

Identification Automatique des Dialectes du Maghreb

**LACHACHI Nour-Eddine &
ADLA Abdelkader**

Université d'Oran 1 Ahmed Ben Bella

Résumé

Le travail de recherche que nous avons réalisé s'intègre dans le domaine l'Identification Automatique des Langues (IAL) et en particulier celle des dialectes du Maghreb au moyen de méthodes formelles de classification en utilisant l'information acoustique et l'information phonotactique. Dans cette étude, nous présentons une expérimentation sur des types de systèmes basés sur une modélisation phonotactique et une modélisation acoustique en utilisant un corpus que nous avons conçu nous-mêmes. Pour la première modélisation, l'étude est basée sur des systèmes de reconnaissance de phonèmes que nous appelons PRLM et PRLM Parallèle. Quant à la deuxième, nous étudions des systèmes basés sur les Machines à Vecteurs Support et les Modèles de Mélanges de lois Gaussiens.

1. Introduction

Le but général du traitement de la parole est d'extraire automatiquement d'un signal numérique de la parole des informations de haut niveau, c'est-à-dire facilement interprétables par l'être humain.

Le signal de la parole contient plusieurs types d'information que ceux portant sur le message, comme l'identité de la langue ou de l'individu qui a prononcé le message. Cette problématique est le centre d'intérêt de la reconnaissance ou l'identification d'un dialecte, domaine qui connaît plusieurs sous-branches selon le contexte applicatif visé. Une formulation du problème, qui a fait l'objet d'une grande partie

des recherches plus récentes dans le domaine, est l'identification automatique des dialectes.

Une langue est définie comme un regroupement de dialectes partageant un vocabulaire similaire et ayant des systèmes phonologiques et grammaticaux similaires. Par exemple, dans le monde arabe on dénombre plusieurs dialectes parmi ces derniers nous nous intéressons aux dialectes du Maghreb.

Le but de l'identification automatique des dialectes est de reconnaître le dialecte parlé par un locuteur inconnu, parmi un ensemble fini de dialectes, pour des énoncés d'une durée limitée (entre 3 et 50 secondes habituellement).

L'identification automatique des dialectes peut permettre de confirmer ou d'infirmer les théories linguistiques portant sur les différences entre les langues. Ces théories reposent souvent sur un nombre limité d'expériences; certaines sont parfois plus basées sur des intuitions que sur des faits réels. Il est important de procéder à des expérimentations sur de plus larges bases de données, tâche rendue possible par le traitement automatique de la parole.

Pour identifier un dialecte, deux types d'information sont les plus fréquemment utilisées : 1) les informations phonétiques et phonotactiques (caractéristique des sons d'un dialecte et règles d'enchaînement de ces sons) et ; 2) les informations acoustiques qui sont est l'une des formes les plus primitives qui peuvent être obtenues durant le processus de paramétrisation à partir d'un signal de parole.

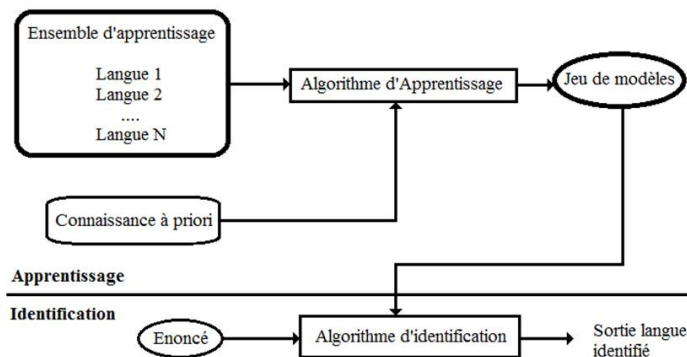


Fig. 2.1: Aspect générale d'un système IAL

2. Structure générale d'un système IAL

Le système général s'apparente à un système statistique classique de reconnaissance des formes utilisant un apprentissage supervisé.

