

Structuration en tours de parole multi-documents

Etudes au LIMSI pour les projets QUAERO et ETAPE

Claude Barras

LIMSI-CNRS, Orsay, France



Journées ETAPE
Rennes, 18 juin 2012

Motivations

Architectures SRL et SRL-X

Validation des approches

Evaluation ETAPE

Conclusions

Tâche de structuration en locuteurs

- Cadre habituel hérité des évaluations NIST RT
- Pas de connaissance a priori des voix des locuteurs
- Chaque émission est traitée indépendamment

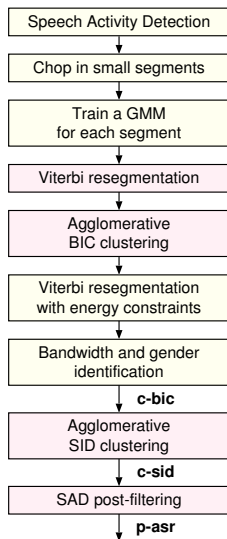
Besoins applicatifs plus larges

- Flux audio : traitement en ligne avec délai
- Suivi de locuteur : modèles de locuteurs connus
- Indexation de collections de documents provenant d'une même source
→ exploré dans le programme QUAERO à partir de 2010

Structuration multi-documents

- Un locuteur donné doit conserver le même identifiant
- Tran et al., *Comparing Multi-Stage Approaches for Cross-show Speaker Diarization*, Interspeech 2011.

Architecture multi-passes standard



*avec Sylvain Meignier et Xuan Zhu,
IEEE Trans. ASLP, 2006.*

- 1ère passe (segments courts)
 - modélisation par une Gaussienne à matrice de covariance pleine
 - classification hiérarchique ascendante
 - regroupement des segments à forte similarité (critère BIC)
- 2ème passe
 - modélisation plus riche par GMMs adaptés d'un modèle générique
 - rapport de log vraisemblance croisée (CLR) entre classes

Schéma 1: BIC + CLR global

- Equivalent à la concaténation de toutes les émissions
- Simple mais souvent trop coûteux en ressources

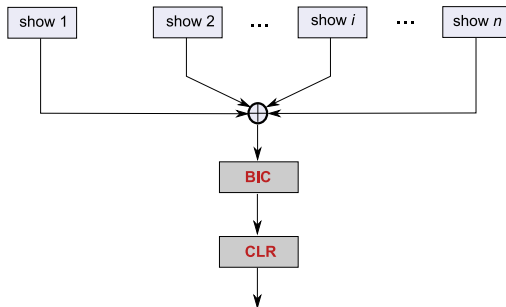


Schéma 2: BIC local + CLR global (hybride)

- Segmentation en tours de parole et classification BIC locale + classification CLR globale à tous les documents
- Réduction significative du coût de calcul

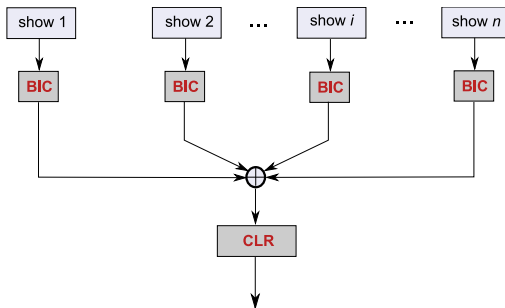
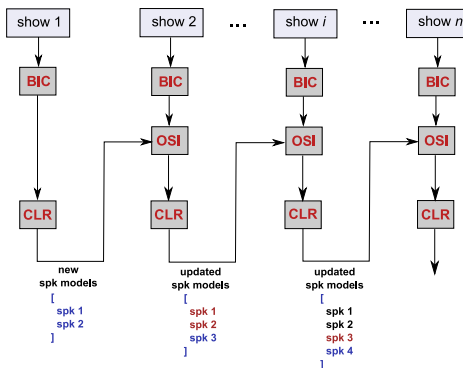


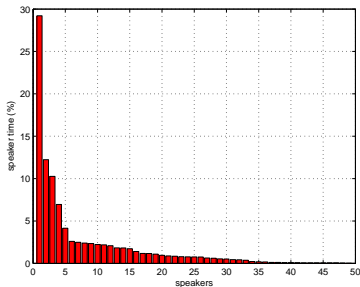
Schéma 3: BIC + CLR local (incremental)

