

Campagne d'Evaluation ETAPE: Segmentation et regroupement en locuteurs

Houssemeddine Khemiri
Maroua Thlithi
Dijana Petrovska-Delacrétaz
Gérard Chollet

Ateliers ETAPE

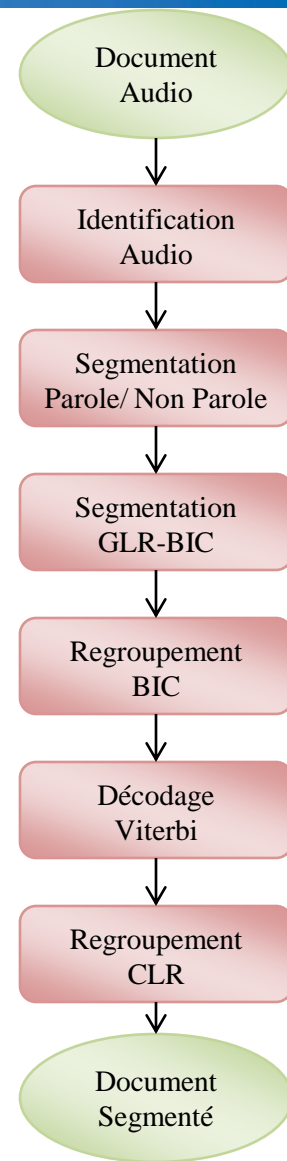
18 Juin 2012

Plan

- Architecture Générale du système proposé
- Identification Audio
- Segmentation Parole /Non Parole
- Segmentation et Regroupement en Locuteurs
- Résultats

Architecture Générale du système proposé

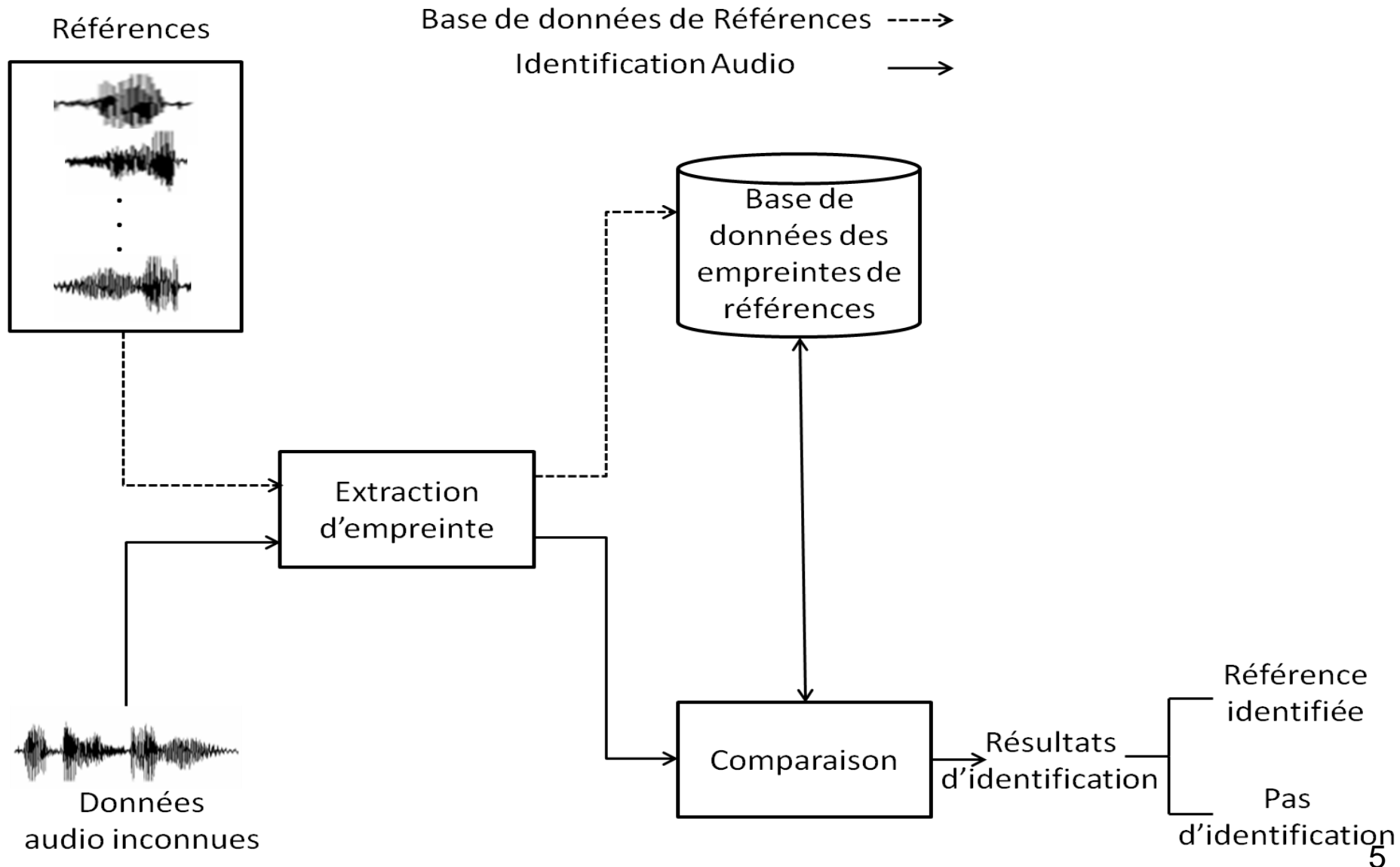
- ❖ Identification audio par extraction d'empreinte basée sur ALISP
- ❖ Segmentation parole/non parole: mise en compétition de deux modèles constitués de mélanges gaussiennes
- ❖ Système classique de segmentation et regroupement en locuteurs basé sur les critères GLR, BIC et CLR avec l'outil «LIUM_SpkDiarization»



