Architecture d'un Système de Vérification Automatique du Locuteur appuyée par la Détection du Genre

Djellali hayet, Radia Amirouche **, Mohamed Tayeb Laskri****

* Département d'informatique
Université Badji Mokhtar Annaba
B. P. 12 Annaba 23200 Algérie.
hayetdjellali2003@yahoo.fr
** Département d'informatique
Université Badji Mokhtar Annaba
B. P. 12 Annaba 23200 Algérie.
amradia2003@yahoo.fr
*** Département d'informatique
Université Badji Mokhtar Annaba
B. P. 12 Annaba 23200 Algérie
mtlaskri@wissal.dz

Abstract

We propose a new approach in Automatic speaker verification ASV based on detection Gender (male,female). We determine with speaker voice his gender. knowing that ,the speaker could be an impostor with opposite gender that he claims. The aim of this work is to experiment if detection gender module can improve speaker verification decision when we compare it with baseline ASV system.

Mots-Cles: *Vérification automatique du locuteur, les modèles de mélange gaussienne GMM, Adaptation MAP, Maximum à posteriori, Détection du genre, Maximum de vraisemblance, fausse acceptation.*

1. Introduction

La Vérification du locuteur VAL en indépendant du texte (pas de contrainte sur le texte à prononcé du locuteur), l'entrée de la VAL est le message vocal et un code d'entré permettant de récupérer la référence du client.

En phase d'apprentissage, une étape de prétraitement et d'extraction de caractéristique est nécessaire puis la modélisation par les modèles de mélanges gaussienne (GMM qui constitue l'état de l'art) de chaque client et des autres locuteurs autre que le client dit modèle du monde (UBM :Universel background Model). Lors de la phase de test, les paramètres du signal de test sont extrait et le calcul du score(tiré à partir du modèle client et du monde) est effectué qui est comparé à un seuil de décision. Le résultat attendu est soit acceptation soit rejet. Un calcul du maximum de vraisemblance est nécessaire afin d'estimer les scores.

Log($p(X \mid \lambda_{client})$) : Score su signal de test par rapport au modèle client

Log ($p(X \mid \lambda_{UBM})$) : Score du signal de test par rapport au modèle du monde.
La formule $\Lambda(X)$ est calculé et comparé à un seuil θ :

$$\Lambda(X) = \log(p(X|\lambda_{client}) - \log(p(X|\lambda_{UBM})) > \theta \qquad (1)$$

Si $\Lambda(X) > \theta$ accès Client
Sinon Imposteur.

Deux approches sont adoptées en modélisation :

- **Modélisation GMM-UBM-ML**

Dans ce cas deux modèles sont crées, le modèle client sur la base de ces données acoustiques et le modèle du monde UBM dont les vecteurs acoustiques sont issues d'une large population de locuteurs autre que nos clients. L'apprentissage des deux modèles de GMM repose sur l'algorithme EM (Estimation-Maximisation). Le modèle GMM-UBM ML, repose sur la l'estimation du Maximum de vraisemblance ML (Maximum Likelihood). Cette approche a cependant l'inconvénient de ne pas bien généraliser face aux faibles données d'apprentissage du client, ainsi il est difficile d'avoir un modèle client robuste au sens du Maximum de vraisemblance (ML) [Ves,2008] [Hau,2008][Hsi,2004].

- **Modélisation par Adaptation MAP**

L'approche MAP (Maximum a Posteriori) règle le problème de généralisation soulevé par ML en modélisation du client. Des connaissances à priori de la distribution des paramètres du modèle sont introduite dans le processus de modélisation, même si les données d'apprentissage du client sont réduites, l'information à priori concernant les paramètres peut aider à dépasser ce problème. Elle consiste à se servir du modèle du monde afin d'estimer le modèle du client sur la base de ces

données d'apprentissage et de l'adaptation par Maximum a posteriori [Pre,2008].

