# CAMPAGNE D'ÉVALUATION ESTER 2005 : TRANSCRIPTION ET SEGMENTATION EN LOCUTEURS

*Paul Deléglise, Yannick Estève, Bruno Jacob, Teva Merlin, Sylvain Meignier*

LIUM-CNRS FRE 2730
Institut d'Informatique Claude Chappe
Université du Maine, Avenue Laënnec
72085 LE MANS CEDEX 9

## 1. INTRODUCTION

Le LIUM participe aux tâches de :

- transcription,
- segmentation en locuteurs (SRL),
- et suivi d'événements sonores (SES).

## 2. TRANSCRIPTION

### 2.1. Système

Le système de transcription du LIUM repose en grande partie sur le toolkit SPHINX [1] :

- version 2 pour les modules d'apprentissage,
- version 3.3 pour le décodeur rapide,
- version 3.0 pour certains utilitaires comme l'alignement temporel.

Des modifications ont été apportées à SPHINX, en particulier pour :

- l'utilisation des accents dans les mots du vocabulaire,
- les techniques d'adaptation MAP, MLLR et CMLLR,
- l'apprentissage MMIE,
- le re-scoring de lattice en 4-grammes.

### 2.2. Paramètres acoustiques

13 MFCC (C0 à C12) sont calculés dans une fenêtre de 10 ms avec un décalage de 25,6 ms. Les delta et delta-delta sont calculés au vol par les outils SPHINX. Il en est de même pour la normalisation du canal (moyenne et écart-type) appliquée à chaque segment. Deux jeux de paramètres sont calculés : un en bande large (BL) et un en bande étroite (BE, filtrage de la bande passante entre 440Hz et 3500 Hz).

### 2.3. Segmentation automatique

L'objectif de la segmentation est de découper le signal sonore afin d'obtenir les segments de parole à transcrire. Deux systèmes de segmentation sont utilisés en parallèle, leurs sorties étant ensuite fusionnées.

La segmentation est utilisée lors de la transcription d'un test mais aussi lors de l'apprentissage des modèles acoustiques par les méthodes SAT et MMIE.

#### *2.3.1. Segmentation acoustique*

Ce premier système repose sur des décodages DAP à partir desquels :

- Les segments de musique seront rejetés.
- Les frontières de segments seront placées sur les fillers (inspiration, musique, silence, bruit).

Ce premier système repose sur deux décodages DAP dépendant de la bande (BL et BE). Le décodage en BL est utilisé pour corriger les erreurs de détection en BE des longs phonèmes "ss" qui sont souvent reconnus comme des inspirations.

Les modèles utilisés sont indépendants du contexte, à 512 gaussiennes pour chaque état. Les DAP sont appliqués dans une fenêtre glissante de 100 secondes avec un décalage de 80 secondes (recouvrement de 20 secondes entre deux fenêtres successives). L'application des DAP dans une fenêtre glissante est due à une limitation de SPHINX ne permettant pas de décoder une heure de parole. La construction d'une segmentation phonétique consiste à fusionner les $n$ segments de 100 secondes en recherchant les zones communes. Pour chaque segmentation phonétique (BE et BL), les phonèmes consécutifs sont regroupés. La segmentation contient alors des zones de phonèmes et des fillers. Les deux segmentations phonétiques sont ensuite alignées en utilisant les inspirations. Les phonèmes "ss" du DAP BL auxquels correspond un filler "inspiration" dans BE sont remplacés

par "ss" dans BE. Par la suite, seule la segmentation BE fait l'objet de traitements.

Un jeu d'heuristiques est appliqué sur la segmentation BE pour supprimer les fillers de très courte durée. Les zones de phonèmes d'une durée supérieure à 0,2 seconde sont alors identifiées comme appartenant à l'une des classes suivantes :

- 0F : parole en BL prononcée par une femme.
- 0M : parole en BL prononcée par un homme.
- 2F : parole en BE prononcée par une femme.
- 2M : parole en BE prononcée par un homme.
- F6 : musique pure.

