Application Forensique à la Reconnaissance Vocale du Locuteur

Ouassila Kenai, Mhania Guerti

الملخّــص

في هذا العمل، نهتم بالتّعرّف الصّوتيّ على المتكلمين باللّغة العربيّة الفصحى في الإطار الإجرامي (RVC)، وبالأخص IVC والمصادقة AVC في وضع مستقلّ عن النص؛ وذلك انطلاقا من الإشارات الصّوتيّة الصّادرة عن أولئك النّاطقين، الذين استخرجت معلومات تتعلّق بهويّاتهم، باستعمال التحليل الطيفي (MFCC)، وتقدير مع هذا الأخير نموذج قوي للمتكلم بما يسمح بالتّعرف عليه وتحديد هويته من قبل نموذج خليط غوسيان GMM. أمّا الآثار الصوتية فقد حلّلت واستخدمت مع GMM في مرحلة المقارنة، مطبقين المقاربة البايزيه LLR لتحديد صوت المجرم والمصادقة عليه.

أظهرت تجاربنا، أنّ نموذج دو 32 غوسان كفيل بالتّعرّف على المتكلّم المجرم لا سيما المكان ووسائل التّسجيل، وقد أعطت التّجارب المشار إليها نتائج جدّ مرضية لهذا النظام المنجز RVC، الذي بإمكانه أن يساعد المحكمة في اتّخاذ قرار لحل المشاكل الإجرامية.

الكلمات المفاتيح : التعرف الصوتي الإجرامي، التحديد الصوتي الإجرامي، المصادقة الصوتية الإجرامية، التحليل الطيفي، نموذج خليط غوسيان، المقاربة البايزيه.

Application Forensique à la Reconnaissance Vocale du Locuteur

Ouassila Kenai, Mhania Guerti

Laboratoire Signal et Communications, Ecole Nationale Polytechnique, Alger, Algérie

wassi_ke@yahoo.fr, mhania.guerti@enp.edu.dz

Résumé

Dans cet article, nous nous intéressons à la Reconnaissance Vocale des Locuteurs Arabophones en vue de la Forensique c'est-à-dire dans le domaine Criminalistique (RVC) en particulier sur les deux tâches majeures l'Identification Vocale d'un Locuteur Arabophone Criminalistique (IVC) et l'Authentification Vocale d'un Locuteur Arabophone Criminalistique (AVC) en mode indépendant du texte.

A partir des signaux vocaux de ces locuteurs des informations relatives à leurs identités sont extraites par l'analyse cepstrale MFCC (Mel Frequency Cepstral Coefficients) avec ces derniers en estimant des modèles GMM (Gaussian Mixture Models) de locuteurs robustes. Ainsi une trace vocale, peut être analysée et par la suite comparée avec les GMM en appliquant l'approche Bayésienne, Likelihood-Ratio (LLR), afin de permettre son identification et son authentification.

Nos expériences réalisées sur le système RVC montrent qu'un GMM composé de 32 Gaussiennes est largement suffisant pour représenter la distribution des vecteurs d'un seul locuteur (le criminel) ainsi que le matériel d'enregistrements qui donne de meilleures performances de ce système élaboré. En effet, nous avons obtenu des résultats satisfaisants. Ceux-ci peuvent aider la justice à prendre une décision afin de résoudre des problèmes criminalistiques.

Mots clés: Reconnaissance Vocale Criminalistique RVC, Identification Vocale Criminalistique IVC, Authentification Vocale Criminalistique AVC, MFCC, GMM et Approche LLR.

1. Introduction

L'expression vocale est une caractéristique propre d'un locuteur. La reconnaissance vocale est un terme générique regroupant les problèmes relatifs à la reconnaissance du locuteur sur la base de l'information contenue dans le signal acoustique de parole [1].

La Reconnaissance Vocale Criminalistique (RVC) est une application semblable à celle de la reconnaissance vocale du locuteur sauf que la RVC est utilisée dans le domaine forensique avec des moyens et des techniques spécifiques. Nous nous intéressons en particulier à l'Identification Vocale Crimina-

listique (IVC) ainsi qu'à l'Authentification Vocale Criminalistique (AVC) qui sont deux tâches majeures de la RVC :

- l'identification Vocale Criminalistique IVC est une tâche qui permet d'identifier un suspect parmi un certain nombre de modèles de différents locuteurs à partir d'un signal de parole, appelé segment de test ;
- l'authentification Vocale Criminalistique AVC est une tâche qui permet de décider à partir d'un signal de parole, appelé trace, et une identité proclamée si la trace provient de l'identité en question ou non.

