Automatic News Transcription System

v3.0

LORIA Nancy
Equipe PAROLE

Participants LORIA

Modèles acoustiques

- Christophe Cerisara
- Dominique Fohr
- Irina Illina
- Odile Mella

Modèles de langage

- Armelle Brun
- David Langlois
- Kamel Smaïli

Schéma général

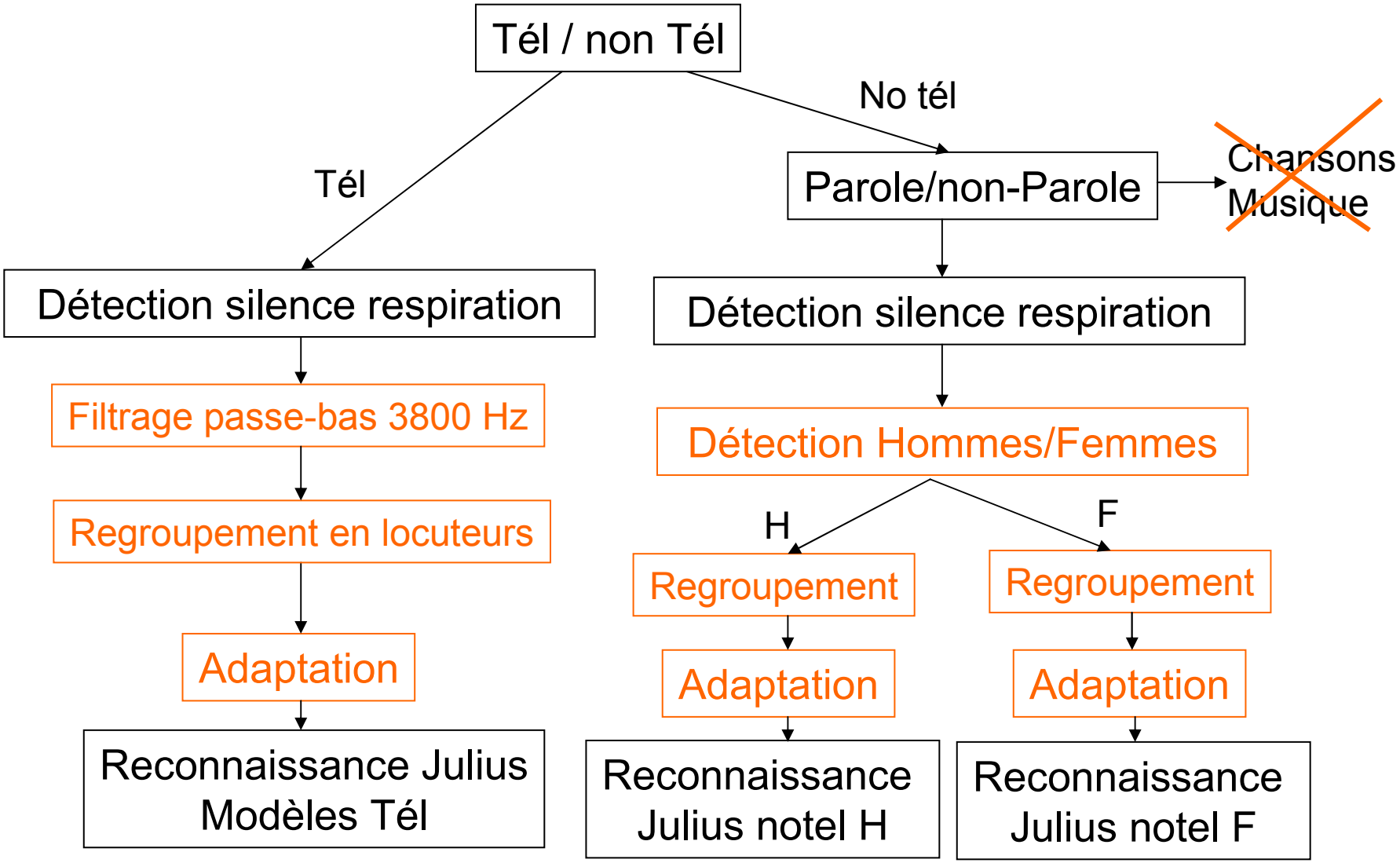
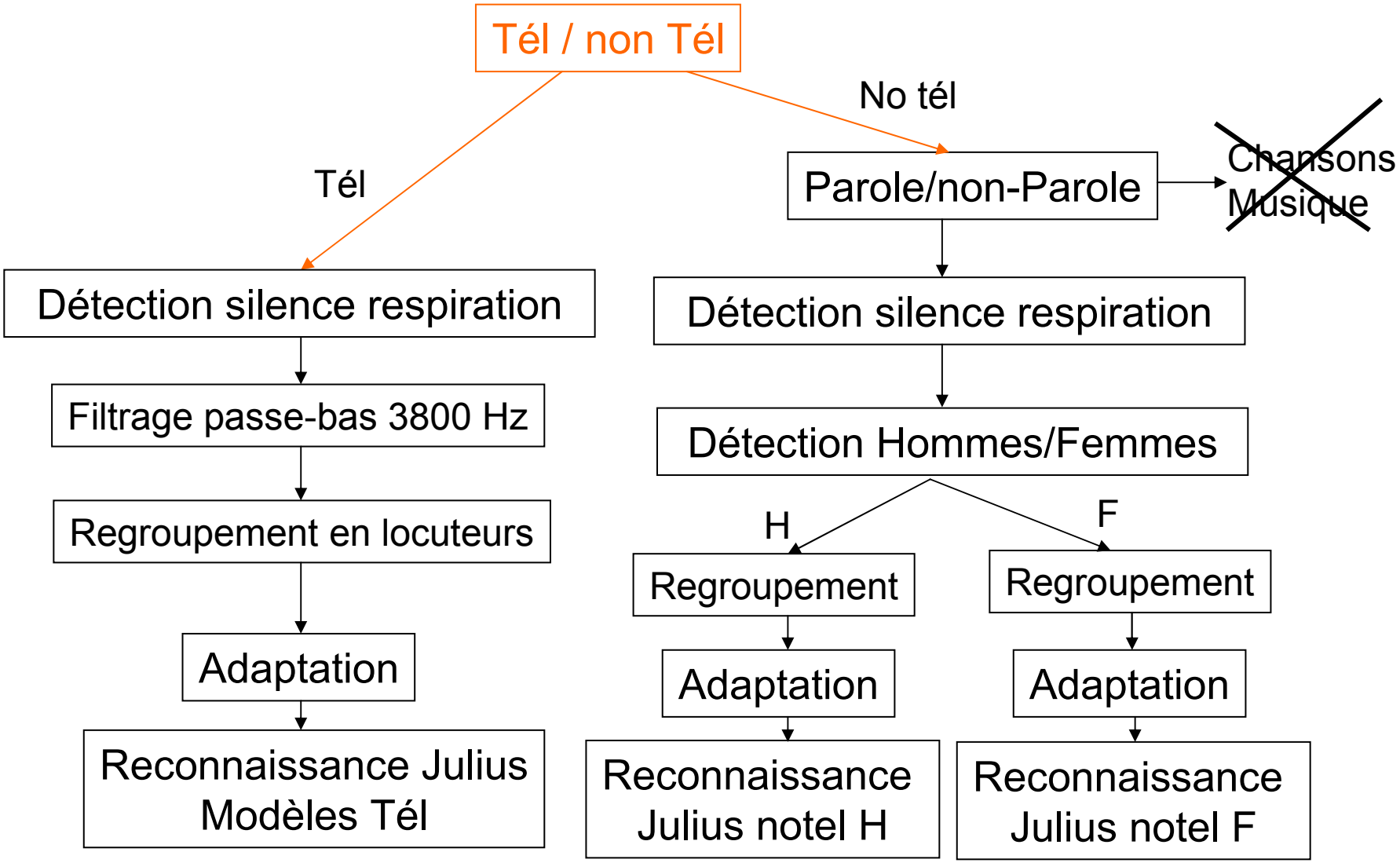