#### *2.3.2. Segmentation en locuteurs*

La segmentation en locuteurs utilise des outils développés par le LIUM. La méthode repose sur un système de classification hiérarchique utilisant la métrique BIC aussi bien comme critère de regroupement que comme critère de sélection des classes à regrouper. Le système se divise en trois étapes:

- segmentation du signal,
- regroupement BIC des segments contigus,
- regroupement hiérarchique.

Les paramètres sont calculés au moyen du toolkit SPRO 4.0. 12 MFCC et la log-énergie sont calculés dans une fenêtre glissante de 20 ms avec un décalage de 10 ms. Le jeu de paramètres SPHINX donne une segmentation en locuteurs moins précise (de l'ordre de 0,5 point sur le corpus de développement de la phase 1).

La première étape consiste à segmenter le signal en segments d'une taille supérieure à 2,5 secondes. Le principe repose sur le calcul pour chaque trame d'un GLR (Generalized Likelihood Ratio [2]) dans une fenêtre de 5 secondes centrée sur la trame en utilisant des mono-gaussiennes à matrice de covariance pleine. A partir de la courbe des scores GLR, les frontières des segments sont déterminées parmi les maxima locaux de telle façon que la taille d'un segment soit supérieure à 2,5 secondes. Une frontière candidate est rejetée s'il existe un score supérieur dans les 2,5 secondes de signal suivantes.

A partir de cette pré-segmentation, un regroupement BIC est appliqué sur les segments consécutifs. Pour tout $i \in [0..n-1]$, le segment $s_i$ est regroupé avec le segment $s_{i+1}$ si $BIC(s_i, s_{i+1}) < 0$. Avec :

$$BIC(s_i, s_{i+1}) = GLR(s_i, s_{i+1}) - \lambda \times P \times log(n_i + n_{i+1})$$

où $n_i$ est le nombre de trames du segment $s_i$ ($n_{i+1}$ respectivement pour $s_{i+1}$), $P$ est la complexité du modèle et $\lambda = 2$ un facteur de normalisation.

À la suite du regroupement BIC, une classification hiérarchique est appliquée. On notera que la mesure et le critère d'arrêt reposent sur la métrique BIC locale [3]. Le facteur de pénalité utilise le nombre de trames des deux segments regroupés plutôt que le nombre de trames du signal à segmenter classiquement employé [2, 4, 5].

Un décodage par Viterbi a été ajouté en post-traitement pour affiner les frontières qui ont été obtenues lors de la première étape (segmentation). Un locuteur est modélisé par un état contenant un GMM à 8 composantes à matrices diagonales appris sur les segments du locuteur au moyen de l'algorithme EM-ML. Une pénalité de changement de modèle (changement d'état) a été fixée expérimentalement à 250.

Note : le système actuel n'utilise :

- ni une détection des zones Parole / Silence / Musique,
- ni une détection en bande large et bande étroite,
- ni une détection des genres des locuteurs.

#### *2.3.3. Fusion de la segmentation acoustique et de la segmentation en locuteurs*

Une étiquette de locuteur issue de la segmentation en locuteurs est affectée à chaque zone de phonèmes issue de la segmentation acoustique. Chaque zone de phonèmes est identifiée en terme de bande et de genre lors de la segmentation acoustique. Le type de bande et le genre de chaque segment sont obtenus par vote pondéré par la durée reposant sur les étiquettes des zones de phonèmes du segment.

Puis les segments de locuteurs d'une durée supérieure à 20 secondes sont découpés en sous-segments de taille inférieure à 20 secondes (la frontière étant placée sur le filler le plus long).

### 2.4. Apprentissage des modèles acoustiques

La méthode d'apprentissage repose sur l'algorithme Baum-Welch (BW) dans les différentes phases. Des modèles dépendants du contexte de la bande (large/étroite) et du genre sont complétés par un modèle SAT.