La fonction d'identification se décompose en deux phases (cf. Fig. 2.1) :

- **Apprentissage:** des paramètres sont extraits pour les signaux de parole de chaque langue. Pour chaque source d'information prise en compte, un modèle spécifique à chaque langue est appris à partir de ces paramètres.
- **Reconnaissance:** les paramètres sont extraits pour un signal de parole d'une langue inconnue. La langue la plus vraisemblable est déterminée en fonction des modèles issus de la phase d'apprentissage. Se greffe un problème de fusion si plusieurs sources d'information sont modélisées engendrant plusieurs scores. Un comparatif des différentes approches est disponible dans [1].

3. Le corpus

Pour notre expérimentation, nous avons utilisé un corpus que nous avons conçu nous-mêmes contenant de la parole spontanée issue des films et des séries télévisées où les conditions acoustiques d'enregistrement ne sont pas similaires pour tous les locuteurs[2].

Pour la conception de ce corpus, nous avons divisé l'espace du Maghreb Arabe en 5 régions sur trois pays : Maroc, Algérie et Tunisie. Le résultat était le corpus Marocain, Oranais, Algérois, Constantinois et Tunisien. Le Tableau suivant définit notre corpus.

Table 3.1: Corpus des dialectes du Maghreb.

<i>Corpus</i>	<i>Durée (apprentissage.)</i>	<i>Durée (Test)</i>
<i>Marocain</i>	44.19 h	10.53 h
<i>Oranais</i>	40.73 h	09.15 h
<i>Algérois</i>	41.32 h	10.25 h
<i>Constantinois</i>	38.18h	08.40 h
<i>Tunisien</i>	42.73h	10.35 h

4. Modélisation phonotactique

Les systèmes donnant les meilleurs résultats sont ceux qui se basent principalement sur l'aspect phonologique et phonotactique. Ils sont les plus usuels dans la littérature. En général, ces systèmes possèdent un ou plusieurs systèmes de reconnaissance de phonèmes (souvent appelés décodeurs acoustico-phonétique) constitués de Modèles de Markov Cachés (**Hidden Markov Models- HMM**). Ces décodeurs acoustico-phonétiques sont le plus souvent spécifiques à l'inventaire phonétique d'une langue spécifique ; c'est pourquoi il faut en employer plusieurs en parallèle afin d'obtenir la meilleure couverture possible de l'ensemble des sons.

Le module de reconnaissance de phonèmes transforme le signal de parole en une suite d'éléments discrets formant des séquences de phonèmes. Ces séquences sont ensuite modélisées à l'aide de modèles probabilistes *N-grammes*. Ainsi, les enchaînements (de 2 ou 3 phonèmes) les plus caractéristiques des langues sont retrouvés.

4.1 Système PRLM

L'idée de base dans ce système d'identification PRLM (cf. Fig. 4.1) est de modéliser une séquence de phonèmes en utilisant un modèle du langage. La séquence de phonèmes est produite par un reconnaiseur de phonèmes constitué d'un ensemble de phonèmes unifiés qui couvre la majorité des phonèmes de tous les langages cibles et est appris à partir d'un corpus conçu d'un mélange de langages. Occasionnellement, un reconnaiseur de phonèmes d'une langue quelconque peut faire l'objet d'un décodeur unifié. Seulement dans ce cas-là, nous ne serons pas sûres que tous les phonèmes des autres langages puissent être correctement reconnus.

4.2 Système PRLM parallèle

L'une des issues clés dans le système PRLM est que certains phonèmes ne peuvent être reconnus quand il n'y a pas une part de modèles de phonèmes dans un reconnaiseur de phonème. Il est vrai que la séquence peut encore être modélisée, mais l'information clé directement associée avec celle des phonèmes spéciaux est inexistante. De là, les systèmes PPRLM (**P**arallel **P**hone **R**ecognition **L**anguage **M**odel) ont été introduits pour palier à ces problèmes. Comme le montre la figure 4.2, de multiples reconnaiseurs de phonèmes sont déployés et chacun d'eux est suivi par un ensemble complet de modèles de langages [1]. La mesure de confiance la plus haute est choisie comme sortie qui vérifie explicitement le reconnaiseur de phonème le plus approprié.

4.3 Expérimentation

Pour notre expérimentation nous avons utilisé les reconnaissseurs de phonèmes Français, Anglais, Allemand, Japonais, Hindou, Mandarin, et Espagnol à partir d'un kit développé par l'université de technologie de Brno. Ces reconnaissseurs sont appris à partir de la base de données multi-langages OGI [3] utilisant une approche hybride basée sur les réseaux de neurones et le décodage de Vitterbi sans modèle de langage [4].