Identification Audio (1)

- ❖ Audio Id : recherche des méta-données par requête audio
- ❖ Robuste aux dégradations du signal
 - compression, filtrage, bruit, «pitching»,...
- ❖ Contraintes de «scalabilité» et complexité
 - une large base de données de références
 - application temps réel
- ❖ Deux techniques
 - audio fingerprinting
 - audio watermarking
- ❖ Applications
 - structuration audio
 - surveillance du flux radio...

Audio Fingerprinting



Audio ID Basée sur ALISP

❖ ALISP (Automatic Language Independent Speech Processing)

- compression de la parole à très bas débit
- vérification de locuteur
- reconnaissance de la langue

❖ Principe général

- acquisition et modélisation des modèles HMM ALISP
- transcription ALISP des données audio
- recherche de similarité dans les transcriptions ALISP

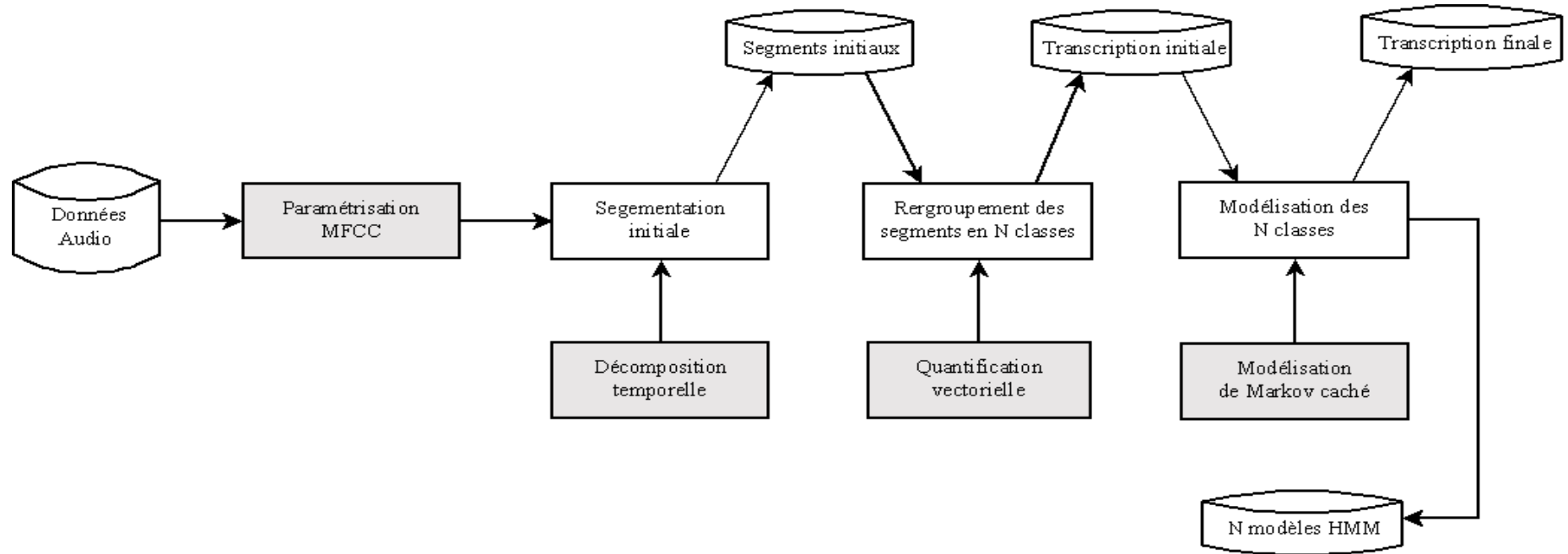
❖ Avantages

- pas de transcriptions textuelles pendant l'étape d'apprentissage
- déploiement sur de nouvelles données, les tâches, ou les langues
- méthode de recherche approchée