Schéma général



Segmentation Tel/notel

- Paramétrisation
  - 13 MFCC MFCC_0 (sans MCR, sans Δ)
  - 10 ms décalage, 32 ms fenêtre, 24 filtres
- Modèles
  - 2 GMM
- Apprentissage des 2 GMMs Tel et Notel
  - Corpus Ester ~30h étiqueté manu par CLIPS (merci!)
  - Logiciel HTK
  - 32 gaussiennes
- Segmentation
  - Concaténation de 50 GMM -> HMM
    - Durée minimum pour un segment : 0.5s
  - Algorithme : Viterbi

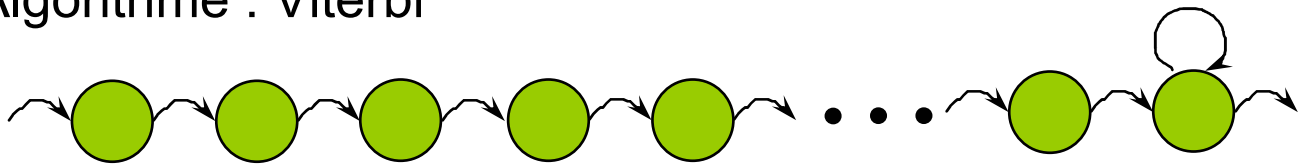
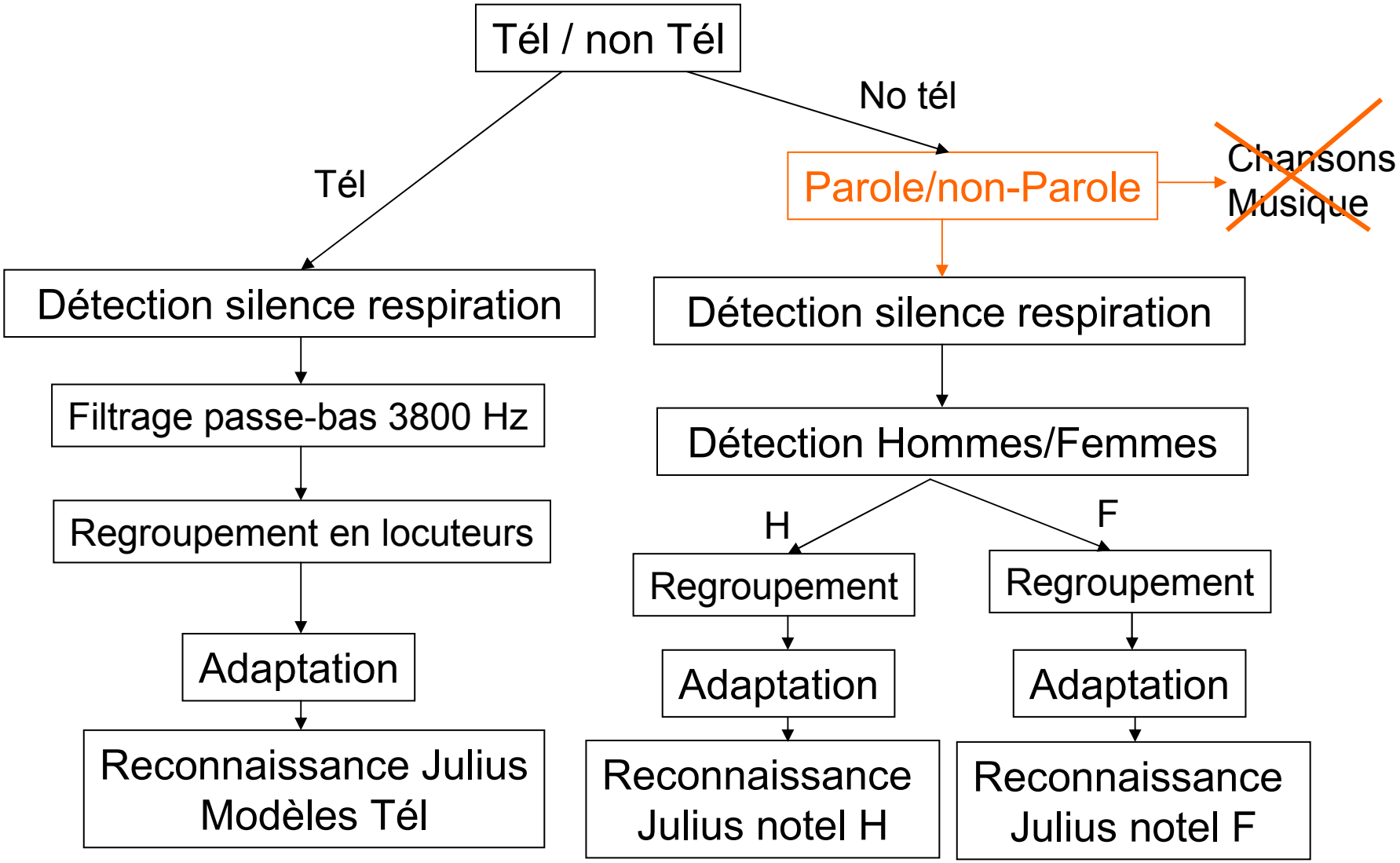


