Automatic News Transcription System

v1.0

LORIA
Equipe PAROLE

Personnels impliqués

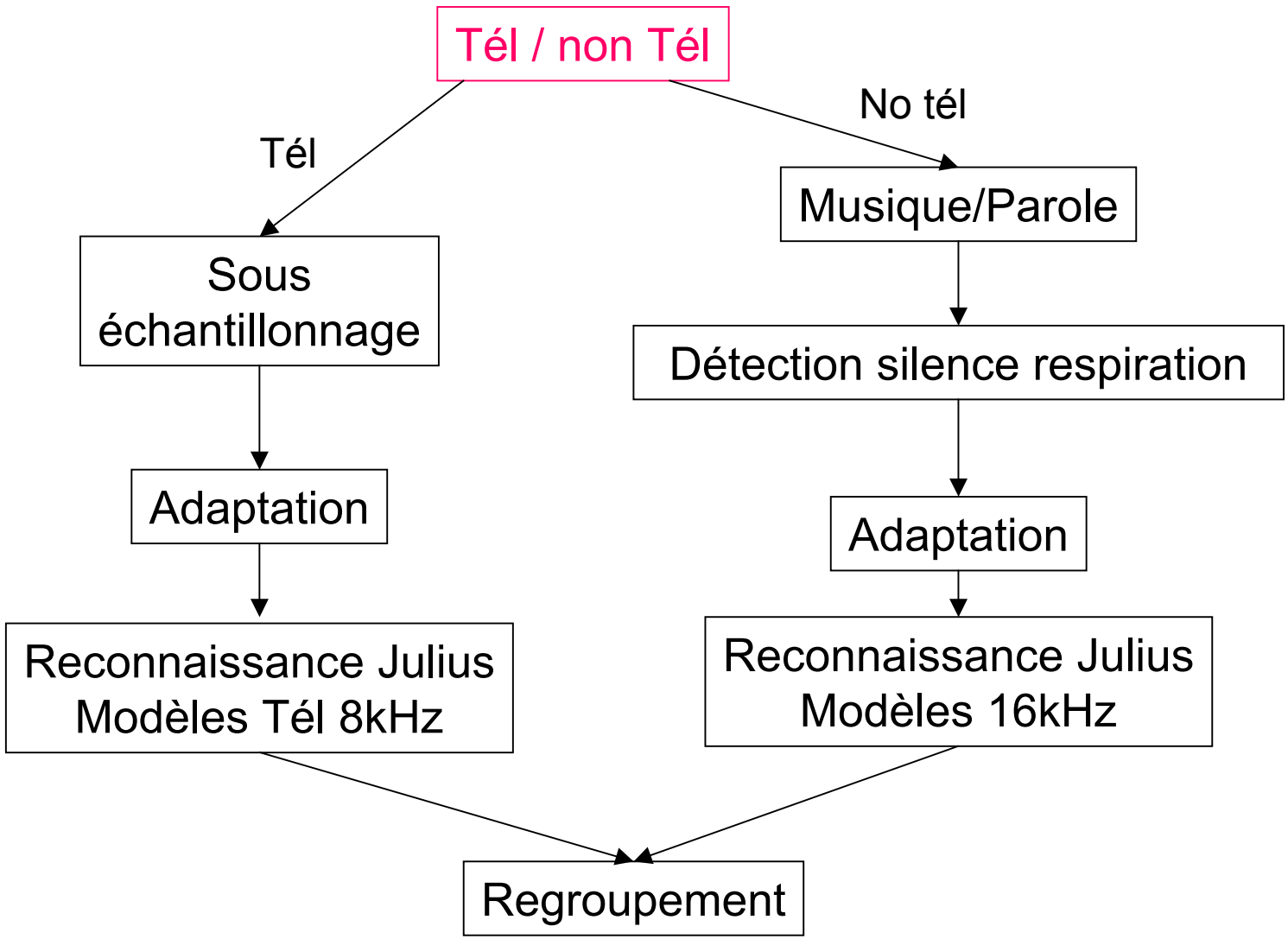
Modèles acoustiques

- Christophe Cerisara (CR)
- Dominique Fohr (CR)
- Irina Illina (MC)
- Odile Mella (MC)