Nous avons mené deux expérimentations pour tenter de classifier les cinq dialectes en utilisant d'abord le système PRLM puis le système PRLM Parallèle. Les données de chaque dialecte sont décodées en utilisant les reconnaissseurs phonétiques tels qu'ils sont définis en figure 4.2 pour générer des séquences de phonèmes 3-grammes. Les perplexités de ces séquences sont calculées afin de permettre la reconnaissance des dialectes adéquats. Dans les reconnaissseurs phonétiques, chaque phonème est modélisé par un HMM à 3 états et, pour chaque état nous avons utilisé une mixture gaussienne d'ordre 6.

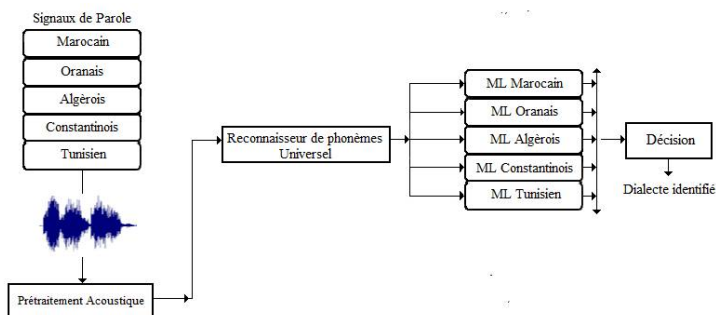


Fig.4.1: Système IAL basé PRLM.

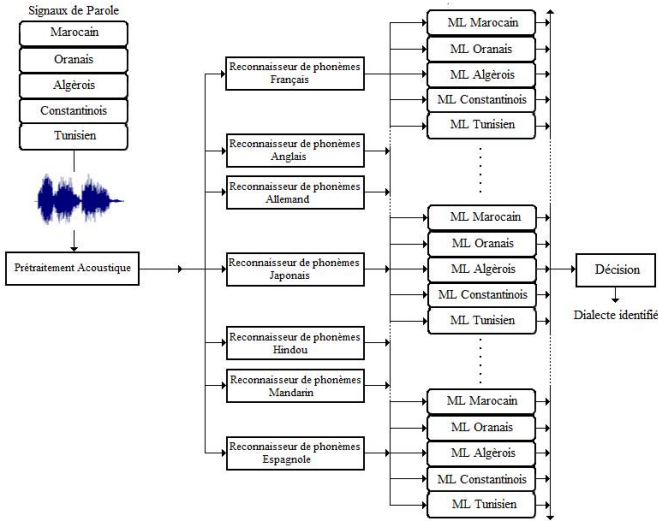


Fig.4.2: Système IAL basé PRLM Parallèle.

4.3.1 Paramétrisation

Nous avons utilisé des trames de 10 secondes pour l'apprentissage et le test. Ensuite, nous avons extrait 12 coefficients MFCC de dimension 39 dérivés de 20 bancs de filtre. Chaque caractéristique est extraite sur des intervalles de 10 millisecondes utilisant un fenêtrage de Hamming de 30 millisecondes sur une bande de parole limitée entre 300 Hz et 3400 Hz.

4.3.2 Résultat de l'expérimentation

<i>Dialecte</i>	Reconnaisseur de phonèmes			
	<i>FRE</i>	<i>ENG</i>	<i>GER</i>	<i>JAP</i>
Marocain	20.83	17.32	11.54	21.19
Oranais	10.56	14.27	08.94	15.15
Algérois	15.34	12.41	10.17	17.73

Constantinois	18.07	15.28	09.39	19.57
Tunisien	21.14	19.73	14.23	23.19
Reconnaisseur de phonèmes				
<i>Dialecte</i>	<i>HIN</i>	<i>MAN</i>	<i>ESP</i>	<i>Univ</i>
Marocain	10.03	13.93	18.32	11.32
Oranais	6.73	10.26	12.13	07.24
Algérois	10.37	13.11	14.19	09.78
Constantinois	10.13	12.61	12.55	10.21
Tunisien	12.66	15.95	20.65	13.04

5. Modélisation Acoustique

5.1 Module de réduction

Les systèmes acoustiques utilisés se basent sur un module de réduction de données. Nous présentons deux approches dont le but est d'éliminer les caractéristiques de paroles non adéquates aux dialectes spécifiques. Ce module de réduction est greffé à deux systèmes d'identification automatique des langues baseline qui fera naître une version améliorée des deux systèmes.

