Robustesse de la vérification du locuteur par mot de passe personnalisé

Bruno JACOB [△], *Johnny MARIETHOZ* [▼], *Guillaume GRAVIER* [▽], *Frédéric BIMBOT* [△]

[△] IRISA (CNRS/INRIA), Campus Universitaire de Beaulieu, 35042 Rennes Cedex, http://www.irisa.fr
[▼] IDIAP, rue du Simplon 4, Case postale 592, CH-1920 Martigny, http://www.idiap.ch
[▽] ENST, 46 rue Barrault - Paris 75634 Cedex 13, http://www.enst.fr

E-mail: {bimbot,bjacob}@irisa.fr - marietho@idiap.ch - gravier@tsi.enst.fr

Abstract

This paper presents a speaker verification approach using customized password (i.e passwords chosen by the client). In this context, the issue consists in estimating a speaker-independent (world) model for the password, using speech utterances from a single speaker. For this purpose, we use a universal (speaker-independent) speech model from which we derive a model of the password on the basis of the decoded (graph of) units. Our results tend to show that the quality of the transcription influences strongly the performance of the system.

1. Introduction

Le travail présenté dans cet article porte sur la vérification du locuteur au téléphone dans le contexte de la sécurisation de transactions ou d'accès à des services à caractère commercial. Il s'inscrit dans le cadre du projet européen Telematics PICASSO [1] [BBB+99].

Pour ce type d'applications, la technologie couramment utilisée repose sur des approches *dépendantes du texte*, opérant la vérification du locuteur sur une séquence de mots *fixe* et invariable (par exemple, une séquence de chiffres correspondant au numéro de compte ou d'abonné du client). Pour réduire les efforts de mémoire de la part de l'utilisateur, ainsi que les risques de fraude, certains profils d'applications (dits à *texte prompté*) font appel à des séquences aléatoires de mots issus d'un petit vocabulaire, commun à tous les clients, que le locuteur doit prononcer en réponse à un *prompt* [Dod85]. Cependant, ces approches sont toutes deux vulnérables à l'utilisation de parole pré-enregistrée, dès lors que l'imposteur est capable de procéder à de la synthèse par concaténation, même si la qualité de la concaténation n'est pas très bonne (voir [LB99] dans le cas de concaténation de mots).

L'approche par *mot de passe personnalisé* (MPP) a pour objectif de remédier, au moins partiellement, aux difficultés qui viennent d'être exposées. Pour ce profil d'application, le client choisit lui même l'énoncé sur lequel s'effectue la vérification. Le premier avantage de cette approche réside en ce qu'elle introduit un niveau supplémentaire de protection car les imposteurs intentionnels potentiels doivent d'abord avoir connaissance du mot de passe avant de pouvoir tenter une imposture (ce qui va de pair avec une impression de sécurité accrue de la part des utilisateurs). Par ailleurs, cette approche est plus ergonomique car elle réduit les difficultés de mémorisation pour l'utilisateur.

Nous présentons dans cet article une série de premiers résultats d'expériences simulant un procédé de vérification du locuteur à partir de mots de passe personnalisés, produits dans le cadre du projet PICASSO. Nous exposons tout d'abord le cadre théorique du problème de la vérification du locuteur. Puis nous introduisons la problématique essentielle de la vérification par mot de passe personnalisé, à savoir l'inférence d'un modèle acoustique *indépendant* du locuteur à partir de l'énoncé d'un *seul* locuteur. Nous décrivons une technique possible pour résoudre cette question, qui utilise un modèle de parole universel (indépendant du texte et du locuteur). Nous présentons finalement une série de résultats expérimentaux qui permettent notamment de comparer l'approche à base de mot de passe personnalisé à une approche dépendante du texte sur les mêmes énoncés.

2. Cadre théorique

2.1. Modèle probabiliste

L'approche utilisée dans l'ensemble de cet article s'appuie sur un formalisme probabiliste du problème de la vérification. Pour un énoncé de test noté Y prononcé par un locuteur proclamant l'identité X, on calcule le logarithme du rapport de vraisemblance :

$$S_X(Y) = log\left(\frac{\hat{\mathcal{P}}(Y|X)}{\hat{\mathcal{P}}(Y|\bar{X})}\right)$$

où $\hat{\mathcal{P}}(Y|X)$ représente la vraisemblance de l'énoncé sous l'hypothèse qu'il a été prononcé par le locuteur proclamé et où $\hat{\mathcal{P}}(Y|\bar{X})$ représente la vraisemblance de l'énoncé sous l'hypothèse qu'il a été prononcé par un autre locuteur.