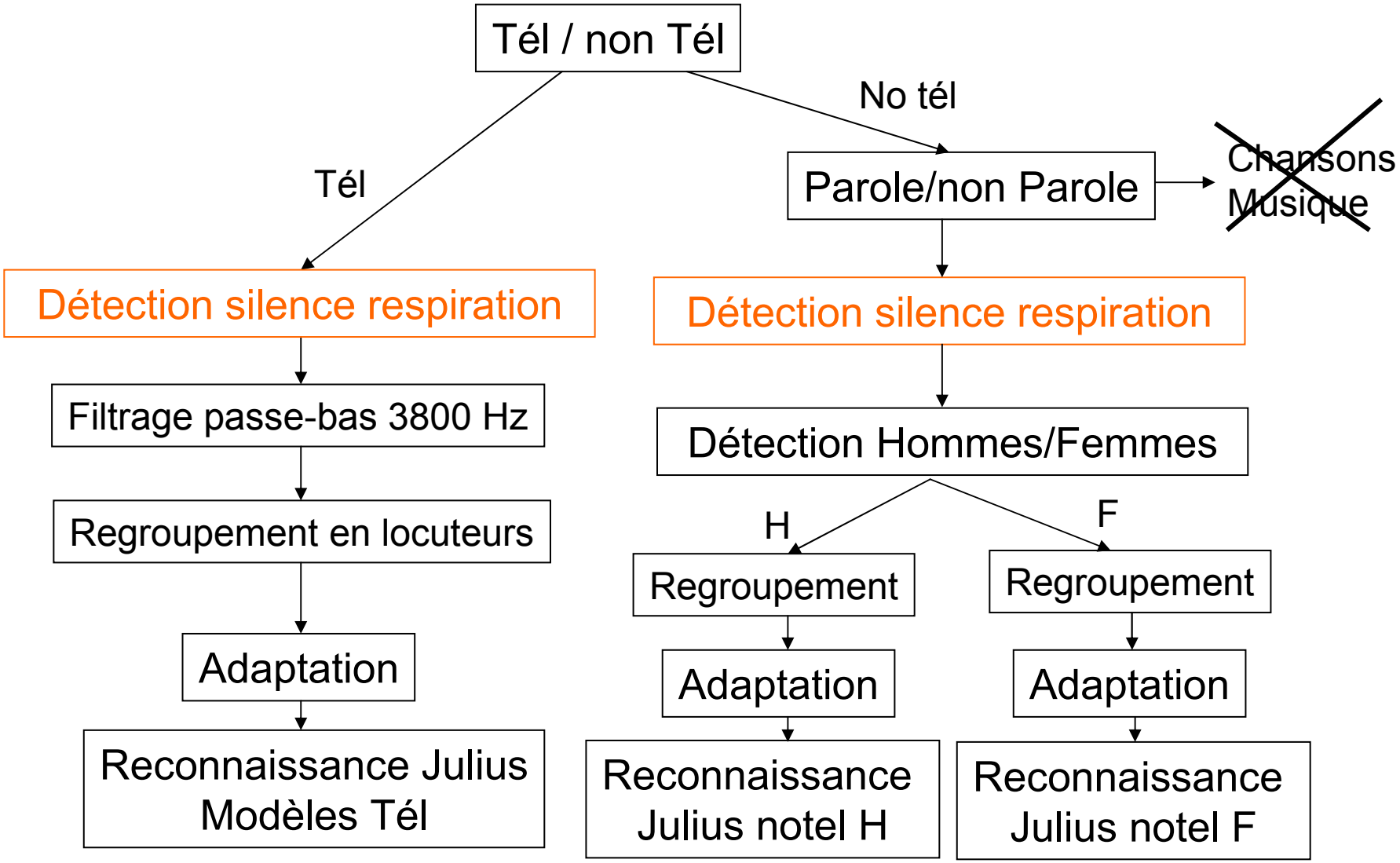
Schéma général



Détection parole/non-parole

- Paramétrisation
  - 13 MFCC + Δ + Δ Δ (MFCC_?_D_A_Z)
  - Fenêtre 32 ms, 24 filtres, 10 ms de décalage
- Modèles 4 GMMs
  - Parole : 8h30 de EsterTrainDry F0
  - Parole sur fond musical : 1H45 EsterTrainDry F3
  - Musique instrumentale: 48 minutes de CD audio
  - Chanson : 1H30 de CD audio
- Segmentation
  - Concaténation de 50 GMM -> HMM
    - Durée minimum pour un segment : 0.5s
  - Algorithme : Viterbi
  - Élimination des segments «Musique instrumentale » et «Chanson »

