Université Lumière Lyon 2
École doctorale : Sciences Cognitives
Institut de Psychologie

Discrmininant Models for Text-independent Speaker Verification / Algorithmes d'apprentissage discriminants en vérification du locuteur

Par Johnny MARIÉTHOZ

Thèse de doctorat de Sciences Cognitives
Mention Informatique
Dirigée par Hélène PAUGAM-MOISY
Présentée et soutenue publiquement le 20 décembre 2006

Devant un jury composé de : Samy BENGIO, Senior Researcher, Idiap Research Institute
Jean-François BONASTRE, Maître de conférences, HDR, Université d'Avignon Yves GRANDVALET, Chargé de recherche, HDR, Université de Compiègne Hélène PAUGAM-MOISY, Professeur des universités, Université Lyon 2

Table des matières

Contrat de diffusion

Résumé

Dans cette thèse le problème de la vérification du locuteur indépendante du texte est abordée du point de vue de l'apprentissage statistique (machine learning). Les théories développées en apprentissage statistique permettent de mieux définir ce problème, de développer de nouvelles mesures de performance non-biaisées et de proposer de nouveaux tests statistiques afin de comparer objectivement les modèles proposés. Une nouvelle interprétation des modèles de l'état de l'art basée sur des mixtures de gaussiennes (GMM) montre que ces modèles sont en fait discriminants et équivalents à une mixture d'experts linéaires. Un cadre théorique général pour la normalisation des scores est aussi proposé pour des modèles probabilistes et non-probabilistes. Grâce à ce nouveau cadre théorique, les hypothèses faites lors de l'utilisation de la normalisation Z et T (T- and Z-norm) sont mises en évidence.

Différents modèles discriminants sont proposés. On présente un nouveau noyau utilisé par des machines à vecteurs de support (SVM) qui permet de traîter des séquences. Ce noyau est en fait la généralisation d'un noyau déjà existant qui présente l'inconvénient d'être limité à une forme polynomiale. La nouvelle approche proposée permet la projection des données dans un espace de dimension infinie, comme c'est le cas, par exemple, avec l'utilisation d'un noyau gaussien. Une variante de ce noyau cherchant le meilleur vecteur acoustique (frame) dans la séquence à comparer, améliore les résultats actuellement connus. Comme cette approche est particulièrement coûteuse pour les séquences longues, un algorithme de regroupement (clustering) est utilisé pour en réduire la complexité.

Finalement, cette thèse aborde aussi des problèmes spécifiques de la vé-ri-fi-ca-tion du locuteur, comme le fait que les nombres d'exemples positifs et négatifs sont très déséquilibrés et que la distribution des distances intra et inter classes est spécifique de ce type de problème. Ainsi, le noyau est modifié en ajoutant un bruit gaussien sur chaque exemple négatif. Même si cette approche manque de justification théorique pour l'instant, elle produit de très bons résultats empiriques et ouvre des perspectives intéressantes pour de futures recherches.

Mots-Clés : Mixture de gaussiennes, machine à vecteurs de support, fonction de coût, vérification du locuteur indépendante du texte, problème déséquilibré, mesure de similarité, noyau de séquences.

Abstract

This thesis addresses text-independent speaker verification from a machine learning point of view. We use the machine learning framework to better define the problem and to develop new unbiased performance measures and statistical tests to compare objectively new approaches. We propose a new interpretation of the state-of-the-art Gaussian Mixture Model based system and show that they are discriminant and equivalent to a mixture of linear classifiers. A general framework for score normalization is also given for both probability and non-probability based models. With this new framework we better show the hypotheses made for the well known Z- and T- score normalization techniques.

Several uses of discriminant models are then proposed. In particular, we develop a new sequence kernel for Support Vector Machines that generalizes an other sequence kernel found in the literature. If the latter is limited to a polynomial form the former allows the use of infinite space kernels such as Radial Basis Functions. A variant of this kernel that finds the best match for each frame of the sequence to be compared, actually outperforms the state-of-the-art systems. As our new sequence kernel is computationally costly for long sequences, a clustering technique is proposed for reducing the complexity.