#### *2.4.1. Topologie des modèles acoustiques*

Un modèle acoustique contient $N$ phonèmes (identiques à ceux utilisés par le LIA) et $M$ fillers ("euh", inspiration, silence, musique). Les HMM gauche-droite de chaque phonème sont composés de 5 états émetteurs et d'un état final non-émetteur (topologie issue de contraintes du décodeur rapide SPHINX 3.3).

*2.4.2. Pré-traitement des transcriptions de référence*

Cette phase repose sur la segmentation manuelle provenant des transcriptions fournies avec le corpus. La segmentation manuelle a été ré-étiquetée de manière semi-automatique.

Les modèles utilisés lors de ce ré-étiquetage ont été obtenus à partir de la segmentation manuelle réalisée au LIA par Corinne Fredouille (un sous-ensemble de fichiers de la phase I). Quatre modèles dépendants de la bande et du genre ont été appris et complétés d'un modèle de musique. À chaque étiquette correspond un HMM à un état avec 512 gaussiennes. Les segments donnés dans les transcriptions ont été étiquetés automatiquement. Les zones de désaccord ont été vérifiées manuellement pour le genre et automatiquement (par un vote majoritaire) pour le type de bande.

*2.4.3. Phases d'apprentissage*

- Phase 1 : les modèles acoustiques indépendants du contexte sont appris :
  - un modèle sur les segments BL,
  - un modèle sur l'ensemble des segments mais avec des paramètres BE.

  Chaque modèle est indépendant du contexte, avec des états contenant une gaussienne à matrice de covariance diagonale.

  Le modèle en bande large (respectivement en bande étroite) à 512 gaussiennes utilisé dans le module de segmentation acoustique est obtenu par *split* des gaussiennes du modèle de la phase 1 en bande étroite (respectivement en bande large).

- Phase 2 : Les modèles acoustiques dépendants du contexte à une gaussienne sont appris par BW.

- Phase 3 : Les modèles acoustiques dépendants du contexte à états partagés sont obtenus par *clustering* des gaussiennes des modèles de la phase précédente. Les modèles résultants contiennent environ 5500 états.

- Phase 4 : Des modèles à 22 gaussiennes sont obtenus par *split* des modèles de la phase 3.

- Phase 5 : Le modèle bande étroite de la phase précédente est adapté aux segments étiquetés bande étroite. L'adaptation repose sur la méthode MAP[6], où les moyennes, les covariances et les poids sont adaptés.

- Phase 6 : En accord avec les étiquettes de la segmentation, le modèle en bande large de la phase 4 et le modèle en bande étroite de la phase 5 sont spécialisés (par MAP) suivant le genre.

- Phase 7 : Des modèles SAT (*Speaker adaptive training*) sont appris. Cette phase utilise la segmentation automatique (*cf.* § 2.3.2) au lieu d'utiliser la segmentation établie en semi-automatique avec des corrections manuelles.

  Pour chaque groupe de segments dépendant du genre et du type de bande, une transformation CMLLR diagonale est calculée pour chaque segment. Après l'application de ces transformations, de nouveaux modèles sont obtenus en appliquant les phases d'apprentissage 1 à 6. Une seule itération SAT est appliquée.

- Phase 8 : Deux jeux de modèles MMIE ont été envisagés, ils sont issus soit de la phase 6, soit des modèles SAT de la phase 7. Deux itérations MMIE sont appliquées en utilisant les résultats du décodage des données d'apprentissage. Le décodage fournit un lattice généré par le modèle acoustique de la phase 6 avant SAT et un modèle linguistique tri-grammes utilisant les données des transcriptions et les données du journal *Le Monde*. Ce lattice est transformé en un réseau de confusion qui est utilisé lors de l'apprentissage BW discriminant.