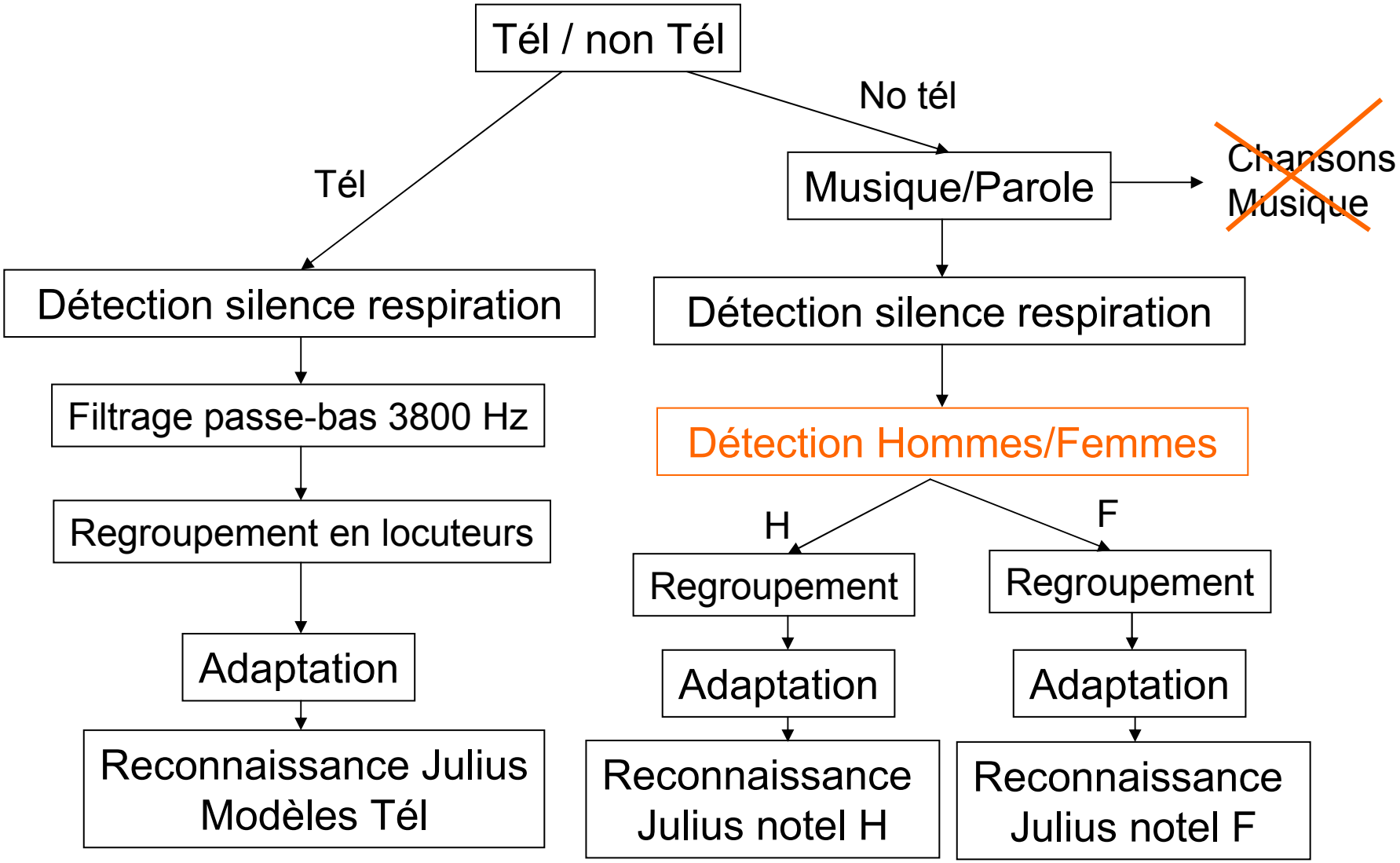
Schéma général



Segmentation Silence/Respiration

- Paramétrisation
  - MFCC 39 coeffs 0_D_A_Z
  - 10 ms décalage, 32 ms fenêtre, 24 filtres
- Modèles
  - HMM 3 états
  - monophones + silence + respiration
- Apprentissage
  - 15 heures EsterTrainDry BL
  - 3 heures EsterTrainDry BE
  - Logiciel HTK
  - 2 gaussiennes par état
- Segmentation
  - Algorithme : Viterbi (grammaire : boucle de phonèmes)
  - Découpage si respi > 0.15s ou silence > 0.3s