We also address in this thesis some problems specific to speaker verification such as the fact that the classes are highly unbalanced. And the use of a specific intra- and inter-class distance distribution is proposed by modifying the kernel in order to assume a Gaussian noise distribution over negative examples. Even if this approach misses some theoretical justification, it gives very good empirical results and opens a new research direction.

Keywords: Gaussian Mixture Models, Support Vector Machines, loss function, cost, text-independent speaker verification, unbalanced class problem, similarity measure, sequence kernel.

Résumé étendu

Dans cette thèse, le problème de la vérification du locuteur indépendante du texte est abordée du point de vue de l'apprentissage statistique. Les modèles de référence en vérification du locuteur sont des modèles non-discriminants, ce qui va à l'encontre de la théorie proposée par les chercheurs en apprentissage statistique. Le but de cette thèse est d'utiliser les cadres théoriques proposés en apprentissage statistique pour mieux définir le problème de la vérification du locuteur, mieux comprendre les modèles de référence, proposer de nouvelles mesures de performance et finalement développer de nouveaux modèles discriminants.

Cette thèse se déroule comme suit. Tout d'abord, la vérification du locuteur et l'apprentissage statistique sont brièvement décrits. Les modèles de référence sont présentés et une nouvelle mesure de performance non biaisée est décrite pour comparer les modèles proposés de la manière la plus objective possible. L'apprentissage statistique nous offre un nouveau regard sur la vérification du locuteur et permet ensuite de développer un nouveau cadre théorique qui peut s'étendre à la normalisation de scores. Finalement, de nouvelles approches utilisant des modèles discriminants sont développées incluant une nouvelle mesure de similarité.

Introduction

Avec l'avènement et l'omniprésence des technologies de l'information, on retrouve de plus en plus de situations dans lesquelles des utilisateurs ont besoin de stocker ou d'échanger des informations personnelles de manière sûre. La solution la plus utilisée consiste à utiliser des cartes d'accès ou des codes personnels. Ces types de systèmes sont communément utilisés dans des applications de type bancaire ou lors de l'utilisation d'un ordinateur. Le désavantage de ces systèmes est qu'une carte ou un code peut facilement être volé ou perdu. Une alternative consiste à utiliser des informations biométriques telles que l'empreinte digitale, le visage, l'iris ou la voix, que l'on espère uniques pour chaque individu. Cette thèse, s'intéresse à l'authentification ou vérification d'identité biométrique basée sur des enregistrements vocaux utilisant des algorithmes provenant du domaine de l'apprentissage statistique.

Qu'est-ce que la vérification du locuteur ?

Un système de vérification du locuteur vérifie l'identité proclamée d'une personne en utilisant un enregistrement vocal de sa voix. Il doit soit l'accepter comme client, soit le rejeter comme imposteur. Différents systèmes sont à considérer :

un système dépendent du texte : le contenu phonétique de l'enregistrement est fixé à l'avance. Par exemple, le système demande à la personne de prononcer une phrase.

un système indépendant du texte : le choix du contenu phonétique est laissé au client.

Le premier a l'avantage d'être robuste aux attaques intentionnelles (lorsqu'un imposteur reproduit une phrase pré-enregistrée du client), mais a l'inconvénient de nécessiter des modèles plus complexes (comme des reconnaisseurs de la parole) et demande à l'utilisateur de répéter exactement la phrase demandée. Dans cette thèse, seule la vérification du locuteur indépendante du texte sera considérée, car elle est la plus utilisée : elle est simple à mettre en oeuvre et ne demande pas de reconnaisseur de la parole. Elle est donc mieux adaptée pour la plupart des applications courantes (téléphone mobile, assistant personnel,...).

