CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

GEOD
Groupe d'Etude sur l'Oral et le Dialogue

LE SYSTEME DE RAP DU CLIPS : EMISSION RADIOPHONIQUE

- **Richard LAMY**
- Daniel MORARU
- Brigitte BIGI
- Laurent BESACIER

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Plan de la présentation

- Plate-forme expérimentale
- Système de base
  - Modélisations
  - Présegmentation
  - Performances
- Système amélioré
  - Adaptation de modèles
  - Segmentation en qualité
  - Segmentation en locuteurs

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Plan de la présentation

- Plate-forme expérimentale
- Système de base
  - Modélisations
  - Présegmentation
  - Performances
- Système amélioré
  - Adaptation de modèles
  - Segmentation en qualité
  - Segmentation en locuteurs

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Outils utilisés

- Janus 3.2 Toolkit
  - Paramétrisation et modélisation acoustique pour la transcription
  - Décodeur
- SRI-LM (Modèles de Langage)
- ELISA
  - Segmentation en locuteurs (avec param. Spro)
- LIA-Phon
- Audioseg (IRISA)

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Ressources utilisées

- Limitées pour l'instant à ce qui est fourni dans ESTER Phase1
- Signal : 30h40 TRAIN (ESTER Phase1)
- Texte
  - Transcriptions TRAIN (ESTER Phase1)
  - Le Monde (87-02)
    - 2.3 Go de données nettoyées
- Dictionnaire phonétique : BDLex
- Réglages sur DEV (ESTER Phase1)

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Plan de la présentation

- Plate-forme expérimentale
- Système de base
  - Modélisations
  - Présegmentation
  - Performances
- Système amélioré
  - Adaptation de modèles
  - Segmentation en qualité
  - Segmentation en locuteurs

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Modélisation du langage(1/2)

- Vocabulaire
  1) Mots issus de TRAIN(Ester) : 22340
  2) Mots issus de TRAIN+DEV(Ester) : 24000

- Modèle de langage
  1) Nettoyage de corpus
  2) Apprentissage
     données supplémentaires : « le monde » de 1987 à 2002.
     interpolation « le monde » et transcription ESTER train.

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Modélisation du langage(2/2)

Influence du vocabulaire et des ML sur les performances de transcription (fichier 19981217_0700_0800_inter seulement : 1h)
LZ/LP=25/6
test sur uttérances manuelles

Vocabulaire et ML	%Err	PPL(DEV)
Vocab. : TRAIN(ester), sans séquences %OOV=4.24 ML : TRAIN(ester)	49.2	245
ML : Le Monde filtré (87_02)	41.7	153
Interpolation (0.7/0.3)	39.6	109
Vocab. : TRAIN(ester), avec séquences %OOV=4.24 ML : TRAIN(ester)	X	250
ML : Le Monde filtré (87_02)	X	145
Interpolation (0.7/0.3)	39.5	111
Vocab. : TRAIN+DEV(ester), avec séquences ; %OOV=0 ML : TRAIN(ester)	X	250
ML : Le Monde filtré (87_02)	X	148
Interpolation (0.75/0.25)	34.7	135

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Modélisation acoustique

Paramétrisation :

- Extraction toutes les 10ms sur fenêtres de 20ms de
- 43 paramètres { (13 MFCC + E) + D(…) + DD(…) + ZCR } puis
- réduction par LDA à 24 paramètres.

Dictionnaire phonétique :

43 classes prédéfinies,

phonétisation : BDLEX + LIAphon + vérification manuelle.

Modèles acoustiques :

- contextuels (allophones), avec partage de gaussiennes
- « multicondition » appris sur tout TRAIN(ESTER),
- HMM à 3états (sauf silence : 4 états), 16 gaussiennes / état.