Le modèle probabiliste correspondant à X (dit *Modèle du Client*) est estimé à partir de données d'apprentissage composées d'énoncés prononcés par X. Le modèle correspondant à $\bar{X}$ (dit *Modèle Non-Client*) est obtenu à partir d'énoncés semblables prononcés par d'autres locuteurs. Habituellement, le modèle du non-client est le même pour tout les clients, et on le désigne couramment par *Modèle du Monde* (noté Ω). Dans le cas du mot de passe personnalisé, le modèle Ω dépend du mot de passe du locuteur et donc du locuteur. Il faut donc disposer d'un *modèle Client* et

1. partenaires : les opérateurs KPN-Telecom (*NL*), Fortis (*NL*), Swisscom (*CH*), les laboratoires de recherche ENST (*F*), IDIAP (*CH*), KPN-Research (*NL*), KTH (*S*), KUN (*N*), les banques de l'UBS-Ubilab (*CH*) et la société de technologie vocale Vocalis (*UK*).

d'un *modèle du Monde* pour chaque locuteur.

Dans les travaux décrits ici, les modèles probabilistes utilisés sont des MMC (Modèles de Markov Cachés) dont les fonctions d'émission des états sont des mélanges de distributions Gaussiennes.

2.2. Décision, types d'erreur et mesure des performances

Dans les applications où il s'agit de prendre une décision binaire d'acceptation ou de rejet de l'identité proclamée, le score S_x est comparé à un seuil de décision choisi de façon à optimiser les performances du système dans une condition de fonctionnement particulière. Cette condition de fonctionnement est spécifiée par le rapport des coûts associés aux deux types d'erreur possibles : *faux rejet*, si un client authentique est rejeté par le système et *fausse acceptation* si un imposteur n'est pas détecté.

Les performances des approches décrites dans cet article sont présentées essentiellement sous forme de courbes DET [MP97] qui indiquent les caractéristiques du système en terme de pouvoir de séparation des clients et des imposteurs : plus la courbe DET est proche de l'origine, meilleure est la séparation apportée par le système. Le point situé à l'intersection de la courbe DET et de la première bissectrice correspond à l'EER (Equal Error Rate) c'est-à-dire le taux d'erreur du système quand les fausses acceptations sont égales aux faux rejets.

3. Problématiques spécifiques au mot de passe personnalisé

La mise en oeuvre du procédé de vérification nécessite d'être en mesure d'estimer le Modèle du Client (X) et le Modèle du Monde (Ω) du mot de passe. Le modèle X caractérise la façon dont le client prononce son mot de passe, ce qui peut être déduit des énoncés d'apprentissage produits pendant la phase d'apprentissage. En revanche, le modèle Ω doit caractériser la façon dont l'ensemble des locuteurs sont susceptibles de prononcer ce même mot alors que l'on ne dispose pas de tels exemples en pratique. Ceci pose donc le problème d'être capable d'inférer un modèle acoustique indépendant du locuteur à partir d'un énoncé produit par un seul locuteur.

Par ailleurs, même si le Modèle du Client peut, pour sa part, être estimé à partir des données d'apprentissage, le volume de données disponibles pour l'estimation est, en pratique, très insuffisant par rapport au volume de données nécessaires pour un apprentissage par maximum de vraisemblance complet.

Le paragraphe suivant expose les réponses que nous apportons à ces deux difficultés.

4. Approche

L'approche adoptée repose sur l'utilisation d'un Modèle de Parole Universel (noté U) représentant les propriétés acoustiques de la langue considérée, indépendamment du locuteur et du texte.

4.1. Modèle de Parole Universel

Le Modèle de Parole Universel U est un réseau ergodique dont les états sont des MMCs gauche-droite, que nous dénoterons par MMC_i. On modélise ainsi un réseau d'unités acoustiques (notées U_i) dont les transitions successives sont régies par un modèle bigramme. Ces unités U_i sont indépendantes du locuteur.

Le modèle U est utilisé pour transcrire en unités de parole U_i l'énoncé d'apprentissage (noté Y_0) du mot de passe produit par le client. Ce modèle U fournit l'ensemble des N-meilleures solutions consistant en N séquences d'unités U_i. Cet ensemble de solutions est ensuite factorisé sous forme d'un graphe minimal gauche-droite : $\mathcal{G}(Y_0)$. Ce graphe peut être compris comme un ensemble d'hypothèses de décodage de l'énoncé Y_0 dans le système de représentation linguistique à base des unités U_i.