Durant les vingt dernières années, le domaine de la vérification du locuteur a bénéficié des résultats de la recherche dans le domaine de l'apprentissage statistique grˆace notamment à l'expansion rapide de la puissance de calcul des ordinateurs.

Qu'est-ce que l'apprentissage statistique ?

L'apprentissage statistique est un domaine à la frontière de l'informatique et des statistiques. Il consiste à développer des algorithmes qui permettent aux ordinateurs "d'apprendre" grˆace à l'expérience. Pour apprendre une solution à un problème, l'algorithme a besoin d'exemples d'apprentissage. Le but est alors de trouver la meilleure

fonction, parmi un ensemble préétabli de fonctions, en minimisant une fonction de coût sur les exemples d'apprentissage. L'ensemble de fonctions choisi au préalable doit être suffisamment riche pour contenir une bonne solution, mais suffisamment simple pour que la solution choisie puisse être généralisée à des exemples jamais vus par le système. La solution trouvée par un algorithme d'apprentissage statistique est appelée modèle. La communauté de l'apprentissage statistique a développé des algorithmes pour résoudre des problèmes variés tels que : la vérification du locuteur, la catégorisation de textes, la vérification d'identité utilisant le visage, etc.

Modèles de référence

Le modèle le plus utilisé en vérification du locuteur est basé sur des mélanges de distributions gaussiennes (GMM) (Reynolds and Rose, 1995) [1] . On commence par entraîner un premier GMM, appelé modèle de monde, en maximisant la vraisemblance des exemples d'enregistrements vocaux venant d'une grande quantité de locuteurs. Plus la diversité des locuteurs est grande, meilleur sera le modèle. Ce modèle représente l'hypothèse qu'un imposteur a prononcé la phrase enregistrée. Par opposition le modèle client représente l'hypothèse que le client a prononcé la phrase enregistrée. Contrairement au modèle de monde, le nombre d'exemples d'entraînement disponibles pour estimer ce modèle est restreint : le client prononce généralement entre une et trois phrases avant d'utiliser le système. Donc, comme peu de données sont disponibles, plutôt que d'apprendre un nouveau GMM avec les données du client, les paramètres du modèle de monde sont adaptés avec ces données. Cette méthode, appelée Maximum A Posteriori (MAP), (Gauvain and Lee, 1994) [2] comporte un paramètre à ajuster qui permet de contraindre le modèle client à rester plus ou moins proche du modèle de monde. Habituellement seules les moyennes des gaussiennes sont modifiées. Finalement lors de la prise de décision, chaque hypothèse est testée en calculant un score, appelé vraisemblance, pour chacun des modèles. Le ratio de ces vraisemblances est comparé à un seuil de décision appris au préalable sur un autre ensemble de clients. Ce seuil est donc indépendant des clients.

Il est à noter que pour obtenir des performances optimales, des modifications ont été apportées de manière empirique par la communauté de la vérification du locuteur. L'utilisation de techniques d'adaptation en est une. Un facteur de normalisation a aussi été ajouté pour rendre le ratio des vraisemblances indépendant de la longueur de la phrase à traiter. De plus, lors de l'estimation du modèle de monde, les variances des gaussiennes sont contraintes à des valeurs minimales souvent comprises entre 10 et 60% de la variance globale des données. Les modèles de référence sont donc a priori non-discriminants, en ce sens que chaque classe est modélisée séparément, ce qui va à l'encontre de la vision de l'apprentissage statistique. De plus les modifications citées plus haut n'ont pas toutes de justification théorique.

[1] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Transactions On Speech and Audio Processing, 3 (1), 1995.

[2] J. L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observation of markov chains. In IEEE Transactions on Speech Audio Processing, volume 2, pages 291–298, April 1994.