Modèles de langage

- Armelle Brun (MC)
- David langlois (MC)
- Kamel Smaili (Prof)

Schéma général



Segmentation Téléphone/non-téléphone

- Basée sur la différence d'énergie haute/basse fréquence

  dif = energie[0…4000] – energie [4000…8000]

- Lissage sur 1 seconde (fenêtre glissante)

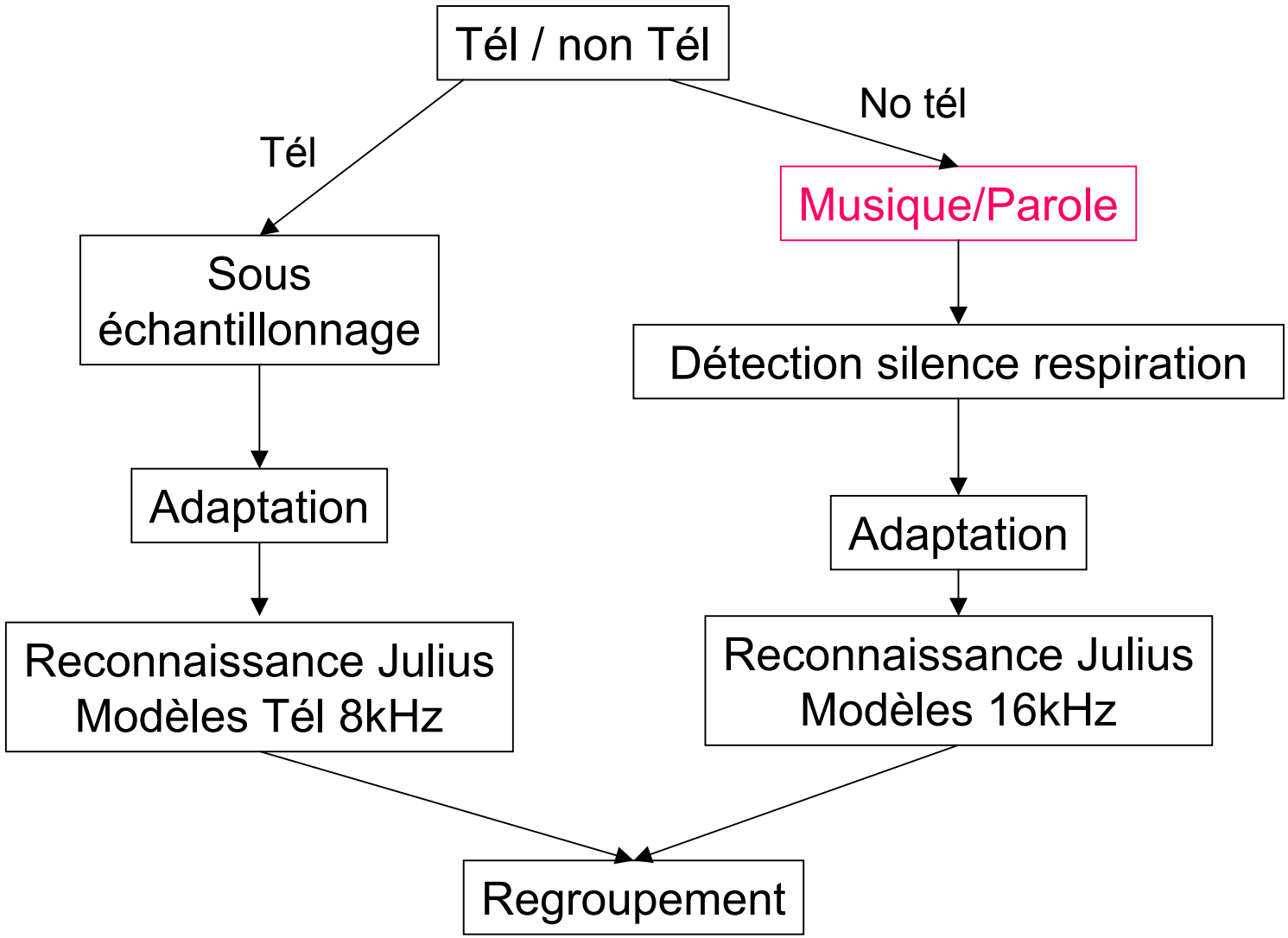
  – $$q(t)=\sum_{i=t-T}^{t} dif(i)-\sum_{i=t}^{t+T} dif(i)$$