Le module de réduction est basée sur une équivalence entre la formulation de la plus petite boule englobante (Minimal Enclosing Ball - MEB) et la formulation de la norme L2-SVM avec un développement spécial. Deux algorithmes issus des deux approches multi-classe ont été développés, une-contre-une et une-contre-toutes[5]. Le but des deux approches est d'adopter la formulation des Machines à Vecteurs Supports Multi-classes et la formulation de la plus petite boule englobante pour réduire les données par élimination de toutes celles qui sont en dehors de cette boule.

L'idée clé de cette méthode de réduction est d'utiliser un espace caractéristique de données S qui est ensuite décomposé en un ensemble de partitions S_p où nous procédons à une

réduction pour chaque partition. La décomposition de S en partitions S_p est réalisée par un algorithme de partitionnement. Pour notre étude, nous avons utilisé les algorithmes k -plus proches voisins et C-Moyennes Floues.

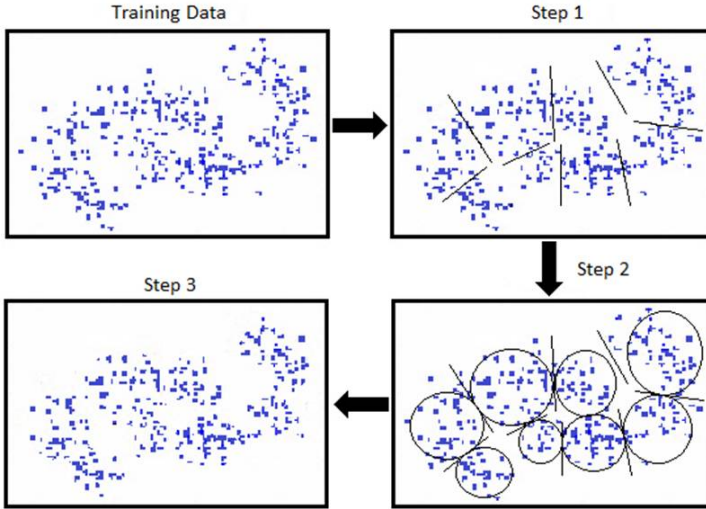


Fig. 5.1: Visualisation du processus d'apprentissage. Réduction globale des partitions en la plus petite boule englobante (MEB).

5.2 Paramétrisation pour les deux systèmes

Nous avons utilisé des trames de 10 secondes pour l'apprentissage et le test. Ensuite, nous avons extrait 12 coefficients MFCC de dimension 39 dérivés de 20 bancs de filtre. Chaque caractéristique est extraite sur des intervalles de 10 millisecondes utilisant un fenêtrage de Hamming de 30 millisecondes sur une bande de parole limitée entre 300 Hz et 3400 Hz. En se basant sur les coefficients MFCCs, nous avons calculé 12 coefficients SDC.

L'apprentissage est basé sur 512 Modèles de Mélanges de lois Gaussiennes (GMMs) avec des matrices de covariance

diagonales pour chacun des cinq dialectes en utilisant l'algorithme EM (Expectation Maximisation). Le noyau utilisé pour les deux approches est le RBF (Radial Basis Function) gaussien avec une valeur fixe de $\sigma = 0.50$.

5.3 Système d'Identification Automatique de Dialectes (IAD) basé sur GMM

Dans ce type de systèmes, nous considérons que chaque vecteur suit hypothétiquement une loi de distribution dont la densité de probabilité est une somme pondérée de lois gaussiennes multidimensionnelles. Plusieurs variantes existent au niveau de l'apprentissage de modèles: le cas le plus classique consiste à apprendre un modèle par langue en utilisant les données d'apprentissages spécifiques à chaque langue [6].