4.2. Modèle du Monde

Le Modèle du Monde est construit à partir du graphe $\mathcal{G}(Y_0)$ en y compilant les MMC_i acoustiques indépendants du locuteur correspondant aux unités U_i dans le graphe. Les états des MMC_i sont représentés par des densités de probabilité multigaussiennes. Le modèle Ω représente donc un graphe d'hypothèses de modèles acoustiques pour le mot Y_0 en mode indépendant du locuteur.

4.3. Modèle du Client

La création du modèle du Client s'effectue en deux phases :

1. Dans une première étape, le Modèle du Client est construit à partir du graphe $\mathcal{G}(Y_0)$ en y compilant les $\mathrm{MMC}_i{}^{mono}$ acoustiques indépendants du locuteur correspondant aux unités U_i dans le graphe. Les états des $\mathrm{MMC}_i{}^{mono}$ sont représentés par des densités de probabilité mono-gaussiennes. On peut voir les $\mathrm{MMC}_i{}^{mono}$ comme des versions simplifiées des MMC_i. Le modèle X_{init} ainsi obtenu sert de modèle d'initialisation au modèle X du Client.
2. On entraîne ensuite (seulement) les moyennes des états du modèle initial X_{init} au sens du maximum de vraisemblance à partir d'une ou de plusieurs répétitions du mot de passe par le client. Les variances du modèle Client restent égales à celles du modèle X_{init}. Une prochaine étape de notre travail consistera à effectuer cet apprentissage par des techniques Bayésiennes [MM99].

La figure 1 illustre les étapes de ce processus.

5. Protocole expérimental et expériences

5.1. Base de données

La base de données sur laquelle nous avons réalisé nos expériences est un sous-ensemble de la base de données Polyvar de l'IDIAP. Elle est composée d'énoncés correspondant à 17 mots de commande d'une application touristique de la ville de Martigny en Suisse : annulation, cinéma, Corso, galerie du Manoir, guide, manifestation, mode d'emploi, précédent, suivant, casino, concert, exposition, Giannada, Louis Moret, message, musée, quitter. Dans cette étude, nous avons sélectionné une population de 19 locuteurs (12 Hommes et 7 Femmes).

L'ensemble d'apprentissage contient les 5 premières répétitions de chaque mot de passe de chaque locuteur.

L'ensemble de test est composé d'environ 18000 énoncés dont 1/3 sont des accès authentiques et 2/3 sont des accès d'imposteurs. Les accès imposteurs sont les mêmes énoncés que les accès clients, mais testés contre d'autres identités.

5.2. Protocole d'évaluation

Dans nos expériences, nous avons considéré qu'un mot de commande était un mot de passe hypothétique. Il faut remarquer que ces mots de passe ne sont formés que d'un très petit nombre de syllabes : la décision d'accepter ou non un locuteur comme client doit donc s'effectuer sur un temps très court et rend la tâche de vérification singulièrement difficile.

Les modèles acoustiques des unités de parole U_i représentent des phonèmes hors contexte appris avec la base PolyPhone Suisse Romand. Chaque phonème est composé de trois états émetteurs. Les lois d'émission sont formées d'un mélange de trois gaussiennes pour les MMC_i, et d'une mono-gaussienne pour les MMC_i^{mono}.

Le graphe $\mathcal{G}(Y_0)$ est obtenu à partir des 3 meilleures solutions pour l'énoncé Y_0.

Pour évaluer cette approche, nous l'avons comparée à une expérience de référence utilisant explicitement des exemples acoustiques de chaque mot de passe. L'approche de référence s'apparente donc à une méthode dépendante du texte. Dans cette configuration, on dispose de données acoustiques multi-locuteurs et on est en mesure d'estimer le modèle du *Monde* à partir de celles-ci. En pratique, les énoncés utilisés pour estimer le modèle du Monde proviennent d'un ensemble de locuteurs distincts des 19 locuteurs utilisés dans nos tests.