*2.4.4. Utilisation des données de la phase III d'ESTER*

Dans une première phase de développement, seules 73 heures issues des données de la phase I et II ont été utilisées lors de l'apprentissage. Dernièrement environ 94 heures ont été transcrites automatiquement et incorporées aux données d'apprentissage. L'ajout des données a été réalisé en trois étapes. Environ 30 heures ont été ajoutées à chaque étape. Les données proviennent des différentes radios (sauf RTM) du mois de décembre 2003 à mars 2004 et du mois de septembre 2004.

*2.4.5. Temps de calcul*

L'exécution des phases d'apprentissage 1 à 6 prend 48 heures en utilisant 4 machines (Pentium 4, 3GHz, 512Mo). Les figures 1, 2 et 3 résument le processus d'apprentissage des modèles acoustiques.

### 2.5. Apprentissage des modèles linguistiques

*2.5.1. Lexique*

- 65530 mots différents complétés des symboles début de phrase, fin de phrase et mot inconnu (<s>, </s>)
- Mots des transcriptions ESTER avec vérification par ispell et vérification manuelle (env. 30K mots).
- Mots les plus fréquents apparaissant dans *Le Monde* de 1987 à 2003.

#### 2.5.2. Phonétisation

Nous utilisons 107227 variantes phonétiques issues de BDLEX [7]. Pour les mots non trouvés dans BDLEX, nous utilisons le logiciel de phonétisation automatique fourni par le LIA : lia_phon [8].

#### 2.5.3. Modèle trigramme

Les modèles de langage sont appris au moyen du toolkit de SRI [9].

- 65533 unigrammes, 18,4 M bigrammes, 25,7 M trigrammes
- Le modèle est une combinaison de 3 modèles de langage : Le Monde de 1987 à 2002 (coefficient 0,49), Le Monde 2003 (coefficient 0,19) et les transcriptions ESTER (89h : coefficient 0,32).
- Cutoffs : pas pour les unigrammes et les bigrammes, cutoff de 2 pour les trigrammes.
- Le discounting utilise la méthode Kneser-Ney modifiée [10].

#### 2.5.4. Modèle quadrigramme

- 65533 unigrammes, 18,4 M bigrammes, 22,2 M trigrammes, 19,7 M quadrigrammes
- Le modèle est une combinaison de 3 modèles de langage : Le Monde de 1987 à 2002 (coefficient 0,49), Le monde 2003 (coefficient 0,19) et les transcriptions ESTER (89h : coefficient 0,32).
- Cutoffs : identique au modèle trigramme et un cutoff de 2 pour les quadrigrammes.
- Le discounting repose sur la méthode Kneser-Ney modifiée.

### 2.6. Test

Les résultats sont résumés dans le tableau 1. Le décodage repose sur le décodeur rapide de SPHINX (version 3.3). Différentes options de décodage ont été envisagées. Le meilleur résultat, 25,3%, est obtenu avec deux passes de décodage (modèles provenant des phases d'apprentissage 6 et 7) et d'un re-scoring du lattice avec un modèle 4-grammes. Sur le corpus de développement, nous avons obtenu de meilleurs résultats avec le modèle MMIE (phase 8.2). Les résultats obtenus après ajout des données de la phase III sont présentés dans le tableau 2.

| système | passe | WER (%) |
|---|---|---|
| MAP+SAT | 2 | 26,2 |
| MAP+SAT | 3 | 25,3 |
| MAP+MMIE+SAT | 2 | 26,6 |
| MAP+MMIE+SAT | 3 | 25,6 |

**Table 1**. Résultats obtenus sur le test

| système | passe | WER (%) |
|---|---|---|
| MAP+SAT | 2 | 25,7 |
| MAP+SAT | 3 | 24,8 |
| MAP+MMIE+SAT | 2 | 25,9 |
| MAP+MMIE+SAT | 3 | 25,0 |

**Table 2**. Résultats obtenus sur le test : apprentissage utilisant les données de la phase III

#### 2.6.1. Passe 1

Nous avons utilisé les modèles contextuels à états partagés dépendants du genre et du type de bande, issus de la phase 6 d'apprentissage. Le modèle de langage utilisé est le modèle tri-grammes.

