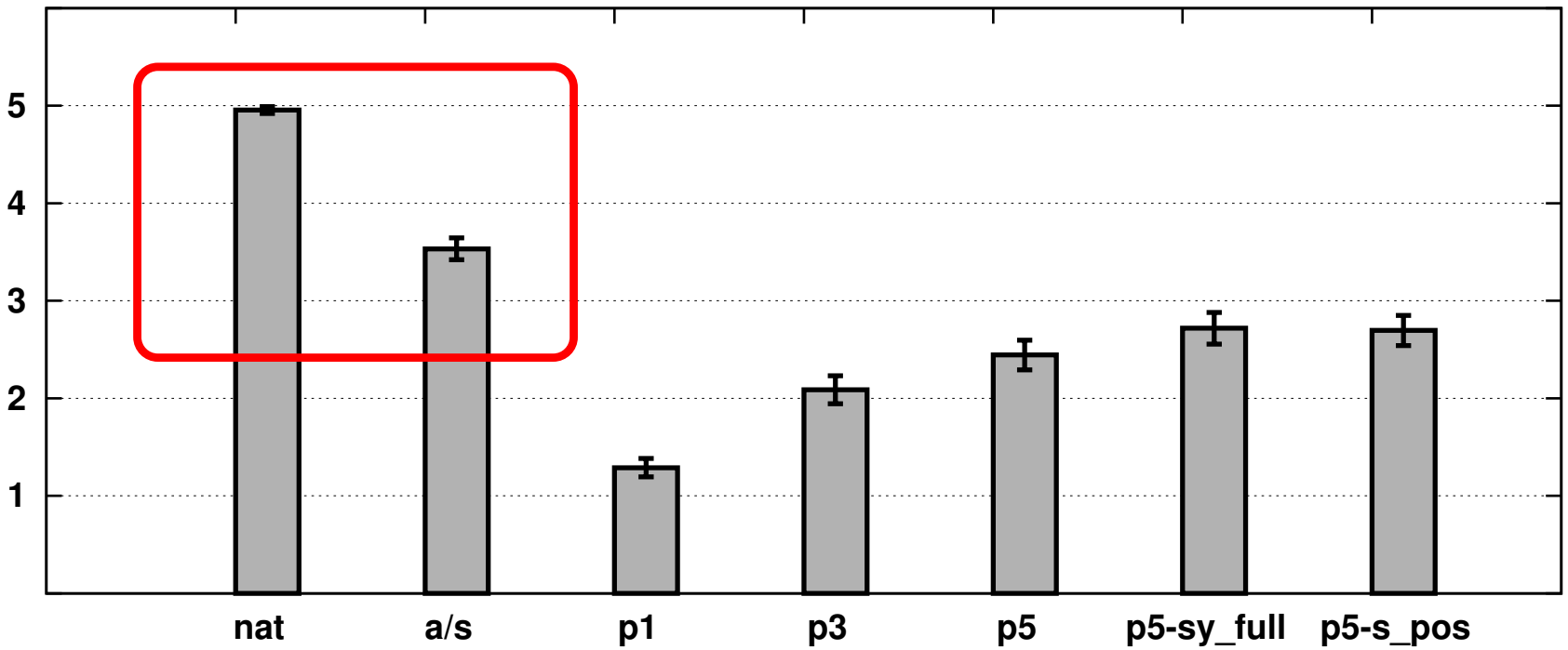
- MOS (de 1. Mauvais à 5. Excellent)
- Protocole de constitution du corpus
  - Énoncés correspondants au corpus de test T
  - Labels obtenus à l'issue d'une segmentation automatique
- Protocole d'évaluation
  - 10 auditeurs (experts)
  - 7 systèmes (nat + a/s + 5 synthèses HTS)
  - 30 échantillons par système (de 2s à 4s)
  - 105 stimuli par auditeur

$\Rightarrow$ Durée totale par auditeur $\simeq$ 30min
$\Rightarrow$ Environ 150 notes pour chaque système