Récemment, Campbell (2002) [3] a proposé l'utilisation d'un modèle discriminant pour la vérification du locuteur qui a donné de bons résultats empiriques. Le noyau proposé permet de calculer une mesure de similarité entre deux enregistrements vocaux, chacun représenté par une séquence de taille variable de vecteurs caractéristiques. Il s'agit de calculer pour chaque vecteur caractéristique une expansion polynomiale de degré trois et de moyenner les vecteurs étendus sur toute la séquence. Le vecteur résultant est utilisé comme entrée d'une machine à vecteurs de support (SVM) avec un noyau linéaire. Même si cette approche semble prometteuse, elle est limitée aux noyaux polynomiaux et manque d'interprétation théorique.

Mesures de performance

Tout au long de cette thèse, différents systèmes sont comparés. Afin que ces comparaisons soient le plus objectives possible, il faut utiliser les méthodes les moins biaisées possible et donner un intervalle de confiance pour chaque taux d'erreur mesuré. Les mesures utilisées en vérification du locuteur proviennent de la combinaison de deux types d'erreurs : les faux positifs (le système accepte un imposteur) et les faux négatifs (le système rejette un client). L'erreur résultante est obtenue en faisant varier le seuil de décision, soit en minimisant une fonction de coût qui dépend du niveau de sécurité voulu, soit en considérant tous les seuils possibles. Dans le premier cas, un nombre est obtenu, dans le second, l'ensemble des nombres est représenté sous la forme d'une courbe appelée courbe DET (Martin et al., 1997) [4] . Lorsque le seuil est estimé sur une population de clients différente (ensemble de développement) de celle utilisée pour estimer la qualité d'un système (ensemble de test) on l'appelle a priori sinon on l'appelle a posteriori. Un seuil a posteriori donne des résultats biaisés de manière optimiste et ne devrait donc pas être utilisé pour comparer différents systèmes.

Malheureusement, en parcourant la littérature, les mesures a posteriori sont souvent utilisées pour comparer des systèmes. En particulier la courbe DET est une courbe a posteriori et donc ne devrait pas être utilisée pour comparer des systèmes. De nouvelles courbes appelées courbes de performances espérées (EPC), incluant l'estimation des seuils sur un ensemble de développement, ont été développées dans le cadre de cette thèse. Ce travail de recherche a été publié dans :

[3] W.M. Campbell. Generalized linear discriminant sequence kernels for speaker recognition. In Proc IEEE International Conference on Audio Speech and Signal Processing, pages 161–164, 2002.

[4] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In Proceedings of Eurospeech'97, Rhodes, Greece, pages 1895–1898, 1997.

CONTRIB S. Bengio, J. Mariéthoz, and M. Keller. The expected performance curve. In *International Conference on Machine Learning, ICML, Workshop on ROC Analysis in Machine Learning*, 2005

et plus spécifiquement pour la vérification du locuteur dans :

CONTRIB S. Bengio and J. Mariéthoz. The expected performance curve: a new assessment measure for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, 2004

De plus comme aucun test statistique, permettant d'estimer un intervalle de confiance, n'est directement utilisable pour la vérification du locuteur, une variante du Z-test, très utilisé, à été proposé dans :

CONTRIB S. Bengio and J. Mariéthoz. A statistical significance test for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, pages 237–240, 2004

Ce test permet de dire si deux systèmes peuvent être considérés comme statistiquement significativement différents ou non avec plus de 95% de confiance.

Toutes les expériences faites dans cette thèse ont été effectuées avec l'aide de trois bases de données :

Switchboard : base de données d'enregistrements téléphoniques américaine utilisée durant les concours NIST effectués chaque année par la communauté de la vérification du locuteur.

Polyvar : base de données d'enregistrements téléphoniques en français enregistrée durant plus d'une année.

Banca : base de données incluant des enregistrements effectués dans des environnements

de qualités variées.