Les deux tâches proposées dans ce travail sont basées sur les modèles GMM, les algorithmes **LBG (L**inde **B**uzo **G**ray) et **EM (E**xpectation **M**aximisation), le LLR comme approches de reconnaissance et l'Arabe Standard pour les enregistrements des différents signaux vocaux. Cette particularité de la langue permet d'aboutir à un système commun à tous les arabophones et ceci à cause de la diversité des dialectes des locuteurs.

Ce travail propose les sections suivantes :

- dans la section 2, nous présentons les notions de base sur la RVC, l'IVC, l'AVC et l'architecture du système RVC ;
- les sections 3 et 4 exposent l'approche Bayésienne, la paramétrisation et la modélisation des MFCC par GMM ainsi leurs apprentissage ;
- la dernière section est consacrée aux expériences et résultats des tests d'évaluation de notre système.

2. Reconnaissance Vocale Criminalistique

La RVC est une application très complexe, car elle nécessite la compréhension de plusieurs disciplines scientifiques y compris, la linguistique, l'acoustique, l'électronique, les statistiques, l'informatique, etc. Avec le développement fulgurant de la téléphonie et l'utilisation de la voix humaine pour commettre des crimes, l'identification des personnes à partir de leur voix est devenue un domaine d'étude, populaire et objectif dans plusieurs centres de recherche dans le monde [2]. La définition de la Reconnaissance Criminalistique d'un Locuteur correspond à l'avis des experts dans un processus légal pour répondre à la question suivante [2] :

Est-ce qu'un ou plusieurs enregistrements vocaux sont générés par le même locuteur ou non ?

L'utilisation de la RVC dans les domaines judiciaires ou criminalistiques. Il s'agit par exemple de rechercher un individu parmi une population de suspects potentiels (tâche d'IVC) ou encore de comparer un enregistrement vocal issu d'une écoute téléphonique de la voix d'un suspect potentiel (tâche d'AVC).

Les techniques de reconnaissance vocale sont basées sur des mesures de ressemblance entre des enregistrements de parole. Ces mesures sont faites sur des paramètres acoustiques extraits de l'analyse du signal de parole. Elles peuvent prendre en compte les informations spécifiques au locuteur, le contenu du message vocal, les informations sur l'environnement et le matériel d'enregistrement [3].

2.1. Identification Vocale Criminalistique

L'identification vocale doit reconnaître

un locuteur dans une population constituée de N locuteurs connus. La réponse donnée par le système correspond à l'identité de la personne dont le signal de parole est le plus proche de celui qui est testé.

2.2. Authentification Vocale Criminalistique

L'authentification du locuteur consiste, après que le locuteur a décliné son identité, à vérifier l'adéquation de son message vocal avec la référence acoustique du locuteur qu'il prétend être.

2.3. Architecture du système RVC

L'architecture du système RVC est décomposée en différents modules de traitement qui sont la :

- paramétrisation (analyse acoustique) : on extrait du signal les paramètres pertinents permettant une discrimination des locuteurs ;
- modélisation (création des références de locuteurs) : à partir des données du locuteur, extraites par le module de paramétrisation, une référence du locuteur est créée. Elle sert comme référence pour la **RVC**, où elle est comparée avec le signal de test ;
- comparaison (tests) : une comparaison est effectuée entre la référence (tâche de vérification) ou les références (tâche d'identification) et le signal de test ;
- décision (résultat final) : à partir du résultat du module précédent, la décision est le nom d'un locuteur en identification, un rejet ou une acceptation en authentification.

3. Reconnaissance Vocale Criminalistique par l'Approche Bayésienne

L'approche bayésienne (Likelihood-Ratio, LR) est parmi les approches les plus utilisées pour interpréter une preuve scientifique [4]. Cette dernière a été établie comme une base théorique de n'importe quelle discipline criminalistique [5]. La structure principale du calcul de la preuve ainsi que son emplacement dans le processus général de reconnaissance et d'interprétation d'une preuve scientifique [6].