#### 2.6.2. Passe 2

Après une adaptation CMLLR à partir du décodage obtenu avec les modèles de la phase 6, un deuxième décodage est mis en œuvre en utilisant soit :

- Le modèle SAT sans MMIE (phase 7).
- Le modèle SAT avec MMIE (phase 8.2).

Le modèle de langage est toujours le modèle trigrammes utilisé dans la passe 1. Un lattice est généré à partir du graphe de décodage.

#### 2.6.3. Re-scoring de lattice

Les lattices sont re-scorés en utilisant un modèle de langage 4-grammes.

## 3. SUIVI D'ÉVÉNEMENTS SONORES

Cette tâche n'a pas fait l'objet de développement spécifique. Dans un premier temps, des segments parole et musique ont été générés à partir de la transcription :

- les segments de parole correspondent aux mots de la transcription,
- les segments de musique correspondent aux fillers musique.

Les segments étiquetés musique lors de la phase de segmentation acoustique (§ 2.3.1) sont ajoutés à ces segments issus de la transcription.

Dans un second temps, les segments adjacents de même nature sont groupés s'ils sont distants de moins de 0,5 seconde.

## 4. SEGMENTATION EN LOCUTEURS

### 4.1. Systèmes constratifs

Le système *constrate-002* de segmentation en locuteurs a été décrit au paragraphe (§ 2.3.2).

Un système rapide de segmentation a été fourni au LIA. Il est identique au système *constrate-002* à l'exception des points suivants :

- le décodage par Viterbi n'est pas appliqué,
- les modèles utilisés pour le calcul des distances BIC sont des mono-gaussiennes diagonales.

Le système *constrate-001* est une extension du système *constrate-002*. La segmentation *constrate-002* est filtrée en utilisant les informations issues de la tâche SES (§ 3). Les zones de musique sont supprimées de la segmentation en locuteurs. Les silences de moins d'une seconde coupant un segment de locuteur sont ignorés.

### 4.2. Système primaire

Le système primaire utilise les résultats du système *constrate-001*. Il repose sur une méthode d'identification du locuteur pour regrouper ou diviser les classes issues du système *constrate-001*. Les outils d'identification reposent sur le toolkit AMIRAL développé au LIA [11].

#### *4.2.1. Paramètres acoustiques*

Le signal est caractérisé par 12 MFCC et leur dérivé calculés dans une fenêtre de 20ms avec un décalage de 10ms. Une bi-gaussienne calculée sur l'énergie du signal permet de supprimer environ 20 à 30 % des trames de moindre énergie. Puis les paramètres sont centrés et réduits.

#### *4.2.2. Modèle UBM*

Des modèles du mondes dépendant du genre (homme, femme) et du canal (téléphone, studio) sont appris à partir d'environ une heure de parole. Ces modèles sont composés d'un GMM à 128 composantes diagonales. Les quatre modèles sont fusionnés pour obtenir un UBM à 512 composantes [12].

| système | WER (%) |
|---|---|
| *contraste-002* | 19.14 |
| *contraste-001* | 17.94 |
| primaire | 17.38 |

**Table 3**. Résultats obtenus sur le test pour la tâche SRL.

#### *4.2.3. Modèle de locuteurs*

Un ensemble de 271 locuteurs a été sélectionné parmi les locuteurs du corpus d'apprentissage. Chaque modèle de locuteur est obtenu par adaptation de la moyenne de l'UBM à deux minutes de parole (méthode MAP).