Expérimentation

Nous avons réalisé une série de 3 expériences : la première a été faite sur le système d'identification de dialectes Baseline sans le module de réduction, la deuxième a été faite sur le système d'identification de dialectes utilisant la réduction des données selon l'approche multi-classe *une-contre-une* avec un partitionnement k-NN et la troisième a été faite sur le système d'identification de dialectes utilisant la réduction des données (cf. Fig. 5.2) selon l'approche multi-classe *une-contre-une* avec un partitionnement C-Moyenne Flou.

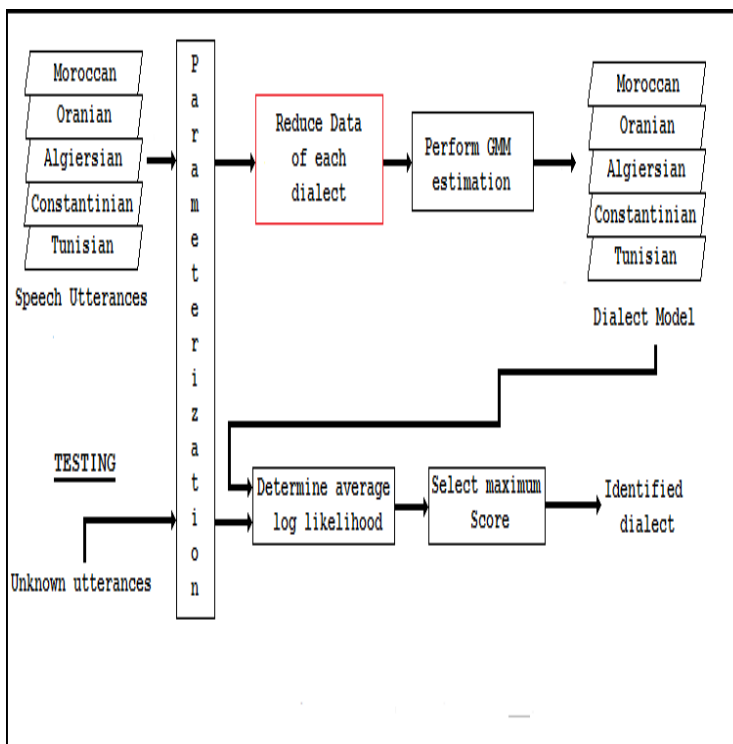


Fig. 5.2: Système d'Identification Automatique de Dialectes basé sur les Modèles de Mélanges de lois Gaussiennes (version améliorée).

Finalement, nous comparons la précision de l'identification de dialecte suivant les tables Table 5.1, Table 5.2 et Table 5.3 et nous donnons la précision en pourcentage pour les cinq dialectes selon différentes mixtures.

Table 5.1: Précision des cinq dialectes sur le système d'identification de dialecte Baseline.

Mixture	4	8	16	32
Marocain	58.07	62.21	64.01	65.56
Oranais	55.37	54.48	61.95	62.85
Algérois	58.03	61.02	61.97	62.48
Constantinois	58.43	65.23	65.53	67.83
Tunisien	66.34	70.07	70.67	73.84
Mixture	64	128	256	512
Marocain	66.93	67.33	68.01	69.33
Oranais	63.22	67.45	67.33	68.13
Algérois	64.07	64.57	65.13	65.63
Constantinois	67.73	69.68	70.96	70.11
Tunisien	75.13	75.76	77.05	77.85

Table 5.2: Précision des cinq dialectes sur le système d'identification

de dialectes basé sur l'approche multi-classe *une-contre-une* utilisant le partitionnement k -NN.

Mixture	4	8	16	32
Marocain	66.82	68.48	71.73	72.11
Oranais	58.93	60.24	64.13	65.72
Algérois	62.42	63.83	64.35	64.93
Constantinois	65.23	67.02	67.79	70.83
Tunisien	68.81	73.48	74.08	75.51
Mixture	64	128	256	512
Marocain	72.38	73.92	74.15	75.52
Oranais	69.48	70.16	71.07	71.96
Algérois	66.97	68.08	69.52	70.61
Constantinois	70.93	72.03	73.26	74.76
Tunisien	76.83	79.74	80.94	81.53

Table 5.3: Précision des cinq dialectes sur le système d'identification de dialectes GMM basé sur l'approche multi-classe *une-contre-une* utilisant le partitionnement C-Moyenne Flou.