Dans cette approche, l'interprétation d'une preuve scientifique doit être faite en considérant au moins deux hypothèses concurrentes [5]. Au lieu de considérer une seule hypothèse, l'échantillon en question vient d'une personne suspecte, l'expert forensique doit considérer la probabilité d'avoir la preuve en donnant au moins une autre hypothèse compétitive, par exemple, l'échantillon en question vient d'une autre personne et par la suite, évaluer la puissance de cette preuve sous ces deux hypothèses. Cette approche montre comment combiner de nouvelles données avec des connaissances préalables pour donner des probabilités postérieures à des problèmes juridiques.

Dans ce cadre bayésien, les rôles des experts et des juges sont clairement séparés, car la justice veut savoir la probabilité de la proposition (C), « le suspect a commis le crime», en donnant les circonstances du cas (I) et les observations faites par les experts (E) [7]. Cette probabilité est donnée par la formule suivante :

$$O(C/E,I) = \frac{P_r(E/C,I)}{P_r(E/C',I)}.O(C/I) \quad (1)$$

Exprimée en mots, la probabilité postérieure est égale au rapport de la vraisemblance multiplié par la probabilité préalable, où la probabilité préalable concerne la justice (des informations relatives au cas qui est de nature indéterminée). Le rapport de vraisemblance est donné par les experts. Ce dernier mesure la puissance d'une preuve scientifique sous une hypothèse donnée.

Par exemple :

Pr (E/C, I) = 0.9, Pr (E/ C/, I) = 0.1;

Donc, LR = 0.9/0.1= 9.

Alors il est neuf fois plus probable d'avoir la preuve sous l'hypothèse C que sous l'hypothèse C/.

4. Paramétrisation et Modélisation

La paramétrisation ou l'analyse acoustique peut être considérée comme un changement de représentation afin de préserver au maximum l'information présente dans le signal d'origine, tout en donnant une description plus compacte.

4.1. Paramétrisation

Dans nos expériences, une analyse est appliquée toutes les 10 ms par glissement et recouvrement sur des fenêtres d'analyse de 20 ms. A chaque trame, un vecteur de représentation acoustique les MFCC sont calculés à partir d'un banc de 24 filtres triangulaires répartis dans l'échelle fréquentielle Mel.

Les MFCC d'une trame de parole sont calculés de la façon suivante :

- après le filtrage de préaccentuation, le signal de parole est d'abord découpé en fenêtres de taille fixe réparties uniformément le long du signal ;
- la FFT (Fast Fourier Transform) de la trame est calculée. Ensuite, l'énergie est calculée en élevant au carré la valeur de la FFT. L'énergie est passée ensuite à travers chaque filtre Mel. Soit Sk l'énergie du signal à la sortie du filtre K, nous avons maintenant mp (le nombre de filtres) paramètres Sk. (Des études ont montré que les 20 premiers paramètres de chaque trame extraits du filtre Mel représentent très bien le locuteur) [8] ;
- le logarithme de Sk est calculé ;
- finalement les coefficients sont calculés en utilisant la IDCT (inverse Discrete Cosinus Transform). Avec la FFT, nous sommes passées à l'échelle fréquentielle et avec la IDCT nous retournons vers le temporel, nous avons utilisé IDCT au lieu de IFFT car IDCT a l'avantage de la décorrélation (c'est-à-dire. une matrice de covariance diagonale) :

$$C_i = \sqrt{\frac{2}{m_p}} \sum_{k=1}^{k=m_p} Log(S_k)\cos(i(k-1/2)\frac{\pi}{m_p}) \quad (2)$$

i = 1,…., N, où N est le nombre des MFCC que nous souhaitons obtenir.

avec : Ci = les MFCC ; Sk = l'énergie du signal à la sortie du filtre k ; mp = nombre de filtres.

4.2. Modélisation des MFCC par les GMM

La reconnaissance par Mélanges de Gaussiennes GMM, consiste à modéliser un locuteur par une somme pondérée de

composantes (densités) gaussiennes. Chaque composante est supposée modéliser un ensemble de classes acoustiques. Une densité de mélange de gaussiennes est donnée par l'équation :

$$p(x) = \sum_{m=1}^{M} \pi_m \, b_m(x) \qquad (3)$$

x : un vecteur aléatoire de dimension D,

$b_m(x)$: les densités de probabilités gaussiennes, paramétrées par le vecteur moyenne μ_m et une matrice de covariance Σm, et π_m représente le poids des mélanges avec :

$$\sum_{m=1}^{M} \pi_m = 1$$

$$b_m(x) = \frac{1}{(2\pi)^{D/2}|\sum_m|^{1/2}} e^{[\frac{1}{2}(x-\mu_m)(\sum_m)^{-1}(x-\mu_m)]} \qquad (4)$$

Un GMM λ à M gaussiennes est défini donc par :

$\lambda = \{\pi_m, \mu_m, \Sigma_m\}$ avec m = 1,…, M.