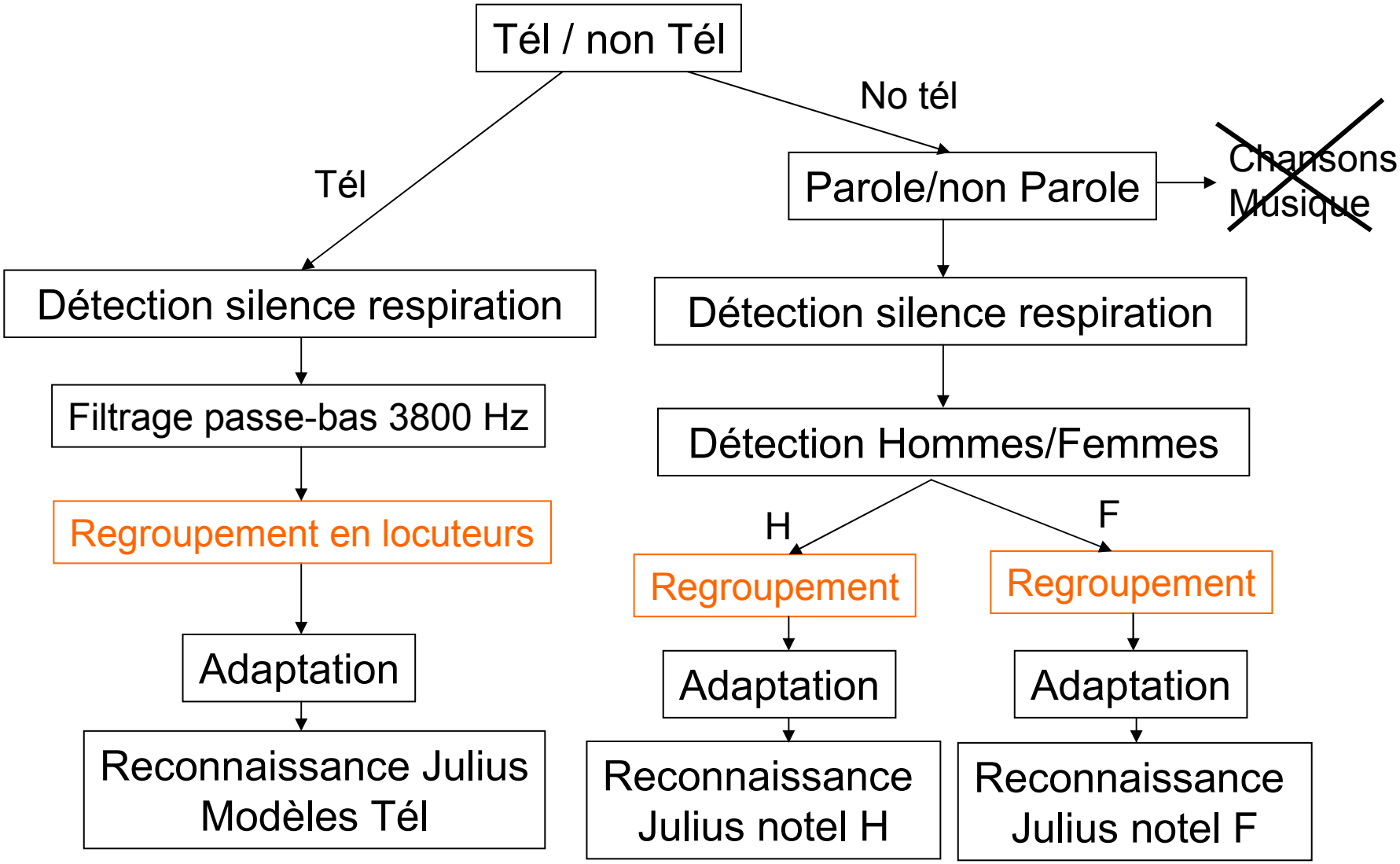
Schéma général



Détection hommes/femmes BL

- Paramétrisation
  - 39 MFCC (E_D_A_Z)
  - 10 ms décalage, 32 ms fenêtre, 24 filtres
- Apprentissage de 2 GMMs Hommes et Femmes
  - Corpus d'apprentissage:
    - H: 33 heures EsterTrain BL
    - F: 16 heures EsterTrain BL
  - 256 gaussiennes
- Segmentation
  - Concaténation de 50 GMM -> HMM
    - Durée minimum pour un segment : 0.5s
  - Viterbi

Schéma général