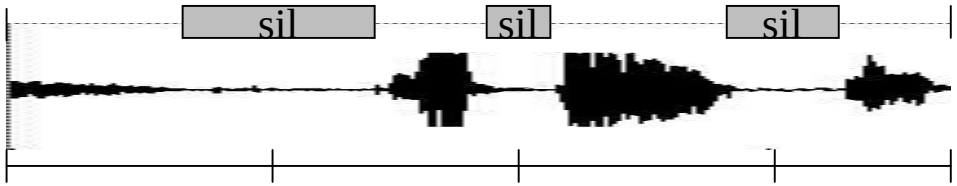
CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Présegmentation (1/2)

Traiter un signal d'une heure ? → NON
Segmentation automatique en morceaux de petites tailles.

⟹ Utilisation de « audioseg » (outils de l'IRISA).

1. Détection de zones de silence d'au moins 0.3s.
2. Découpage du signal en utterances : une zone de silence détectée est un séparateur.



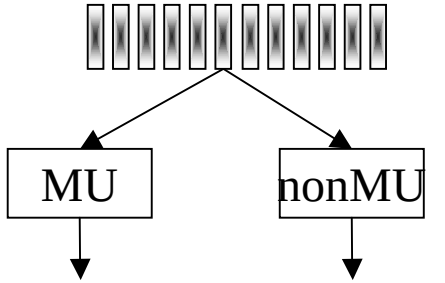
Nombre d'utterances obtenu automatiquement cohérent avec le nombre d'utterances obtenu dans les références (segmentation manuelle).

Durée des segments < 1mn.

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

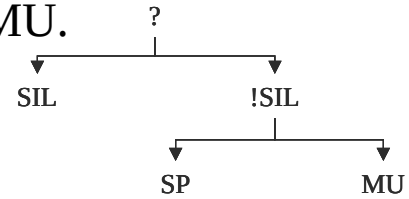
Présegmentation (2/2)

Segmentation en qualité par quantification vectorielle
Quantification de l'espace par algorithme des kmeans par éclatement binaire.
Modélisation des classes audios par une cartographie.



Nous appelons BE : bande étroite, BL : bande large, SP : parole, MU : musique seule.

- Détection MU/nonMU : 2,5% erreur de segmentation.
- Gain avec présegmentation Sil/nonSil : 1,4% err pour MU/nonMU.



CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Performances du système de base

- Influence du découpage automatique sur les performances en transcriptions : (test sur un seul fichier DEV France-inter 1h)

type de découpage	manuel (références)	automatique (avec MU)	automatique sans MU)
%Erreur	33.9%	33.6%	33.5%

- Performances globales en transcription avec et sans segments MU.

	DEV		TST	
	Avec MU	Sans MU	Avec MU	Sans MU
taux d'erreur %	41.6	41.3	46.0	45.8

Système ~15 RT

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Plan de la présentation

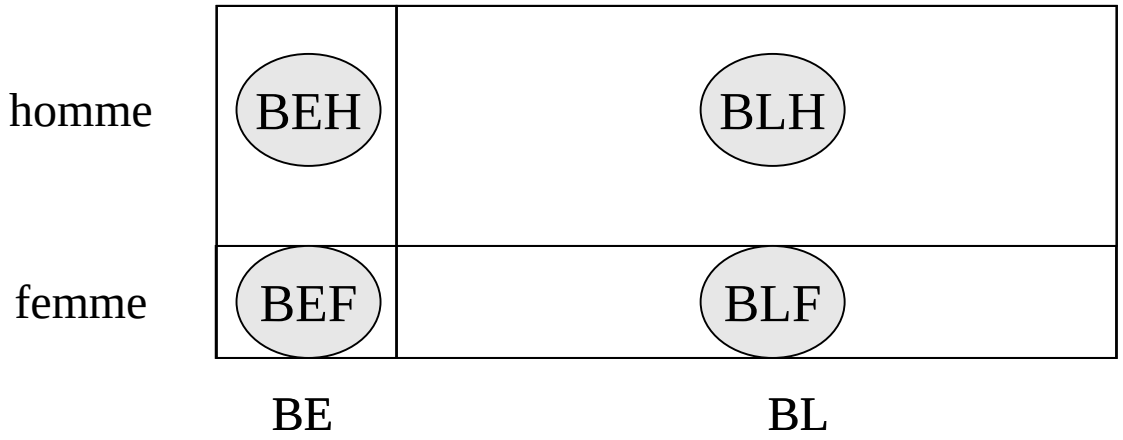
- Plate-forme expérimentale
- Système de base
  - Modélisations
  - Présegmentation
  - Performances
- Système amélioré
  - Adaptation de modèles
  - Segmentation en qualité
  - Segmentation en locuteurs