- Détection de pics
  - Maximum
  - Vallées suffisantes

Segmentation Téléphone/non-téléphone

- Résultats
  - Sur un fichier RFI de 60 minutes
    - 225024 trames (16 ms)
    - 338 trames mal classées (0.15% soit 5s)
      - 258 trames notel classées en tel (4 s)
      - 80 trames tel classées en notel (1s)
  - Sur un fichier France-Inter de 14 minutes
    - 54953 trames
    - 237 trames mal classées (0.43% soit 3s)
      - 234 trames notel classées en tel (3s)
      - 3 trames tel classées en notel

Schéma général



Segmentation Parole/Musique

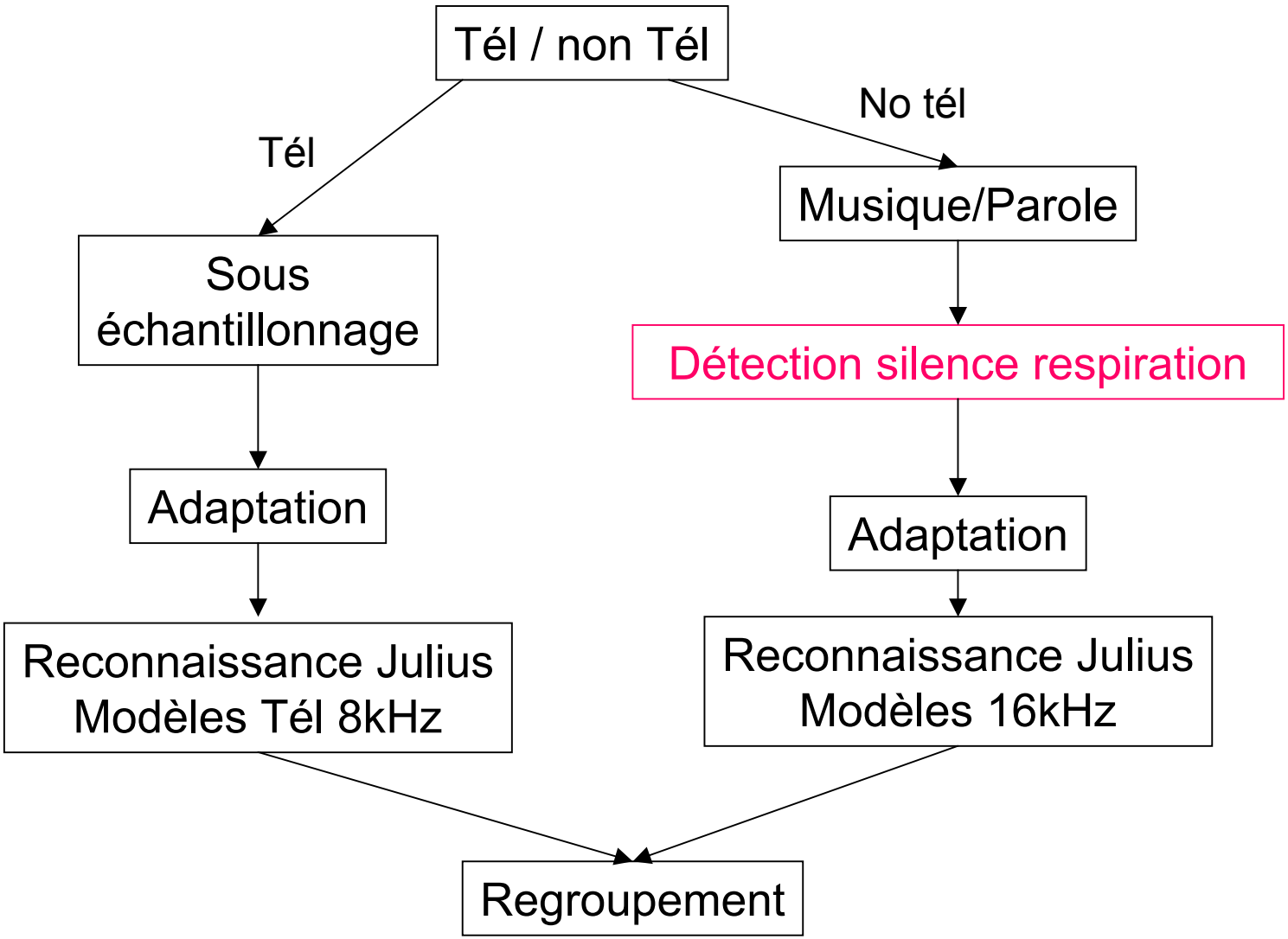
- 5 Modèles GMM à 16 gaussiennes
  - Parole studio
  - Parole téléphonique
  - Musique instrumentale
  - Chansons
  - Parole sur fond musical
- Apprentissage
  - Uniquement sur Train-Ester sauf CD audio pour musique et chanson
- Paramétrisation
  - 12 MFCC + Δ + $\Delta\Delta$