2. Détection du Genre

Presque toute l'information comprise dans la parole est comprise entre 200hz et 8Khz. Les humains arrivent à distinguer les hommes des femmes selon leurs fréquences. Les femmes ont une fréquence fondamentale plus élevé que les hommes. L'adulte masculin a entre 50Hz et 250hz avec une valeur moyenne de 120Hz. Pour un adulte de sexe féminin elle est plus grande. En analysant la moyenne du pitch, on peut utiliser un algorithme de détection du genre. Dans le domaine fréquentiel, la fréquence du signal est calculée et l'information est extraite du spectrum[Chi,2002].

3. Architecture du Système de Vérification Automatique du Locuteur

Nous proposons un système de VAL basé sur GMM-MAP aidé par la détection du genre (Figure 1), juste après l'extraction des caractéristiques. Afin de permettre de discerner entre un locuteur imposteur masculin qui prétend être un client féminin et le contraire également, le but est de réorienter le module de calcul des scores vers le modèle du monde qui correspond au sexe détecté et non pas déclaré et être sur au moins qu'on comparé notre locuteur par rapport à son vrai modèle du monde dépendant du genre et son modèle dérivé par adaptation MAP. Nous décrirons les différents modules de notre architecture de VAL qui comporte :

3.1 Phase d'Apprentissage

❖ **Prétraitement et Extraction des caractéristiques**

Elle comporte:

- **Détection du silence DS :**

L'énoncé de parole d'un locuteur comporte des trames de parole et de silence ou de bruit, il est primordial de supprimer les trames de silence ou de bruits qui diminuent les performances des systèmes de val. Le critère d'énergie est utilisé pour sélectionner les trames de paroles (haute énergie) et éliminer les trames de silence (énergie faible).

- **Extraction des caractéristiques EC :**

L'analyse cepstrale est très utilisé en val due à sa robustesse d'éstimation des signaux bruités [Preti,2008]. On a extrait les 13 coefficients cepstraux avec leurs dérivés et dérivé seconde toutes les 10ms (hypothèse de pseudo-stationnarité) calculé sur une fenêtre d'analyse de type hamming de 25ms.

- **Normalisation CMS :**

Afin d'atténuer l'influence du canal de communications la normalisation par soustraction de la moyenne cepstrale est appliquée (Cepstral Mean Substraction) qui consiste à retirer la moyenne de la distribution de chacun des paramètres cepstraux.

❖ **La Modélisation :**

- **Modélisation GMM-UBM-ML**

On élabore deux modèles UBM dépendant du genre , on procède à l'apprentissage du modèle UBM masculins avec les Modèles de Mélanges gaussiennes (échantillon de parole homme) et l'autre UBM féminin(extrait de parole féminin). Les paramètres du modèle (moyenne, covariance, poids des gaussienne) sont appris par l'algorithme EM (Estimation-Maximisation).