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Adaptation de modèles acoustiques

Apprentissage de modèles acoustiques spécialisés. Pourquoi ?

1. Environ 20% de signal = interviews téléphoniques, signal « bande étroite ».
2. Environ 70% de parole voix homme, 30% de parole voix femme.



Nécessité d'une présegmentation automatique en genre et en qualité.

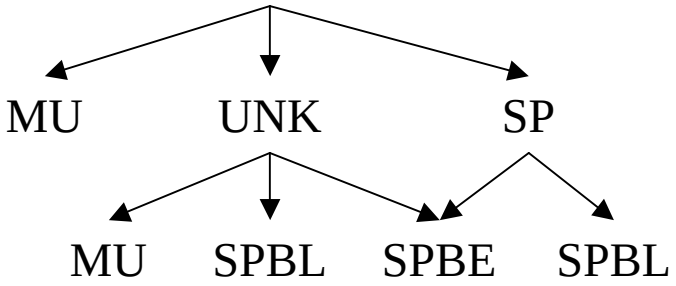
CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Segmentation en qualité (1/2)

Modélisation de classes audios par cartographies (dictionnaires de 1024 vect.).

Décision en deux passages :

- Étiquette de la trame i décidée sur la fenêtre de taille N centrée sur i.
- Premier passage donne trois étiquettes : MU, UNK ou SP
- Second passage sur parties SP : BE ou BL.
- Second passage sur parties UNK : BE ou BL ou MU.
- Lissage des labels obtenus : élimination des segments < X ms.



CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Segmentation en qualité (2/2)

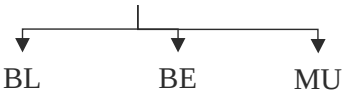
Test sur DEV avec comme référence : étiquettes du LIA obtenues automatiquement et vérifiées manuellement.

- Résultats de segmentation BE,BL et MU :

error diarization	% de parole BE bien détectée	% de parole BL bien détectée	% de musique bien détectée
1,9%	98,0%	98,4%	92,4%

