



MINISTÈRE DE L'ÉDUCATION
NATIONALE, DE L'ENSEIGNEMENT
SUPÉRIEUR ET DE LA RECHERCHE

MAISON DE LA
RECHERCHE

AVIS DE SOUTENANCE DE THESE DE DOCTORAT

Monsieur Anthony LARCHER soutiendra une thèse
le 24 septembre 2009 à 14h

amphithéâtre du CERI - Agroparc

SPÉCIALITÉ : INFORMATIQUE ED 166

Titre de la thèse : Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée.

Membres du jury :

ANDRE-OBRECHT Régine, PR Informatique, Université de Toulouse,
CERNOCKY Jan, PR Informatique, DCGM FIT BUT, Brno, République Tchèque,
MARCEL Sébastien, Chercheur Informatique, IDIAP Research Institute, Martigny, Suisse
GRAVIER Guillaume, CR1-HDR Informatique, IRISA Rennes,
VERLINDE Patrick, PR Informatique, École Royale Militaire, Bruxelles, Belgique,
MASON John S. D. PR Informatique, Université de Swansea, Royaume Uni,
BONASTRE Jean-François, PR Informatique, Université d'Avignon et des Pays de Vaucluse.

Résumé de la thèse :

La vérification automatique du locuteur est une tâche de classification qui vise à confirmer ou infirmer l'identité d'un individu d'après une étude des caractéristiques spécifiques de sa voix. L'intégration de systèmes de vérification du locuteur sur des appareils embarqués

impose de respecter deux types de contraintes, liées à cet environnement

- les contraintes matérielles, qui limitent fortement les ressources disponibles en termes de mémoire de stockage et de puissance de calcul disponibles;
- les contraintes ergonomiques, qui limitent la durée et le nombre des sessions d'entraînement ainsi que la durée des sessions de test.

En reconnaissance du locuteur, la structure temporelle du signal de parole n'est pas exploitée par les approches état-de-l'art. Nous proposons d'utiliser cette information, à travers l'utilisation de mots de passe personnels, afin de compenser le manque de données d'apprentissage et de test.

Une première étude nous a permis d'évaluer l'influence de la dépendance au texte sur l'approche état-de-l'art GMM/UBM. Nous avons montré qu'une contrainte lexicale imposée à cette approche, généralement utilisée pour la reconnaissance du locuteur indépendante du texte, permet de réduire de près de 30% (en relatif) le taux d'erreurs obtenu dans le cas où les imposteurs ne connaissent pas le mot de passe des clients. Dans ce document, nous présentons une architecture acoustique spécifique qui permet d'exploiter à moindre coût la structure temporelle des mots de passe choisis par les clients. Cette architecture hiérarchique à trois niveaux permet une spécialisation progressive des modèles acoustiques. Un modèle générique représente l'ensemble de l'espace acoustique. Chaque locuteur est représenté par une mixture de Gaussiennes qui dérive du modèle du monde générique du premier niveau. Le troisième niveau de notre architecture est formé de modèles de Markov semi-continus, qui permettent de modéliser la structure temporelle des mots de passe tout en intégrant l'information spécifique au locuteur, modélisée par le modèle GMM du deuxième niveau. Chaque état du modèle SCHMM d'un mot de passe est estimé, relativement au modèle indépendant du texte de ce locuteur, par adaptation des paramètres de poids des distributions Gaussiennes de ce GMM. Cette prise en compte de la structure temporelle des mots de passe permet de réduire de 60% le taux d'égaux erreurs obtenu lorsque les imposteurs prononcent un énoncé différent du mot de passe des clients. Pour renforcer la modélisation de la structure temporelle des mots de passe, nous proposons d'intégrer une information issue d'un processus externe au sein de notre architecture acoustique hiérarchique. Des points de synchronisation forts, extraits du signal de parole, sont utilisés pour contraindre l'apprentissage des modèles de mots de passe durant la phase d'enrôlement. Les points de synchronisation obtenus lors de la phase de test, selon le même procédé, permettent de contraindre le décodage Viterbi utilisé, afin de faire correspondre la structure de la séquence avec celle du modèle testé. Cette approche a été évaluée sur la base de données audio-vidéo MyIdea grâce à une information issue d'un alignement phonétique. Nous avons montré que l'ajout d'une contrainte de synchronisation au sein de notre approche acoustique permet de dégrader les scores imposteurs et ainsi de diminuer le taux d'égaux erreurs de 20% (en relatif) dans le cas où les imposteurs ignorent le mot de passe des clients tout en assurant des performances équivalentes à celles des approches état-de-l'art dans le cas où les imposteurs connaissent les mots de passe. L'usage de la modalité vidéo nous apparaît difficilement conciliable avec la limitation des ressources imposée par le contexte embarqué. Nous avons proposé un traitement simple du flux vidéo, respectant ces contraintes, qui n'a cependant pas permis d'extraire une information pertinente. L'usage d'une modalité supplémentaire permettrait néanmoins d'utiliser les différentes informations structurelles pour déjouer d'éventuelles impostures par play-back.

Ce travail ouvre ainsi de nombreuses perspectives, relatives à l'utilisation d'information structurelle dans le cadre de la vérification du locuteur et aux approches de reconnaissance du locuteur assistée par la modalité vidéo.

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE
COLLEGE DES ETUDES DOCTORALES
CASE 20
74 rue Louis Pasteur
84029 AVIGNON CEDEX 1
<http://www.univ-avignon.fr>
tél : +33(0)4 90 16 25 29
fax : +33(0)4 90 16 27 44
joelle.derbaisse@univ-avignon.fr