Regroupement de locuteurs

première ébauche

- Découpage par genre
- Découpage en segments de 15 s max
- Calcul des Paramètres MFCC_D_A
- Initialisation 1segment = 1 cluster
- Pour chaque cluster: une gaussienne
- Pour chaque couple de cluster: BIC
- Si tous les BIC négatifs FIN
- Sinon Couple qui maximise BIC :
  - Regroupement des deux clusters
  - Calcul de la nouvelle gaussienne

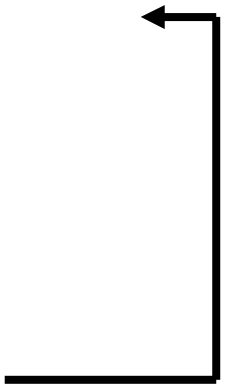
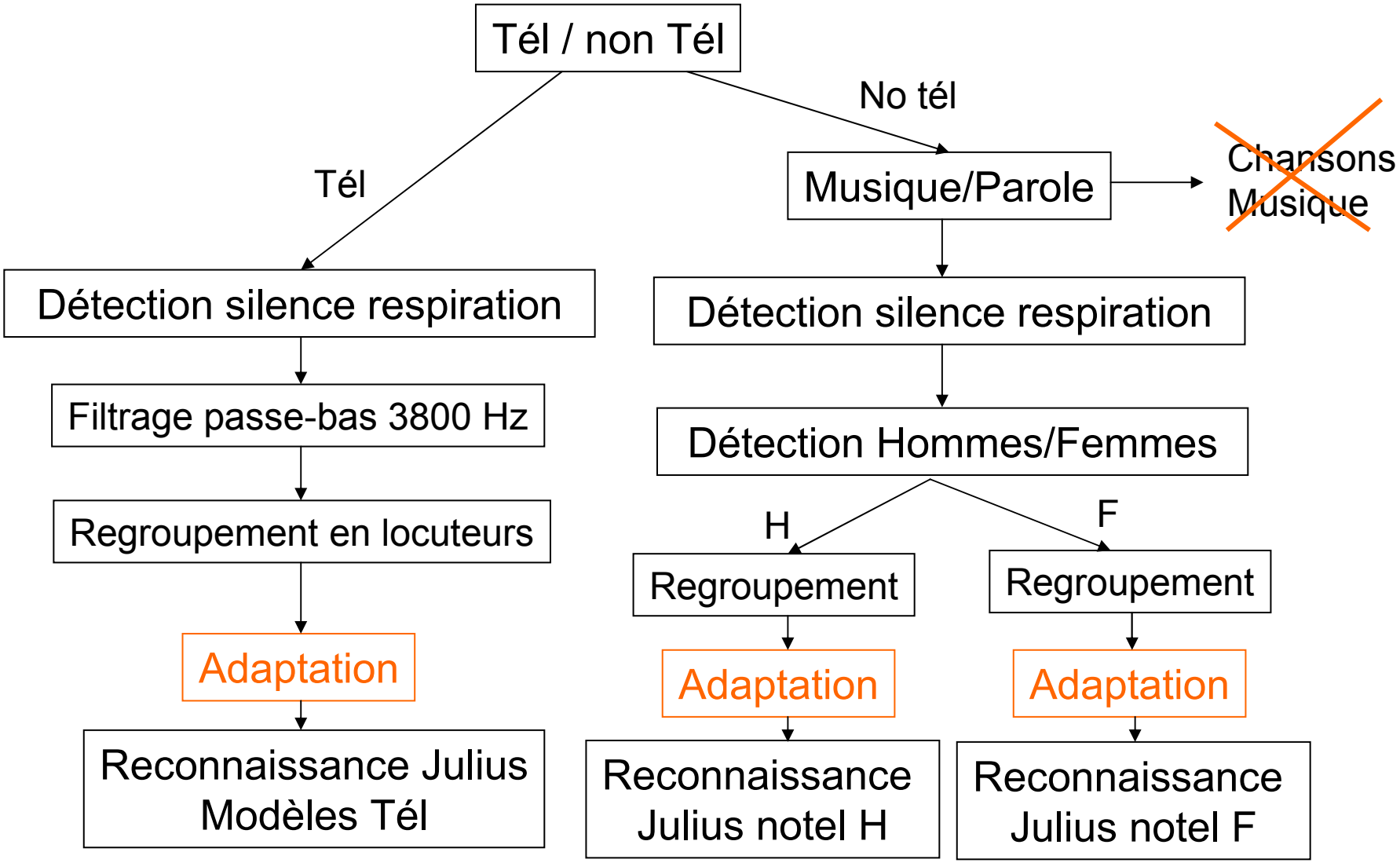