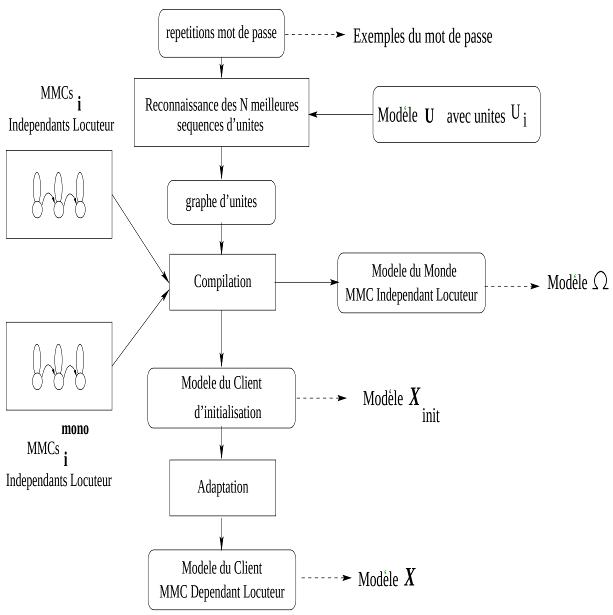

FIG. 1 – *Synopsis de la création du mot de passe personnalisé*

5.3. Résultats

Comme on l'observe sur les courbes DET de la figure 2 et dans la table 1, les performances de l'approche par mot de passe personnalisé (MPP) s'améliorent avec le nombre de répétitions d'apprentissage du mot de passe, mais elles plafonnent au-delà de 4 répétitions.

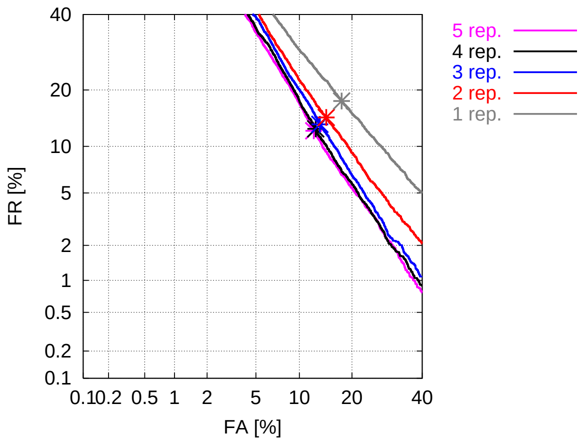

FIG. 2 – *Résultats de la méthode MPP en fonction du nombre de répétitions.*

TAB. 1 – *EER de la méthode MPP en fonction du nombre de répétitions.*

Nb répétitions	1	2	3	4	5
EER %	17,7	14,5	13,4	12,6	12,3

La figure 3 permet de comparer l'approche à base de mot de passe personnalisé (3 répétitions) et l'approche dépendante du texte (de référence) sur les mêmes mots (3 répétitions également). On constate assez naturellement une performance moindre de la méthode MPP par rapport à la méthode de référence avec un EER de 13,4% pour la méthode testée contre 9,6% pour la méthode de référence. Il semble que le facteur

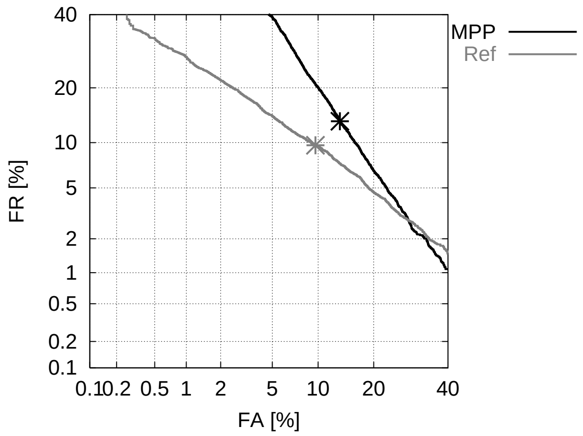

FIG. 3 – *Comparaison de la méthode de référence et de la méthode MPP.*

limitant des deux méthodes provienne beaucoup de la brièveté des mots de passe utilisés.

Au moins deux raisons peuvent expliquer la différence de performances observées pour les deux méthodes :

1. La plus ou moins bonne adéquation de la *topologie* du modèle Ω obtenu par transcription automatique. En d'autres termes, la pertinence des unités U_i.
2. La meilleure qualité des *lois d'émission* du modèle Ω lorsqu'elles sont estimées à partir de données réelles (dans l'expérience de référence).

C'est pourquoi, nous avons réalisé une expérience complémentaire, où l'étape de transcription automatique dans la méthode MPP est remplacée par une transcription *manuelle* du mot de passe (méthode MPP-TM).

La figure 4 montre que l'utilisation d'une transcription manuelle apporte des amélorations très notables à la méthode par mot de passe personnalisé, conduisant à des performances équivalentes (voire meilleures) qu'avec la méthode de référence (EER de 9,2% contre 9,6% respectivement).