- Importance de l'arbre de décision

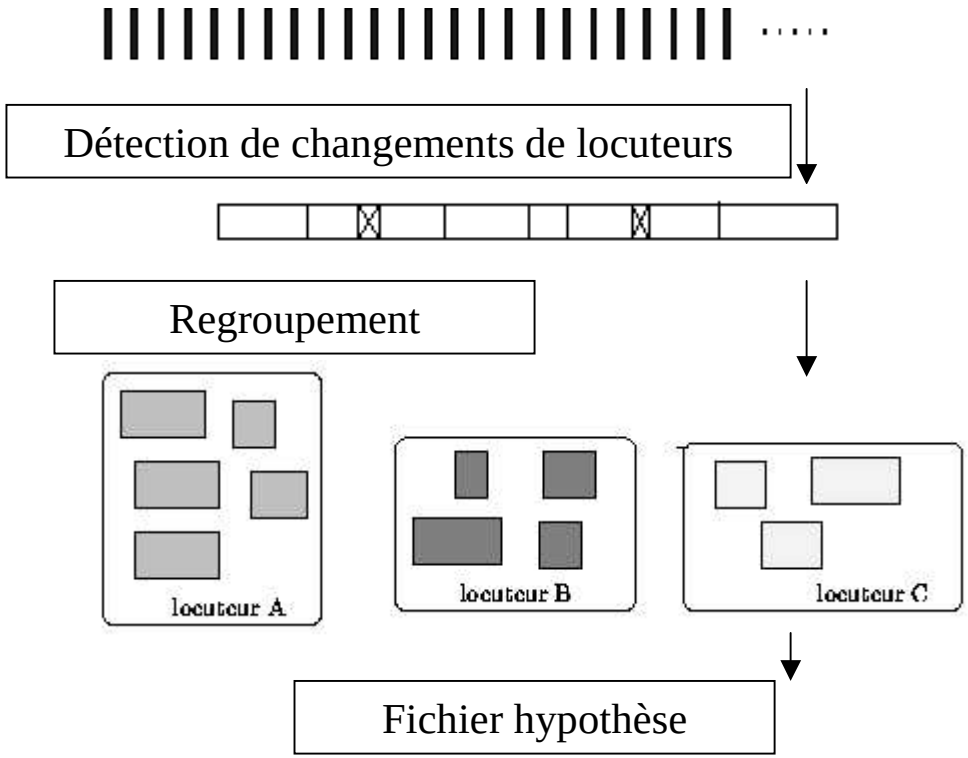
Arbre N-aire



error diarization	% de parole BE bien détectée	% de parole BL bien détectée	% de musique bien détectée
4.8%	86.3%	97.8%	90.6%

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Segmentation en locuteurs (1/2)



- Stratégie ascendante
- 16 MFCC + E
- Sans présegmentation acoustique
- Signal divisé en segments uni-locuteur par détection de ruptures
- Regroupement hiérarchique des segments par locuteurs
- Estimation du nombre de locuteur à l'aide du critère BIC pénalisé

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Segmentation en locuteurs (2/2)

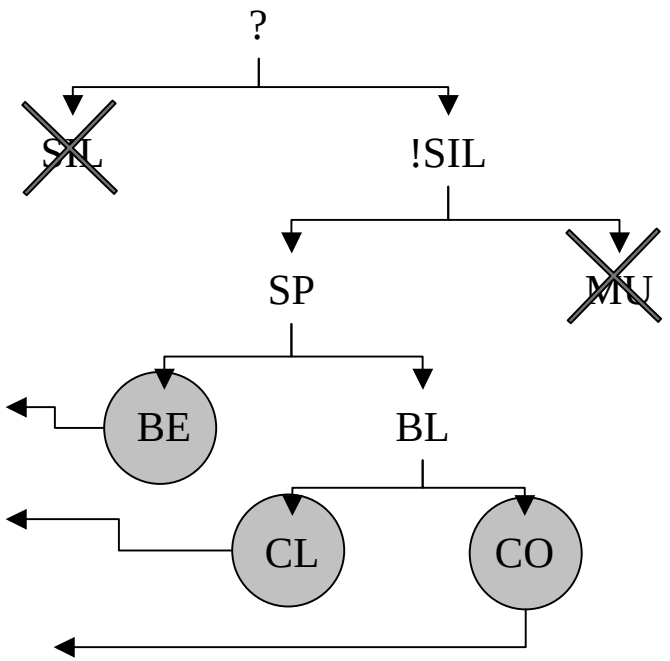
Résultats segmentation:

	% err seg DEV	%err seg TST
1217(18)_0700_0800_inter	12.0	16.9
1217(18)_0800_0900_inter	11.3	22.4
622(624)_1900_1920_inter	6.9	12.7
623(625)_1900_1920_inter	8.0	6.4
907(908)_0930_1030_rfi	13.9	13.1
907(908)_1130_1230_rfi	23.1	25.1
TOTAL	**13.8**	**17.7**

différences entre TST et DEV due à la quantité de signal non-parole (musique, pub, etc.) et à l'augmentation du nombre de locuteurs à faible temps de parole (ex: 1 sec / 1heure)

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Système de RAP amélioré : bilan provisoire



Reconnu par modèles acoustiques « bande étroite »

Reconnu par modèles acoustiques « bande large propre »

Adapté, reconnu par modèles acoustiques « bande large propre »

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Merci…