L'élément qui permet d'expliquer le succès des GMM est l'existence d'un outil très puissant pour l'estimation des paramètres qui leur sont associés : l'algorithme EM qui permet d'avoir un maximum local, il maximise de façon itérative la fonction de vraisemblance P(X / λ) (où X présente l'ensemble de données à modéliser et λ présente le modèle GMM) [3].

4.3. Apprentissage du modèle

Il s'agit, lors de la phase d'apprentissage, d'estimer l'ensemble λ des paramètres d'un modèle GMM de locuteur. La méthode conventionnelle est celle du Maximum de Vraisemblance, dont le but est de déterminer les paramètres du modèle qui maximisent la vraisemblance des données d'apprentissage (N vecteurs d'apprentissages ; X = x_1, …, x_N suffisamment indépendants).

L'apprentissage se fait en deux étapes :

- la 1ère concerne l'initialisation des paramètres du modèle en utilisant l'algo-rithme LBG ;
- la 2ème cible l'optimisation des paramètres obtenus en utilisant l'algo-rithme EM [9-10].

5. Expériences et résultats

Notre système de **RVC** par les modèles GMM a été évalué à l'aide d'un corpus que nous avons réalisé.

5.1. Corpus d'évaluation

Le corpus utilisé pour ces expériences est composé de phrases en Arabe Standard. Il est construit par des adultes. Le critère de choix de ces locuteurs vise à couvrir certaine gamme d'âges et de sexes. Les enregistrements vocaux sont réalisés dans des salles équipées par un matériel professionnel (microphones, consoles de mixages numériques).

Le corpus utilisé pour les expériences est structuré comme suit :

- nous avons effectué trois enregistrements vocaux, un enregistrement pour construire deux corpus d'apprentissage (imposteurs et suspects) et un enregistrement pour le corpus de test ;
- chaque locuteur prononce 5 phrases en mode indépendant du texte. Le choix de ces dernières est spontané mais similaire à celles prononcées par les suspects et les imposteurs dans le cas d'une enquête policière (Table .1) ;
- les corpus d'apprentissage contient

respectivement : 17 locuteurs pour les imposteurs, 10 pour les suspects et tous les autres locuteurs pour les tests ;

- par contre, le corpus de test pour la deuxième expérience contient un seul locuteur (le criminel en question).

Corpus d'apprentissage	ج 1- انفي كل ما ينتسب إلي و أطالب بالتعويض المعنوي و المادي. ج 2 - هي تهمة باطلة و لا أساس لها من الصحة. ج 3 - اطلب التعويض المعنوي جراء ما نسب إلي. احترم أوقات العمل ج 4 -دائما. ج 5 - رأيت المسؤول وهو يغادر مكتبه في منتصف النهار.
Corpus de test	ج - موعدنا اليوم على الساعة الواحدة زوالا.

Table 1. Exemples de quelques phrases utilisées dans le Corpus (pour un seul suspect)

5.2. Mesure d'évaluation

Les performances du système de reconnaissance sont évaluées en termes de Taux d'Identification (TI) pour la tâche d'IVC et de variations des moyennes μ_k^* en fonction du nombre de gaussiennes pour la tâche d'AVC. Ces taux sont définis par :

TI= TIC+TII= 100%

$$\mathrm{TIC} = \frac{\mathrm{NTCI}}{\mathrm{NTT}}$$ avec : NTCI =Nombre de Tests Correctement Identifiés et NTT=: Nombre Total de Tentatives ;

$$\mathrm{TII} = \frac{\mathrm{NTMI}}{\mathrm{NTT}}$$ avec NTMI =Nombre de test mal identifiés et NTT=: Nombre Total de Tentatives ;

$$\text{et } \mu_{\mathrm{k}}^* = \frac{\sum_{n=1}^{T} P_{nk} \vec{X}_n}{\sum_{n=1}^{T} P_{nk}} \qquad (5)$$

avec :

$\vec{X}_n$: Vecteurs acoustiques ; $n=1, \ldots, T$;

P_{nk} : Probabilités générées par k ;

k : Nombre de gaussiennes.