Pour chacune de ces bases, les résultats des modèles de référence obtenus correspondent à ceux trouvés dans la littérature. Un protocole expérimental a été créé pour l'utilisation de modèles discriminants. La base de donnée Banca ainsi que son protocole ont été publiés dans :

CONTRIB E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 625–638. Springer-Verlag, 2003

La base de donnée Polyvar et son protocole ont été décrits dans :

CONTRIB 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 *6th European Conference on Speech Communication and Technology — Eurospeech'99*, volume 5, pages 1963–1966, Budapest, Hungary, september 1999

Vérification du locuteur du point de vue de l'apprentissage statistique

Un problème habituel en apprentissage statistique est de classer des exemples en deux catégories; c'est ce qu'on appelle un problème de classification supervisée à deux classes (Bishop, 1995) [5] . Les modèles utilisés généralement pour résoudre cette tˆache sont soit discriminants (ils cherchent un hyperplan qui sépare le mieux les deux classes), soit génératifs (ils estiment indépendamment la distribution de chacune des deux classes et utilisent la règle de Bayes pour prendre une décision). Selon Vapnik (2000) [6] , il ne faudrait pas essayer de résoudre un problème plus difficile que la tˆache qui est assignée. Donc les modèles discriminants devraient être préférés aux modèles non-discriminants pour des tˆaches de classification. Dans cette thèse on considère le problème de la vérification du locuteur comme un problème de classification à deux classes pour chaque client.

Lorsque l'on parcourt la littérature de la vérification du locuteur, il est intéressant de

[5] C. Bishop. Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1995.

[6] V. N. Vapnik. The nature of statistical learning theory. Springer, second edition, 2000.

noter que le modèle de référence dominant ne semble pas discriminant. En fait, tout se trouve dans les détails : la communauté a rajouté empiriquement au cours des années différentes modifications qui permettent d'atteindre de très bonnes performances. Il est possible de montrer, en utilisant un algorithme d'alignement synchrone, que ce modèle basé sur des GMM utilisant des techniques d'adaptation est devenu discriminant avec l'apport de modifications empiriques. Il peut être notamment interprété comme un mélange d'experts linéaires.

L'algorithme d'alignement synchrone a été publié dans :

CONTRIB J. Mariéthoz, Dominique Genoud, Frédéric Bimbot, and Chafik Mokbel. Client / world model synchronous alignement for speaker verification. In *6th European Conference on Speech Communication and Technology — Eurospeech'99*, Budapest, Hungary, September 1999

La comparaison entre différentes méthodes d'adaptation a été publiée dans :

CONTRIB J. Mariéthoz and S. Bengio. A comparative study of adaptation methods for speaker verification. In *International Conference on Spoken Language Processing ICSLP*, pages 581–584, Denver, CO, USA, September 2002. IDIAP-RR 01-34

Comme le modèle basé sur des GMM est devenu discriminant, il est intéressant de considérer directement d'autres modèles discriminants comme les SVM. Il faut tout d'abord généraliser le cadre théorique utilisé pour les GMM aux modèles discriminants. Le développement de ce cadre théorique a été originalement présenté dans :

CONTRIB J. Mariéthoz and S. Bengio. A kernel trick for sequences applied to text-independent speaker verification systems. IDIAP-RR 77, IDIAP, 2005

Une extension de ce cadre théorique permet de généraliser les techniques standards de normalisation de scores. Ce travail a été publié dans :

CONTRIB J. Mariéthoz and S. Bengio. A unified framework for score normalization techniques applied to text independent speaker verification. *IEEE Signal Processing Letters, Volume 12*, 12, 2005. IDIAP-RR 04-62

L'idée principale est de rendre les modèles robustes aux changements de conditions

d'enregistrement. Le nouveau cadre théorique permet de mieux comprendre les hypothèses faites lors de l'utilisation des méthodes de référence telles que la Z-norm et la T-norm. De plus elle permet le développement de nouvelles techniques de normalisation estimant n'importe quelle forme de distribution de scores alors que les méthodes de référence sont limitées à une distribution gaussienne.