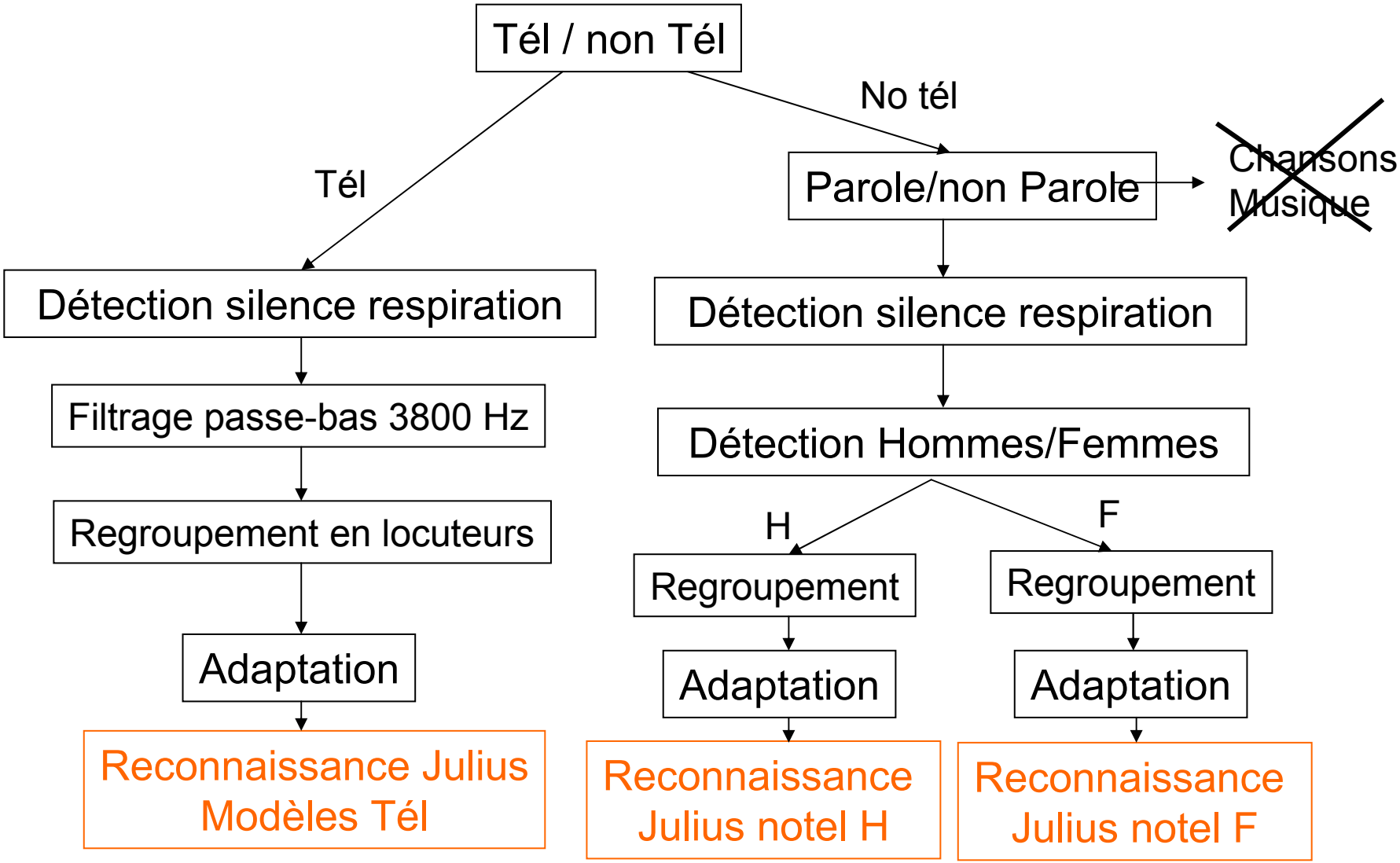
Schéma général



Adaptation au locuteur

- Pour tous les segments d'un locuteur
  - Reconnaissance avec Julius
    - Vocabulaire : 59549 mots
    - HMM: triphones
    - Résultat : suite de mots => suite de phonèmes
  - Adaptation
    - SMLLR 3 block-diagonal (HEAdapt)

Schéma général



Moteur de reconnaissance

- Julius:
  - *Open-Source Large Vocabulary CSR Engine*
  - Développé par LEE Akinobu *Continuous Speech Recognition Consortium*
  - Modèles acoustiques :
    - triphones (cross-word context)
    - Format HTK
  - Lexique qui autorise plusieurs prononciations
  - Modèle de langage:
    - Format CMU/SRI arpa