<i>Mixture</i>	4	8	16	32
Marocain	70.11	70.89	74.42	75.33
Oranais	60.23	62.73	66.28	68.02
Algérois	63.27	66.64	67.13	68.83
Constantinois	65.15	71.43	72.16	72.84
Tunisien	71.97	74.06	76.16	79.23
<i>Mixture</i>	64	128	256	512
Marocain	76.03	76.57	77.84	78.12
Oranais	72.04	72.94	73.91	74.63
Algérois	69.72	70.61	71.09	71.81
Constantinois	73.63	74.19	76.21	77.13
Tunisien	81.17	82.84	84.61	85.67

5.3 Système d'Identification Automatique de Dialectes (IAD) basé sur GMM-UBM

Ce système se base sur un apprentissage d'un modèle dit «**du monde**» (Universal **B**ackground **M**odel - **UBM**) comprenant toutes les données de toutes les langues. Il s'agit par la suite d'adapter les paramètres de ce modèle afin de créer des modèles spécifiques à chaque langue[7].

Expérimentation

Nous avons réalisé une série de 2 expériences : la première a été faite sur le système d'identification de dialectes GMM-UBM Baseline sans le module de réduction, la deuxième a été faite sur le système d'identification de dialecte GMM-UBM utilisant la réduction des données (cf. Fig. 5.3) selon l'approche multi-classe *une-contre-une* avec un partitionnement C-Moyenne Flou.

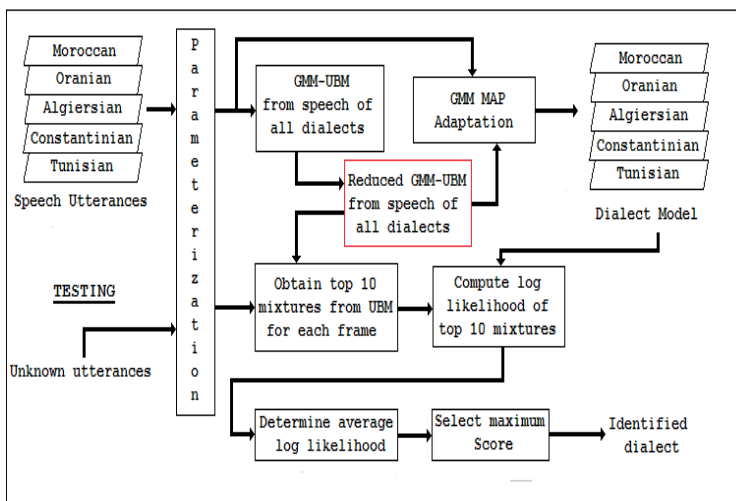


Fig. 5.3: Système d'Identification Automatique de Dialecte basé sur la réduction des Modèles de Mélanges de lois Gaussiennes et modèle du monde.

Finalement, nous comparons la précision de l'identification de dialectes suivant les tables Table 5.4 et Table 5.5. et nous donnons la précision en pourcentage pour les cinq dialectes selon différentes mixtures.

Table 5.4: La précision des cinq dialectes sur le système d'identification de dialecte basé sur les Modèles de Mélange de lois Gaussiennes et modèle du modèle du monde Baseline

<i>Mixture</i>	4	8	16	32
Marocain	63.11	65.88	70.08	68.87
Oranais	55.23	54.93	63.17	65.33
Algerois	59.13	62.73	64.83	64.98
Constantinois	60.34	67.27	67.93	69.01
Tunisien	68.25	72.33	72.91	76.16
<i>Mixture</i>	64	128	256	512
Marocain	70.21	70.33	71.43	71.67

Oranais	65.15	69.53	69.93	70.54
Algerois	66.94	67.12	67.78	67.85
Constantinois	69.41	71.83	72.16	72.18
Tunisien	76.22	80.91	81.39	81.95

Table 5.5: La précision des cinq dialectes sur le système d'identification de dialecte basé sur les Modèles de Mélanges de lois Gaussiennes et modèle du monde par réduction des données selon l'approche multi-classe *une-contre-une*.