Modèles discriminants simples

Dans une première approche, un modèle discriminant simple a été proposé. L'idée principale est de remplacer la fonction de décision d'un modèle basé sur des GMM. Elle peut être vue comme une fonction linéaire de deux vraisemblances produites par le modèle client et le modèle de monde de pente 1. Il est possible d'utiliser à la place un modèle discriminant, par exemple une SVM, prenant comme entrée les deux vraisemblances. Ce travail à été publié dans :

CONTRIB S. Bengio and J. Mariéthoz. Learning the decision function for speaker verification. In *IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP*, Salt Lake, City, USA, 2001. IDIAP-RR 00-40

En dehors de ces deux vraisemblances, d'autres valeurs peuvent être utilisées par une SVM. Il est possible par exemple d'enrichir cette représentation avec des ratios de vraisemblances calculés pour chaque gaussienne du modèle de référence dans le but d'augmenter la dimension et la richesse du vecteur d'entrée. Après avoir analysé les résultats, il semble que le fait d'avoir un modèle discriminant commun à tous les clients soit une limitation. L'utilisation d'un modèle discriminant comme fonction de décision utilisant un grand vecteur de ratios de vraisemblances à été présentée dans :

CONTRIB J. Mariéthoz and S. Bengio. An alternative to silence removal for text-independent speaker verification. IDIAP-RR 51, IDIAP, Martigny, Switzerland, 2003

Noyaux de séquences

Des problèmes propres à la vérification du locuteur rendent l'utilisation de modèles discriminants difficile. Tout d'abord, chaque enregistrement vocal est transformé en séquence de vecteurs de taille variable qui dépend du contenu phonétique de la phrase et du débit de parole propre à chaque locuteur. Malheureusement, la plupart des modèles

discriminants fonctionnent uniquement avec des vecteurs de taille fixe. Dans la section précédente, le problème était résolu en utilisant des GMM. Une alternative consiste à utiliser des SVM avec un noyau particulier qui traite des séquences. Habituellement, chaque exemple est un vecteur de taille fixe et le noyau calcule une mesure de similarité entre deux exemples dans un espace de projection.

Afin de pouvoir traiter des séquences, le noyau proposé estime la moyenne des valeurs calculées par un noyau local entre toutes les combinaisons possibles de paires de vecteurs des deux séquences à comparer. Il est possible de montrer que cette approche généralise le modèle de référence proposé par Campbell (2002) [7] et est équivalent à ce modèle si le noyau local est de forme polynomiale. Il est intéressant de noter qu'avec l'approche proposée n'importe quel noyau standard peut être utilisé comme noyau local. Ceci est vrai aussi pour des noyaux de dimension infinie comme le noyau gaussien.

Il semble cependant contre-intuitif de comparer tous les vecteurs caractéristiques d'une séquence avec tous les vecteurs caractéristiques d'une autre séquence. En effet, ils représentent une sorte de sous-unité phonétique et donc il semble raisonnable de vouloir comparer des vecteurs caractéristiques représentant le même sous-phonème. Partant de cette idée, un autre noyau de séquences a été développé. Il cherche, pour chaque vecteur caractéristique d'une première séquence, son plus proche voisin dans une deuxième séquence. Cette approche améliore de manière significative les résultats. Il est aussi possible de régulariser la recherche du meilleur vecteur caractéristique en appliquant une fenêtre glissante sur la séquence. Empiriquement, cette approche donne de très bons résultats et suggère de poursuivre la recherche dans cette direction. Les noyaux de séquences ont été publiés dans :

CONTRIB J. Mariéthoz and S. Bengio. A kernel trick for sequences applied to text-independent speaker verification systems. In *Second Workshop on Multimodal User Authentication, MMUA*, 2006. IDIAP-RR 05-77

Une version étendue à été soumise au journal Pattern Recognition. Malheureusement, les noyaux de séquences proposés sont relativement coûteux en temps de calcul pour des séquences longues de plusieurs minutes comme c'est le cas pour la base de donnée Switchboard (NIST). Une méthode basée sur des algorithmes de regroupement est proposée pour en réduire la complexité.