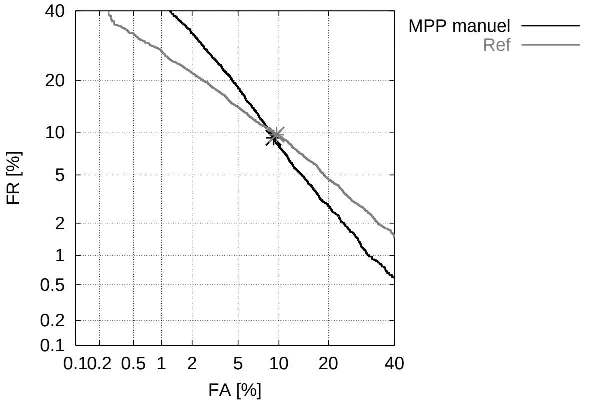

FIG. 4 – *Comparaison de la méthode de référence et de la méthode MPP-TM.*

A l'inverse, nous avons évalué les performances d'une méthode pour laquelle la transcription en unités est effectuée automatiquement mais pour laquelle on procède à la ré-estimation des lois d'émission du modèle Ω à partir de données acoustiques multi-locuteurs (méthode MPP-AC). Cette configuration, sans intérêt pratique, permet néanmoins de mesurer l'importance des données acoustiques.

La figure 5 compare l'approche MPP-AC avec la méthode de référence. Les résultats montrent que la topologie de Ω inférée par décodage des N meilleures solutions n'est pas rédhibitoire, car on peut obtenir des performances comparables à l'expérience de référence si le réapprentissage du Modèle du Monde s'effectue sur des données réelles (EER=10% contre 9,6% respectivement). Cependant, les performances observées sont sensiblement moins bonnes que lorsque l'on dispose de la transcription, même en l'absence de données acoustiques (méthode MPP-TM).

6. Conclusions

Dans une application de vérification du locuteur par mot de passe personnalisé, l'absence de données acoustique pour estimer le Modèle du Non-Client (du Monde) associé au mot de passe du client, constitue un handicap sérieux. Nos travaux indiquent l'importance cruciale de la qualité du décodage en unités (ici phonétiques) pour envisager d'atteindre des performances comparables á celles des approches dépendantes du texte. Parmi les perspectives d'améliorations, il convient de tester l'utilisation de phonèmes contextuels ou d'unités de parole plus longues que le phonème.

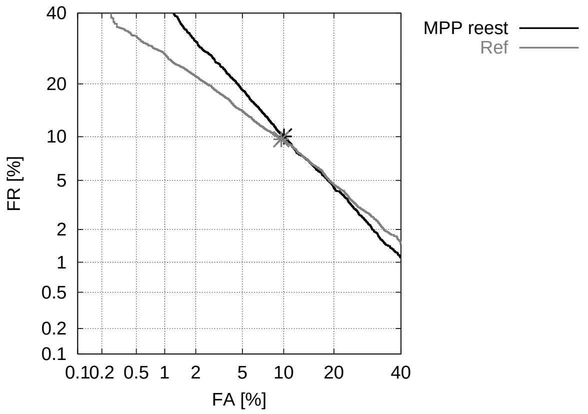

FIG. 5 – *Comparaison de la méthode de référence avec la méthode MPP-AC.*

Références

[BBB+99] F. Bimbot, M. Blomberg, L. Boves, G. Chollet, C. Jaboulet, B. Jacob, J. Kharroubi, J. Koolwaaij, J. Lindberg, J. Mariéthoz, C. Mokbel, and H. Mokbel. An Overview of the PICASSO Project Research Activities in Speaker Verification for Telephone Applications. In *Eurospeech*, volume 5, pages 1963–1966, 1999.

[Dod85] G.R. Doddington. Speaker Recognition – identifying people by their voices. In *Proc. IEEE*, volume 73, pages 1651–1664, March 1985.

[LB99] J. Lindberg and M. Blomberg. Vulnerability in Speaker Verification - A study of Technical Impostor Techniques. In *Eurospeech*, volume 3, pages 1211–1214, 1999.

[MM99] J. Mariéthoz and C. Mokbel. Polyvar Protocol. In *Rapport Interne IDIAP*, 1999.

[MP97] A. Martin and M. Przybocki. The DET curve in assesment of detection task performance. In *Eurospeech*, volume 4, pages 1895–1898, 1997.