Application aux Evaluations ETAPE

- ❖ Origine des documents audio des bases d'apprentissage, développement et évaluation (TV/Radio Show) est identique
- ❖ Identifier les segments audio similaires entre les bases d'apprentissage et de développement et la base d'évaluation
- ❖ Principe général
 - acquisition et modélisation des modèles HMM ALISP à partir de la base d'apprentissage
 - transcription ALISP de la base des références et la base de test
 - rechercher de similarité dans les transcriptions ALISP

Acquisition des modèles ALISP



- Segmentation du flux audio en unités pseudo-phonétiques appelées unité ALISP
- Utilisation de la décomposition temporelle, quantification vectorielle et la modélisation HMM

Acquisition des modèles ALISP

❖ Décomposition temporelle

- décompose la matrice des MFCC en vecteurs cibles et des fonctions d'interpolation
- représente les parties stationnaires de l'évolution spectrale

❖ Quantification vectorielle

- une classification non-supervisée des segments initiaux
- la taille du codebook définit le nombre des unités ALISP

❖ Modélisation HMM

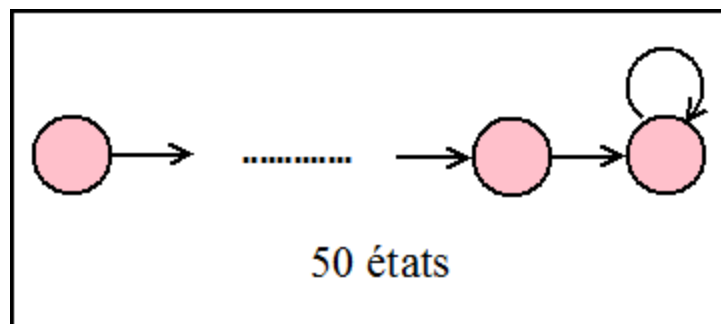
- initialisation des HMM avec les transcriptions initiales
- apprentissage sans contexte et en contexte des modèles HMM
- raffinement des modèles
- Incrémentation dynamique du nombre de gaussiennes

Recherche de similarité

- ❖ Transcription ALISP de l'ensemble des références
- ❖ Transcription ALISP du document audio à segmenter
- ❖ Chercher dans les transcriptions ALISP des références la séquence ALISP la plus proche à celle du flux radio
- ❖ Un segment audio est identifiée si la distance de Levenshtein entre sa transcription ALISP est celle de la référence est inférieure à un certain seuil

Segmentation Parole/Non Parole

- ❖ Etape indispensable dans la segmentation et le regroupement en locuteurs
- ❖ Mise en compétition de 2 modèles constitués de mélanges de 64 gaussiennes
- ❖ Une durée minimale d'une demi-seconde pour chaque segment reconnu a été imposée



Segmentation et Regroupement en locuteurs

❖ Segmentation GLR/BIC

- première passe: calcul d'une distance entre deux fenêtres adjacentes de taille 1.5s toutes les 0.16s
- deuxième passe: fusionner les segments voisins prononcés par le même locuteur

❖ Regroupement BIC

- une classification hiérarchique bottom-up basée sur la métrique BIC
- à chaque itération les deux clusters les plus proches sont fusionnés

❖ Décodage de Viterbi

- raffinement des frontières des segments obtenus
- chaque cluster (locuteur) est modélisé par un GMM à 8 composantes à matrices diagonales appris sur les segments du cluster

❖ Regroupement CLR

- normalisation des coefficients MFCC
- apprentissage d'un modèle UBM à 512 composantes gaussiennes
- adapter le modèle UBM à chaque cluster
- fusionner les cluster qui maximisent le NCLR

Résultats

Fichier	DER
BFMBFMStory	15.87
LCPCaVousRegarde	12.60
LCPEntreLesLignes2011-05-06-192800	17.31
LCPEntreLesLignes-2011-05-13-192800	18.48
LCPPilesEtFace	19.76
LCPTopQuestions-2011-05-18-000400	29.55
LCPTopQuestions-2011-05-25-213800	2.44
TV8LaPlaceDuVillage-2011-05-03-201300	22.27
TV8LaPlaceDuVillage-2011-05-12-172800	20.40
EST2BCFREFR-20100208-1000	13.75
EST2BCFREFR-20100208-1750	22.93
EST2BCFREFR-20101007-2152	27.43
EST2BCFREFR-20101014-2152	23.93
EST2BCFREFR-20101018-0910	8.26
EST2BCFREFR-20101024-2004	15.48
Global	16.23

Interprétations des résultats

- ❖ Pour LCPTopQuestions-2011-05-18-000400 le fichier «uem» contient une partie qui n'est pas prise en compte par la vérité terrain
- ❖ Temps de calcul nécessaire pour segmenter 60 secondes du signal audio est de 80 secondes avec une machine 3.00GHz Intel Core 2 Duo 4 Go de RAM
- ❖ La recherche des segments récurrents avec le système d'identification audio est exhaustive (pour 60 secondes su signal 30 secondes de traitement)