Mesures de similarité

Un autre problème spécifique de la vérification du locuteur est le fort déséquilibre qui

[7] W.M. Campbell. Generalized linear discriminant sequence kernels for speaker recognition. In Proc IEEE International Conference on Audio Speech and Signal Processing, pages 161–164, 2002.

existe entre le nombre de données d'entraînement du client et celui des imposteurs. Le client prononce habituellement entre une et trois phrases, alors que les phrases imposteurs viennent d'une large population de locuteurs (souvent plusieurs centaines). Il semble donc important de tenir compte de ce déséquilibre lors de l'apprentissage. Dans le cas des SVM, il existe des critères qui tiennent compte de ce phénomène, (Lin et al., 2002) [8] . Une approche basée sur une interprétation probabiliste des SVM pour résoudre ce problème

a aussi été proposée dans :

CONTRIB Y. Grandvalet, J. Mariéthoz, and S. Bengio. A probabilistic interpretation of svms with an application to unbalanced classification. In *Advances in Neural Information Processing Systems, NIPS 15*, 2005. IDIAP-RR 05-26

Empiriquement, ces deux approches n'ont donné aucun résultat probant pour le cas de la vérification du locuteur. Cela peut être expliqué par le fait que la fonction optimale trouvée par la SVM sépare parfaitement les données. Cela veut dire que tous les exemples d'apprentissage ont été correctement classés par le modèle. Dans ce cas, la modification de la fonction de coût ne sert à rien.

En fait, le déséquilibre lui-même n'est plus vraiment important, car la SVM va considérer uniquement les exemples proches de la surface de séparation. Par contre, si le déséquilibre des données ne semble pas un problème, il reste que la distribution des imposteurs est peu représentative. En effet, si les exemples d'apprentissage d'un client doivent couvrir la variabilité d'un seul locuteur, les exemples d'un locuteur utilisés comme donnée d'imposture doivent couvrir sa propre variabilité, mais aussi celle d'éventuels futurs imposteurs. Un bon modèle devrait donc tenir compte de cette particularité. Vapnik (2000) [9] propose une méthode appelée minimisation du risque de proximité (vicinal risk minimization) qui considère des distributions plutôt que des points comme exemples d'apprentissage. Partant de cette idée, un bruit gaussien est rajouté sur chaque exemple d'imposteur. Si le noyau de la SVM est gaussien, la solution est analytique. Afin d'obtenir de très bons résultats empiriques, des simplifications ont été nécessaires. Même si finalement, la justification théorique reste incomplète, les résultats obtenus orientent la recherche en ce sens.

Conclusion

[8] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. Machine Learning, 46:191–202, 2002.

[9] V. N. Vapnik. The nature of statistical learning theory. Springer, second edition, 2000.

De manière générale, dans cette thèse, le cadre théorique de l'apprentissage statistique est utilisé afin de développer une bonne méthodologie et un bon cadre théorique pour la vérification du locuteur. Différents modèles discriminants ont été proposés. Ils améliorent la performance (en terme d'erreur) des systèmes de référence, mais surtout ils augmentent la compréhension de ces modèles.