<i>Mixture</i>	4	8	16	32
Marocain	68.32	71.54	76.19	78.03
Oranais	65.27	67.94	68.15	72.73
Algérois	62.41	66.17	68.73	69.37
Constantinois	63.28	70.39	72.57	73.13
Tunisien	71.73	76.23	79.19	81.66
<i>Mixture</i>	64	128	256	512
Marocain	78.93	82.32	82.55	83.92
Oranais	73.26	75.13	75.19	76.22
Algérois	72.11	72.19	74.27	77.67
Constantinois	74.61	76.55	77.38	78.58
Tunisien	82.95	83.65	85.85	86.07

Conclusion

Le traitement automatique des langues consiste à construire les modèles représentant les langues. La connaissance des propriétés linguistiques des langues permet d'orienter la recherche de paramètres et de modélisation. La quantité et qualité des données disponibles pour chaque langue sont également déterminantes: un système très complexe peut nécessiter des ressources coûteuses pour chaque langue ; ce qui rend leur généralisation à un grand nombre de langues difficiles. Par exemple, un système utilisant des modélisations acoustiques peut nécessiter un étiquetage phonétique manuel pour pouvoir estimer ses modèles, ou bien un système faisant

appel à des données lexicales doit disposer de corpus de très gros volumes pour pouvoir disposer d'une représentation exhaustive de la langue. En fonction des besoins, les contraintes sont plus ou moins fortes sur les données disponibles.

Les dialectes du Maghreb ne se distinguent pas de manière tranchée, et bien que les dialectes éloignés soient assez différents, il existe des continuums passant de l'un à l'autre. Ainsi les dialectes Marocain, Algérien et Tunisien sont trois langues assez proches, mais vis-à-vis des autres dialectes que regroupent les pays arabes, ces derniers sont très différents.

Dans cette étude, nous avons développé deux nouveaux systèmes d'identification de dialectes : l'un basé sur les Modèles de Mélange de lois Gaussiennes (GMM) et le second est fondé sur une combinaison d'un modèle du monde et des Modèles de Mélanges de lois Gaussiennes (UBM-GMM). Des détails ont été définis sur ces systèmes afin de montrer toute la différence entre un système baseline et les deux systèmes que nous avons proposés. L'expérimentation a montré que la réduction des données basée sur l'équivalence entre la formulation L2-SVM et la plus petite boule englobante (MEB) donne de meilleurs résultats où cette dernière joue vraiment un rôle de filtre sur les données ; ce qui explique les bons résultats que nous avons obtenus.

Références

- ZISSMAN, M.A. 1996. *Comparison of Four Approaches to Automatic Language Identification of Telephone Speech*. IEEE Transactions of Speech and Audio Processing, 4(1), 1996. 4.2, 5.2, 5.3, 8.4
- LACHACHI Nour-Eddine, ADLA Abdelkader. 2012. *Reduced universal background model for speech recognition and identification system*. Pattern Recognition, Lecture Notes in computer science, volume 7329, pp 303-312, Springer-Verlag.

- MUTHUSAMY, Y.K., COLE, R.A. & OSHIKA B.T. 1992. *The OGI Multi-Language Telephone Speech Corpus*. In Proceedings of ICSLP'92, 1992. 5.4
- MATEJKA, P. SCHWARZ, P., CERNOCKY, J. & CHYTIL, P. 2005. *Phonotactic Language Identification Using High Quality Phoneme Recognition*. In Proceedings of Eurospeech'05, 2005. 5.4, 10.3.2.2
- LACHACHI Nour-Eddine, ADLA Abdelkader. 2014. *Reduced Data Based Improved MEB/L2-SVM Equivalence*. Pattern Recognition, Lecture Notes in computer science, volume 8495, pp 1-10, Springer-Verlag.
- LACHACHI Nour-Eddine, ADLA Abdelkader 2016. *Two Approaches for Dialect Identification Based on L2-SVMs reduced to MEB problems*. International Journal of Computational Vision and Robotics, Volume 6, N° 1/2, pp. 1-18. Inderscience Publishers.
- LACHACHI Nour-Eddine, ADLA Abdelkader 2015. *GMM-Based Maghreb Dialect Identification System*. Journal of Information Processing System (JIPS), Volume 11, N° 01, pp. 22-33, Edition KIPS.