- Présentation incrémentale des documents
- Génération de modèles pour tous les locuteurs des documents précédents
- Phase intermédiaire d'identification ouverte du locuteur (OSI)

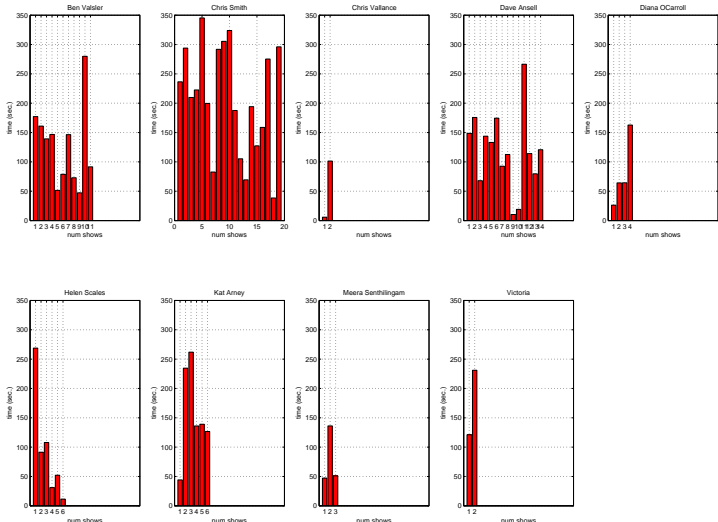


Podcasts “The Naked Scientists” (anglais britannique)

- 23 extraits d'environ 10 mn (total 4 h)
- annotés en 2010 pour le programme Quaero
- 49 locuteurs dont 9 sur plusieurs émissions



Distribution des locuteurs multi-documents



Système	DER mono-doc	DER multi-doc
Standard	6.9	54.7
Concaténé	8.2	15.2
Hybride	8.1	15.4
Incrémental	$\mu = 7.6 \sigma = 0.5$	$\mu = 17.3 \sigma = 2.4$

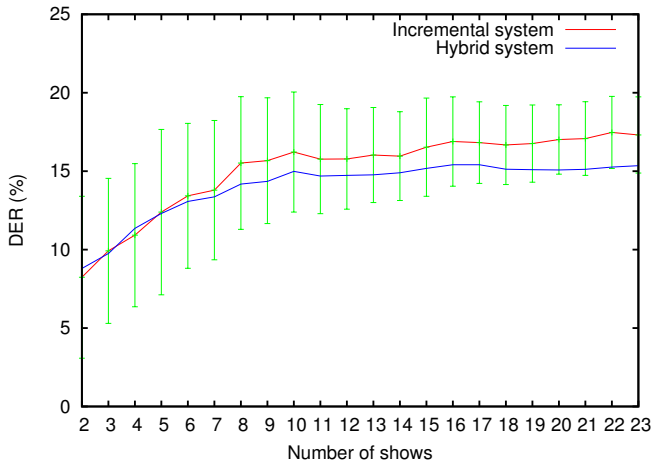
Hybride

- 10x plus rapide que concaténé, performance comparable

Incrémental

- Testé sur 25 permutations des documents
- La performance est sensible à l'ordre

Systèmes incrémental et hybride



- Evolution du taux DER multi-show en fonction du nombre croissant de documents sur 25 permutations

SRL

- Ensemble de développement limité aux données ELDA (alignement forcé réalisé au LNE)
- Pas d'intégration avec le système SES-2 (manque de temps!)
- Système standard avec choix du seuil CLR sur le dev

SRL-X

- Peu d'émissions par source, l'approche incrémentale n'est pas indispensable
- Approche hybride choisie

DER sur le dev. ELDA (hors parole superposée)