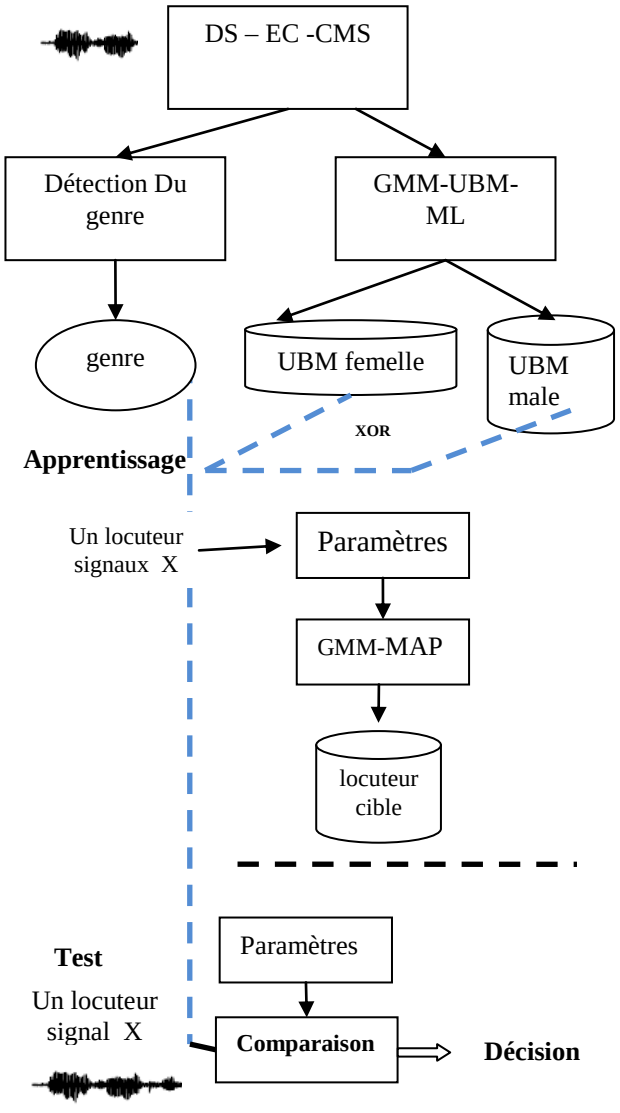

Figure 1 Architecture du Système de VAL Indépendant du Texte

Le modèle du client est dérivé du modèle du monde par adaptation de ces paramètres GMM (moyenne, covariance, les poids) qui sont estimés. Or,

expérimentalement, seuls les moyennes des GMM est adapté en VAL [Preti,2008]. La maximisation du paramètre de moyenne est formulé ainsi :
Pour une gaussienne i du GMM, s'exprime sous la forme :

$$\tilde{\mu}_i = \alpha_i \ \mu_i^c + (1-\alpha_i)\,\mu_i^w \qquad (2)$$

μ_i^c : la moyenne i du GMM client
μ_i^w : la moyenne i du GMM du monde UBM

α : est un coefficient de pondération qui permet d'affecter plus ou moins de poids, aux paramètres à priori par rapport aux paramètres estimées sur les données d'apprentissage.
Il est défini par :

$$\alpha_i = \frac{ni}{ni+\tau} \qquad (3)$$

avec ni: le nombre de trames associé à la gaussienne i.

$\boldsymbol{\tau}$ **:** Le relevance factor, il contrôle le degré d'adaptation de chaque gaussienne en termes de trames attribuées.

❖ Détection du Genre

Le Module de détection du genre est inspiré par un programme sous matlab basé sur le calcul du pitch moyen. Le pitch est quant à lui calculé par autocorrélation [Kam,2006]. Le principe de l'algorithme est le suivant :
Extraction du pitch : La parole est divisé en segments de 60 ms ,chaque segment est extrait tous les 50ms d'intervalle donc le recouvrement est de 10ms. Chaque segment appelle une fonction « pitch autocorrélation » afin d'éstimer la fréquence fondamentale de ce segment. Ensuite ,un filtre médian est appliqué tous les trois segments donc moins affecté par le bruit. Enfin, la moyenne de toutes les fréquences fondamentales est retournée[Chu,2000].

3.2 Phase de Test

- **Paramétrisation**

Les mêmes traitements que lors de l'apprentissage sont éffectués en phase de test pour les données acoustiques du locuteur.

- **Décision et Calcul du Score**

Le score sera calculé de la manière suivante :
Log(p(X | λ client) : Score du modèle client

Log (p(X | λ_{UBMF})) : Score par rapport au modèle du monde féminin.

Log (p(X | λ_{UBMM})) : Score par rapport au modèle du monde Masculin :

$\Lambda(X)$ est calculé et comparé à un seuil θ :

$\Lambda(X) = \Lambda1(X)$ Si le module détection du genre Désigne un homme
$\Lambda(X) = \Lambda2(X)$ Sinon

Sachant que $\Lambda1(X)$ et $\Lambda2(X)$ sont calculées ainsi :

$$\Lambda1(X) = \log(p(X|\lambda_{client}) - \log(p(X|\lambda_{UBMM})) \qquad (4)$$

$$\Lambda2(X) = \log(p(X|\lambda_{client}) - \log(p(X|\lambda_{UBMF})) \qquad (5)$$

Si $\Lambda(X) > \theta$ accès client Sinon imposteur.