5.3. Configuration du système

Les systèmes de reconnaissance utilisent souvent les MFCC qui permettent une parfaite déconvolution de la contribution du conduit vocal et celle de la source d'excitation [11-12].

Dans les expériences, 12 MFCC avec leurs dérivées premières et secondes : Δ et $\Delta\Delta$, sont utilisés et l'approche Bayésienne LLR est appliquée. Les modèles statistiques GMM appris sur le corpus d'apprentissage par algorithmes LBG et EM d'un modèle GMM indépendant du sexe. La durée des données d'apprentissage de chaque modèle GMM varie entre 15 s et 40 s sachant que la durée totale du corpus d'apprentissage (imposteurs et suspects) est environ 20 mn.

5.4. Résultats du système RVC

Dans les parties suivantes, nous allons présenter les résultats de la tâche IVC et la tâche AVC de notre système RVC.

5.4.1. Identification Vocale Criminalistique (IVC)

L'histogramme de TIC des suspects en fonction du nombre de gaussiennes pour un système d'IVC. L'objectif de cette expérience est d'évaluer l'influence du nombre de gaussiennes sur la performance du système d'IVC mis en œuvre, en variant le nombre de gaussiennes de 1 jusqu'à 64 (Figure 1) [13].

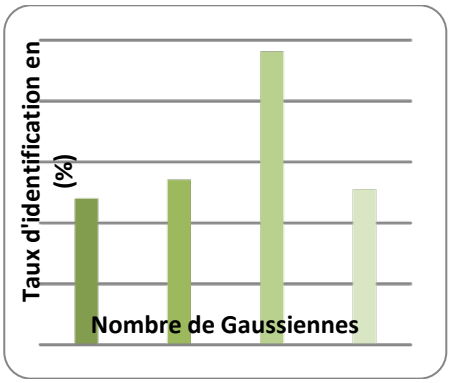

Figure 1 : Histogrammes de TI en fonction du nombre de gaussiennes

Les résultats montrent que le TIC augmente avec le nombre de gaussiennes, il est de 96,33% pour un nombre égal à 32, et se dégrade légèrement de 51,1% au-delà de 32 gaussiennes. L'architecture optimale de notre système correspond à 32 gaussiennes (Table .2).

	Nb de Gaussiennes			
	8	**16**	**32**	**64**
Taux Iden.	48%	54.2%	**96.33%**	51.1%

Taux Ident : Taux d'Identification

Table 2. Résultats des tests d'évaluation

Le meilleur TIC est obtenu pour un nombre de gaussiennes égal à 32. Par conséquent, un modèle de 32 gaussiennes suffit pour représenter convenablement locuteur.

5.4.2. Authentification Vocale Criminalistique (AVC)

Les variations des moyennes μ_k^* de la trace et des suspects en fonction du nombre de gaussiennes pour un système d'AVC. L'objectif de cette expérience est d'étudier l'influence du nombre de gaussiennes sur la performance du système mis en œuvre. Des variations de 1 jusqu'à 64 gaussiennes ont été faites (Figure 2).
Les résultats montrent que les moyennes augmentent avec le nombre de gaussiennes et se dégradent légèrement au-delà de 32 gaussiennes, pour tous les suspects. L'architecture optimale de notre système correspond à un nombre égal à 32.
Comme nous avons constaté que toutes les courbes des suspects sont loin de la trace sauf la courbe d'un seul suspect qui est proche, ce qui montre que ce suspect est à l'origine de la trace.
Dans notre cas, les expériences et les tests dépendent du matériel d'enregistrements ainsi que de l'état émotionnel et du sexe du locuteur. S'ils sont bien respectés, nous obtenons de bons résultats c'est- à - dire de bonnes performances du système, Dans le cas inverse, ces performances peuvent chuter considérablement.

6. Conclusions

L'objectif principal de ce travail est d'évaluer notre système de RVC appliqué à l'Arabe Standard et particulièrement dans le domaine forensique.