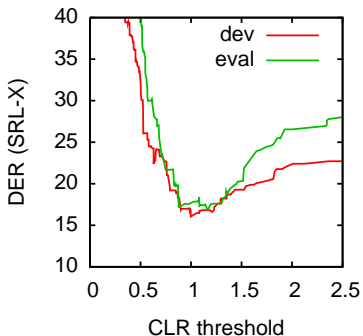
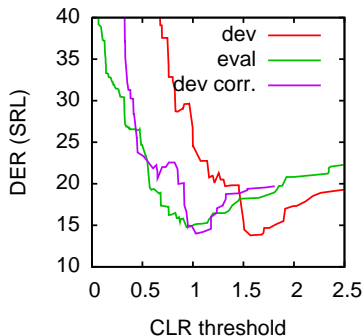
Emission	Match error	Missed speech	F. alarm speech	DER
BFMTV_BFMStory_2011-03-17_175900	2.2	1.0	1.7	4.9
LCP_CaVousRegarde_2011-02-17_204700	3.3	0.0	2.8	6.1
LCP_EntreLesLignes_2011-03-18_192900	9.6	0.0	0.6	10.2
LCP_EntreLesLignes_2011-03-25_192900	10.3	0.0	0.9	11.2
LCP_PileEtFace_2011-03-17_192900	2.1	0.0	1.1	3.2
LCP_TopQuestions_2011-03-23_213900	5.3	1.0	4.4	10.7
LCP_TopQuestions_2011-04-05_213900	19.6	1.1	2.2	22.9
TV8_LaPlaceDuVillage_2011-03-14_172834	6.8	0.1	4.1	11.0
TV8_LaPlaceDuVillage_2011-03-21_201334	27.9	0.3	10.1	38.3
Toutes (meilleur seuil global =1.6)	10.7	0.4	2.7	13.8
Cross-show	12.0	0.4	2.7	15.1

DER sur le test (hors parole superposée)

Emission	Match error	Missed speech	F. A. speech	best DER	actual DER
BFMTV_BFMStory_2011-05-31_175900	13.6	0.3	1.5	15.3	16.0
LCP_CaVousRegarde_2011-05-12_235900	2.4	0.0	1.5	3.9	10.9
LCP_EntreLesLignes_2011-05-06_192800	5.6	0.0	3.1	8.7	15.7
LCP_EntreLesLignes_2011-05-13_192800	11.3	0.0	3.0	14.3	14.3
LCP_PileEtFace_2011-05-26_192800	3.6	0.0	1.7	5.3	11.7
LCP_TopQuestions_2011-05-18_000400	0.1	0.0	12.0	12.1	12.1
LCP_TopQuestions_2011-05-25_213800	0.6	0.0	2.1	2.7	4.7
TV8.LaPlaceDuVillage_2011-05-03_201300	11.7	0.0	7.9	19.7	34.9
TV8.LaPlaceDuVillage_2011-05-12_172800	21.4	0.1	6.7	28.2	31.5
20100208_1000_FINTER_DEBATE	4.6	0.2	4.5	9.2	18.1
20100208_1750_FINTER_DEBATE	34.3	0.0	17.2	51.5	56.5
20101007_2152_FINTER_DEBATE	21.6	2.3	13.6	37.5	38.4
20101014_2152_FINTER_DEBATE	23.3	0.2	13.5	37.0	40.6
20101018_0910_FINTER_DEBATE	1.7	0.0	5.5	7.1	12.4
20101024_2004_FINTER_DEBATE	16.0	0.0	6.4	22.4	22.5
all (best global threshold)	10.1	0.1	4.6	14.8	18.3
Cross-show	12.2	0.1	4.6	16.9	17.4

Evaluation ETAPE - Bilan

- Perte de 3,5% DER sur le système SRL
problème de calibration du seuil CLR (bug!)
- Difficulté plus importante des débats France Inter
(fausses alarmes de parole + diarization)
- Résultat conforme aux attentes pour SRL-X



Structuration cross-show

- Essaimage réussi des évaluations initiées avec le LNE dans le cadre du programme QUAERO
- Intérêt et participation de plusieurs équipes

Participation du LIMSI

- SRL : pas de modifications au système 'standard', une dérive observée du seuil de regroupement CLR entre le dev et le test (erreur de système)
- SRL-X : pas assez de documents pour le mode incrémental, choix de l'approche hybride plus robuste
- Détection de parole superposée pas (encore) intégrée