4. Protocole Expérimentale

- **Base de Donnée**

Les Données sont enregistrées sous goldwave à fréquence 16KHz pour une durée de 90s pour chaque locuteur lors de l'apprentissage et 30s en phase de test. La population du modèle du monde masculin est de 15 hommes et féminine de 15 femmes. Deux sessions sont enregistrées pour chaque locuteur à intervalle de 1 mois entre eux. Dix 10 clients sont enregistrées dans la base (5 hommes et 5 femmes).

- **Système de référence**

Le système de VAL de référence GMM-UBM-IG indépendant du genre est adopté obtenue par fusion des deux modèles (masculin et féminin), donc de 30 locuteurs. L'égalité des deux populations dans le modèle GMM-UBM-IG est une contrainte afin d'éviter que le modèle finale soit biaisé [Rey,2000].

Table 1. Population choisie

	Effectifs	Apprentissage	test
GMM-UBM IG	30	2sessions 90s	30s
GMM-UBMM	15	2sessions 90s	30s
GMM-UBMF	15	2sessions 90s	30s
GMM-MAP	10	2sessions 90s	30s

IG : indépendant du genre

L'intérêt majeur pour nous est de déterminer si la détection du genre permet d'apporter une meilleure prise de décision et ainsi améliorer le taux de fausse acceptation FA.

- **Détection du Genre**

Cet algorithme a été testé sur des échantillons de parole sur une centaine de personne de sexes différent à effectif égale (50 hommes et 50 femmes) de durée de 2 à 5s échantillonné à16khz. Les erreurs sont de 15%. Il est cependant nécessaire de préciser que lors du test de cet algorithme la base de donnée n'était pas complète, c'est pourquoi nous avons utilisé une centaine de fichiers sons(wav) déjà acquis enregistrés sous goldwave .

5. Conclusion

Certes la partie expérimentale est en cours de réalisation pour l'obtention des résultats. Toutefois, nous espérons que le module de détection du genre permettra d'apporter un plus et diminuer le taux d'égal erreur EER(Equal Error Rate).

Cependant, il devra être amélioré de telle manière que la confiance soit meilleure que d'opter pour un modèle UBM unique obtenu par combinaison de deux modèles l'un masculin et l'autre féminin. Le taux d'erreur ne devra pas dépasser pas les 2% comme nouvelle exigence.

Référence

[Ves,2008] Bôstjan Vesnicer,France Mihelic. 'The likelihood ratio decision criterion for nuisance attribute projection in GMM speaker Verification', Hindawi Publishing Corporation Eurasip Journal On advances in Signal Processing volume (2008).

[Pre,2008] Alexandre Preti, Thése 'Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur. Académie d'Aix Marseille ,Laboratoire d'informatique d'Avignon (2008).

[Hau,2008] Ville Hautomaki,Ismo Karkkainen,Tomi Kinnunen, "Maximum a posteriori adaptation of the centroid model for speaker verification", IEEE Signal Processing letters volume 15 (2008).

[Kam,2006] Kamran Mustapha,I,C, Bruce, "Robust formant tracking for continous speech with speaker variability", IEEE Transaction on Speech and Audio Processing",(2006).

[Hsi,2004] Chu song Chen, Hsin Min Wang,Yi ping hung,"An efficient Approach to multimodal Person Identity verification by fusing face and voice information",Taiwan (2004).

[Rey,2000] Douglas Reynolds, Thomas F. Quateri, Robert B Dunn,"Speaker verification using Gaussian mixture models, Digital Signal processing 10.19.41. (2000).

[Chu,2000] Chui Ling Lay, N. Hian James, "Gender classification from speech" (2000). http://sg.geocities.com/nghianja/ CS5240.doc