Résultats (taux d'erreur)

	Inter 7h-8h	Inter 8h-9h	Inter 20mn	Inter 20mn	RFI 9h30	RFI 11h30	Moy
SES	0.1%	0.5%	6.0%	4.6%	11.5%	23.8%	**7.8%**

- 23.8% principalement du à Parole reconnue comme Parole sur Musique

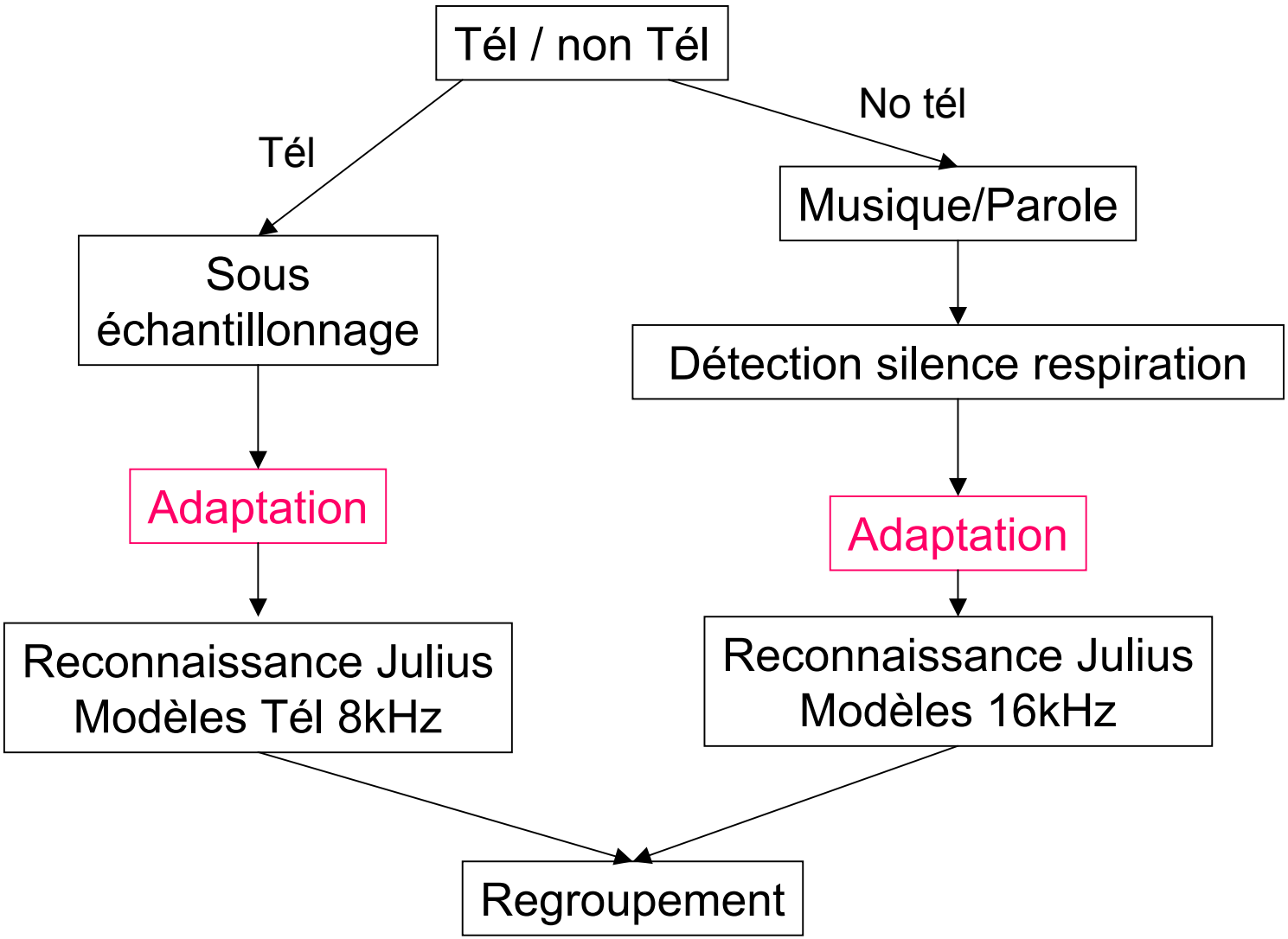
Schéma général



Détection silence respiration

- Reconnaissance phonétique (HTK)
  - 36 Modèles de phonèmes (3 états)
  - 1 Modèle de respiration
  - 2 Modèles de bruits
  - 1 Modèle de silence
- Coupure sur respirations et silences

Schéma général



Apprentissage des modèles acoustiques

- Modèles monophones 256 gaussiennes
  - 13 MFCC + Δ + $\Delta\Delta$
  - MCR
- 36 modèles de phonèmes
  - 3 « é », 3 « o », 3 « eu »
- Parole téléphonique
  - Apprentissage sur SpeechDat1000 (80h)
- Parole non téléphonique
  - Apprentissage sur Train-Ester (7h)