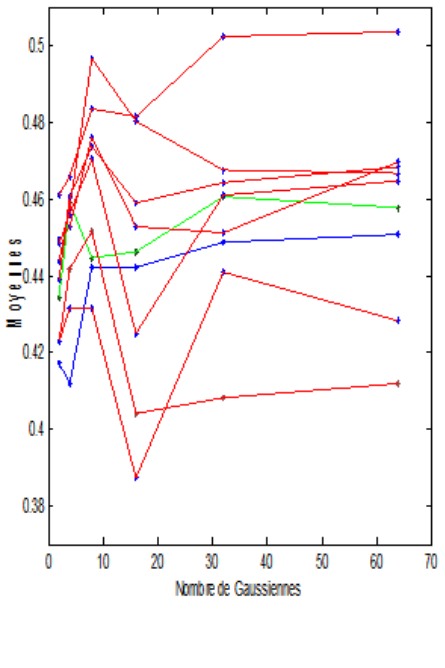

Figure 2: Variation des moyennes en fonction du nombre de gaussiennes

Le corpus que nous avons élaboré nous a permis de faire plusieurs tests d'évaluation sur les deux tâches principales de ce système. En plus, nous avons créé des modèles GMM qui représentent des locuteurs arabophones en utilisant les deux techniques LBG et EM. L'approche Bayésienne, offre un moyen très puissant qui permet d'analyser une preuve scientifique et d'estimer les résultats de notre système. Cette dernière sera adressée au juge qui peut, par la suite, être aidé pour faire son jugement. Le bon choix de l'ordre à 32 du modèle GMM est largement suffisant pour représenter un seul locuteur.

Les conditions d'enregistrements (le matériel professionnel et le bon état physique du locuteur) ont un grand effet sur la performance de notre système.

Dans nos expériences, nous pouvons dire que le GMM est très puissant et peut représenter des distributions aléatoires très complexes d'une manière très fidèle. En effet, si nous choisissons un petit ordre, nous pouvons avoir une grande perte de données et par conséquent, une dégradation de performance. Dans le cas contraire, si nous choisissons un grand ordre, nous pouvons avoir le pro blème de surapprentissage du GMM, c'est-à-dire, présenter des données qui n'existent pas dans l'espace des para mètres acoustiques du locuteur en question ; ainsi que la durée des enregistrements utilisés lors de la phase d'apprentissage du modèle.

7. Références

[1] Mami Y., "Reconnaissance de locuteur par localisation dans un espace de locuteurs de référence", Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, 2003.

[2] Bonastre J-F., Bimbot F., Boë L-J, Campbell J.J, Reynolds D.A. and Magrin-Chagnolleau I., "Authentification des personnes par leur voix: un nécessaire devoir de précaution", AFCP, 1-5, 2003.

[3] Maaradji A. and Mecheri I., "Système de vérification du locuteur pour une application d'accès biométrique", Institut National de formation en Informatique (I.N.I), PFE, Alger, 2006.

[4] Alexander A., "Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions", Thèse de Doctorat, Ecole Polytechnique Fédérale de Lausanne, Suisse, 2005.

[5] Evett I. W., "Towards a Uniform Framework for Reporting Opinions in Forensic

Science Casework", Science & Justice, 38(3), 198-202, Suisse 1998.
[6] Andrzej D., Meuwly D. and Alexander A., "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition", Speech Processing Group, the Forensic Science Service.
[7] Gonzales-Rodriguez J., Ortega-Garcia J. and Sanchez-Bote J., "Forensic Identification Reporting Using Automatic Biometric Systems", Speech and Signal Processing Group (ATVS), DIAC, Spain 2005.
[8] Kanungo T. and Mount D. M., "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE transactions on pattern analysis and machine intelligence, Vol. 24, N° 7. July 2002.
[9] Dempster A. P., Laird N-M and Rubin D-B., "Maximum-likelihood from incomplete data via the EM algorithm", J. Royal Statist. Soc. Ser. B, 39, 1997.
[10] Bilmes J., Gentle, A. "Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", Report, University of Berkeley, ICSI-TR-97-021, 1997.
[11] Burget L., Matejka P., Schwarz P., Glembek O. and Cernocky J-H., "Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System", IEEE Transactions on Audio, Speech, and Language Processing, Vol 15, N° 7, 2007.
[12] Amrous A-I., Debyeche M. and Krobba A., "Coopération de connaissances dans les modéles de Markov cachés pour la reconnaissance de mots isolés arabes", 1st International Conference on Image and Signal Processing and their Applications, Mostaganem, Algérie 2009.
[13] Kenai O. and Guerti M., "Identification d'un Locuteur Arabophone en vue la Criminalistique", Third International Conference on Industrial Engineering & Manufacturing, Université de Batna, 4-5, 2013.