Cela ouvre différentes directions de recherche. Par exemple, le cadre théorique proposé pour la normalisation des scores permet l'utilisation de nouvelles procédures basées sur des estimations de distributions de scores nongaussiennes. L'utilisation de fenêtres de lissage pour les noyaux de séquences suggère de développer de nouvelles contraintes temporelles pour ces noyaux. Il semble aussi prometteur d'inclure un bruit sur les exemples d'imposteurs afin de couvrir les imposteurs de test. Ce type d'approche peut aussi être utilisé pour compenser des variations de conditions d'enregistrement. Un autre problème général est que, dans les applications réelles, personne ne sait ce que peut être un imposteur, et quel genre de stratégie peut être mise en oeuvre pour percer ces systèmes. Ce problème est d'autant plus difficile à définir qu'un bon imitateur professionnel n'arrive pas à confondre un système automatique, (Mariéthoz and Bengio, 2005) et inversement les êtres humains sont plus performants pour la vérification d'enregistrements en environnements bruités. En terme d'applications, il est évident que les besoins s'orientent de plus en plus vers des applications nomades et donc les futurs systèmes devront être robustes à des environnements très bruités. Même s'il existe déjà des solutions pour des niveaux de bruit raisonnable, pour que ces systèmes soient capable de traiter des enregistrements faits en tous lieux, il faudra sûrement spécialiser les microphones, par exemple par l'utilisation de groupes de microphones (microphone array). Une autre approche consiste à utiliser d'autres modalités biométriques telles que le visage, le suivi du mouvement des lèvres, etc. Des approches existantes combinent les scores de systèmes appris indépendamment sur chaque modalité, mais il serait plus élégant d'apprendre ces modèles de manière conjointe.

Autres contributions

Tous les algorithmes développés dans cette thèse sont basés sur une bibliothèque informatique d'apprentissage statistique appelée Torch. Elle est largement utilisée par les chercheurs de ce domaine et est disponible sur http://www.torch.ch. L'auteur est un des principaux contributeurs de cette bibliothèque.

Durant le déroulement de cette thèse, d'autres contributions scientifiques ont été publiées mais n'ont pas été traitées dans ce doucement. En voici la liste :

CONTRIB S. Marcel, J. Mariéthoz, Y. Rodriguez, and F. Cardinaux. Bi-modal face and speech authentication: a biologin demonstration system. In *Workshop on Multimodal User Authentication (MMUA)*, 2006. IDIAP-RR 06-18

CONTRIB Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24(8):882–893, 2006

CONTRIB M. Liwicki, A. Schlapbach, H. Bunke, S. Bengio, J. Mariéthoz, and J. Richiardi. Writer identification for smart meeting room systems. In *Seventh IAPR Workshop on Document Analysis Systems, DAS*, 2006

CONTRIB J. Mariéthoz and S. Bengio. Can a professional imitator fool a GMM-based speaker verification system? IDIAP-RR 61, IDIAP, 2005

CONTRIB J. Mariéthoz and S. Bengio. A new speech recognition baseline system for numbers 98 version 1.2 based on torch. IDIAP-RR 16, IDIAP, 2004

CONTRIB Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Estimating the quality of face localization for face verification. In *IEEE International Conference on Image Processing, ICIP*, 2004

CONTRIB C. Sanderson, S. Bengio, H. Bourlard, J. Mariéthoz, R. Collobert, M.F. BenZeghiba, F. Cardinaux, and S. Marcel. Speech & face based biometric authentication at idiap. In *International Conference on Multimedia and Expo, ICME*, 2003

CONTRIB S. Bengio, C. Marcel, S. Marcel, and J. Mariéthoz. Confidence measures for multimodal identity verification. *Information Fusion*, 3(4):267–276, 2002

Thesis in PDF format

Contents

mariethoz_these_tdm.pdf

1 Introduction

mariethoz_these_1intro.pdf

2 Text-Independent Speaker Verification Systems 5

mariethoz_these_chap2.pdf

3 Performance Measures for Speaker Verification 21

mariethoz_these_chap3.pdf

4 Experimental Methodology

mariethoz_these_chap4.pdf

5 Text-Independent Speaker Verification: a Machine Learning Perspective

mariethoz_these_chap5.pdf

6 GMMs and Discriminant Models

mariethoz_these_chap6.pdf

7 Sequence Kernel Based Speaker Verification

mariethoz_these_chap7.pdf

8 A New Perspective: Working on the Distance Measure

mariethoz_these_chap8.pdf

9 Conclusion

mariethoz_these_9conclu.pdf

Bibliography

mariethoz_these_biblio.pdf