Reconnaissance

- Moteur Julius 2 passes
  - 1ère passe :
    - avant
    - trame synchrone
    - bigrammes
  - 2ème passe :
    - arrière
    - A* (estimé=score de première passe)
    - trigrammes reverse
- 20 fois temps réel

Modèles acoustiques BL

- Sélection des phrases BL à l'aide du GMM tel/notel
- Étiquetage phonétique du corpus BL
  - Transcription -> prononciations possibles
  - Alignement forcé
- Apprentissage
  - Paramétrisation :
    - MFCC E_D_A_Z 39 coefficients
  - Modèles appris sur 62 heures de EsterTrain BL
  - 36 phones + 1 short pause + bb + bruit + respiration
    - 3 « é »
    - 3 « eu »
    - 3 « o »
  - Triphones
    - 15 gaussiennes par état (3 états sans saut)
    - 89490 gaussiennes au total
  - HTK toolkit HERest
  - MAP pour s'adapter au genre

Modèles acoustiques BE (tel)

- Étiquetage phonétique de tout le corpus EsterTrain
  - Transcription -> prononciations possibles
  - Alignement forcé
- Apprentissage
  - Paramétrisation :
    - Filtrage passe bas 3800 Hz
    - MFCC E_D_A_Z 39 coefficients
  - Modèles appris sur tout EsterTrain (BL+BE) filtré
  - 36 phones + 1 short pause + bruit bouche + bruit + respiration
  - Triphones
    - 20 gaussiennes par état (3 états sans saut)
    - 115200 gaussiennes au total
  - HTK toolkit HERest
  - Pas d'adaptation au genre

Lexique

- 5000 mots +fréquents EsterTrain+Dev
- 55000 mots +fréquents du journal Le Monde
- Phonétisation
  - Noms communs
    - Bdlex
  - Noms propres
    - Phonétiseur
    - Correction manuelle
- 112000 prononciations au total

Modèles de langage

- Corpus de développement
  - 1/3 FranceInter, 1/3 FranceInfo, 1/3 RFI de EsterDev
  - Détermination pour chaque modèle de :
    - La méthode de *discounting* optimale (tests sur *linear, witten-bell, good-turing, absolute*)
    - Le *cut-off* optimal (test sur 0,1, …, 10)
- Interpolation entre :
  - ML « Le Monde » trigrammes
    - w=0.55
    - Méthode de *discounting* : « *witten-bell* »
    - *Cutt-off*=1 (pour que ça tienne en mémoire)
  - ML « Ester » trigrammes
    - pas de *cut-off*
    - Méthode de *discounting* : «*absolute* »
    - w=0.45
- Avec UNK
- 7.4 M bigrammes 25.4 M trigrammes

Cas particulier RTM

- Modèles adaptés pour RTM
  - Adaptation Map sur Train RTM notel Hommes
  - Adaptation Map sur Train RTM notel Femmes

Résultats sur 10 min FranceInfo (sans tel)

- Influence du nombre de gaussiennes

Nb total gaussiennes	17898	29830	41762	59660	89490	119276
Nb Moyen de gaussienne par état	3g	5g	7g	10g	15g	20g
WER	24.2%	24.3%	22.4%	21.0%	21.1%	21.7%

Problème au-delà de 60000 gaussiennes ?

Résultats sur 10 min FranceInfo (sans tel)

- Modification de « search » dans le Julius A*
- Tri de la pile des hypothèses
  - Temps de début de l'hypothèse
  - Score en cas d'égalité sur le temps de début
- Résultats
  - A* standard : WER=20.4%
  - A* largeur d'abord : WER=18.4%

Résultats sur 10 min FranceInfo (sans tel)

- Influence du « Fudge factor »
  - 1ère passe: Log(V)= Score _ac + (Score_ML * K_{11} + K_{12})
  - 2e passe : Log(V)= Score _ac + (Score_ML * K_{21} + K_{22})

1ère passe		2ème passe		WER
K_{11}	K_{12}	K_{21}	K_{22}	
8	-8	8	-8	19.5%
12	-8	14	-12	26.6%

Regroupement/Adaptation

- 1h France Inter
- Sans : 26.8%
- Avec regroupement et/adaptation 25.5%

Tâche Temps réel

Peu de gaussiennes large faisceau	⟺	plus de gaussiennes faisceau étroit

Taille du faisceau	Nb total de gaussiennes	Nb moyen de gaussiennes par état	WER
600	47728	8g	21.8
800	35796	6g	20.4
1000	29830	5g	20.7

Temps de calcul (1h audio) : Tel/notel : 1m, respi : 1m15, musique/parole : 1 m, H/F : 1m, MFCC : 1m, reste : calculs reco

Tâche Temps-réel

- Même système que précédemment sauf:
  - Julius est compile en mode “Fast”
  - l’algorithme de recherché de deuxième passe est sans modification.
  - le faisceau de recherche pour la première passe est plus fin
  - ni regroupement ni d’adaptation au locuteur
  - seulement 41762 gaussiennes au total pour les modèles « bande large »
  - seulement 23040 gaussiennes au total pour les modèles « bande étroite »