#### *4.2.4. Identification en regroupement*

Chaque segment et chaque classe issus de la segmentation en locuteurs du système *constrate-001* sont identifiés en utilisant le critère du maximum de vraisemblance. La vraisemblance est calculée en utilisant les 20 top gaussiennes [12]. La pureté des classes, pondérée par la longueur des segments, est utilisée comme critère de fusion ou de division des classes. Les scores tels que la vraisemblance, le rapport de vraisemblance ou les scores TNorm semblent être moins pertinent sur le corpus de développement.

Pour une pureté de classe supérieure à 80%, l'identité du locuteur est utilisée comme identifiant de la classe. Il est possible que deux classes reçoivent la même identité de locuteur, *ie* ces classes sont fusionnées.

Pour une pureté de classe inférieure à 10%, l'étiquette des segments de la classe est composée de l'identifiant donné lors de la segmentation et de l'identité de locuteur du segment (qui peut être différente de l'identité de la classe). La classe est donc divisée suivant l'identité des locuteurs attribués aux segments.

Pour une pureté comprise entre 10 et 80%, l'étiquette donnée lors de la segmentation en locuteurs est conservée.

Ce système est le système primaire du LIUM pour la tâche SRL.

### 4.3. Résultats et temps de calcul

Les résultats sont donnés dans le tableau 3. Les temps de calcul n'ont pas été mesurés avec exactitude. Sur un Pentium 4 à 3GHz, nous obtenons :

- *constrate-002* : environ 0,2 RT.
- *constrate-001* et SES : environ 0,2 RT, le filtrage étant négligeable.
- *constrate-002* : environ 1 RT pour la phase d'identification du locuteur.

## 5. REFERENCES


[1] K. Lee, H. Hon, R. Reddy, An overview of the SPHINX speech recognition system, IEEE Transactions on Acoustics, Speech, and Signal Processing 38 (1) (1990) 35–45.

[2] M.-H. Siu, G. Yu, H. Gish, Segregation of speakers for speech recognition and speaker identification, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 91), Toronto, Canada, 1991, pp. 873–876.

[3] C. Barras, X. Zhu, S. Meignier, J.-L. Gauvain, Improving speaker diarization, in: DARPA RT04 Fall, Palisades, NY, USA, 2004.

[4] S. Chen, P. Gopalakrishnan, Speaker, environment and channel change detection and clustering via the bayesian information criterion, in: DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, USA, 1998.

[5] Y. Moh, P. Nguyen, J.-C. Junqua, Towards domain independent speaker clustering, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2003), Hong Kong, 2003.

[6] J.-L. Gauvain, C. H. Lee, Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains, IEEE Transactions on Speech and Audio Processing 22 (1994) 291–298.

[7] G. Pérennou, M. D. Calmès, BDLEX lexical data and knowledge base of spoken and written French, Edinburgh, Ecosse, 1987, pp. 393–396.

[8] F. Béchet, LIA_PHON un système complet de phonétisation de texte, Traitement Automatique des Langues 42 (1) (2001) 47–68.

[9] A. Stolcke, SRILM-an extensible language modeling toolkit, in: Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2002), Vol. 2, Denver, Colorado, USA, 2002, pp. 901–904.

[10] J. Chen, S. F.and Goodman, An empirical study of smoothing techniques for language modeling, in: A. Joshi, M. Palmer (Eds.), Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, Morgan Kaufmann Publishers, San Francisco, USA, 1996, pp. 310–318.

[11] T. Merlin, AMIRAL, une plateforme générique pour la reconnaissance automatique du locuteur — de l'authentification à l'indexation, Ph.D. thesis, Université d'Avignon et des Pays de Vaucluse (Novembre 2004).

[12] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted gaussian mixture models, Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3) (2000) 19–41.