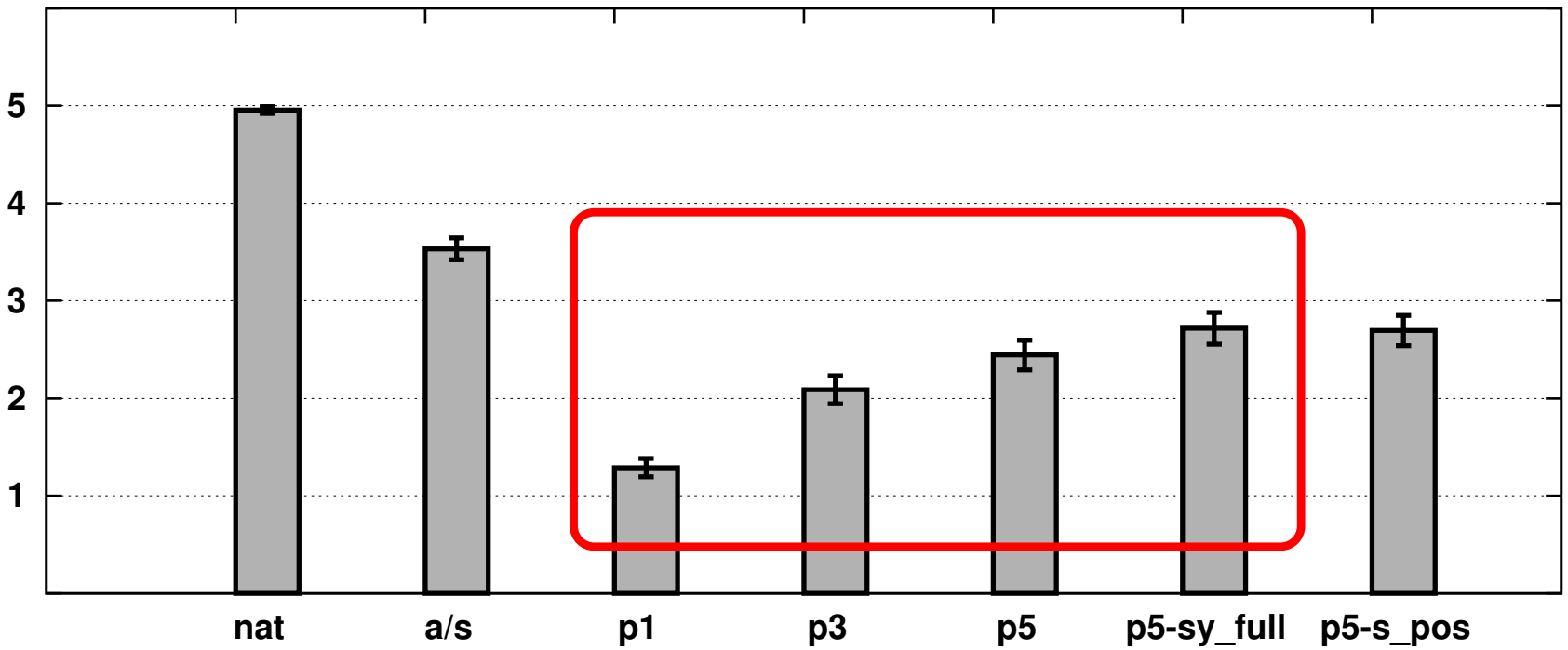
Résultats



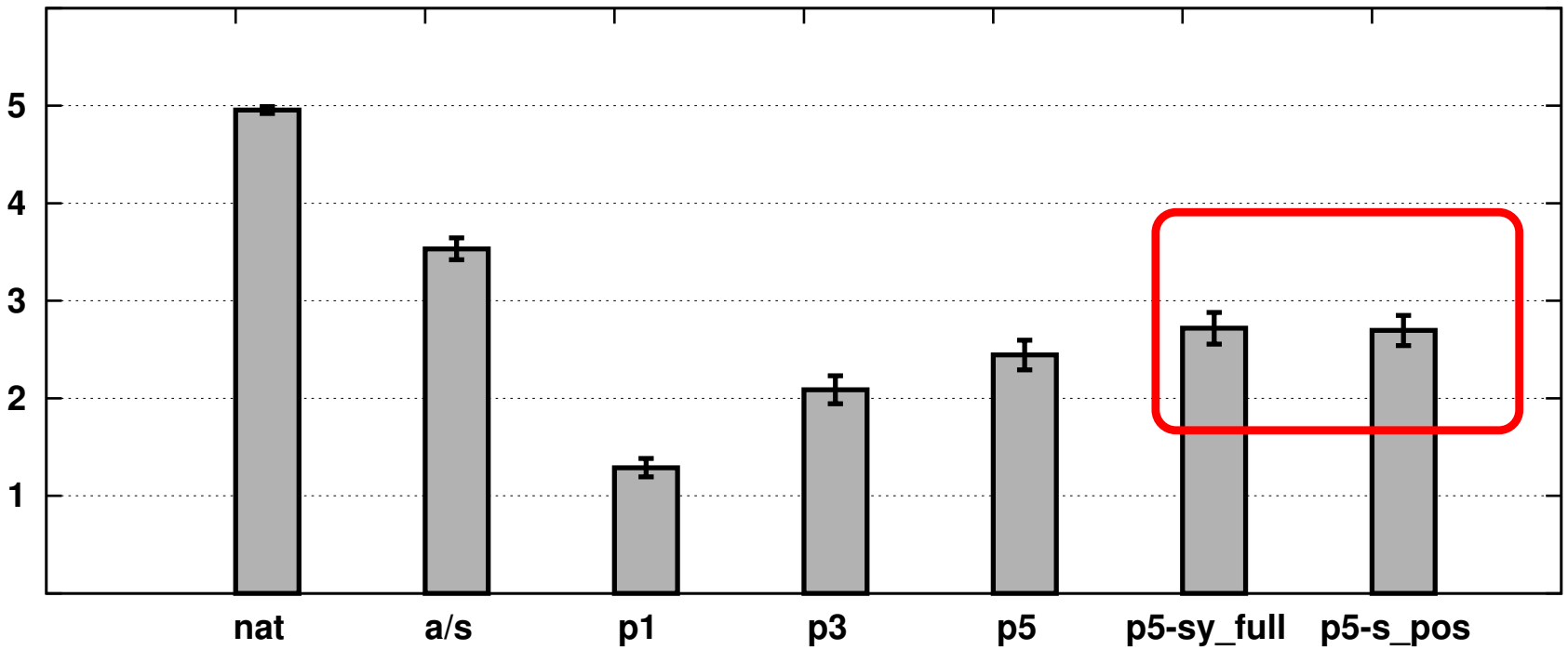
Résultats



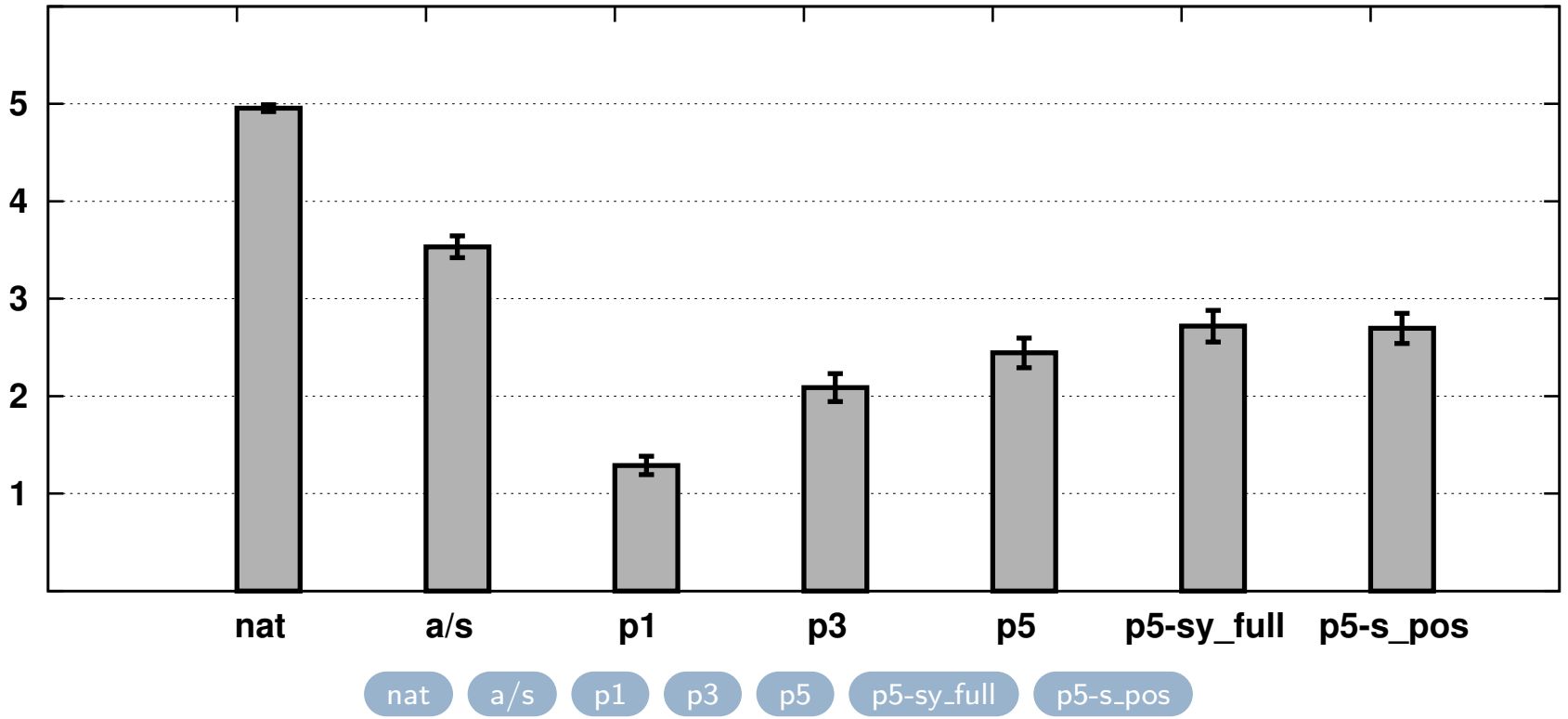
Résultats



Résultats



Résultats


5
4
3
2
1
nat
a/s
p1
p3
p5
p5-sy_full
p5-s_pos
nat
a/s
p1
p3
p5
p5-sy_full
p5-s_pos

Overview

Conclusion

Problématique

Quel est l'impact du choix des descripteurs sur la qualité du signal produit dans le cadre du français ?

- Objectif double
  1. Mise au point du système HTS pour la synthèse du français.
  2. Analyse de l'influence des descripteurs sur la qualité de la synthèse.

Conclusion - Apports

- Protocole complet d'évaluation objective de la synthèse HTS
  - Évaluation globale par GMM
  - Évaluation locale basée sur des écarts
- Évaluation subjective réalisée
- Application dans le cadre du français
  - Jeu optimal = `p5-sy_full` (de 44 à 20 desc.)
  - Divergence de qualité de synthèse entre modèles $\neq$ descripteurs
  - Concordance des résultats obtenus par évaluations objectives et par évaluation subjective

Conclusion - perspectives

- Évaluation complète du système HTS
  - Autres paramètres ?
  - Autres langues ? Effets de voix ?
  - Comparaison des techniques de modélisation
- Comparaison HTS/SPC
  - Corpus identique

NAT HTS SPC

- Utilisation du système HTS hors TTS
  - Étude complète du phénomène de la parole (production, perception)
  - Étude des troubles de la parole

Merci pour votre attention

UMR IRISA


Robert Donovan.
Trainable speech synthesis.
PhD thesis, Cambridge University, 1996.


Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai.
An adaptative Algorithm for mel-cepstral analysis of speech.
In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 137–140, 1992.


I. Ipsic and S. Martincic-Ipsic.
Croatian hmm-based speech synthesis.
Journal of Computing and Information Technology, 14(4) :307–313, 2006.


Hideki Kawahara, Ikuyo Masuda-katsuse, and Alain De Cheveigné.
Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-

based F0 extraction : Possible role of a repetitive structure in sounds 1.
Speech Communication, 27 :187–207, 1999.


Dennis H Klatt.
Software for a cascade/parallel formant synthesizer.
the Journal of the Acoustical Society of America, 67 :971, 1980.


S. Krstulovic, A. Hunecke, and M. Schröder.
An hmm-based speech synthesis system applied to german and its adaptation to a limited set of expressive football announcements.
In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 7. Citeseer, 2007.


J. Lienard, D. Teil, C. Choppy, G. Renard, and J. Sapaly.
Diphone synthesis of french : Vocal response unit and automatic prosody from the text.

In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '77.*, volume 2, pages 560–563, 1977.


J. Olive.
Rule synthesis of speech from dyadic units.
In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '77.*, volume 2, pages 568–570, 1977.


Douglas A Reynolds.
Speaker identification and verification using Gaussian mixture speaker models.
17 :91–108, 1995.


Yoshinori Sagisaka.
Speech synthesis by rule using an optimal selection of non-uniform synthesis units.

In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 679–682, 1988.

Koichi Shinoda and Takao Wanabe.
MDL-based context-dependent subword modeling for speech recognition.
Acoustical Science and Technology (AST), 21(2) :79–86, 2000.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg.
Tobi : A standard for labeling English prosody.
In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 867–870, 1992.

Y. Stylianou, O. Cappe, and E. Moulines.
Continuous probabilistic transform for voice conversion.
Speech and Audio Processing, IEEE Transactions on, 6(2) :131–142, mar 1998.

Tomoki Toda and Keiichi Tokuda.

Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis.
In *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pages 2801–2804, 2005.


Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai.
Speech parameter generation from HMM using dynamic features.
In *Proceedings of the International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pages 660–663, 1995.


Keiichi Tokuda, Heiga Zen, and Alan W Black.
An hmm-based speech synthesis system applied to english.
In *Proceedings of the Speech Synthesis Workshop (SSW)*, pages 2–5, 2002.


Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura.

Speech parameter generation algorithms for hmm-based speech synthesis.
In *Proceedings of the International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pages 1315–1318, 2000.


Oliver Watts, Junichi Yamagishi, and Simon King.
The Role of Higher-Level Linguistic Features in HMM-Based Speech Synthesis.
In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 841–844, 2010.


Shuji Yokomizo, Takashi Nose, and Takao Kobayashi.
Evaluation of Prosodic Context Factors for HMM-Based Speech Synthesis.
In *proceedings of Interspeech*, pages 430–433, 2010.


Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura.

Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis.
In *Proceedings of the European Conference on Speech Communicationand Technology (Eurospeech)*, volume 5, pages 2347–2350, Budapest, Hungary, 1999.

Steve J Young, Julian J Odell, and Phil C Woodland.
Tree-based state tying for high accuracy acoustic modelling.
In *Proceedings of the workshop on Human Language Technology (HLT)*, pages 307–312, Morristown, New Jersey, USA, 1994. Association for Computational Linguistics.

Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura.
Hidden semi-markov model based speech synthesis.
In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 1397–1400, 2004.

Heiga Zen and Tomoki Toda.

An overview of Nitech HMM-based speech synthesis system for blizzard challenge 2005.
In *Proceedings of the 9th European Conference on Speech Communicationand Technology (Eurospeech)*, Lisbon, Portugal, 2005.

H. Zen, K. Tokuda, and A.W. Black.
Review : Statistical parametric speech synthesis.
Speech Communication, 51(11) :1039–1064, 2009.