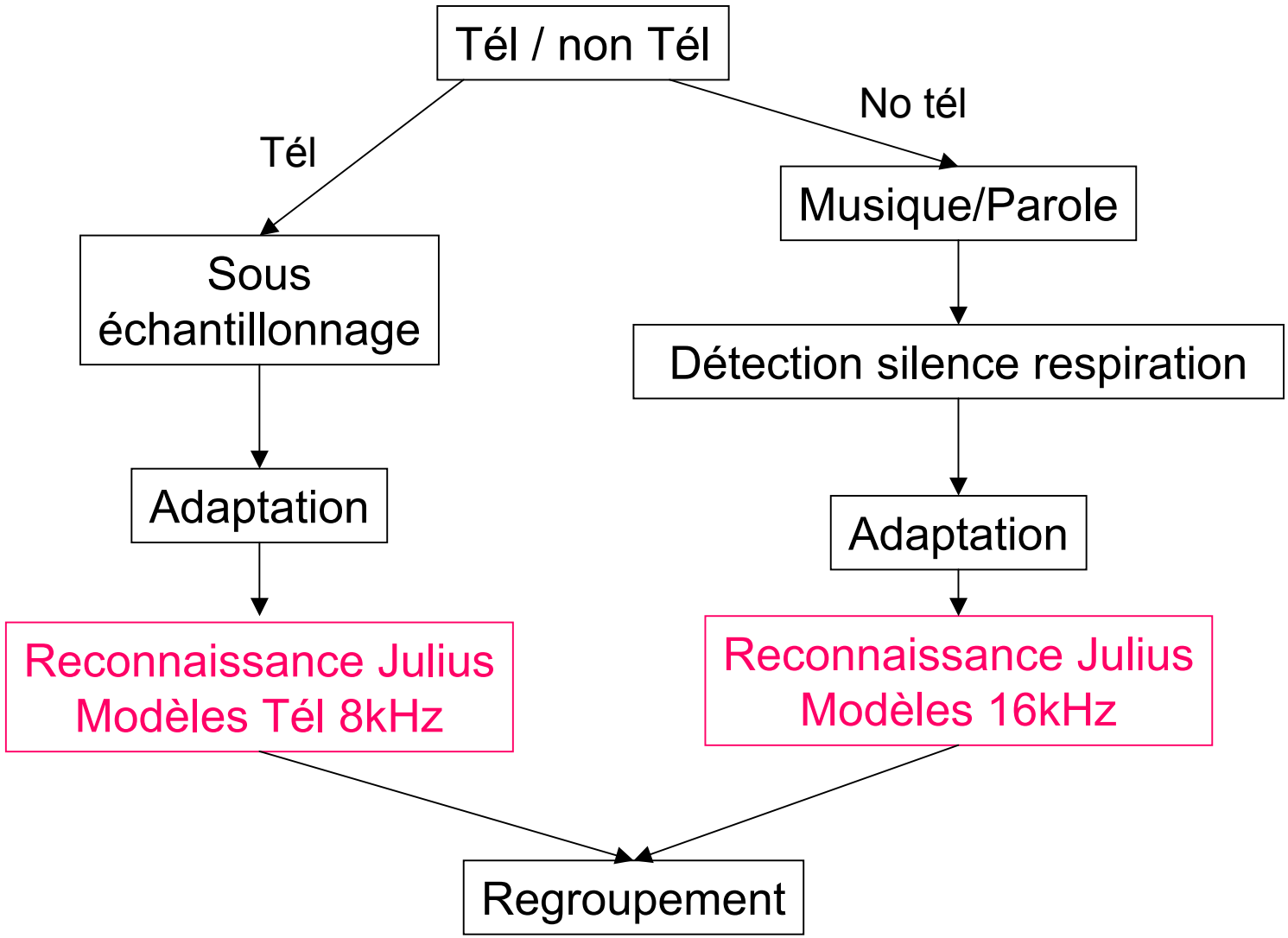
Apprentissage des modèles acoustiques

- Résultats au niveau phonèmes notél
  - 256 gaussiennes par état
  - Accuracy phonétique **69.9%**
    - Pas de grammaire
    - insertions=omissions
    - Uniquement sur des phrases non téléphoniques

Adaptation

- Reconnaissance phonétique avec HTK
- Adaptation MLLR
  - « En aveugle » : sur 3 segments
    - le précédent
    - l'actuel
    - le suivant
  - Une seule matrice de transformation
  - Matrice block diagonale (3)

Schéma général



Reconnaissance

- Moteur
  - Julius
  - 2 passes
    - Viterbi trame-synchrone, Bigramme
      - treillis de mots
    - Algorithme à pile, trigramme
- Modèles acoustiques non téléphoniques
  - monophones 256 gaussiennes
  - 13 MFCC + Δ + $\Delta\Delta$ et MCR
  - Apprentissage sur Train-Ester (7h)

Reconnaissance

- Lexique
  - 55000 mots
    - Mots les plus fréquents « Le Monde » 1992-2002
    - mots de Train-Ester apparus au moins 3 fois
- Modèle de langage
  - Bigramme et trigramme avec CMU Toolkit
    - 10 ans du Monde + 10x Train-Ester
    - 2.5 millions bigrammes
    - 5.8 millions trigrammes

Résultats (taux d'erreur)

- Sur un PC 3.0 GHz Linux avec 1Go Ram

	Inter 7h-8h	Inter 8h-9h	Inter 20mn	Inter 20mn	RFI 9h30	RFI 11h30	Moy
20xTR	25.7%	31.7%	37.7%	35.2%	46.9%	41.8%	**36.0%**
1xTR	38.5%	39.6%	46.0%	41.1%	51.4%	51.9%	**44.5%**

Système temps réel

- Pas de détection respirations/silences
- Modèles phonétiques
  - 256 -> 128 gaussiennes pour notel
  - 256 -> 64 gaussiennes pour tel
- Modèles de langage
  - Bigrammes 2.5M -> 1.5M
  - Trigrammes: 5.8M -> 2.8M
- Pruning plus important pour Julius

Perspectives pour v2.0

- Triphones
- Modèles homme/femmes
- Segmentation en locuteurs
- Adapter au locuteur
- Amélioration lexique et modèles de langage avec les nouvelles données