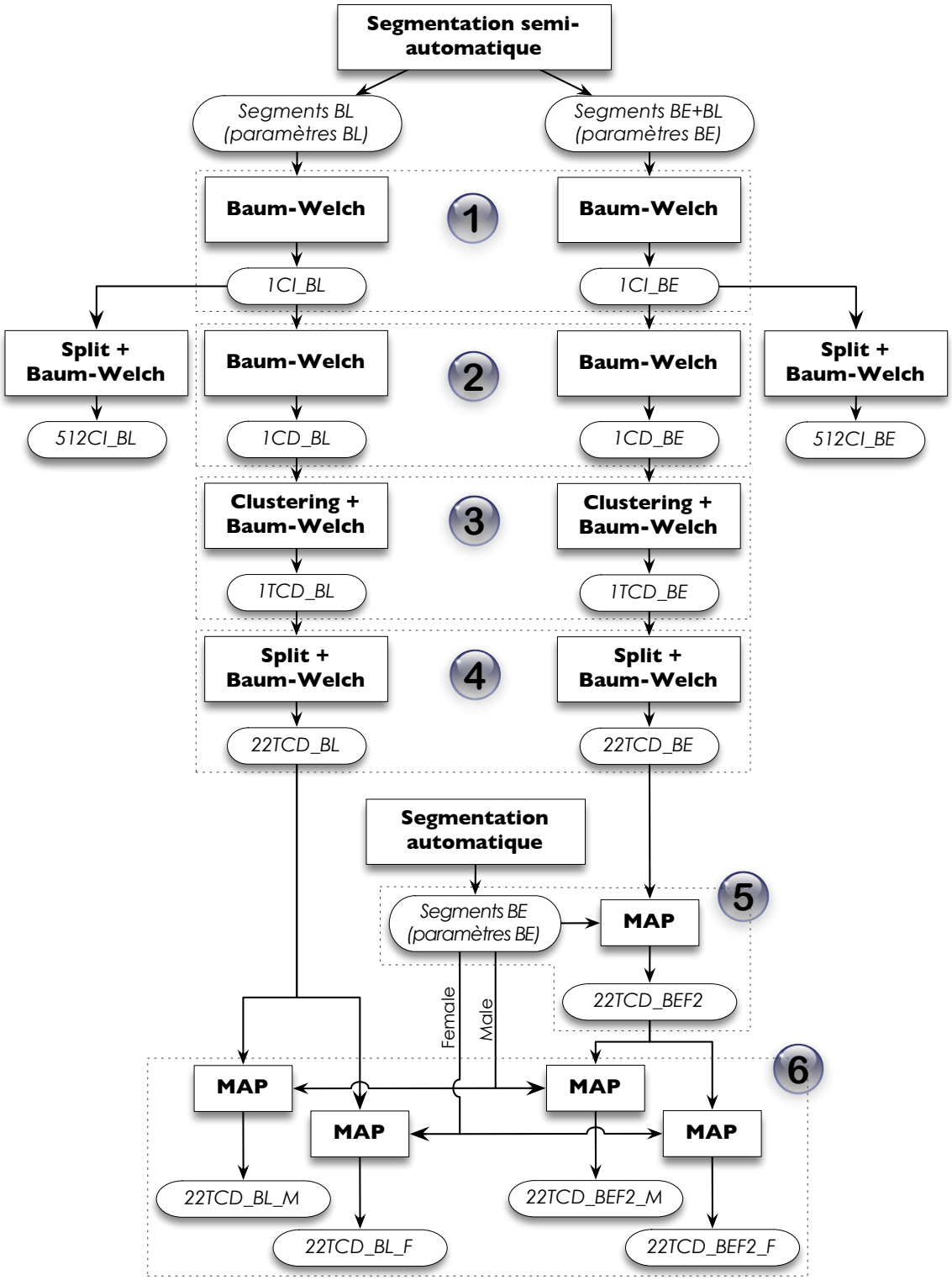


**Fig. 1.** *Apprentissage des modèles acoustiques*

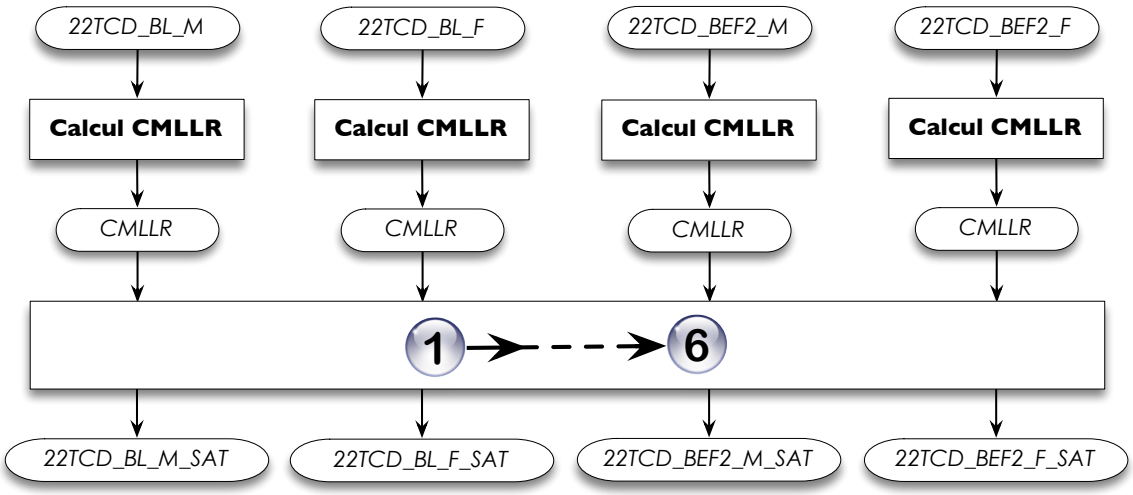


**Fig. 2**. *Apprentissage des modèles acoustiques – Phase 7 (SAT + CMLLR)*

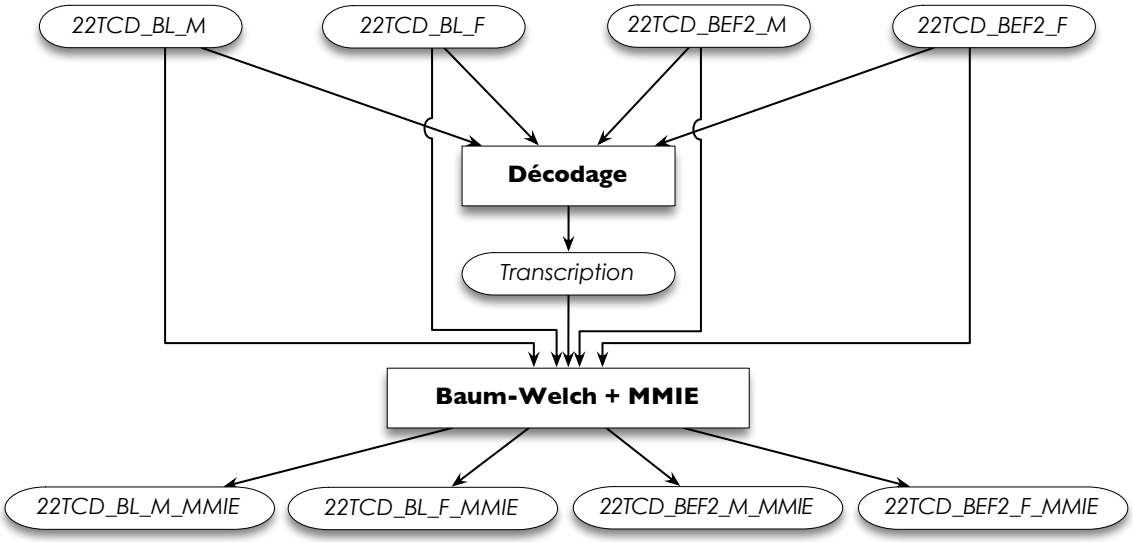


**Fig. 3**. *Apprentissage des modèles acoustiques – Phase 8 (MMIE)*